université
PARIS-SACLAY

Mean-field methods and algorithmic perspectives for high-dimensional machine learning

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 564
École Doctorale Physique en Île-de-France (EDPIF)
Spécialité de doctorat: Physique
Unité de recherche: Université Paris-Saclay, CNRS, CEA
Institut de physique théorique
91191, Gif-sur-Yvette, France.
Référent: Faculté des sciences d'Orsay

Thèse présentée et soutenue en visioconférence totale, le 16/12/2020, par

Benjamin AUBIN

Composition du jury:

Romain COUILLET Professeur, Centrale-Supélec Université Paris-Saclay	Président
Sundeep RANGAN Professeur, directeur associé NYU Wireless	Rapporteur & Examinateur
David SAAD Professeur Aston University	Rapporteur & Examinateur
Marc MEZARD Directeur de recherche CNRS École Normale Supérieure	Examinateur
Alberto ROSSO Directeur de recherche CNRS Université Paris-Saclay	Examinateur
Lenka ZDEBOROVA Directrice de recherche CNRS EPFL	Directrice de thèse
Florent KRZAKALA Professeur EPFL	Invité

Thèse de doctorat

NNT: 2020UPASP083

MEAN-FIELD METHODS AND ALGORITHMIC PERSPECTIVES FOR HIGH-DIMENSIONAL MACHINE LEARNING

BENJAMIN AUBIN

Institut de Physique Théorique
CEA & Université Paris-Saclay

December 16, 2020

The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particules, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other.

More is different— P. W. Anderson (1972)

ACKNOWLEDGEMENTS

First of all, I would like to warmly thank Sundeep Rangan, David Saad, Romain Couillet, Marc Mézard and Alberto Rosso for accepting to read and review this Ph.D manuscript.

Ensuite, je souhaite remercier chaleureusement Lenka et Florent pour m'avoir donné la chance de passer ces trois années à leur contact et l'occasion de voyager aux quatre coins du monde. Membre à part entière de leur grande famille, ils m'auront donné goût à la recherche fondamentale et énormément appris autour de nombreuses discussions et moments de convivialité. Naturellement je remercie tous les membres des groupes de recherche *Sphinx & Smile* pour leur bonne humeur et enthousiasme: Alia, Marylou, Antoine M, Stefano et Francesca avec qui j'ai partagé (à l'occasion) mon bureau à l'IPhT, et tous les autres: Federica, Jonathan, Cédric, Ruben, Hugo, Maria, Paula, Luca, Gabriele, Laura et Alejandro pour leur contact quotidien. Et surtout, un immense merci à Sebastian, Bruno, Christian, Antoine B, Stéphane et Pierre pour la relecture de ce manuscrit, et à Alaa, Thibault, André et Levent pour leurs nombreux conseils avisés. Je souhaite évidemment remercier Jean, Nicolas, Will et Yue pour avoir contribué à démontrer certains résultats de ce manuscript. Une pensée particulière à Felix pour avoir partagé notre bureau, nos repas, nos débats socio-politiques et pour avoir bravé les vagues au milieu des "requins"; et enfin à Marco qui a su nous faire profiter d'une éclipse surréaliste au milieu de la Death Valley. Merci à Tristan, Clément, Louise, José et Dhruv pour avoir égayé les couloirs de l'ENS; à Léo et Alexandre pour notre expérience de crypto-trading; à Samuel Kindermann et Velten Doering pour leurs suggestions musicales; et à Stefano et Riccardo pour m'avoir appris quelques "rudiments" d'italien. Merci à Pierfrancesco pour avoir répondu à toutes mes questions techniques, à Guilhem pour sa gentillesse et son humour détonant et à Giulio pour ses conseils et pour m'avoir laissé la possibilité d'étudier la physique tout en pratiquant la planche à voile et le catamaran, et en suivant la coupe du monde de football. Enfin un grand merci à Laure, Sylvie et Carine pour leur gentillesse, leur sens du détail et le travail administratif colossal qu'elles m'auront aidé à surmonter. Une attention toute particulière à G. Montambaux et J. P. Bouchaud qui m'ont introduit et enseigné la physique statistique des systèmes complexes à l'X, et sans qui mon parcours aurait été probablement très différent. Merci à Marc Goerbig pour son amitié et son accueil chaleureux au LPS d'Orsay et à Léon Bottou pour sa bienveillance au sein de FAIR.

Enfin merci à Albane pour son soutien dans les moments de doute et surtout pour avoir toléré mon ordinateur allumé lors de nos soirées films! Pour finir, j'éprouve une reconnaissance toute particulière envers mes parents sans qui je ne serais jamais arrivé aussi loin.

TABLE OF CONTENTS

I AN INTRODUCTION AT THE CROSSROADS OF MACHINE LEARNING AND STATISTICAL PHYSICS

II MAIN CONTRIBUTIONS

II A. BAYES-OPTIMAL, EMPIRICAL RISK MINIMIZATION AND WORST-CASE ANALYSIS IN SIMPLE FEED-FORWARD NEURAL NETWORKS

II B. Theory for the statistical estimation with random multi-layer neural networks generative priors

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AMP	Approximate Message Passing
ANN	Artificial Neural Networks
BBP	Baik, Ben Arous and Péché
BP	Belief Propagation
CLT	Central Limit Theorem
CNN	Convolutional Neural Networks
CPU	Central Processing Units
CS	Compressed Sensing
CSP	Constraints Satisfaction Problem
DAG	Directed Acyclic Graphs
dAT	de Almeida Thouless
DL	Deep Learning
DNN	Deep Neural Networks
EA	Edwards-Anderson
EC	Expectation Consistency
EP	Expectation Propagation
f1RSB	frozen 1-step Replica Symmetry Breaking
FRSB	Full Replica Symmetry Breaking
GAMP	Generalized Approximate Message Passing
GAN	Generative Adversarial Network
GD	Gradient-Descent
GLM	Generalized Linear Model
GPU	Graphics Processing Units
i.i.d	independent and identically distributed
IT	Information Theory
JPD	Joint Probability Distribution
KL	Kullback-Leibler
LAMP	Linearized Approximate Message Passing
LSTM	Long Short-Term Memory
MAP	Maximum A Posteriori
MC	Monte-Carlo
MCMC	Markov-Chain Monte-Carlo
ML	Machine Learning
MLE	Maximum Likelihood Estimator
MMSE	Minimum Mean Squared Error
MRF	Markov Random Field
MSE	Mean Squared Error
NLP	Natural Language Processing
PAC	Probably Approximately Correct

PCA	Principal Component Analysis
PR	Phase Retrieval
rCSP	random Constraints Satisfaction Problem
ReLU	Recitfied Linear Unit
RFIM	Random Field Ising Model
RL	Reinforcement Learning
RNN	Recurrent Neural Network
RS	Replica Symmetry
RSB	Replica Symmetry Breaking
RI	Rotationally Invariant
RV	Random Variable
SE	State Evolution
SGD	Stochastic Gradient-Descent
SI	Statistical Inference
SK	Sherrington-Kirkpatrick
SVD	Singular Value Decomposition
SVM	Support Vector Machines
T-S	Teacher-Student
TAP	Thouless-Anderson-Palmer
VAE	Variational Auto-Encoder
VAMP	Vector Approximate Message Passing
VC	Vapnik-Chervonenkis

LIST OF SYMBOLS

Algebra

a_i	A scalar
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$\mathbf{a} \cdot \mathbf{b}$	Scalar product of the vectors **a** and **b**
$\mathbf{A}^\intercal$	Transpose of the matrix **A**
$\mathbf{A} \otimes \mathbf{B}$	Tensorial product of **A** and **B**
$\mathbf{A} \times \mathbf{B}$	Hadamard product of **A** and **B**
$\mathbf{I}_d$	Identity matrix of size $d \times d$
$\mathbf{J}_d$	Matrix full of ones of size $d \times d$
$\mathbf{1}_d$	Vector of ones of size d
$\det(\mathbf{A})$	Determinant of **A**
$\mathrm{Tr}(\mathbf{A})$	Trace of **A**

Probabilities

$\mathrm{X}\|\mathrm{Y}$	The random variable X knowing the variable Y
$\mathbb{P}(\mathrm{X})$	The probability of the random variable X, shorthand for $\mathbb{P}(\mathrm{X} = x)$
$\mathrm{X} \sim \mathrm{P}_\mathrm{x}(.)$	X is distributed according to the distribution P_x
$\mathrm{p}_\mathrm{x}(.)$	Density of $\mathrm{P}_\mathrm{x}(.)$: $\mathrm{dP}_\mathrm{x}(x) = \mathrm{p}(x)\,\mathrm{d}x$
$\mathbb{E}_\mathrm{X}[x]$	Expectation of the random variable X
$\mathrm{Var}(\mathrm{X})$	Variance of the random variable X
$\mathcal{N}_\mathbf{x}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution of the vector **x** with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathrm{D}\mathbf{x}$	Gaussian measure: $\mathrm{D}\mathbf{x} = \mathrm{d}\mathbf{x}\, \mathcal{N}_\mathbf{x}(\mathbf{0}, \mathbf{I})$

Calculus and functions

$\log$	Natural logarithm
$\equiv$	Defined as

$\simeq$	Equal to, up to negligible terms
$f : \mathbb{A} \mapsto \mathbb{B}$	A function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f(\mathbf{x}; \boldsymbol{\theta})$	A function of $\mathbf{x}$ parametrized by $\boldsymbol{\theta}$
$f \circ g$	Composition of the functions f and g
$\mathbb{1}[x]$	Indicator function equals to 1 if x is true and 0 otherwise

Sets, graphs and indexing

$\mathbb{A}$	A set
$\mathbb{A} \setminus \mathbb{B}$	A set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$
$\mathbb{R}, \mathbb{Z}, \mathbb{N}$	The sets of real numbers, integers and positive integers
$[\![n]\!]$	Integers interval between 1 and n
$[a; b]$	Reals interval including a and b
$\{x_i\}_{i=1}^n = \{x_1, \cdots, x_n\}$	A set containing the elements x_i for $i \in [\![n]\!]$
$< ij >$	All neighbouring pairs (i, j)
∂_i	All neighbours of the node i
$\partial_i \setminus j$	∂_i except the node j

Physics

$\mathcal{H}_d$, $\mathcal{Z}_d$	Hamiltonian and partition function of d variables
Φ, φ	Free entropy and free energy
H, $\mathcal{I}$	Entropy and mutual information
β	Inverse temperature
$\langle . \rangle_\beta$	Gibbs/Boltzmann average at inverse temperature β

FOREWORD

At a time when the use of data has reached an unprecedented level, the access to large datasets precipitated their intense use to train machine learning models. The corresponding algorithms essentially aim to detect and make use of structured informations within excessively large datasets. Specifically, after many twists and turns, the celebrated, now ubiquitous, deep-learning models, based on artificial neural networks architectures, brought important numerical progresses in this direction. Overtaking other existing models from the mid-2000s, they became, in just a few years, indispensable in many industrial applications such as image classification, speech recognition, text mining, (LeCun et al., 2015) or time series prediction, object detection for face recognition, natural language processing, medical diagnosis, etc.

However, understanding most of the practical *gradient-based* algorithms used to train these oversized and complex networks, which contain up to millions of parameters, remains empirical and challenging to analyze theoretically. The main issue arising with *deep* and *wide* neural network architectures lies essentially in the succession of numerous layers through non-linear operations that make the space of optimization very *high-dimensional* and *complex*. Handling and visualizing this large collection of parameters is the central mathematical difficulty in most of state-of-the-art machine learning models and algorithms. This lack of theoretical understanding raises many questions about their efficiency and potential risks in many areas. As a result, establishing theoretical foundations on simple models and providing numerical prescriptions on which to base and explain empirical observations have become one of the fundamental challenges of the research community.

In this manuscript, we investigate these burning questions, arising in machine learning, through the lens of statistical physics of disordered systems. This singular transversal approach to computer science problems has a long and rich history (Engel et al., 2001; Mézard et al., 2009; Grassberger et al., 2012; Zdeborová et al., 2016a; Advani et al., 2017), that we revisit in the *high-dimensional regime* by focusing especially on modern algorithmic considerations and rigorous justifications. Specifically in the context of oversized neural networks, for which the number of parameters explodes, exact analytical solutions are unknown most of the time and numerical computations are ruled out. Techniques from statistical physics have been precisely designed to infer the macroscopic behavior of such a large collection of *particles* from the microscopic description of their elementary interactions. They offer a suitable set of approximations, called *mean-fields methods*, that are simple enough to be computationally tractable and rich enough to capture and reproduce interesting features of the system. Moreover, in this *thermodynamic limit*, physicists experienced that macroscopic behaviors are typically described correctly by a set of a few *order parameters*.

Applied to machine learning theory, which precisely lacks such techniques, we believe that statistical physics insights may contribute in identifying the set of relevant observables that control the large-scale properties of the system, and provide a powerful framework to analyze such complex artificial neural networks. Unfortunately, even though very powerful and believed to lead to the correct result in many situations, these techniques were derived historically without rigorous foundations. Therefore, this work is part of the current momentum of the mathematical physics community that focuses on proving former results obtained heuristically in the 90's. Additionally, while these former statistical analysis were not discussing computational perspectives, we revisit this approach by focusing on the potential *algorithmic phase transitions*.

At the heart of this work, we strongly capitalize on a probabilistic Bayesian reasoning, which contrasts with the traditional optimization approach. Moreover, we make an intense use of the deep connection between the *replica method* and *approximate message passing* algorithms to elicit the phase diagrams of simple theoretical models, which reveal nonetheless interesting features. By revisiting the *teacher-student* paradigm, that allows to create synthetic, but tractable, tasks, we focus our attention on emphasizing the potential gaps between *statistical* and *computational* thresholds.

We illustrate the efficiency of these mean-field methods on various poorly understood machine learning models. We essentially focus on synthetic tasks and data generated in the *teacher-student* paradigm, and we contribute to their understanding by describing their rich phase diagrams. First, we start by presenting the *Bayes-optimal analysis* of committee machines that reveals the existence of large computational-to-statistical gaps. Next, in a *worst-case* analysis, we bring to light a strong connection between the Rademacher generalization bound from statistical learning theory, and the storage capacity and ground state energies from the statistical physics literature, which allows us to explicitly compute the Rademacher complexity of perceptrons. We finally complete the picture by analyzing the intensively used *empirical risk minimization* of generalized linear models and we compare it to the previous *Bayes-optimal* and *worst-case analysis*. In another research direction, we define a general procedure to combine elementary models already analyzed to build up more complex and structured architectures. In this way, we develop a framework that overcomes in particular the standard *separable* prior assumption and makes possible to analyze estimation models, such as low-rank matrix estimation, phase retrieval or compressed sensing, with deep generative priors based on random weights.

ORGANIZATION OF THE MANUSCRIPT

As my work of Ph.D lies at the crossroads of machine learning and statistical physics of disordered systems, in Part I I take the opportunity to pedagogically present the basic, yet essential, theoretical concepts to follow the rest of the manuscript. In Chap. 1, I propose a high-level overview of the field of machine learning with a focus on its tortuous history, basic concepts and current challenges. Chap. 2 covers the basic tools of statistical physics that are relevant to understand the original approach we employ to tackle machine learning problems. These two first chapters are devoted to readers unfamiliar with one or the other background and can be skipped by experts. In Chap. 3, we provide a selection of important historical references to understand how these two fields are intertwined for over thirty years. It is also the occasion to review a selection of the current research axes of the statistical physics approach in artificial neural networks. Finally, we introduce the crucial Bayesian probabilistic framework and its crucial connection with statistical physics. This constitutes the cornerstone of our approach which allows us to analyze simultaneously statistical inference and random constraint satisfaction problems. In Chap. 4, we propose a methodological review of selected fundamental mean-field inference methods, originally motivated in the spin glass literature (Mézard et al., 1987), that are mainly used in the second part of the dissertation. Specifically, we remind the details of the derivations of the replica method and message passing algorithms on the class of *generalized linear models*, as a core example throughout this manuscript. Moreover, by highlighting their complementarities, we attempt to clarify how the methods are related and allow to reveal rich statistical and algorithmic phase transitions.

The Part II of this manuscript is devoted to cover the works I have contributed as a Ph.D student from October 2017 to December 2020, at *Institut de Physique Théorique in CEA-Saclay* under the supervision of Lenka Zdeborová and Florent Krzakala. The contents of the articles have already been published in a series of works which can be found online in their original format. They have been revised in order to standardize the notations of this manuscript. In particular, for the sake of clarity and conciseness, some of the lengthy proofs and calculations to which I have not directly contributed are not reported in this manuscript and can be found in the original publications listed in Sec. 3.

CONTRIBUTIONS

Part II, which brings together my main contributions, is separated in two sub-parts corresponding to parallel axes of research. In order to best reflect my work, I will detail my personal contributions to the various co-signed articles in which I participated.

In Part II. A, we discuss the complementary analysis of the *Bayes-optimal* and *worst-case* scenarios and *empirical risk minimization* of simple feed-forward neural networks with separable prior distributions. In Chap. 5, we first present the *Bayes-optimal* approach on committee machines, that provides an information theoretical lower-bound perspective. Next, we describe the analysis of the storage capacity problem in Chap. 6 and related ground state energies, within a generic random constraint satisfaction problem framework. In Chap. 7, we show that these quantities turn out to be closely related to the *worst-case* Rademacher complexity generalization error upper bound. Finally in Chap. 8, we investigate the *practical* case with the analysis of *empirical risk minimization* which is performed in practice with gradient-descent algorithms.

1. *'The committee machine: Computational to statistical gaps in learning a two-layers neural network'*. Aubin, Maillard, Barbier, Krzakala, Macris, and Zdeborová (2018b)
   Presented in Chap. 5.

   Summary: Heuristic tools from statistical physics have been used in the past to locate the phase transitions and compute the optimal learning and generalization errors in the teacher-student scenario in multi-layer neural networks. In this contribution, we provide a rigorous justification of these approaches for a two-layers neural network model called the committee machine, under a technical assumption. We also introduce a version of the approximate message passing (AMP) algorithm for the committee machine, that allows to perform optimal learning in polynomial time for a large set of parameters. We find that there are regimes in which a low generalization error is information-theoretically achievable while the AMP algorithm fails to deliver it; strongly suggesting that no efficient algorithm exists for those cases, and unveiling a large computational gap.
   Personal contributions: I have developed and implemented the AMP algorithm and its state evolution to depict the corresponding phase diagrams.

2. *'Storage capacity in symmetric binary perceptrons'*. Aubin, Perkins, and Zdeborová (2019d)
   Presented in Chap. 6.

   Summary: We study the problem of determining the capacity of the binary perceptron for two variants of the problem where the corresponding constraint is symmetric. We call these variants the rectangle-binary-perceptron (RPB) and the $u-$function-binary-perceptron (UBP). We show that, unlike for the usual step-function-binary-perceptron, the critical capacity in these symmetric cases is given by the annealed computation in a large region of parameter space, for all rectangular constraints and for narrow enough $u-$function constraints, $K < K^*$.

We prove this result, under two natural assumptions, using the first and second moment methods. We further use the second moment method to conjecture that solutions of the symmetric binary perceptrons are organized in a so-called frozen-1RSB structure, without using the replica method. We then use the replica method to estimate the capacity threshold for the UBP case when the $u-$function is wide $K > K^*$. We conclude that full-step-replica-symmetry breaking would have to be evaluated in order to obtain the exact capacity in this case.
Personal contributions: I have analyzed the RSB Ansätze of the replica free entropy to study the Gardner capacity and the configuration space geometry. I also contributed to the first and second moments proofs.

3. *'Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning'*. Abbara, Aubin, Krzakala, and Zdeborová (2020)
Presented in Chap. 7.

Summary: Statistical learning theory provides bounds of the generalization gap, using in particular the Vapnik-Chervonenkis dimension and the Rademacher complexity. An alternative approach, mainly studied in the statistical physics literature, is the study of generalization in simple synthetic-data models. Here we discuss the connections between these approaches and focus on the link between the Rademacher complexity in statistical learning and the theories of generalization for *typical-case* synthetic models from statistical physics, involving quantities known as *Gardner capacity* and *ground state energy*. We show that in these models the Rademacher complexity is closely related to the ground state energy computed by replica theories. Using this connection, one may reinterpret many results of the literature as rigorous Rademacher bounds in a variety of models in the high-dimensional statistics limit. Somewhat surprisingly, we also show that statistical learning theory provides predictions for the behavior of the ground-state energies in some full replica symmetry breaking models.
Personal contributions: I derived and evaluated the ground state energies and draw the connection with the Rademacher complexity.

4. *'Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization'*. Aubin, Krzakala, Lu, and Zdeborová (2020a)
Presented in Chap. 8.

Summary: We consider a commonly studied supervised classification of a synthetic dataset whose labels are generated by feeding a one-layer neural network with random iid inputs. We study the generalization performances of standard classifiers in the high-dimensional regime where $\alpha = n/d$ is kept finite in the limit of a high dimension d and number of samples n. Our contribution is three-fold: First, we prove a formula for the generalization error achieved by ℓ_2-regularized classifiers that minimize a convex loss. This formula was first obtained by the heuristic replica method of statistical physics. Secondly, focusing on commonly used loss functions and optimizing the ℓ_2 regularization strength, we observe that while ridge regression performance is poor, logistic and hinge regression are surprisingly able to approach the Bayes-optimal generalization error extremely closely. As $\alpha \to \infty$ they lead to Bayes-optimal rates, a fact that does not follow from predictions of margin-based generalization error bounds. Third, we design an optimal loss and regularizer that provably leads to Bayes-optimal generalization error.

Personal contributions: I conducted the theoretical analysis and the numerical evaluations. I also contributed to the proofs based on the Gordon min-max theorem.

In Part II. B, we present a line of research conducted in parallel that investigates different kinds of prior informations for estimation problems, such as the spiked matrix model presented in Chap. 9 or compressed sensing and phase retrieval detailed in Chap. 10. Specifically, we compare the statistical-to-algorithmic gaps for sparse separable priors and structured deep generative priors with random weights.

5. *'The spiked matrix model with generative priors'*. Aubin, Loureiro, Maillard, Krzakala, and Zdeborová (2019b)
   Presented in Chap. 9.

   Summary: Using a low-dimensional parametrization of signals is a generic and powerful way to enhance performance in signal processing and statistical inference. A very popular and widely explored type of dimensionality reduction is sparsity; another type is generative modeling of signal distributions. Generative models based on neural networks, such as GANs or variational auto-encoders, are particularly performant and are gaining on applicability. In this paper we study spiked matrix models, where a low-rank matrix is observed through a noisy channel. This problem with sparse structure of the spikes has attracted broad attention in the past literature. Here, we replace the sparsity assumption by generative modelling, and investigate the consequences on statistical and algorithmic properties. We analyze the Bayes-optimal performance under specific generative models for the spike. In contrast with the sparsity assumption, we do not observe regions of parameters where statistical performance is superior to the best known algorithmic performance. We show that in the analyzed cases the approximate message passing algorithm is able to reach optimal performance. We also design enhanced spectral algorithms and analyze their performance and thresholds using random matrix theory, which was performed by collaborators, showing their superiority to the classical principal component analysis. We complement our theoretical results by illustrating the performance of the spectral algorithms when

the spikes come from real datasets.
Personal contributions: I co-developed the plug-in framework to combine the replica free entropies and the AMP algorithms of sub-models, in close collaboration with B. Loureiro. I also co-developed the LAMP spectral method and conducted the numerical implementations.

6. *'Exact asymptotics for phase retrieval and compressed sensing with random generative priors'*. Aubin, Loureiro, Baker, Krzakala, and Zdeborová (2020b)
   Presented in Chap. 10.

   Summary: We consider the problem of compressed sensing and of (real-valued) phase retrieval with random measurement matrix. We derive sharp asymptotics for the information-theoretically optimal performance and for the best known polynomial algorithm for an ensemble of generative priors consisting of fully connected deep neural networks with random weight matrices and arbitrary activations. We compare the performance to sparse separable priors and conclude that in all cases analyzed generative priors have a smaller statistical-to-algorithmic gap than sparse priors, giving theoretical support to previous experimental observations that generative priors might be advantageous in terms of algorithmic performance. In particular, while sparsity does not allow to perform compressive phase retrieval efficiently close to its information-theoretic limit, it is found that under the random generative prior compressed phase retrieval becomes tractable.
   Personal contributions: I conducted the numerical analysis and evaluation of the phase transitions.

Finally, I have contributed to an additional work, not covered in this dissertation, that introduces a modular python implementation of compositional inference on tree-structured inference models. However, in line with previous works, in Sec. 10.3 we present simple applications of the algorithm to estimation problems with generative priors trained on real datasets.

7. *'TRAMP: Compositional Inference with TRee Approximate Message Passing'*. Baker, Aubin, Krzakala, and Zdeborová (2020)

   Summary: We introduce tramp, standing for *TRee Approximate Message Passing*, a python package for compositional inference in high-dimensional tree-structured models. The package provides an unifying framework to study several approximate message passing algorithms previously derived for a variety of machine learning tasks such as generalized linear models, inference in multi-layer networks, matrix factorization, and reconstruction using non-separable penalties. For some models, the asymptotic performance of the algorithm can be theoretically predicted by the state evolution, and the measurements entropy estimated by the free entropy formalism. The implementation is modular by design: each module, which implements a

factor, can be composed at will with other modules to solve complex inference tasks. The user only needs to declare the factor graph of the model: the inference algorithm, state evolution and entropy estimation are fully automated. The source code is publicly available at *https://github.com/sphinxteam/tramp* and the documentation is accessible at *https://sphinxteam.github.io/tramp.docs*.
Personal contributions: I contributed to implement some parts of the source code and developed entirely the online documentation. I mainly investigated the reconstruction performances of the package for generative priors with weights trained with different GAN and VAE architectures.

AVANT-PROPOS

À une époque où l'utilisation des données a atteint un niveau sans précédent, l'accès à ce grand nombre de données a précipité leur utilisation intensive afin d'entrainer des modèles d'apprentissage automatique. Les algorithmes correspondants visent essentiellement à détecter et à utiliser des informations structurées au sein d'ensembles de données extrêmement volumineux. Plus précisément, après de nombreux rebondissements, les modèles d'apprentissage profond, basés sur des architectures de réseaux de neurones artificiels, ont apporté d'importants progrès numériques dans cette direction et sont désormais omniprésents. Leurs performances dépassant de loin celles des autres modèles existants, ils sont devenus à partir du milieu des années 2000, et en quelques années à peine, indispensables dans de nombreuses applications industrielles telles que la classification d'images, la reconnaissance vocale, l'analyse de texte (LeCun et al., 2015) ou la prédiction de séries temporelles, la détection d'objets et la reconnaissance faciale, le traitement du langage naturel, le diagnostic médical, la robotique, etc. Cependant, la compréhension de la plupart des algorithmes, basés sur la descente de gradient d'une fonction de coût, utilisés en pratique pour entraîner des réseaux surdimensionnés et complexes, qui contiennent jusqu'à des millions de paramètres, reste essentiellement empirique et difficile à analyser en théorie. Le principal problème qui se pose avec les architectures de réseaux de neurones profonds réside essentiellement dans la succession de nombreuses couches constituées d'opérations non-linéaires qui rendent l'espace d'optimisation très complexe et de haute dimension. L'analyse et la visualisation de cette vaste collection de paramètres constituent la principale difficulté mathématique dans la plupart des modèles et algorithmes d'apprentissage automatique de pointe. Ce manque de compréhension théorique soulève de nombreuses questions sur leur efficacité et les risques potentiels dans de nombreux domaines d'application. En conséquence, établir des fondements théoriques sur des modèles simples et fournir des prescriptions numériques sur lesquelles fonder et expliquer les observations empiriques sont devenus l'un des défis fondamentaux de la communauté scientifique.

Dans ce manuscrit, nous étudions ces questions d'envergure, soulevées par la récente utilisation intensive de l'apprentissage automatique, à travers le prisme de la physique statistique des systèmes désordonnés. Transverse et singulière, cette approche des problèmes d'informatique par la physique a une longue et riche histoire (Engel et al., 1993; Mézard et al., 2009; Grassberger et al., 2012; Zdeborová et al., 2016b; Advani et al., 2017), que nous revisitons dans le régime de haute dimension, en nous concentrant essentiellement sur des considérations algorithmiques modernes confortées par des preuves rigoureuses. Spécifiquement, dans le contexte de réseaux de neurones surdimensionnés, pour lesquels le nombre de paramètres explose, les solutions analytiques exactes sont la plupart du temps inconnues et les simulations numériques, quant à elles, très coûteuses. La plupart des techniques issues de la physique statistique ont été précisément conçues pour déduire le comportement macroscopique d'une aussi grande collection de particules à partir de la description microscopique de leurs interactions élémentaires. Ainsi, elles forment un ensemble d'approximations de choix, appelées méthodes à champ moyen, qui sont suffisamment simples pour être calculables et suffisamment riches pour décrire et reproduire les caractéristiques intéressantes du système. De plus, dans cette limite thermodynamique, les physiciens ont constaté que les comportements macroscopiques sont typiquement décrits correctement par seulement quelques paramètres d'ordre. Appliquées à la théorie de l'apprentissage automatique, qui manque cruellement de telles techniques, nous pensons que les connaissances et techniques de la physique statistique peuvent contribuer à identifier cet ensemble d'observables pertinentes qui contrôlent les propriétés à grande échelle du système et fournissent un cadre puissant pour analyser ces réseaux de neurones artificiels complexes. Malheureusement, même si elles sont très puissantes et supposées conduire à des résultat corrects dans de nombreuses situations, ces techniques ont été utilisées historiquement sans fondement rigoureux. Par conséquent, ce travail fait partie de la dynamique actuelle de la communauté de physique-mathématique à démonter d'anciens résultats obtenus de manière heuristique dans les années 90. De plus, alors que ces analyses statistiques antérieures ne discutaient pas des considérations algorithmiques, nous revisitons cette approche en nous concentrant principalement sur ces potentielles transitions de phase algorithmiques.

Au coeur de ce travail, nous capitalisons fortement sur un raisonnement probabiliste Bayésien, qui contraste avec l'approche d'optimisation traditionnelle. De plus, nous utilisons intensément la connexion profonde entre la méthode des répliques et les algorithmes de passage de messages pour obtenir les diagrammes de phase de modèles théoriques simplifiés, qui révèlent néanmoins des caractéristiques intéressantes. En revisitant le paradigme *enseignant-élève*, qui permet de créer des tâches synthétiques et analysables théoriquement, nous concentrons notre attention sur la mise en évidence des écarts potentiels entre les seuils statistiques et algorithmiques. Nous illustrons l'efficacité de ces méthodes à champ moyen sur divers modèles d'apprentissage automatique qui restent mal compris. Nous nous intéressons essentiellement à des tâches synthétiques avec des données générées dans le paradigme enseignant-élève, et nous contribuons à leur compréhension en décrivant leurs riches diagrammes de phases. Tout d'abord, nous commençons par présenter l'analyse Bayes-optimale dans des machines à comité qui révèle l'existence de grandes lacunes algorithmiques par rapports aux seuils statistiques. Ensuite, dans une analyse pessimiste du pire scénario possible, nous mettons en évidence un lien fort entre la complexité de Rademacher, qui fournit une borne supérieure de l'erreur de généralisation et est liée à la théorie de l'apprentissage statistique, et la capacité de stockage et l'énergie de l'état fondamental abordés dans la littérature de physique statistique. Cela nous permet en particulier de calculer explicitement la complexité de Rademacher dans le cas des perceptrons. Nous complétons enfin le tableau en analysant la minimisation du risque empirique dans le cas des modèles linéaires généralisés, qui est intensivement utilisée en pratique, et nous la comparons aux précédentes analyses Bayes-optimales et du pire scénario. Dans une autre direction de recherche, nous définissons une procédure générale pour combiner des modèles élémentaires déjà analysés, afin de construire des architectures plus complexes et structurées. De cette manière, nous développons un cadre qui surmonte en particulier l'hypothèse standard de séparabilité et permet d'analyser des modèles d'estimation, tels que la factorisation matricielle avec un faible rang, la récupération de phase ou la détection compressée, avec des informations à priori fournies par des réseaux génératifs profonds avec des poids aléatoires.

ORGANISATION DU MANUSCRIT

Comme mon travail de doctorat se situe au croisement de l'apprentissage automatique et de la physique statistique des systèmes désordonnés, je profite de l'occasion pour présenter pédagogiquement dans la Partie I les concepts théoriques de base, mais essentiels pour suivre le reste du manuscrit. Dans le Chap. 1, je propose une vue d'ensemble du domaine de l'apprentissage automatique en mettant l'accent sur son histoire tortueuse, ses concepts de base et ses défis actuels. Le Chap. 2 couvre les outils de base de la physique statistique qui sont pertinents pour comprendre l'approche originale que nous employons pour résoudre les problèmes d'apprentissage automatique. Ces deux premiers chapitres sont consacrés aux lecteurs qui ne connaissent pas l'un des deux domaines et peuvent être donc ignorés par les experts. Dans le Chap. 3, nous proposons une sélection de références historiques importantes pour comprendre comment ces deux domaines sont liés depuis plus de trente ans. C'est aussi l'occasion de passer en revue une sélection des axes de recherche actuels auxquels s'intéresse la communauté de physique statistique des réseaux de neurones artificiels. Enfin, nous introduisons le cadre probabiliste Bayésien et son lien crucial avec la physique statistique. Ceci constitue la pierre angulaire de notre approche qui nous permet d'analyser simultanément les problèmes d'inférence statistique et de satisfaction de contraintes aléatoires. Dans le Chap. 4, nous proposons une revue méthodologique de certaines méthodes fondamentales d'inférence à champ moyen, motivées à l'origine dans la littérature des verres de spin (Mézard et al., 1987), qui sont principalement utilisées dans la seconde partie de la thèse. Plus précisément, nous rappelons en détails la méthode des répliques et les algorithmes de passage de messages que nous présentons et illustrons sur la classe des *modèles linéaires généralisés*, qui sert d'exemple de base tout au long de ce manuscrit. De plus, en mettant en évidence leurs complémentarités, nous tentons de clarifier comment les méthodes sont étroitement liées et permettent de révéler de riches transitions de phase statistiques et algorithmiques.

La partie II de ce manuscrit est consacrée à couvrir les travaux auxquels j'ai contribué en tant que doctorant d'Octobre 2017 à Décembre 2020, à l'*Institut de Physique Théorique du CEA-Saclay* sous la direction de Lenka Zdeborová et Florent Krzakala. Le contenu des articles a déjà été publié dans une série d'ouvrages qui peuvent être trouvés en ligne dans leur format original. Ils ont été révisés afin d'uniformiser les notations de ce manuscrit. En particulier, dans un souci de clarté, certaines des longues preuves et calculs auxquels je n'ai pas directement contribués ne sont pas rapportés dans ce manuscrit et peuvent être trouvés dans les publications originales.

CONTRIBUTIONS

La partie II, qui rassemble mes principales contributions, est séparée en deux sous-parties correspondant à des axes de recherche menés en parallèle. Dans la partie II. A, nous discutons de l'analyse complémentaire des scénarios *Bayes-optimal*, du *pire cas*, et de la *minimisation du risque empirique* dans le cadre de réseaux de neurones simples avec des distributions à priori séparables. Dans le Chap. 5, nous présentons tout d'abord l'approche *Bayes-optimale* sur les machines à comité, qui fournit une analyse des bornes inférieures d'un point de vue de la théorie de l'information. Ensuite, dans le Chap. 6 nous décrivons l'analyse du problème de la capacité de stockage et des énergies de l'état fondamental associées, dans le cadre générique des problèmes de satisfaction de contraintes aléatoires. Dans le Chap. 7, nous montrons que ces quantités s'avèrent être étroitement liées à la complexité de Rademacher, connue pour être une borne supérieure de l'erreur de généralisation. Dans le Chap. 8, nous étudions le cas le plus utilisé en pratique avec l'analyse de la minimisation du risque empirique qui est souvent réalisée grâce à des algorithmes de descente de gradient.

1. *'The committee machine: Computational to statistical gaps in learning a two-layers neural network'*. Aubin, Maillard, Barbier, Krzakala, Macris, and Zdeborová (2018)
   Presenté dans le Chap. 5.

   Résumé: Des outils heuristiques issus de la physique statistique ont été utilisés dans le passé pour localiser les transitions de phase et calculer les erreurs d'apprentissage et de généralisation optimales de réseaux de neurones multicouches dans le scénario enseignant-élève. Dans cette contribution nous fournissons, sous une hypothèse technique, une justification rigoureuse de ces approches pour un modèle de réseau de neurones à deux couches, appelé machine à comité. Nous introduisons également une version de l'algorithme de passage de messages approximatifs (AMP) pour la machine à comité, qui permet d'effectuer un apprentissage optimal en temps polynomial pour une

grande région de paramètres. Nous constatons cependant qu'il existe des régimes dans lesquels une faible erreur de généralisation est théoriquement réalisable alors que l'algorithme AMP ne parvient pas à l'atteindre; suggérant fortement qu'aucun algorithme efficace n'existe dans cette région, ce qui met en évidence un grand écart entre seuils statistique et algorithmique.
Contributions personnelles : J'ai développé et implémenté l'algorithme AMP et son évolution d'état afin de représenter et analyser les diagrammes de phase correspondants.

2. *'Storage capacity in symmetric binary perceptrons'*. Aubin, Perkins, and Zdeborová (2019b)
   Presenté dans le Chap. 6.

   Résumé: Nous étudions le problème du calcul de la capacité de stockage du perceptron binaire pour deux variantes du problème, dans lesquelles la contrainte correspondante est symétrique. Nous appelons ces variantes le perceptron binaire rectangulaire (RPB) et le perceptron binaire u (UBP). Nous montrons que, contrairement au perceptron binaire habituel avec une fonction marche, la capacité de stockage dans ces alternatives symétriques est donnée par le calcul recuit dans une grande région d'espace de paramètres, i. e. pour toutes les contraintes rectangulaires et pour des contraintes de fonction u assez étroites pour $K < K^*$. Nous prouvons ce résultat, sous deux hypothèses naturelles, en utilisant la méthode des premier et second moments. Nous utilisons en outre la méthode du second moment pour conjecturer que les solutions des perceptrons binaires symétriques sont organisées dans une configuration gelée dite 1RSB, et ce sans utiliser la méthode des répliques. Nous utilisons ensuite cette méthode des répliques pour estimer la capacité de stockage dans le cas UBP lorsque la fonction $u-$ est large avec $K > K^*$. Finalement, nous concluons que dans ce cas la rupture totale de la symétrie des répliques devrait être évaluée pour obtenir la capacité exacte.
   Contributions personnelles : J'ai analysé l'entropie des modèles sous différents Ansätze pour étudier la capacité de stockage de Gardner et la géométrie de l'espace de configuration. J'ai également contribué aux preuves en utilisant la méthode des premier et second moments.

3. *'Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning'*. Abbara, Aubin, Krzakala, and Zdeborová (2020)
   Presenté dans le Chap. 7.

   Résumé: La théorie de l'apprentissage statistique fournit des bornes sur l'erreur de généralisation en utilisant en particulier la dimension de Vapnik-Chervonenkis et la complexité de Rademacher. Une approche al-

ternative, principalement étudiée dans la littérature de physique statistique, est l'étude de la généralisation dans des modèles de données synthétiques simples. Nous discutons donc des liens entre ces approches et nous nous concentrons sur le lien entre la complexité de Rademacher en apprentissage statistique et la théorie de la généralisation pour des modèles synthétiques dans le *cas typique* étudié en physique statistique. Cela implique notamment des quantités connues sous le nom de *capacité de stockage de Gardner* et *de l'énergie de l'état fondamental* du modèle. Nous montrons que dans ces modèles, la complexité de Rademacher est étroitement liée à l'énergie de l'état fondamental calculée par la méthode des répliques. En utilisant cette connexion, on peut dès lors réinterpréter de nombreux résultats de la littérature comme des bornes de Rademacher rigoureuses dans une variété de modèles et dans le régime de haute dimension. De manière assez surprenante, nous montrons également que la théorie de l'apprentissage statistique fournit des prédictions sur le comportement des énergies de l'état fondamental dans certains modèles présentant une rupture totale de la symétrie des répliques.
Contributions personnelles : J'ai calculé et évalué les énergies de l'état fondamental et établi leur lien avec la complexité de Rademacher.

4. *'Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization'.* Aubin, Krzakala, Lu, and Zdeborová (2020a)
   Presenté dans le Chap. 8.

   Résumé: Nous considérons une tâche de classification supervisée d'un ensemble de données synthétiques dont les étiquettes sont générées en alimentant un réseau de neurone à une couche avec des entrées iid aléatoires. Nous étudions les performances de généralisation de classificateurs standards dans le régime de haute dimension dans lequel $\alpha = n/d$ est maintenu fini dans la limite d'une dimension d et d'un nombre d'échantillons n infinis. Notre contribution est triple : Premièrement, nous prouvons une formule donnant l'erreur de généralisation obtenue par des classificateurs qui minimisent une fonction de coût convexe avec un terme de régularisation ℓ_2. Cette formule a été obtenue initialement et de façon heuristique par la méthode des répliques de la physique statistique. Deuxièmement, en nous concentrant sur des fonctions de coût couramment utilisées et en optimisant l'amplitude de la régularisation ℓ_2, nous observons que même si les performances de la régression Ridge sont médiocres, en outre les régressions logistique et Hinge sont étonnamment capables d'approcher de très près l'erreur de généralisation Bayes-optimale. Dans le régime où $\alpha \to \infty$, ces régressions conduisent à des taux de généralisation Bayes-optimaux, ce qui, cependant, ne découle pas des prédictions asymptotiques de l'erreur de généralisation basées sur les marges. Troisièmement, nous concevons une fonction de coût et un terme de régularisation optimaux

qui conduisent de manière asymptotique et rigoureuse à l'erreur de généralisation Bayes-optimale.
Contributions personnelles : J'ai réalisé l'analyse théorique et les évaluations numériques. J'ai également contribué aux preuves basées sur le théorème de Gordon.

Dans la partie II. B , nous présentons une ligne de recherche menée en parallèle qui étudie plusieurs types d'informations à priori pour différents problèmes d'estimation, comme le modèle de factorisation de matrice présenté dans le Chap. 9 ou la détection compressée et la récupération de phase, détaillées dans le Chap. 10. Plus précisément, nous comparons les écarts statistiques et algorithmiques entre d'une part des informations à priori séparables et parcimonieuses, et d'autre part des informations à priori produites par des modèles génératifs profonds et structurés avec des poids aléatoires.

5. *'The spiked matrix model with generative priors'.* Aubin, Loureiro, Maillard, Krzakala, and Zdeborová (2019a)
Presenté dans le Chap. 9.

Résumé: L'utilisation d'une paramétrisation de faible dimension des signaux est un moyen générique et puissant pour améliorer les performances de traitement du signal et d'inférence statistique. Un type de réduction de dimension très populaire et largement exploré est la parcimonie; une autre méthode plus récente est la modélisation générative de la distribution de signaux. Les modèles génératifs basés sur des réseaux de neurones, tels que les GAN ou les auto-encodeurs variationnels (VAE), sont particulièrement performants et gagnent notamment en applicabilité. Dans cette contribution, nous étudions les modèles matriciels à pics, où une matrice de faible rang est observée à travers un canal potentiellement bruité. L'étude de ce problème avec une structure parcimonieuse a attiré une large attention dans la littérature. Ici, nous remplaçons l'hypothèse de parcimonie par un modèle génératif, et nous étudions les conséquences sur les propriétés statistiques et algorithmiques. Nous analysons les performances Bayes-optimales sous l'hypothèse spécifique de modèles génératifs pour les pics. En contraste avec l'hypothèse de parcimonie, nous n'observons pas de régions de paramètres où les performances statistiques sont supérieures aux performances algorithmiques du meilleur algorithme connu. Nous montrons que dans les cas analysés, l'algorithme de passage de messages est capable d'atteindre ces performances optimales. Nous concevons également de nouveaux algorithmes spectraux et analysons leurs performances et leurs seuils statistiques en utilisant la théorie des matrices aléatoires, qui a été réalisée par mes collaborateurs. Nous montrons leur supériorité par rapport à l'analyse classique de la composante principale (PCA). Nous complétons nos résultats théoriques avec l'illustration des performances des algorithmes spectraux dans le cas où les pics sont générés par des données réelles.
Contributions personnelles : En étroite collaboration avec B. Loureiro,

j'ai co-développé le cadre théorique pour combiner les entropies des répliques et les algorithmes de passage de messages AMP de sous-modèles. J'ai également co-développé la méthode spectrale LAMP et réalisé les implémentations numériques.

6. *'Exact asymptotics for phase retrieval and compressed sensing with random generative priors'*. Aubin, Loureiro, Baker, Krzakala, and Zdeborová (2020b)
   Presenté dans le Chap. 10.

   Résumé: Nous considérons le problème de détection compressée et de récupération de phase (à valeurs réelles) pour une matrice de mesure aléatoire. Nous calculons précisément le comportement asymptotique des performances optimales, au sens de la théorie de l'information, et celles du meilleur algorithme polynomial connu, dans le cas d'informations à priori génératives provenant de réseaux de neurones profonds entièrement connectés par des matrices de poids aléatoires et des activations arbitraires. Nous comparons ces performances à celles obtenues pour des informations à priori séparables parcimonieuses et nous concluons que dans tous les cas analysés les informations à priori génératives présentent un écart statistique-algorithmique bien plus petit que pour des à priori parcimonieux, ce qui confirme théoriquement les observations expérimentales antérieures selon lesquelles les à priori génératifs pourraient être bien plus avantageux en terme de performances algorithmiques. En particulier, alors que la parcimonie ne permet pas d'effectuer efficacement une récupération de phase compressive proche de sa limite théorique, nous constatons qu'en utilisant un à priori génératif aléatoire, la récupération de phase devient possible.
   Contributions personnelles : J'ai réalisé l'analyse numérique et l'évaluation des transitions de phase.

Enfin, j'ai contribué à un travail supplémentaire, qui n'est pas présenté dans ce manuscrit, qui introduit une implémentation modulaire en python de l'inférence compositionnelle de modèles graphiques, structurés en arbres. Cependant, dans la lignée des travaux précédents, nous présentons dans la Sec. 10.3 des applications simples de l'algorithme à des problèmes d'estimation avec des à priori génératifs entraînés sur des données réelles.

7. *'TRAMP: Compositional Inference with TRee Approximate Message Passing'*. Baker, Aubin, Krzakala, and Zdeborová (2020)

   Résumé: Nous introduisons **tramp**, pour *TRee Approximate Message Passing*, un code python pour l'inférence compositionnelle dans des modèles structurés en arbre et en grande dimension. Le logiciel unifie et fournit un cadre pour étudier plusieurs algorithmes de passage de messages approximatifs et qui s'appliquent à une variété de tâches d'apprentissage automatique, telles que les modèles linéaires généralisés, l'inférence dans les réseaux multicouches, la factorisation

matricielle et la reconstruction à l'aide de pénalités non séparables. Pour certains modèles, la performance asymptotique de l'algorithme peut être théoriquement prédite par l'évolution d'état et un formalisme d'entropies libres. L'implémentation est modulaire par construction: chaque module, qui implémente un facteur du modèle graphique, peut être composé à volonté avec d'autres modules pour résoudre des tâches d'inférence complexes. L'utilisateur n'a qu'à déclarer le modèle graphique: l'algorithme d'inférence, l'évolution d'état et l'estimation de l'entropie sont entièrement automatisés. Le code source est accessible au public à *https://github.com/sphinxteam/tramp* et la documentation est accessible à *https://sphinxteam.github.io/tramp.docs*.
Contributions personnelles: J'ai contribué à implémenter certaines parties du code source et développé entièrement la documentation en ligne. J'ai principalement étudié les performances de reconstruction du package pour des à priori génératifs avec des poids entraîné avec différentes architectures GAN et VAE sur des données réelles.

LIST OF PUBLICATIONS

Abbara, Alia, Benjamin Aubin, Florent Krzakala, and Lenka Zdeborová (2020). 'Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning.' In: *Mathematical and Scientific Machine Learning*. PMLR, pp. 27–54. eprint: *https://arxiv.org/abs/1912.02729*.

Aubin, Benjamin, Florent Krzakala, Yue M Lu, and Lenka Zdeborová (2020a). 'Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization.' In: *Advances in Neural Information Processing Systems 33*. eprint: *https://arxiv.org/abs/2006.06560*.

Aubin, Benjamin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová (2020b). 'Exact asymptotics for phase retrieval and compressed sensing with random generative priors.' In: *Mathematical and Scientific Machine Learning*. PMLR, pp. 55–73. eprint: *https://arxiv.org/abs/1912.02008*.

Aubin, Benjamin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová (2019a). 'The spiked matrix model with generative priors.' In: *Advances in Neural Information Processing Systems 32*, pp. 8366–8377. eprint: *https://arxiv.org/abs/1905.12385*.

Aubin, Benjamin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová (2018). 'The committee machine: Computational to statistical gaps in learning a two-layers neural network.' In: *Advances in Neural Information Processing Systems 31*, pp. 3223–3234. eprint: *https://arxiv.org/abs/1806.05451*.

Aubin, Benjamin, Will Perkins, and Lenka Zdeborová (2019b). 'Storage capacity in symmetric binary perceptrons.' In: *Journal of Physics A: Mathematical and Theoretical* 52.29, p. 294003. eprint: *https://arxiv.org/abs/1901.00314*.

Baker, Antoine, Benjamin Aubin, Florent Krzakala, and Lenka Zdeborová (2020). 'TRAMP: Compositional Inference with TRee Approximate Message Passing.' In: *arXiv preprint arXiv:2004.01571. Submitted to Journal of Machine Learning Research*. eprint: *https://arxiv.org/abs/2004.01571*.

Part I

AN INTRODUCTION AT THE CROSSROADS OF MACHINE LEARNING AND STATISTICAL PHYSICS

1

A SHORT INTRODUCTION TO MACHINE LEARNING

Current Machine Learning (ML) techniques vastly rely on Deep Neural Networks (DNN) and pioneered unprecedented advances in various fields of Artificial Intelligence (AI). Despite how recently it gained popularity, Deep Learning (DL) in fact has a long story starting in the 40's. The field of AI was known under different names along its history depending on the most influential research directions and perspectives. Even though the recent progresses might seem very promising, many theoretical challenges on the theoretical foundations of the current DNN-based methods remain unanswered. In this first chapter, we start by describing a few breakthroughs in AI and ML in Sec. 1.1 to provide some context for the recent developments of the field. In Sec. 1.2, we provide a short and comprehensive review of modern machine learning basics, so that the unfamiliar reader may correctly follow the rest of the manuscript. The aim is not to provide a fully thorough description, but a qualitative introduction; the interested reader may find more furnished details in reference books such as (Murphy, 2012; Mohri et al., 2012; Shalev-Shwartz et al., 2014; Goodfellow et al., 2016). Finally in Sec. 1.3, in order to fully grasp the scope and the motivations of this work, we take advantage of the opportunity to review the current challenges and fundamental questions which remain unanswered and that statistical physics may contribute to solve.

1.1 A BRIEF HISTORICAL REVIEW OF ARTIFICIAL INTELLIGENCE

We present a short selection of some key steps in the development of AI, from the early Artificial Neural Networks (ANN) of the 1950's to the modern DNN used since 2010's. For a more detailed historical overview please refer to (Ganascia, 1993; Hutchins, 2001; Schmidhuber, 2015; Lazard et al., 2016; Goodfellow et al., 2016; Sejnowski, 2018; Skansi, 2018).

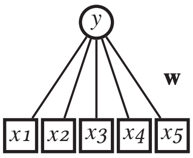
1.1.1 THE FIRST ARTIFICIAL INTELLIGENT MACHINES: 1940-1980

Symbolists vs Connectionists AI is a wide field whose goal is to design intelligent programs. Inside this field, two main intellectual currents emerged. On one hand *rule* or *knowledge*-based *symbolists* whose pioneers are for instance J. McCarthy, M. L Minsky or J. Von Neumann, and in the other hand *learning-based connectionists*. While symbolists aim to simulate intelligence through a succession of predefined rules, connectionists investigate instead the possibility that a computer may learn a solution directly from examples and handle complex edge-cases by itself. In other words, *symbolism* refers to feature *engineering* and *connectionism* to feature *learning*. Jumping ahead to modern AI, *machine learning* refers to a connectionist kind of AI. Notice that in early stages of AI, symbolism and connectionism were two different approaches that many researchers tried to bring together, see (Dreyfus et al., 1984), while nowadays these two approaches have become quite orthogonal.

Modern AI started with the emergence of computers and the *Turing test* invented in 1950 by A. Turing (Turing, 2009) to determine whether a computer may "think" like a human. During this same period, the first ANN was developed, initially designed to model the biological learning of the human brain.

The beginning of neural networks: 1940-1960 A significant advance in ANN came with the work of W. McCulloch and W. Pitts (1943) (McCulloch et al., 1943) who created the first simplified mathematical model of the human brain, an interconnected circuit of binary units, called formal neurons, and demonstrated that it was equivalent to a universal Turing machine. A few years later, D. Hebb (1949) (Hebb, 1962) reinforced the concept of neurons and pointed out that neural pathways are strengthened each time they are used, introducing for the first time the concept of *plasticity*. Later on in 1955, A. Samuel invented a computer program able to play checkers, combining connectionist and symbolist approaches with a tree search on weights learned with *temporal-difference* adjusted according to the number of errors. Such early experiments, together with the first machine translation results, lead to the Dartmouth conference in 1956 where important figures of the field such as J.McCarthy, M.Minsky and C.Shannon declared the birth of AI and provided a boost to both AI and ANN. Finally, the first battle horse of the connectionist empire was introduced in the 1960's by psychologist F. Rosenblatt (1958) (Rosenblatt, 1958) : the perceptron, an improved version of the McCulloch and Pitts units. Though very simple, such machines are the basic units of what we call today *deep-learning*. The main innovation was to try to simulate the behaviour of biological neurons. In this perspective Rosenblatt added continuously adjustable valued connections, called today weights $\mathbf{w}$, to enable plasticity of the unit. The weights are then trained in a supervised manner minimizing the number of mistakes with respect to the desired output y for a particular input pattern $\mathbf{x}$. More significantly, he gave the first convergence proof of the perceptron algorithm, stating that after training the perceptron would perfectly memorize the training set. Later, B. Widrow and M. Hoff (Widrow et al., 1960) developed the first ANN to be applied in a real-world problem : (M)ADALINE for (Multiple) ADAptive LINear Elements, for echo suppression on phone lines.

A neural-network is a simple supervised model with learnable parameters $\mathbf{w}$ *in which the output* y *is a linear/non-linear transformation of an input vector* $\mathbf{x}$: $y = \varphi\left(\frac{1}{\sqrt{d}}\mathbf{w}\cdot\mathbf{x}\right)$.



The first AI winter: the quiet decade 1965-1976 The *quiet decade* refers to W. J Hutchins (Hutchins, 2001) formulation about the fact that discoveries and progresses in machine translation stalled. At the end of the 1960's there was no more hope in machine learning translation and research fundings in this direction were deeply cut. In 1965 AI was soon compared to *alchemy* (Dreyfus, 1965) because the early successes held only on very simple tasks and led only to disenchantment in complex tasks.

Thus in the 1970's the future of the connectionist AI turned dark. The godfathers of AI themselves, M. Minksy and S. Papert (Minsky et al., 1969), showed that perceptrons, which are stuck in the realm of linear models, are limited to very simple tasks and moreover are hard to train. They proposed a harsh critique of perceptrons by proving that they could only be trained to solve linear separable problems and fail to learn non-linearly separable rules such as the XOR function, such that $y = 1$ for $\mathbf{x} \in \{(0,1),(1,0)\}$ and $y = 0$ for $\mathbf{x} \in \{(0,0),(1,1)\}$. In addition they stated a number of fundamental problems with the neural network research program and they argued that despite being an interesting subject to study, perceptrons were a sterile direction of research. This was the first big hit to connectionism. This lead a few years later to the Lighthill report in 1973 (Lighthill, 1973) which came to the conclusion that the early promises of AI, especially in machine translation, were overstated and fundings were accordingly drastically reduced. Study of neural networks thus fell into a quick decline in the late 1960's due to Minsky and Papert's campaign and AI research fundings were turned towards other AI projects such as Bobrow's STUDENT program (Bobrow, 1964), Evan's Analogy program (Evans, 1964) and the Quillian's semantic memory program Teachable Language Comprehender (Quillian, 1969). Note that in spite of their harsh criticism, M. Minksy and S. Papert continued contributing to neural network research. Yet these events put an end to the first phase of connectionist research, see (Hecht-Nielsen, 1989).

1.1.2 FROM EXPERT SYSTEMS TO MACHINE LEARNING: 1980-2007

During the 1970's, most of AI research focused on the symbolist approach. But ANN oriented research continued with a series of works (Kohonen et al., 1977; Grossberg, 1976) and a deeply philosophical study by Anderson (Anderson, 1972) on the nature of complex systems[1]. Again unfulfilled claims led to a slowdown in funding in AI and ANN research until early 1980's.

The realm of expert systems In the early 1980's, large conferences instigated a rapid increase in interest from industries and governments, showing a renewed interest and hope in AI and expert systems. In particular, the focus was shifted towards commercial products with applications in financial prediction, geological exploration, medical diagnosis or microelectronic circuit

[1] This work was largely influential and is the cornerstone of modern statistical physics.

design. Instead of being based on neural networks, this AI era was the climax of the symbolist AI approach, during which it was believed that the best approach to perform AI was top-down with handcrafted knowledge-based systems with huge expertise. But as the hype increased, the field started fearing another winter and a corresponding dry up in funding if AI was to disappoint expectancies.

The second AI winter This fear became true. In the following years, the claims of what AI was capable of had to face reality. *Expert systems* were at the heart of the AI revolution and faced many issues. In particular J. McCarthy strongly criticized them as lacking common sense and knowledge about their limitations. Indeed predictions in medicine based on these systems would have killed many patients and many tasks such as vision or speech recognition were still too complicated for engineers to design handcrafted rules that contain all potential edge cases. To conclude, the success of expert systems at that time was very limited and failed to reach the broader goal at which these initial AI successes seemed to lead. Therefore mid-1990's, again the activity and publications in AI research largely dropped and conferences did not attract that much anymore, leading naturally to another a decrease of fundings.

Machine learning developing in the shadow Fortunately, research continued in the shadow during the second AI winter and surprisingly significant advances were made. After a decade of interruption, connectionist research was back on stage as a significant driving force. In 1982, J. Hopfield (Hopfield, 1982) proposed an analysis of the collective behaviors of physical neural networks. In 1986, G. Hinton demonstrated that energy-based neural network could be trained efficiently by *back-propagation* (Rumelhart et al., 1986b). This simple algorithm is still the dominant approach for training deep learning model nowadays. It also provides interesting distribution representations (McClelland et al., 1986; Hinton et al., 1986), stating that inputs can be represented by many features. This lead to the emergence of a second wave of neural network oriented research. AI started evolving towards a new approach, the so-called ML, based on feature learning, with the first access to datasets. Connectionist advances held strong with Y. LeCun who successfully trained a convolutional ANN to recognize handwritten zip code digits using back-propagation (LeCun et al., 1989). In 1997, Long Short-Term Memory (LSTM) recurrent neural networks were developed to model long sequences such as text (Hochreiter et al., 1997). In 1998, a gradient-based learning method was applied to document recognition (LeCun et al., 1998). In parallel, kernel methods and Support Vector Machines (SVM) (Cortes et al., 1995; Burges, 1998; Scholkopf et al., 1999) were developed and quickly displayed impressive performances in mainstream tasks. They rapidly took over the ML community and delayed the ANN climax until 2007. During the early 2000s, the volume of available data was already strongly increasing as well as the range of data sources and types. This marks the beginning of

The back-propagation technique is based on a simple chain rule computation: $\partial_{\mathbf{w}}\varphi(\mathbf{w}\cdot\mathbf{x}) = \varphi'(\mathbf{w}\cdot\mathbf{x})\times\partial_{\mathbf{w}}(\mathbf{w}\cdot\mathbf{x}) = \varphi'(\mathbf{w}\cdot\mathbf{x})\cdot\mathbf{x}$

the onslaught of *big data*, but still ANN are not yet democratized because of practical reasons.

1.1.3 THE REALM OF DEEP LEARNING: 2007-TODAY

Big data age and GPUs Fifty years after the introduction of the perceptron, ANN finally stroke back with a third wave. They were mostly inactive due to practical issues: the computational power was until then insufficient to train large DNN and there was not enough data available to train them. Early 2009 the open-source ImageNet database (Deng et al., 2009) was released and set the cat among the pigeons. The dataset contained over 14 million *labeled images* and solved the first technical issue. It was followed by CIFAR-10 (Krizhevsky et al., 2010). With this first essential ingredient, the ML, and soon DL, revolution was on its way.

Thus ML drastically changed with the confirmation that *big data* helps and started driving the field. Since then gathering data became easier and easier with social and professional networks, and data became a valuable resource. To give an example, there was 5 exa-bytes (10^{18} bytes) created data per year in 2002 against 10 zeta-bytes (10^{21} bytes) in 2019, a factor of 2000 increased ! However the revolution was mainly made possible thanks to another major technological novelty. Computers started becoming faster and faster at processing data with Central Processing Units (CPU) and in parallel another type of processing units called Graphics Processing Units (GPU) was developed in the late 1990's. While a CPU contains a few large *cores*, a GPU[2] contains thousands of cores which are particularly suited to perform small tasks in parallel. Therefore GPU are particularly suited to the training of ANN that contain millions of parameters and require millions of simultaneous operations.

Neural networks gaining in popularity As a result ANN started competing with SVM provided better performances even though they were slower. But very interestingly, the performances of ANN continued improving with the number of training data, so that entering the age of the *big data* made it suitable for the climax of the DL and large neural networks. This third ANN-oriented research wave started with the breakthrough of (Hinton et al., 2006) that showed that deep belief networks could be trained efficiently using a greedy layer-wise pertaining strategy. G. Hinton had also the idea to mimic the human brain by increasing the network capacity and therefore increasing the number of layers (Bengio et al., 2007). Minsky and Papert (Minsky et al., 1969) already knew that multiple layers would be able to solve the perceptron limitations. But at that time there was no practical algorithm to train such large networks. Thus it took 17 years for back-propagation to become

DL is the field of ML based on deep and wide ANN architectures.

[2] GPU are originally used for computer graphics, image and video processing and gaming.

popularized (Rumelhart et al., 1986b). And GPU power increased by a factor 1000 over ten years, it allowed to finally train large neural networks.

By 2011, the speed of GPU has increased to enable training of large Convolutional Neural Networks (CNN) for vision recognition, and marks the beginning of the DL age. The revolution of DL came from the fact that large neural networks were able to be trained with the use of GPU, whose initial graphical goal had been diverted to linear parallel computing, and the access to large datasets with the emergence of internet. With the increasing computational power, DNN (LeCun et al., 2015) such as AlexNet started rising in international pattern recognition competitions. They outperformed the classical feature engineering approach, and the community started believing that the next revolution would be carried by supervised DL. In particular deep CNN succeeded the ImageNet challenge in 2012 (Krizhevsky et al., 2012) and the year after the challenge was strikingly dominated by neural networks methods. The power of DL methods compared to symbolist approaches lies in the fact that connectionist were simply asking a computer to minimize an energy-based model to learn automatically the features that symbolists were trying to design by hand.

Explosion of Deep Learning DL became very popular in particular thanks to its wide practical successes, even though the early beginning started in the 1950's. However even though ANN were inspired by biological models, the connection between DL and neuroscience is becoming increasingly narrow. Indeed the lack of understanding of the human brain does not drive DL anymore. And the hope of understanding the human brain from shedding light on the learning processes in ANN is still present but weaker and weaker as DL started becoming a standalone discipline. Therefore DL went beyond its biological inspiration and appeals instead to a more general principle of learning hierarchical representations.

Among DL successes, we may cite machine learning translation also called Natural Language Processing (NLP) in which great progresses were been made in recent years. The Google translation engine (Wu et al., 2016) based on LSTM, which are a special case of Recurrent Neural Network (RNN), Sequence to Sequence models and Transformers (Vaswani et al., 2017) out-performed state of the art machine learning translation. In computer vision, progresses have been made in many applications such as lip reading (Chung et al., 2017), visual reasoning (Santoro et al., 2017) or face recognition (Taigman et al., 2014; Parkhi et al., 2015). In Generative Adversarial Network (GAN) (Goodfellow et al., 2014), which allow to generate fake images, it is now possible to synthetize them directly from text sentences (Reed et al., 2016) or even to transform images with Image-to-image generation (Isola et al., 2017) or image processing (Ulyanov et al., 2018). Works on adversarial attacks (Madry et al., 2017; Zhu et al., 2017a) opened a new research direction to build ML models more robust to changes in data distribution. In reinforcement learning (Sutton et al., 2000), after the victory of the DeepBlue computer program against chess champion Kasparov in 1996 (Campbell et al., 2002), it has been generalized to more complicated games such as Atari (Mnih et al., 2013) and more recently Alpha Go beat the world champion Sedol at Go in 2016 (Silver et al., 2016). Of course this list is not exhaustive and many fields are currently moving to DL methods used in various applications. Among the most recent, we may cite self-driving cars and healthcare.

This concludes the non-exhaustive historical overview of ML and DL. As illustrated, DL applies to various domains with complex network architectures and led to considerable successes in AI. Yet despite their wide range of application, high-performances and popularity, many theoretical questions about the efficiency of DL models and algorithms remain unanswered. To fully grasp these burning challenges, in the next section we propose a technical introduction to the ML basics.

1.2 MACHINE LEARNING BASICS

As explained in the last section, the great successes of ML — whether in the *supervised, unsupervised or reinforcement learning* setting — rely on DL and DNN. This section is devoted to accustom the unfamiliar reader to the essential and basic concepts in ML and ANN so that he/she may apprehend correctly the rest of the manuscript and the connection with the statistical physics approach introduced in Sec. 3.2.

For a more detailed introduction to ML, let us mention a few classical references (Bishop, 2006; Murphy, 2012; Shalev-Shwartz et al., 2014) and a more recent and comprehensive reference on DL (Goodfellow et al., 2016) which can be completed by perspectives from different fields (Carleo et al., 2019; Mehta et al., 2019).

1.2.1 THE MACHINE LEARNING WORKFLOW

"Data is the new oil"
Clive Humby

One of the main reasons why ML flew the nest in the recent years lies in the ubiquity of internet, which allows to collect large amount of data that naturally became an essential resource. In this context ML refers essentially to a branch of applied statistics that makes use of a large amount of data to estimate complex functions. In other words, a ML "algorithm" is nothing but a computer program able to solve a given task from such a set of data as formulated by (Mitchell, 1997; Goodfellow et al., 2016): "*A computer program is said to learn from* ***experience*** $\mathscr{E}$ *with respect to some* ***class of tasks*** $\mathscr{T}$ *and* ***performance measure*** $\mathscr{P}$*, if its performance at tasks in* $\mathscr{T}$*, as measured by* $\mathscr{P}$*, improves with experience* $\mathscr{E}$*.*" This formulation can be schematically represented by a ML workflow in Fig. 1 and we briefly detail each of its elements in the next sections. In more details, a ML program aims to solve a certain task $\mathscr{T}$, see Sec. 1.2.2 for a variety of examples, based on observation of a dataset $\mathbb{D}$ within a certain experience $\mathscr{E}$. ML is commonly divided in three main kinds of experiences $\mathscr{E}$ qualified of *supervised*, *unsupervised* and

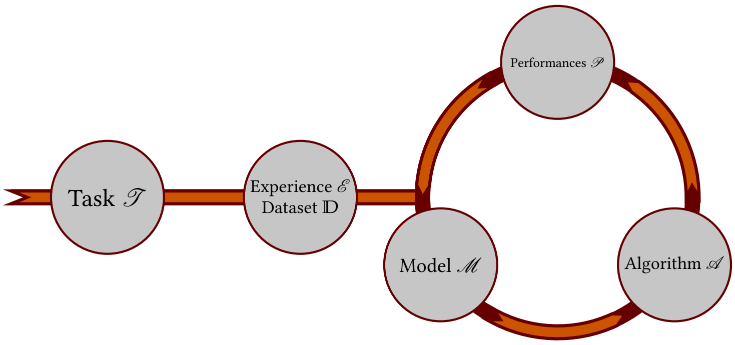


Figure 1: A typical machine learning workflow considers a given task $\mathscr{T}$ to be solved from the experience $\mathscr{E}$ of a dataset $\mathbb{D}$. The task is eventually solved by a model $\mathscr{M}$ trained with an algorithm $\mathscr{A}$, which accuracy is measured by the performance measure $\mathscr{P}$.

reinforcement learning whose frameworks are briefly presented in Sec. 1.2.3. To characterize the underlying rule of the task $\mathscr{T}$, one proposes a probabilistic model $\mathscr{M}$, very often qualified of *parametric* as assumed to depend on a set of parameters $\boldsymbol{\theta}$, see Sec. 1.2.4. We review the most common kind of models $\mathscr{M}$ in Sec. 1.2.9 and discuss the main difficulties encountered when choosing a model class in Sec. 1.2.6-1.2.7. In classical statistics we distinguish two central approaches to estimate the parameters $\boldsymbol{\theta}$ of the model $\mathscr{M}$: *frequentist* and *Bayesian* estimators, introduced in Sec. 1.2.8. Finally the parameters $\boldsymbol{\theta}$ of this estimator can be computed with an algorithm $\mathscr{A}$ using the collected dataset $\mathbb{D}$. We recall the most common algorithms such as gradient-descent algorithm or sampling methods in Sec. 1.2.10. Finally, the accuracy of the predicted model $\mathscr{M}$ is measured by a performance measure $\mathscr{P}$, whose variants are presented in Sec. 1.2.5, and the model is adjusted accordingly.

1.2.2 VARIOUS MACHINE LEARNING TASKS

"You can have data without information, but you cannot have information without data." Napoléon Bonaparte

The task $\mathscr{T}$ denotes the ultimate goal for which the ML algorithm is designed. For a general task, the ML program aims to recover a hidden underlying structure in a dataset $\mathbb{D}$ containing n *observations*. Each *observation*, also called *example*, represents a collection of *features*, that the program must exploit, either directly if features are meaningful or after processing them to obtain a better features representation, to solve the task $\mathscr{T}$. Depending on the kind of *task* $\mathscr{T}$ and *experience* $\mathscr{E}$, each observation may be either a vector $\mathbf{x}$, a tensor $\mathbf{X}$ or an input-output pairs $(y, \mathbf{x})$. There exists a wide range of specific tasks and we will not present an exhaustive list. Instead, we focus on a series of simple tasks considered later in the contribution part Part II such as regression, classification and inverse problems, even though current ML enables to handle tasks with increasing difficulty that a human being would not be able to tackle.

1.2.2.A REGRESSION AND CLASSIFICATION

The simplest and most common tasks in ML are classification and regression. In these tasks, the goal is to predict a function $f : \mathbb{X} \mapsto \mathbb{Y}$ that maps a given input vector $\mathbf{x} \in \mathbb{X}$ to a numerical value $y \in \mathbb{Y}$. The only difference between classification and regression is the output space $\mathbb{Y}$. In the case of regression, the space $\mathbb{Y}$ is continuous, while for classification $\mathbb{Y}$ is discrete and finite so that each output value in $\mathbb{Y}$ is called a *class*. Regression can be applied to various applications such as time series prediction in biology, finance, price prediction, but also to predict the total energy of a molecule. For the sake of the illustration, let us recall the one-dimensional *linear regression* toy example illustrated in Fig. 2 (**Left**) . Observing the input-output pairs $\{x_\mu, y_\mu\}_{\mu=1}^n$ (green dots), the simplest ML task consists in finding the best linear fit whether data is linear or not (orange line). While a human is able to easily find a solution to this one-dimensional task, regression becomes harder and harder with increasing problem dimension d while ML algorithms can handle this easily.

The trendiest example of *classification* is certainly the image recognition task, where one needs to classify pictures of handwritten digits from a dataset such as the MNIST dataset (LeCun et al., 2010), or classify pictures of cat and dogs from CIFAR-10 (Krizhevsky et al., 2010), as illustrated in Fig. 2 (**Right**) . Similar tasks consist in recognizing objects from the ImageNet (Deng et al., 2009) or Fashion-MNIST (Xiao et al., 2017) datasets. Object recognition is particularly well accomplished with CNN particularly suited to treat images and that allow for instance face recognition, self-driving cars or robots captors, tumor detection, and many other classification tasks.

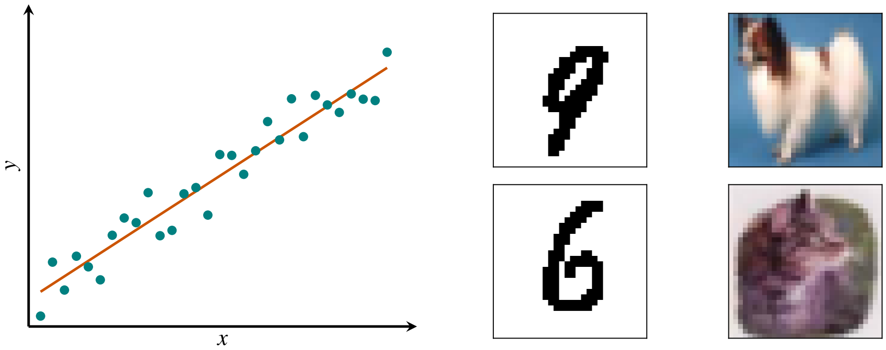


Figure 2: (**Left**) Illustration of one-dimensional linear regression with $d = 1$ and $n = 30$. (**Right**) Images of digits from MNIST and images of a cat and a dog from CIFAR10 to be classified by a machine learning algorithm.

1.2.2.B INVERSE PROBLEMS

In the field of communications and information theory, we are very often interested in a wide class of *inverse problems* where one receives a corrupted signal $\mathbf{y}$ generated from a target signal $\mathbf{x}$, that we aim to reconstruct, through a noisy channel φ_{out}. Observing the output of the channel $\mathbf{y} = \varphi_{\text{out}}(\mathbf{x})$, the goal of the ML program is to reconstruct the input $\mathbf{x}$ signal or equivalently to predict the conditional probability $\mathrm{P}(\mathbf{x}|\mathbf{y})$. Applications vary according on the form of the noisy channel φ_{out} and the signal dimensions.

Denoising and inpainting The simplest case is when an *additive noise* has been added to the signal $\mathbf{x}$, equivalent to a channel $\varphi_{\mathrm{out}}(\mathbf{x}) = \mathbf{x} + \boldsymbol{\xi}$. The goal of the *denoising* task is therefore to *filter* the noise to reconstruct $\mathbf{x}$. Note this denoising task may be extended to multiplicative noise.

The channel may as well corrupt a few entries of the input vector. The computer program must retrieve the missing entries of the input. For instance, the channel may modify an input image $\mathbf{x} = (x_1, x_2, \cdots, x_{d-1}, x_d)$ by removing some pixels x_1, x_{d-1} resulting in an observation $\mathbf{y} = (0, x_2, \cdots, 0, x_d)$. The task to recover the corrupted pixels is known as an *inpainting*. Both tasks are illustrated in Fig. 3.

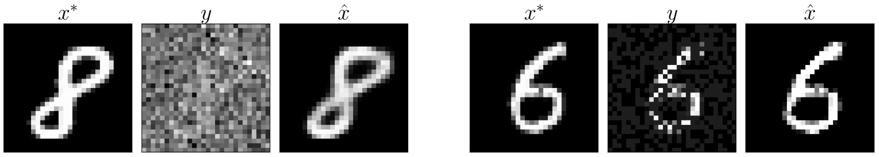


Figure 3: A ground truth image $\mathbf{x}^\star$ is corrupted and results in an observation $\mathbf{y}$ for (**Left**) a denoising task and (**Right**) an inpainting task, from (Baker et al., 2020). The goal is to reconstruct the ground truth signal $\mathbf{x}^\star$ from the observation of $\mathbf{y}$. As an illustration, $\hat{\mathbf{x}}$ may be the output of a machine learning reconstruction.

Compressed sensing and phase retrieval In many applications, the channel involves a multiplication by a known rectangular matrix $\mathbf{A}$ which applies a linear transformation to the initial signal. This is the case of *compressed sensing* (Donoho, 2006) with $\varphi_{\mathrm{out}}(\mathbf{x}) = \mathbf{A}\mathbf{x}$. We may add and extra difficulty on top of that by adding a non-linearity such as an absolute value $\varphi_{\mathrm{out}}(\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|$. Depending if the matrix and the vector belong to $\mathbb{R}$ or $\mathbb{C}$, it refers to real or complex *phase retrieval.* The phase retrieval task is relevant to many real-life settings in which a detector is only able to capture the amplitude of the signal, for instance in electron microscopy, astronomy, crystallography, optics, etc.

Low-rank matrix factorization Another classical task considered in this work is *low-rank matrix factorization*, used in practice for recommendation systems. The channel is the simple matrix multiplication of rank-k matrices according to $\mathbf{Y} = \mathbf{U}\mathbf{V} + \boldsymbol{\xi}$, with $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times d}$. Observing the matrix product $\mathbf{Y}$ that contains a table of users and movies preferences, the aim is to infer separately the latent vectors coding for the users $\mathbf{U}$ and the movie preferences $\mathbf{V}$.

1.2.2.C MANY OTHERS

With recent progresses in ML, practical tasks handled in industry are becoming more and more complex than the simple tasks presented above such as the transcription of unstructured representation of some data into discrete textual form such as *optical character recognition* or *speech recognition*. The latter are used by large technological companies to process images, videos or audio recordings, or annotate or describe input data. Another useful application is *machine translation* in which the algorithm must translate sentences from a language to another and is referred to NLP (Collobert et al., 2011). These fields have been the subject of many important advances especially because of the recent use of DL models (Sutskever et al., 2014; Graves et al., 2013). Let us briefly mention that trying to solve many tasks at the same time is known as *multi-task learning* (Caruana, 1997). While learning a given task and trying to apply it to another task, possibly similar enough, refers to *transfer learning* or *domain adaptation* (Pan et al., 2010).

1.2.3 SUPERVISED, UNSUPERVISED AND REINFORCED EXPERIENCES

ML is typically divided in three kinds of paradigms or experiences $\mathscr{E}$: *supervised*, *unsupervised* and *reinforcement* learning. Let us present the different frameworks, even though we will focus on the simplest supervised learning case in most of the manuscript. In all these different frameworks, the experience $\mathscr{E}$ consists in observing a *dataset* $\mathbb{D}$ made of n *samples*, also called *examples* or *observations*, each being a collection of *features* that the algorithm must process, denoted in full generality by a vector of size d, $\mathbf{x} = \{x_i\}_{i=1}^{d}$.

1.2.3.A SUPERVISED LEARNING

The particularity of supervised learning algorithms lies in the fact that each sample in the dataset is made of a pair of an input features vector $\mathbf{x}$ and a label or target value $\mathbf{y}$, so that the dataset $\mathbb{D} = \mathbb{X} \times \mathbb{Y}$ contains a collection of input-output pairs $\{\mathbf{x}_\mu, \mathbf{y}_\mu\}_{\mu=1}^{n}$ and where each input $\mathbf{x}_\mu \in \mathbb{R}^{d_x}$ and $\mathbf{y}_\mu \in \mathbb{R}^{d_y}$. In most of the cases under investigation, we consider scalar outputs, i. e. $d_y = 1$ and we use the shorthand $d_x = d$. In this case, as each sample has the same dimension, we may introduce a *design matrix* $\mathbf{X} \in \mathbb{R}^{n \times d}$ that contains features in columns and different samples in rows. We assume that the examples are independent and identically distributed (i.i.d) drawn from the joint distribution $\mathrm{P}(\mathbf{x}, y)$. Finally, to fix ideas, the input-outputs pairs may represent coordinates $(\mathbf{x}, y)$ in linear regression as illustrated in Fig. 2 or (image of a digit, class of the digit) in an image recognition task. Having access to the true labels

$\mathbf{y}$ associated to an input matrix $\mathbf{X}$, the algorithm must simply estimate a *mapping* function

$$f : \begin{cases} \mathbb{X} \mapsto \mathbb{Y} \\ \mathbf{x} \mapsto \mathbf{y} \end{cases}$$

that connects inputs to outputs. Equivalently, this can be understood as estimating the probability distribution $\mathrm{P}(\mathbf{y}|\mathbf{x})$. However, we will see later in Sec. 1.2.6 that a *good* function f shall not interpolate and memorize every point in the dataset in order to be robust and predict correctly new data-points. More formally, the goal of supervised learning is to predict future outputs y from observations of unseen inputs vector $\mathbf{x}$, called the *generalization problem*. To fix ideas, in the case of a classification task, we provide examples of images cat and dogs with distinct labels $y = \pm 1$, and the supervised learning algorithm shall separate the feature space according to the labels as illustrated in Fig. 4.

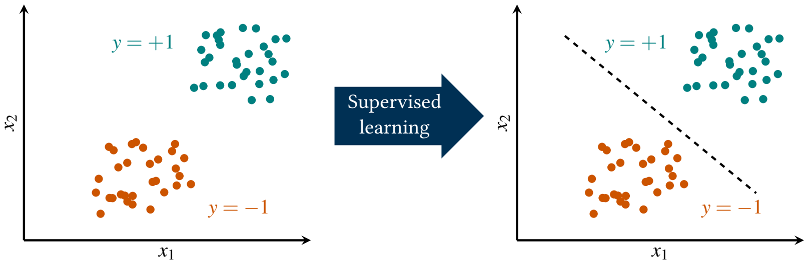


Figure 4: Illustration of a classification supervised dataset. It contains two clouds of points with different labels $y = \pm 1$ and the algorithm must learn a rule to separate cat images ($y = +1$) from images of dogs ($y = -1$).

This setting is called *supervised* learning in the sense that the labels have been provided by a *teacher* who shows a few examples to an algorithm that aims to understand correctly the underlying rule from them to generalize correctly on unseen cases. Unfortunately this ML setting is very expensive as in a way or another a human intelligence shall assign the labels y to the corresponding input vectors $\mathbf{x}$. Even though the collection process of data to create ML datasets was incredibly facilitated with the usage of the internet and social networks, yet this reflects the lack of *intelligence* of supervised algorithms. This remains a strong limitation and is the main reason why the community already opened the door to the *unsupervised* learning framework.

1.2.3.B UNSUPERVISED LEARNING

In contrast with supervised learning, unsupervised learning involves a collection of a random vectors $\{\mathbf{x}_\mu\}_{\mu=1}^n$ and consists in learning interesting quantities related to the probability distribution $\mathrm{P}(\mathbf{x})$ by observing this dataset. While in supervised learning the algorithm observes both label y and input $\mathbf{x}$ and estimates the conditional distribution $\mathrm{P}(\mathbf{y}|\mathbf{x})$, in this more involved setting there is no *teacher* to help the algorithm learning a rule: an unsupervised ML algorithm must make sense of the unstructured data and extract structure from data by itself. Again for the sake of clarity, this situation is analogous to a baby who still does not understand human language and is able anyway to classify cats and dogs when he/she meets them, even though he/she does not literally know what a dog or a cat means.

As a summary, the specificity of unsupervised learning is that it experiences only features vector without supervision labels: it aims to extract useful informations from a distribution that do not require human labor to annotate examples. The core difficulty is to find a *simple* and *compressed representation* which conserves however as much as information as possible of the distribution $P(\mathbf{x})$. Finding such *dimensionality reduction* is fundamental in ML as it provides a powerful and meaningful representation to make sense and process the data. It can be reduced to three approaches: attempting to compress the information in *lower-dimensional representation* by *selecting* only a reduced number of the initial features, or embedding the dataset into a higher-dimensional *sparse* representation whose entries contains mostly zeros to *extract* new features from the original ones, or finally finding an *independent representation* to attempt to disentangle underlying features of the data distribution.

For the sake of conciseness we present the simplest examples: *clustering*, Principal Component Analysis (PCA) and *density estimation* and we refer the interested reader to (Goodfellow et al., 2016) for more details and other applications.

Clustering The *clustering* approach consists in learning the structure of the dataset by trying to separate the dataset in meaningful unlabelled subgroups whose features are close to each other. This method is in particular used for medical imaging, image segmentation, social network analysis, search result grouping, etc. In the absence of labels, the main difficulty is to find a simple representation of the data to appreciate its structure. After being processed, the dataset is split in different *clusters* corresponding to classes defined by the algorithm itself. The battle horse to perform clustering, illustrated in Fig. 5, is the k-means algorithm that divides the dataset into k-clusters, where k is an hyper-parameter that shall be tuned carefully. However, the clustering task is inherently ill-posed as there is no single criterion to obtain a good clustering. As a consequence, separating the dataset may be done in several distinct ways and leads to different clusterings. See (Kaufman et al., 2009) for more details on clustering techniques.

Dimensionality reduction and PCA In order to compress data in a meaningful way, we would like to find a *basis* in which the data can be represented in lower dimensionality than the original input, with statistically independent components. This kind of *dimensionality reduction* can be performed for instance with the so-called PCA method. In the manner of the eigenvalues decomposition of a symmetric positive matrix, PCA is a generalization to any rectangular matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. The idea of PCA is to identify patterns in data by linear transformations such as rotating and

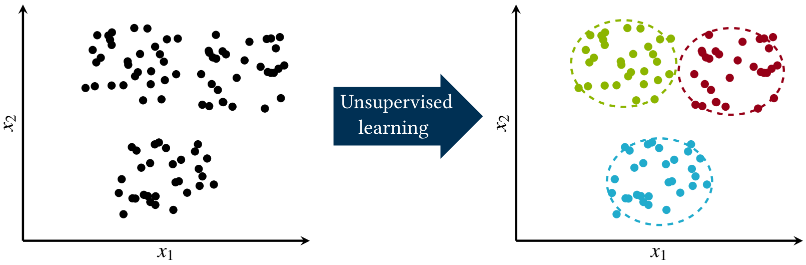


Figure 5: Illustration of an unsupervised clustering task: the algorithm observes a large cloud of points without labels. The k-means algorithm should decide by itself that this large cloud is made of three distinct clusters and assign them different classes.

projecting the matrix in a lower-dimensional subspace whose basis has orthogonal directions, called the *principal components*. Therefore, it builds new independent features that are linear combinations of the initial features. In other words it finds the directions of maximum variance in high-dimensional data and projects it in a lower-dimensional sub-space to keep the maximum of essential data in a smaller space. First, the data matrix may be centered by removing its potential mean $\mathbf{X} \leftarrow \mathbf{X} - \mathbb{E}[\mathbf{X}]$, and the principal components are computed as the eigenvectors of the symmetric covariance matrix $\mathbf{X}^\intercal \mathbf{X}$. Indeed, the Singular Value Decomposition (SVD) of the data matrix yields $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}$, with rotationally invariant matrices $\mathbf{U} \in \mathbb{R}^{n \times k}, \mathbf{V} \in \mathbb{R}^{k \times d}$, $\mathbf{U}^\intercal \mathbf{U} = \mathbf{I} = \mathbf{V}\mathbf{V}^\intercal$. The diagonal matrix $\boldsymbol{\Sigma}$ contains k singular values $\{\boldsymbol{\Sigma}_i\}_{i=1}^k$, such that the covariance matrix decomposes as $\mathbf{X}^\intercal \mathbf{X} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^\intercal$. Rotating the data $\mathbf{X}$ with the rotation matrix $\mathbf{V}$, the covariance matrix becomes diagonal, so that in this basis the components are mutually uncorrelated as illustrated in Fig. 6. More details on PCA may be found in (Jolliffe, 1986; Goodfellow et al., 2016).

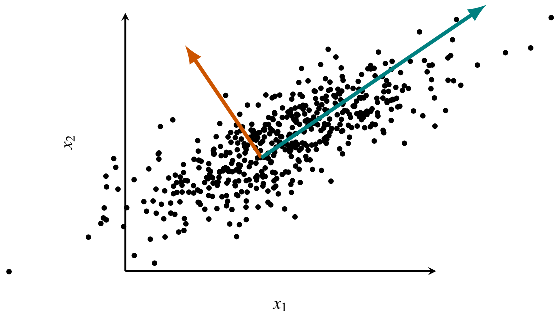


Figure 6: Illustration of Principal Component Analysis of a cloud of random Gaussian matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ for $d = 2$, $n = 500$. The orange and green vectors represent the principal components of the observed dataset.

Density estimation and generative modeling Most unsupervised ML adopts a probabilistic approach known as *density estimation*. It consists in approaching the true probability density $\mathrm{p}(\mathbf{x})$, from which the dataset $\mathbf{X} = \{\mathbf{x}_\mu\}_{\mu=1}^n$ has been drawn, by an approximate density $\hat{\mathrm{p}}$. This density may be parametrized according to its hypothesis class as discussed in Sec. 1.2.4. We consider therefore a set of parametric densities $\mathbb{K}_{\boldsymbol{\theta}} = \left\{ \mathrm{p}_{\boldsymbol{\theta}}\left(\mathbf{X}\right), \boldsymbol{\theta} \in \mathbb{R}^{d_\theta} \right\}$ such that the set of d_θ estimated parameters $\hat{\boldsymbol{\theta}}$ are learned. The corresponding approximate density $\mathrm{p}_{\hat{\boldsymbol{\theta}}}$ captures the ground truth distribution and can be used therefore to generate new samples of the distribution, hence the terminology *generative modeling*. The training of such density estimation method is performed by maximizing the log-likelihood of the observed dataset $\mathbb{D}$, or equivalently the Kullback-Leibler divergence from the approximate distribution $\mathrm{P}_{\boldsymbol{\theta}}$ to the empirical distribution $\mathrm{P}_{\mathbb{D}}(\mathbf{x}) = \frac{1}{n}\sum_{\mu=1}^{n} \delta\left(\mathbf{x} - \mathbf{x}_\mu\right)$

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \sum_{\mu=1}^{n} \log \mathrm{P}_{\boldsymbol{\theta}}\left(\mathbf{x}_\mu\right) \Leftrightarrow \hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \mathrm{KL}\left(\mathrm{P}_{\mathbb{D}} | \mathrm{P}_{\boldsymbol{\theta}}\right) . \tag{1}$$

However, expressing and computing in practice the log-likelihood in high-dimensions is very complex and often intractable. Sampling the density in high-dimensions, with for instance a Monte-Carlo (MC) method, is very costly and becomes slower and slower with the problem dimension. To circumvent these high-dimensional difficulties, new DNN-based models called *deep generative models* have been recently introduced.

Deep generative models Instead of maximizing the above likelihood (1), alternative strategies based on DNN came to light in the recent years such as GAN and Variational Auto-Encoder (VAE), that became very popular thanks to the amazing improvements they brought to the *density estimation* field. Indeed relying on a large amount of data and DNN, they have shown an incredible expressivity and ability to *approximate complex densities* to produce highly realistic images, texts or even sounds. These techniques may be used either to generate new contents or to use as a complex and structured *prior-knowledge* to solve inverse problems, see Sec. 1.2.2.b. The core idea of both VAE or GAN relies on an architecture made of an *encoder* that compresses the data in a low-dimensional representation, and of a *decoder* that tries to decompresses it. Such systems are trained to minimize the difference between the encoded and decoded signals in an unsupervised manner. This situation is typically referred to an *information bottleneck* (Tishby et al., 2000) because the encoder must learn an efficient compression of the data into this lower-dimensional space.

Both VAE and GAN make use of DNN to parametrize the *encoder* and the *decoder*, called the *discriminator* and the *generator* in GAN language. Indeed *high-capacity* DNN are of considerable interest in this task in the sense their wide expressivity allows to approximate any complex density. Also their architecture modularity allows to easily reduce the dimension of the data in a low-dimension latent space. And finally, they have the strong advantage they can be trained and optimized very efficiently using back-propagation algorithm, as specifically presented in Sec. 1.2.10.a.

- *Variational Auto-Encoders*

VAE have been introduced in (Kingma et al., 2013; Rezende et al., 2014) and are a *regularized* version of the classical Auto-Encoders (Vincent et al., 2010). These generative models are nowadays commonly used to approximate a probability distribution $P(\mathbf{x})$ from a dataset, in the perspective to generate new samples from it. The distribution can be reformulated as the marginalization over some latent variables $\mathbf{z}$ as follows

$$P(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}\,. \tag{2}$$

The idea is thus to infer the latent distribution $p(\mathbf{z})$ using the conditional density $p(\mathbf{z}|\mathbf{x})$, which is however also unknown. Therefore, to use it we should instead approximate this density by using a variational principle. Even though we would have access to an approximation $\hat{p}(\mathbf{z}|\mathbf{x})$, we still need to perform the multidimensional integral (2) that is often intractable analytically and hard to evaluate numerically. To make this problem tractable, VAE are essentially made up of an *encoder* $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ parametrized by some parameters $\boldsymbol{\phi}$ that compresses the input data $\mathbf{x}$ in a latent representation $\mathbf{z}$. Yet its particularity lies in the fact that the encoder is regularized during the training in order to ensure that the latent space has *good properties*, that allows to generate appropriate new samples. As illustrated in Fig. 7, the encoder is followed by a *decoder* $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ parametrized by some parameters $\boldsymbol{\theta}$ that tries to maximize the likelihood with the input data.

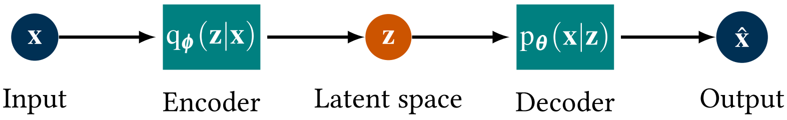

x
$q_\phi(\mathbf{z}|\mathbf{x})$
z
$p_\theta(\mathbf{x}|\mathbf{z})$
$\hat{\mathbf{x}}$
Input
Encoder
Latent space
Decoder
Output

Figure 7: Illustration of a VAE: An encoder $q_{\boldsymbol{\phi}}$ maps the input data into a latent space. The decoder $p_{\boldsymbol{\theta}}$ tries to decode the latent distribution by maximizing the likelihood between the decoded representation $\hat{\mathbf{x}}$ and the original input $\mathbf{x}$.

The VAE objective can be simply derived as a variational approximation $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ of the intractable posterior distribution $p(\mathbf{z}|\mathbf{x})$, see Sec. 4.2 for more details on variational approximations. The Kullback-Leibler (KL) divergence defined in Sec. 4.2.1.b yields

$$\begin{aligned}
&\mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x})\right) = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) - \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \log p(\mathbf{z})\right] + \log p(\mathbf{x}) \\
&\Rightarrow \log p(\mathbf{x}) - \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})\right) \\
&\qquad\qquad = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right)\right],
\end{aligned}$$

so that the VAE objective $\mathscr{L}(\boldsymbol{\phi},\boldsymbol{\theta};\mathbf{x})$ is given by maximizing the variational likelihood *lower bound*

$$\begin{aligned}
\mathscr{L}(\boldsymbol{\phi},\boldsymbol{\theta};\mathbf{x}) &= \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) - \mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right)\right] \\
&\leq \log p(\mathbf{x})\,.
\end{aligned}$$

The first term $\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ represents the *reconstruction* process of the decoder that should minimize the difference between the decoded signal $\hat{\mathbf{x}}$ and the initial input data $\mathbf{x}$ density, or equivalently maximize the likelihood $\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. The second term $\mathrm{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})|p(\mathbf{z})\right)$ should be minimized so that the encoder density closely approaches the latent distribution $p(\mathbf{z})$. In practice this latent distribution is fixed and very often chosen to be Gaussian normal $\mathcal{N}_{\mathbf{z}}(\mu(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x}))$ so that the encoder is trained to return only the two first moments of the Gaussian parametrization. The latent variable is therefore *sampled* and this key step is called the *reparameterization trick*. This trick looks like a *regularization* procedure of the latent space, so that VAE can be simply thought as regularized and probabilistic versions of classical Auto-Encoders. Moreover, it makes the computation of the KL divergence explicitly tractable and the optimization possible with for instance classical gradient-descent algorithms. More details may be found in (Doersch, 2016; Kingma et al., 2019). Notice that the decoder $p_{\boldsymbol{\theta}}$ is very often taken as a DNN which is very expressive but also costly to train. Once trained in this *variational* and *unsupervised* fashion, splitting the encoder from the decoder, the later $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ allows to generate new samples $\mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ of impressive realism, starting from a simple Gaussian noise.

- *Generative Adversarial Networks*

Another kind of common deep generative models are GAN that have enjoyed tremendous success since their introduction (Goodfellow et al., 2014). The idea of GAN is similar to VAE in the sense it exploits a random latent representation. Its conceptual idea is enlightening by its simplicity and allowed to take a leap forward for generative modeling and density estimation. Again the idea is to compare an approximation of the dataset distribution with the true distribution, which is unknown. The brilliant idea of GAN consists in replacing this direct comparison by two *indirect* ones called *generation* and *discrimination*. The GAN architecture, represented in Fig. 8, is therefore made up of a parametric *discriminator* $d_{\boldsymbol{\phi}}$ that takes samples of some true and fake generated data and tries to classify them as well as possible. On the other hand, a *generator* $g_{\boldsymbol{\theta}}$ is trained to generate fake samples to fool the discriminator. Therefore, generator and discriminator have *adversarial* missions. The goal of the generator $g_{\boldsymbol{\theta}}$ is to fool the discriminator $d_{\boldsymbol{\phi}}$, so

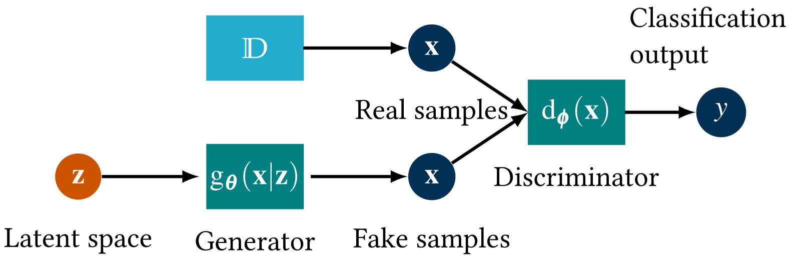


Figure 8: Illustration of a GAN: A discriminator $d_{\boldsymbol{\phi}}$ tries to classify real and fake samples generated from a generator $g_{\boldsymbol{\theta}}$ that tries to fool the discriminator.

the generator computes the probability of samples of belonging to the real dataset $\mathbb{D}$ rather than being fake. It is trained to maximize the classification error between real and fake samples. In contrast, the goal of the discriminator is to detect fake generated data, so that it is trained to minimize the final classification error. Therefore, during the training process, the generator promotes the increase of the classification error whereas the discriminator tries to decrease it. This competition can be thought as a mini-max problem and is translated by the GAN adversarial objective

$$\mathscr{L}(\mathbf{x}) = \min_{\boldsymbol{\phi}} \max_{\boldsymbol{\theta}} \mathbb{E} \log \mathrm{d}_{\boldsymbol{\phi}}(\mathbf{x}) + \log\left(1 - \mathrm{d}_{\boldsymbol{\phi}}(\mathrm{g}_{\boldsymbol{\theta}}(\mathbf{z}))\right) .$$

In practice and as already stressed for VAE, both generator $\mathrm{g}_{\boldsymbol{\theta}}$ and discriminator $d_{\boldsymbol{\phi}}$ are commonly chosen as DNN for their wide expressivity and also because they can be easily jointly trained. GAN are currently used for a variety of tasks such as high quality image or video generation, even though these techniques are not flawless as they can suffer from *mode collapse* issues and raises questions about their ability to really learn the target distribution (Arora et al., 2018b).

1.2.3.c REINFORCEMENT LEARNING

Reinforcement Learning (RL) is the last and more recent class of ML experiences $\mathscr{E}$. The main specificity of RL is that it interacts with an environment so that there is a feedback loop between the learning system and its actions. Qualitatively an agent interacts with the environment so that it dynamically learns and decides what actions to take. In more details, the agent takes some actions a_t at time t that lead to a new state of the agent in the environment s_{t+1} and a corresponding reward r_{t+1} whose value depends on the impact of the action on the environment. It is simply illustrated in Fig. 9. Training such setting to obtain a performant policy $\pi(a,s) = \mathbb{P}(a_t = a | s_t = s)$ is largely beyond the scope of this manuscript. Please refer to (Sutton et al., 1998; Sutton et al., 2000; Mnih et al., 2013) for additional technical details.

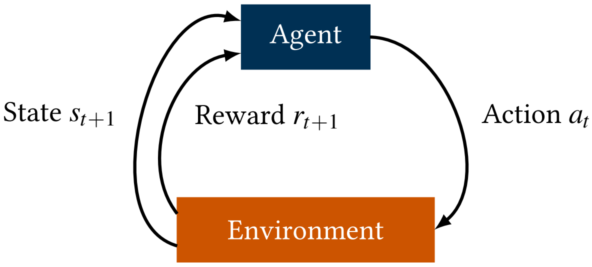


Figure 9: Illustration of reinforcement learning.

In the rest of this manuscript, we will principally focus on the simple *supervised learning* type of experience, and only in Part II we will consider some generative priors generated by deep-generative models such as GAN and VAE, with random weights or trained on real data.

1.2.4 STATISTICAL MODELING

For concreteness, let us summarize the ML workflow in Fig. 1: we have in hand a task $\mathscr{T}$ that we want to solve, for example the classification of images, within an experience $\mathscr{E}$, say supervised such that we have access to a dataset $\mathbb{D} = \{\mathbf{X}, \mathbf{y}\}$ of input images and corresponding labels. The next step consists in *modeling* mathematically the underlying rule observed through the dataset and is referred to as *statistical modeling*. Statistical modeling and learning from data is the subject of a wide literature and is developed for instance in (Cherkassky et al., 2007).

Ground truth assumption and dataset In practice this dataset $\mathbb{D}$ has been collected without any specification on how samples were generated. Yet in the perspective of developing an analysis, it is of practical and theoretical interest to assume that some *oracle* or *teacher* knows the generative process of the dataset, even though most of the time it is not available in real industrial applications. In particular, it has the advantage to allow for measuring the model reconstruction performances as explained in Sec. 1.2.5. But of course, for fairness the generative process should be hidden from the algorithm $\mathscr{A}$ during the learning process, and we introduce it only for a theoretical purpose.

In more details, we assume that there exists either a ground truth function $f^\star$ or equivalently a joint probability $\mathrm{P}^\star(y, \mathbf{x}) = \mathrm{P}^\star(y|\mathbf{x})\mathrm{P}^\star(\mathbf{x})$ accounting for the information contained in the data. The dataset $\mathbb{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ is composed of i.i.d samples such that $\forall \mu \in [\![n]\!], y_\mu = f^\star(\mathbf{x}_\mu)$ or equivalently $y_\mu \sim \mathrm{P}^\star(.|\mathbf{x}_\mu)$. In the case where the generative process is explicitly known and accessible, the ground truth density $\mathrm{p}^\star(y|\mathbf{x})$, which can be can be simply designed by hand in simple theoretical models, is used to generate conveniently new *synthetic datasets*. To conclude, as we have a direct access to the ground truth solution, this setup is very close of the *teacher-student* scenario in planted *spin-glass* models discussed in more details in Sec. 3.2.4, and promotes our statistical physics approach.

Under this assumption, ML aims ultimately to select a *model* $\mathscr{M}$ that *estimates* correctly the underlying data distribution $\mathrm{P}^\star(y|\mathbf{x})$.

1.2.4.A HYPOTHESIS CLASS

To make the estimation problem of the *target function* $f^\star$, or equivalently the *target distribution* $\mathrm{P}^\star(y|\mathbf{x})$, tractable we shall consider models $\mathscr{M}$ in an appropriate *hypothesis class* $\mathbb{H}$. This is the realm of *statistical modeling* that consists in restricting the whole solution space to a smaller set of hypothesis functions $\mathbb{H} = \{f : \mathbb{X} \mapsto \mathbb{Y}\}$, from the *input space* $\mathbb{X}$ to the *target space* $\mathbb{Y}$. This shall be performed carefully such that the hypothesis class $\mathbb{H}$ is rich enough to be contained in the *target class* $\mathbb{H}^\star$, to which $f^\star$ belongs. In this way a function $f \in \mathbb{H}$ may approximate correctly the *target function* $f^\star \in \mathbb{H}^\star$. As an illustration to capture the data-points in Fig. 2, we may consider the set of simple *linear models* parametrized by some weights $\{\mathbf{w}, w_0\} \in \mathbb{R}^{d+1}$: $\mathbb{H}_{\text{linear}} = \{f_{\mathbf{w}} : \mathbb{X} \subseteq \mathbb{R}^d \mapsto \mathbb{Y} : f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + w_0 \text{ with } \mathbf{w} \in \mathbb{R}^d\}$.

As a remark, notice that finding a good statistical model is at the crossroad of two fields of research: the classical *approximation theory* and the modern *machine learning*. Their discriminating difference lies mainly in the input space dimensionality and the features that are engineered in the first and learned from data in the second.

1.2.4.B PARAMETRIC ESTIMATION

The dimension $n_{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$ is not specified as it strongly depends on the model.

Just as the above linear models class example, we often consider *parametric estimation* by restricting statistical models to parametric hypothesis space $\mathbb{H}_{\boldsymbol{\theta}}$ that depend on a collection of parameters $\boldsymbol{\theta} \in \mathbb{R}^{n_{\boldsymbol{\theta}}}$. Estimating the model $f_{\boldsymbol{\theta}}$ is therefore reduced to computing the parameters $\boldsymbol{\theta}$. In general, it denotes the set of parameters of the statistical model that could represent either a scalar, a vector or a set of matrices. In particular, in the neural networks language, the parameters are called instead *weights* and will be denoted $\mathbf{W}$ in the following to represent rectangular matrices. As a remark, there exists also *non-parametric* estimation methods such as nearest neighbors regression or decision trees. Being beyond the scope of this work, we do not cover non-parametric estimation in this manuscript. Refer to (Tsybakov, 2008; James et al., 2013) for an introduction.

1.2.5 MEASURING THE PERFORMANCE

Once the model $\mathscr{M}$ corresponding to an hypothesis class $\mathbb{H}$ has been selected we must introduce a set of tools to measure its validity. In a synthetic dataset setting, in which the ground truth is available, the *reconstruction performance* of the parametric model can be directly measured by the Mean Squared Error (MSE) between parameters $\boldsymbol{\theta}$ of the model $f_{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^\star$ the ones of the target function $f^\star = f_{\boldsymbol{\theta}^\star}$. Otherwise, we need to introduce other statistical tests to measure the model performances.

1.2.5.A RECONSTRUCTION MEASURE: THE MEAN SQUARED ERROR

Whenever the ground truth parameters $\boldsymbol{\theta}^\star$ are available, the performance of the parametric model $\boldsymbol{\theta}(\mathbb{D})$, estimated on the dataset $\mathbb{D}$, can be measured by a direct comparison. The *reconstruction performance* of the estimator is commonly quantified by the MSE between the parameters $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}(\mathbb{D})$ averaged over all potential dataset and ground truth parameters:

$$\text{MSE}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}^\star, \mathbb{D}} \left[\| \boldsymbol{\theta}^\star - \boldsymbol{\theta}(\mathbb{D}) \|_2^2 \right] . \quad (3)$$

This is valid only if the parameters $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}$ have the same dimensions, that is if the target models and statistical models belong to the same hypothesis class $\mathbb{H} = \mathbb{H}^\star$. In our theoretical analysis of the simple models, we will make use of this reconstruction measure. However, in practice the MSE is rarely used because the ground truth parameters $\boldsymbol{\theta}^\star$ are not directly available.

1.2.5.B OBJECTIVE, RISKS AND ERRORS

As an alternative to this reconstruction measure, which is well suited only in the synthetic setting, most ML tasks are instead formulated as the minimization of a *risk* function, also called *objective* or *error* function. This objective depends on a *criterion* or *loss function* $\ell : \mathbb{X} \times \mathbb{Y} \mapsto \mathbb{R}$, whose choice specifically depends on the task $\mathscr{T}$ and experience $\mathscr{E}$. The validity of the statistical model $f_{\boldsymbol{\theta}} \in \mathbb{H}$ is thus appreciated from the value of the risk function: achieving a low risk value advocates for a good statistical model.

Population risk and generalization error The learning objective is commonly chosen as the *population risk* $\mathscr{R}$ defined as

$$\mathscr{R}\left(f_{\boldsymbol{\theta}};\ell\right) = \mathbb{E}_{(\mathbf{x},y)\sim \mathrm{P}(\mathbf{x},y)}\left[\ell\left(y, f_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right)\right] . \tag{4}$$

This is also called the *generalization error* in the ML community and it will be equivalently denoted $e_{\mathrm{gen}}\left(f_{\boldsymbol{\theta}};\ell\right)$. The *loss* function ℓ measures pointwise the error between the target value y and the prediction of the model, $f_{\boldsymbol{\theta}}(\mathbf{x})$. The population risk is simply its average over *all possible* examples drawn from the joint distribution $\mathrm{P}(\mathbf{x}, y)$. Achieving a low population risk defines a strong criterion of validity of the model and allows for model selection. In fact, the optimal model parameters $\hat{\boldsymbol{\theta}}$ would be selected by directly minimizing the population risk $\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}} \mathscr{R}\left(f_{\boldsymbol{\theta}};\ell\right)$. Unfortunately, the population risk and the corresponding minimization program are intractable as the average over the high-dimensional joint distribution $\mathrm{P}(\mathbf{x}, y)$ is very often complex or unknown. Nonetheless, in this theoretical manuscript, we will be able to compute the generalization error in particular cases with synthetic datasets coming from simple joint distributions $\mathrm{P}(\mathbf{x}, y)$.

Empirical risk, training error and training set In general, we do not have knowledge of the generative process and the distribution $\mathrm{P}(\mathbf{x}, y)$. Instead, we only have access to a *finite training set* of n examples $\mathbb{D}_{\mathrm{train}} = \mathbb{X}_{\mathrm{train}} \times \mathbb{Y}_{\mathrm{train}}$. Even if it is very large, $n \gg 1$, this discrete dataset cannot account for the whole continuous and infinite joint distribution $\mathrm{P}(\mathbf{x}, y)$. As a result the intractable *population average* $\mathbb{E}_{(\mathbf{x},y)\sim \mathrm{P}(\mathbf{x},y)}$ is replaced by an *empirical average* over the training set. And consequently the intractable population risk is replaced by the *empirical risk*, also called the *training error* e_{train}, that serves as a proxy of the population risk:

$$\hat{\mathscr{R}}\left(f_{\boldsymbol{\theta}};\ell,\mathbb{D}_{\mathrm{train}}\right) = \frac{1}{n}\sum_{\mu=1}^{n} \ell\left(y_\mu, f_{\boldsymbol{\theta}}\left(\mathbf{x}_\mu\right)\right) . \tag{5}$$

The population risk gives indications along the training of the model validity. For instance, this criterion gives a practical procedure for many ML algorithms, such as Empirical Risk Minimization (ERM), that minimize the empirical risk $\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \hat{\mathscr{R}}\left(\boldsymbol{\theta}; \ell, \mathbb{D}_{\text{train}}\right)$, but only as a proxy of the population risk $\mathscr{R}\left(\boldsymbol{\theta}; \ell\right)$. However, minimizing the empirical risk does not guarantee at all a good *generalization* performance of the estimator on unseen data. Indeed, in high-dimensions the empirical and true underlying distributions can be very different, and thus minimizing the population and the empirical risks do not lead to similar results. Addressing this issue and trying to control their difference $|\mathscr{R}(\boldsymbol{\theta}) - \hat{\mathscr{R}}(\boldsymbol{\theta}; \ell, \mathbb{D}_{\text{train}})|$ is at the heart of modern ML and *statistical learning theory*, as illustrated in Sec. 1.2.7.

Test set and error The purpose of ML is essentially to robustly predict the outcomes of unseen data. Thus, it would make little sense to check the validity of the model on data that have been seen and used to estimate the same model. Therefore, we must allocate a part of the dataset for testing its validity, so that the dataset $\mathbb{D} = \mathbb{D}_{\text{train}} \times \mathbb{D}_{\text{test}}$ is split in a *training set* $\mathbb{D}_{\text{train}}$ that contains observations the algorithm $\mathscr{A}$ may use to estimate the model parameters $\boldsymbol{\theta}\left(\mathbb{D}_{\text{train}}\right)$, and a *testing set* $\mathbb{D}_{\text{test}}$ on which the validity of the model is assessed . Indeed, as suggested in many works such as (Zhang et al., 2016), recent ML models can *perfectly* minimize the empirical risk, meaning that the training error is zero and the model has perfectly *memorize* the training set $\mathbb{D}_{\text{train}}$. As a consequence, reaching zero training error does not ensure the validity of the model, that should be attested instead on the separated *test set* $\mathbb{D}_{\text{test}}$. The error measured on this set is called the *test error*

$$e_{\text{test}}\left(f_{\boldsymbol{\theta}}; \ell, \mathbb{D}_{\text{test}}\right) = \hat{\mathscr{R}}\left(f_{\boldsymbol{\theta}}; \ell, \mathbb{D}_{\text{test}}\right) ,$$

and serves as a finite-size surrogate for the ideal but intractable population risk $\mathscr{R}\left(f_{\boldsymbol{\theta}}; \ell\right)$ and generalization error.

Hyper-parameters and validation test In addition, as illustrated in Sec. 1.2.10, most of current ML algorithms depend on some *hyper-parameters*. These latter are settings that we can use to control the algorithm and must be fixed in some way. However, the hyper-parameters cannot be learned during the algorithm learning procedure, because it would constantly select high-capacity models that easily fit the training set. To circumvent this difficulty, this is often done by introducing a third set, called a *validation set*, that the algorithm does not observe during the training phase and that is used to select good hyper-parameters. Therefore, the dataset $\mathbb{D} = \mathbb{D}_{\text{train}} \times \mathbb{D}_{\text{val}} \times \mathbb{D}_{\text{test}}$ is finally decomposed in train/validation/test sets allocated approximately to $70/10/20\%$ of the total size. In the case of small datasets, where the statistical significance drastically decreases, an alternative approach, called *cross-validation*, is often used. It consists in a *leave-on-out* strategy of repeating and averaging the training and testing operations on randomized sets. See for instance (Goodfellow et al., 2014) for an extended discussion.

1.2.5.c CHOOSING A LOSS FUNCTION

The loss function ℓ is strongly task-dependent, and we review the classical choices of loss functions used in the literature and in this work. In general, in order to minimize the empirical risk (5), with *gradient-based* algorithms, we should prefer smooth loss functions such that the gradients exist, are easy to compute and not too small. The simplest loss is the squared loss $\ell^{\mathrm{l2}}(y, \hat{y}) = (y - \hat{y})^2$ particularly suited for real-valued regression tasks, as well as the absolute loss $\ell^{\mathrm{l1}}(y, \hat{y}) = |y - \hat{y}|$. For classification, where output values are discrete, and very often ± 1, we often use either the hinge loss $\ell^{\mathrm{hinge}}(y, \hat{y}) = \max(0, 1 - y\hat{y})$, the logistic loss $\ell^{\mathrm{logistic}}(y, \hat{y}) = \log(1 + \exp(-y\hat{y}))$, the binary cross entropy loss $\ell^{\mathrm{bce}}(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$ or the hard error-counting loss $\ell^{\mathrm{hard}}(y, \hat{y}) = \mathbb{1}[y \neq \hat{y}]$, even though it is not differentiable.

1.2.6 MODEL COMPLEXITY, LIMITATIONS AND OVERFITTING

The ultimate goal of ML is to predict the output of unseen data of the model $f_{\boldsymbol{\theta}}$ *trained* from the observation of a training set $\mathbb{D}_{\mathrm{train}}$. To do so, most of ML algorithms $\mathscr{A}$ minimize the empirical risk so that many ML problems may be reformulated as an *optimization* problem. Yet the main difference between ML and standard optimization fields is that we require instead the algorithm to not find *any* minima, but a minima that *generalizes* correctly on unseen data. In other words, we require that the generalization error (or simply its finite-size estimation, the test error) remains low, as well as the training error optimized during the training. This is called the *generalization* problem.

1.2.6.a UNDER/OVER FITTING

As the test set is drawn before any learning process, the expected test error e_{test} will be therefore greater or equal than the training error e_{train}. Though, in practice a ML algorithm minimizes the training error as a proxy for minimizing the ultimate generalization error, so that we require the *generalization gap* between the test and training errors $e_{\mathrm{test}} - e_{\mathrm{train}}$ to be as small as possible. The trade-off between this two conditions may lead to key and burning challenges in the ML community: *underfitting and overfitting*. In one hand *underfitting* refers to a model with large training error and therefore a large test error, while *overfitting* occurs when the generalization gap is too large.

1.2.6.b MODEL CAPACITY

Underfitting and overfitting phenomena are closely related to the choice of the hypothesis space $\mathbb{H}$ and in particular its *capacity*. The capacity of a model refers to its ability to fit a wide range of functions. For example linear models cannot fit non-linearly separable data, while high-degree polynomials can. In general, low capacity models may struggle to fit the dataset, while in contrast high-capacity models can easily *memorize* (and not *learn*!) the dataset, so that they will completely *overfit* at test time. These limiting situations are illustrated in Fig. 10 (**Left**) and (**Right**). A *good* statistical model should strike a balance between high-capacity and small test error: the model capacity should be large enough to solve complex tasks resulting in a low training error, but small enough to not perfectly fit the training set and fail in the test set with high test error as illustrated in Fig. 10 (**Center**) . Therefore the capacity of the model should be adapted to the task $\mathscr{T}$ difficulty and the size of the dataset $\mathbb{D}$.

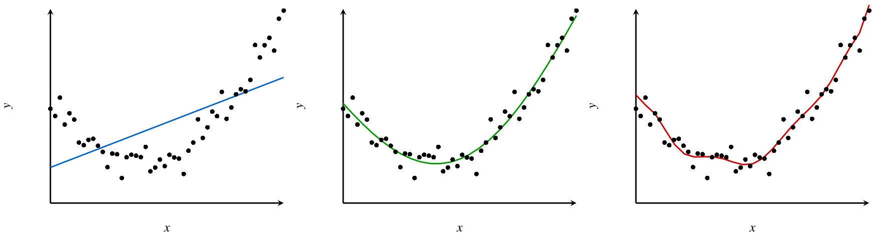


Figure 10: Model complexity illustration on a regression task. Input-output example pairs (x,y) of the training set are shown with black points. (**Left**) A linear model cannot fit the training set and leads to a high training error. (**Center**) An intermediate complexity model yields a good estimator with low training error and low test error. (**Right**) A large complexity model interpolates the training points and achieves almost zero training error. But it completely overfits the training set and does not generalize correctly, resulting in a high test error.

Typically the generalization gap behavior between the test and training errors is summarized with the U-shaped curve in Fig. 11. It can be understood from the *Occam's razor principle* that states that among competing hypotheses that explain a set of observations equally well, we should prefer the hypothesis with the smallest capacity to avoid overfitting. Hence by choosing a statistical model, we shall keep in mind that as soon they have small training error, small capacity functions are more likely to generalize correctly.

1.2.6.c THE BIAS-VARIANCE TRADE- OFF

The illustration in Fig. 10 raises the question of how to properly choose the hypothesis class $\mathbb{H}$ to not be threaten by overfitting. This is formalized by the evolution of the generalization gap with the model complexity described in Fig. 11. In fact, this non-monotonic behavior is traditionally understood from the bias-variance decomposition. Indeed, bias and variance measure two different sources of errors of a given estimator $\boldsymbol{\theta}$, as illustrated by the decomposition of the MSE reconstruction error:

$$
\begin{aligned}
\mathrm{MSE}(\boldsymbol{\theta}) &= \mathbb{E}\left[(\boldsymbol{\theta}^\star - \boldsymbol{\theta})^2\right] = \mathbb{E}\left[(\boldsymbol{\theta}^\star - \mathbb{E}[\boldsymbol{\theta}] + \mathbb{E}[\boldsymbol{\theta}] - \boldsymbol{\theta})^2\right] \\
&= \mathbb{E}\left[(\boldsymbol{\theta}^\star - \mathbb{E}[\boldsymbol{\theta}])^2\right] + \mathbb{E}\left[(\mathbb{E}[\boldsymbol{\theta}] - \boldsymbol{\theta})^2\right] \\
&\qquad + 2\mathbb{E}\left[\boldsymbol{\theta}^\star - \mathbb{E}[\boldsymbol{\theta}]\right]\mathbb{E}\left[\mathbb{E}[\boldsymbol{\theta}] - \boldsymbol{\theta}\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\boldsymbol{\theta}^\star - \mathbb{E}[\boldsymbol{\theta}]\right]^2 + \mathbb{E}\left[(\mathbb{E}[\boldsymbol{\theta}] - \boldsymbol{\theta})^2\right] \equiv \mathrm{Bias}\left[\boldsymbol{\theta}\right]^2 + \mathrm{Var}\left[\boldsymbol{\theta}\right],
\end{aligned}
$$

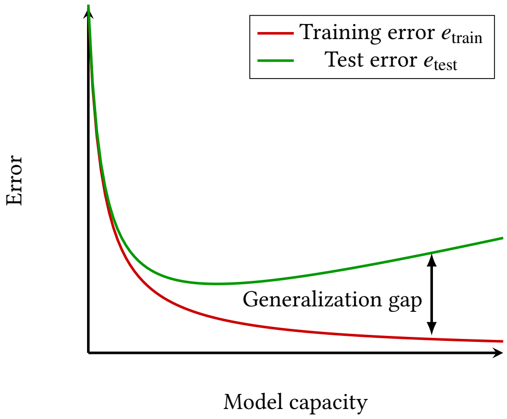


Figure 11: Illustration of the training and test errors as the function of the model capacity. For small capacity models both the training and test errors are high and fall in the underfitting regime. As the capacity grows, the training error eventually decreases to zero, while the test error reaches a minimum at optimal capacity, before growing again and fall in the overfitting regime. The generalization gap is the difference between the test and training errors $e_{\text{test}} - e_{\text{train}}$

where the *bias* of the estimator is the expected deviation from the ground truth value $\text{Bias}\left[\boldsymbol{\theta}\right] \equiv \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\boldsymbol{\theta}^\star - \mathbb{E}\left[\boldsymbol{\theta}\right]\right]$ and and the *variance* is the deviation from the expected estimator value $\text{Var}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\mathbb{E}\left[\boldsymbol{\theta}\right] - \boldsymbol{\theta}\right)^2\right]$. As the model capacity increases, the prediction accuracy increases so that the bias term decreases whereas the variance term increases. Summing these two terms leads to a U-shaped curve similar to the one in Fig. 11 and this decomposition is traditionally used to explain underfitting and overfitting behaviors illustrated in Fig. 10. However, we will describe later that this traditional argument fails explaining the behavior of DNN as suggested in (Zhang et al., 2016).

1.2.6.D REGULARIZATION

As suggested by the above analysis, in order to control the generalization gap, we should act on the model capacity. However in practice, we may prefer to use a fixed model with high-capacity to be able to fit various datasets. Thus in order to avoid overfitting of this high-capacity model, we shall reduce the *effective capacity* of the hypothesis class $\mathbb{H}$. This can be done by *promoting* or *biasing* the training algorithm towards particular solutions. In other words, this means that all potential functions $f_{\boldsymbol{\theta}}$ in the hypothesis class $\mathbb{H}$ are eligible, but a few of them are more likely and have a highest preference. This is commonly done by adding a *regularization term* to the empirical risk (5):

$$\hat{\mathscr{R}}\left(f_{\boldsymbol{\theta}};\ell,\mathbb{D}_{\text{train}}\right) \leftarrow \hat{\mathscr{R}}\left(f_{\boldsymbol{\theta}};\ell,\mathbb{D}_{\text{train}}\right) + \lambda\Omega(\boldsymbol{\theta})\,. \tag{6}$$

λ is called the regularization strength and we can choose different forms for the regularization term such as the classical ℓ_p-norm $\Omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_p$, which promotes sparse or small-weights solutions for $p = 1,2$. Yet depending on the task $\mathscr{T}$ and how one wants to restrict the hypothesis class, more complex reguralization terms can be designed. Minimizing the *regularized* empirical risk (6) results in a trade-off between *fitting the training set* and *satisfying the regularization constraint*, e. g. *keeping small parameters* $\boldsymbol{\theta}$ in the case of a ℓ_2 regularization. This avoids the high-capacity model to fully release its expressivity and overfit the training set. To summarize, regularization refers to any modification made to the learning problem in order to reduce its generalization gap and is essentially at the heart of ML practical challenged.

Since most practical algorithms minimize the (regularized or not) empirical risk, from a theoretical point of view, it would be of great interest to have some *uniform convergence* guarantees that the algorithm simultaneously minimizes the population risk. Bounding the *generalization gap* is a burning challenge widely studied in the statistical learning community and briefly reviewed in the next section.

1.2.7 GENERALIZATION ERROR BOUNDS

The population risk being out of reach, we are reduced to use the empirical risk as a surrogate. Unfortunately, in high dimensions changing the population average by the empirical average may have strong and damaging consequences. First of all, minimizing the empirical risk $\hat{\mathscr{R}}$ eq. (5) is not at all guaranteed to provide the same result than minimizing the true ideal population risk $\mathscr{R}$ eq. (4). This would be correct if we would have a *uniform convergence* theorem that would assert that the *generalization gap* $\|\mathscr{R} - \hat{\mathscr{R}}\|$ decreases quickly with the input dimension d and the number of samples n. This question is part of the realm of *statistical learning* theory, pioneered in (Vapnik et al., 2015; Blumer et al., 1989; Vapnik et al., 1994), and the Probably Approximately Correct (PAC) framework, introduced in (Valiant, 1984), nicely reviewed in (Mohri et al., 2012; Murphy, 2012). Statistical learning theory provides various tools to quantify the model capacity such as the Vapnik-Chervonenkis (VC) dimension d_{vc} or the Rademacher complexity $\mathfrak{R}_n$. Measuring the model capacity allows therefore to bound more finely the generalization gap, i. e. the discrepancy between training error and generalization error.

The goal of the next results is to introduce the main quantities that allow to bound the *generalization gap*. First results have been obtained in the case of classification in (Vapnik, 2013). The first simple result is that any target function $f^\star$ is learnable using a finite hypothesis set $\mathbb{H}$ as soon as $f^\star \in \mathbb{H}$. This result is proven with Hoeffding's inequality (Hoeffding, 1994) and the union bound argument that states $\forall \delta > 0$ with probability $1 - \delta$,

$$\forall f \in \mathbb{H}, \|\mathscr{R}(f) - \hat{\mathscr{R}}(f, \mathbb{D})\| \leq \sqrt{\frac{\ln(|\mathbb{H}|) + \ln(2/\delta)}{2n}}.$$

The underlying union bound argument is responsible for the presence of the cardinality of the hypothesis class $|\mathbb{H}|$ on the right-hand side. Unfortunately, the generalization gap bound becomes vacuous in the case of interest for an infinite class $|\mathbb{H}| = \infty$. This issue is circumvented with the use of finer and tighter bounds such as the VC dimension (Vapnik, 2013) for classification and more recent distribution-dependent Rademacher complexity (Bartlett et al., 2002).

1.2.7.A VC DIMENSION

The idea of the VC dimension, which is restricted to classification tasks, is to count only hypotheses that provide different labelings of the dataset. This can be formalized with the notion of *dichotomies* that is exactly the number of ways of classifying differently the points of the dataset $\mathbb{D}$. To obtain a measure of the richness of the hypothesis class $\mathbb{H}$, we introduce the *growth function* $\Delta_{\mathbb{H}}(n)$ which is the maximum number of dichotomies in which the n points of the dataset can be classified using hypotheses $f \in \mathbb{H}$:

$$\Delta_{\mathbb{H}}(n) = \max_{\{\mathbf{x}_1,..,\mathbf{x}_n\} \subseteq \mathbb{X}} |\{(f(\mathbf{x}_1),..,h(\mathbf{x}_n)) : f \in \mathbb{H}\}| \leq 2^n ,$$

It finally leads to a refinement of the generalization gap bound, called the VC inequality, that states that with probability $1 - \delta$,

$$\forall f \in \mathbb{H}, \|\mathscr{R}(f) - \hat{\mathscr{R}}(f, \mathbb{D})\| \leq \sqrt{\frac{8}{n} \ln(4\Delta_{\mathbb{H}}(2n)/\delta)} .$$

To conclude with this generalization bound, we shall compute the growth function $\Delta_{\mathbb{H}}(n)$, that is unfortunately often intractable. Instead we introduce an alternative measure of the hypothesis class complexity: the VC dimension d_{vc} which is a combinatorial quantity much easier to compute. It is defined as the size of the largest set that can be fully shattered

$$d_{\text{vc}} \equiv \max\{n : \Delta_{\mathbb{H}}(n) = 2^n\} .$$

A set of n points is said to be shattered by a hypothesis set $\mathbb{H}$ when $\mathbb{H}$ realizes all possible dichotomies: $\Delta_{\mathbb{H}}(n) = 2^n$.

From Sauer's lemma (Sauer, 1972; Shelah, 1972), we can show that as soon the VC dimension is finite the growth function verifies $\Delta_{\mathbb{H}}(n) \leq \sum_{i=0}^{d_{\text{vc}}} \binom{n}{i} \leq \Theta\left((ne/d_{\text{vc}})^{d_{\text{vc}}}\right)$ so that $\log \Delta_{\mathbb{H}}(n) = \Theta(\log n)$. Thus, the above generalization bound vanishes with an infinite number of samples. Finally, we obtain the fundamental theorem of statistical learning which states that as soon the VC dimension of hypothesis class $\mathbb{H}$ is *finite*, the target function class $\mathbb{H}^\star$ is PAC learnable. See (Mohri et al., 2012) for an extended derivation.

1.2.7.B RADEMACHER COMPLEXITY

The PAC framework is too restrictive in the sense that it requires the strongest worst-case bound working for any dataset $\mathbb{D}$. To relax this strong hypothesis, a more recent generalization bound has been introduced: the Rademacher complexity (Bartlett et al., 2002), which explicitly depends on the data distribution. The Rademacher complexity captures the richness of the family

$\mathbb{H}$ of functions by measuring the degree to which a hypothesis class can fit random noise. The empirical Rademacher complexity $\hat{\mathfrak{R}}_{\mathbb{D}}(\mathbb{H})$ is defined by

$$\hat{\mathfrak{R}}_{\mathbb{D}}(\mathbb{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathbb{H}} \frac{1}{n}\sum_{\mu=1}^{n} \sigma_\mu f(\mathbf{x}_\mu))\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sup_{f\in\mathbb{H}} \boldsymbol{\sigma}\cdot f(\mathbf{X})\right],$$

where $\boldsymbol{\sigma} = \{\pm 1\}^n$ is a uniform Rademacher random variable with probability $\frac{1}{2}$. The main classical result states that the empirical Rademacher complexity provides a uniform convergence bound. Informally, for any $\delta > 0$, with probability $1-\delta$

$$\sup_{f\in\mathbb{H}} \|\mathscr{R}(f) - \hat{\mathscr{R}}(f,\mathbb{D})\| \leq 2\hat{\mathfrak{R}}_{\mathbb{D}}(\mathbb{H}) + \Theta\left(\sqrt{\frac{\ln(2/\delta)}{n}}\right). \tag{7}$$

Notice that using the Massart's lemma (Massart, 2000) both the growth function and the Rademacher bound may be reconciled as it follows

$$\hat{\mathfrak{R}}_{\mathbb{D}}(\mathbb{H}) \leq \sqrt{\frac{2\ln\Delta_{\mathbb{H}}(n)}{n}} \leq \Theta\left(\sqrt{\frac{d_{\text{vc}}}{n}}\right),$$

so does the VC dimension. To better understand the notion of Rademacher complexity, it is fruitful to notice that it simply measures, on average, the correlation between the prediction of the estimator f and random labels $\boldsymbol{\sigma}$, which are uncorrelated from the inputs examples $\mathbb{X}$. To conclude this section, let us mention that the mathematical and statistical learning community largely focussed on such uniform convergence generalization bounds. However, we will discuss that this kind of *worst-case* scenario bounds are believed to be over-pessimistic and fail, therefore, to capture the generalization behavior of practical model classes such as DNN.

1.2.8 STATISTICAL ESTIMATION

Once we have chosen a parametric model $f_{\boldsymbol{\theta}} \in \mathbb{H}$ within an certain hypothesis class, or equivalently a parametric family of probability distributions $\text{P}_{\boldsymbol{\theta}}(\mathbf{x})$, we shall discuss how to *estimate* in statistics, or equivalently *learn* in ML, the model parameter $\boldsymbol{\theta}$. *Statistical estimation* of the parameters is divided in two ways of thinking: *frequentist* versus *Bayesian*. These approaches undergo long conflicts and the literature is full of debates among statisticians to build proper estimators (Aldrich et al., 2008). In this section, we simply review the two approaches and the most common estimators used in the applications Part II. We standardly denote $\hat{\boldsymbol{\theta}}$ the output of the different estimators.

"Ignorance is preferable to error and he is less remote from the truth who believes nothing than he who believes what is wrong. Thomas Jefferson (1781)"

1.2.8.A FREQUENTIST APPROACH

The frequentist approach assumes that making use of any *a priori* distribution would be misleading. In order to not bias the estimation in the wrong way, frequentists prefer to make no assumption on the a priori probability distributions. The central object of study is the likelihood and follows the work of (Fisher, 1925).

Likelihood Let us consider a set of observations $\mathbf{X} = \{\mathbf{x}_\mu\}_{\mu=1}^n$ drawn i.i.d from an underlying data distribution $\mathrm{P}(\mathbf{x})$ and a family of distributions parametrized by $\boldsymbol{\theta}$, $\mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x}) \equiv \mathrm{P}(\mathbf{x}|\boldsymbol{\theta})$ that models it. We define the *likelihood* function $\mathscr{L}$ or respectively the *log-likelihood* L according to the data $\mathbf{X}$ by

$$\mathscr{L}(\boldsymbol{\theta}|\mathbf{X}) : \boldsymbol{\theta} \mapsto \mathrm{P}(\mathrm{X} = \mathbf{X}|\boldsymbol{\theta}) \,, \quad L(\boldsymbol{\theta}|\mathbf{X}) : \boldsymbol{\theta} \mapsto \log \mathrm{P}(\mathrm{X} = \mathbf{X}|\boldsymbol{\theta}) \tag{8}$$

that both measure the probability of obtaining observations $\mathbf{X}$ for a given value of the model parameters $\boldsymbol{\theta}$. This likelihood function does not assume any *prior knowledge* on the parameter space and is considered by frequentists to contain all relevant information for statistical inference.

Maximum Likelihood Estimation Based on the *log-likelihood* $L(\boldsymbol{\theta}|\mathbf{X})$, the simplest and most common estimator consists in maximizing the probability of observing the data $\mathbf{X}$. The Maximum Likelihood Estimator (MLE) estimator $\hat{\boldsymbol{\theta}}_{\text{mle}}$ is defined as

$$\hat{\boldsymbol{\theta}}_{\text{mle}}(\mathbf{X}) \equiv \operatorname{argmax}_{\boldsymbol{\theta}} \{L(\boldsymbol{\theta}|\mathbf{X})\} = \operatorname{argmin}_{\boldsymbol{\theta}} \{-L(\boldsymbol{\theta}|\mathbf{X})\} \,, \tag{9}$$

that can be, equivalently, simply written as a minimization problem. The MLE can be interpreted as a way of matching the empirical distribution of the data and the model distribution:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{mle}}(\mathbf{X}) \equiv \operatorname{argmax}_{\boldsymbol{\theta}} \{L(\boldsymbol{\theta}|\mathbf{X})\} &= \operatorname{argmax}_{\boldsymbol{\theta}} \left\{ \frac{1}{n} \sum_{\mu=1}^n \log \mathrm{P}(\mathbf{x}_\mu \,|\boldsymbol{\theta}) \right\} \\ &= \operatorname{argmax}_{\boldsymbol{\theta}} \{\mathbb{E}_{\mathbf{x}\sim\hat{\mathrm{P}}} \log \mathrm{P}(\mathbf{x}\,|\boldsymbol{\theta})\}\end{aligned}$$

with the empirical data distribution $\hat{\mathrm{P}}(\mathbf{x}) = \frac{1}{n}\sum_{\mu=1}^n \delta(\mathbf{x} - \mathbf{x}_\mu)$. Indeed the KL divergence serves as a distance within probability densities (see Sec. 4.2.1.b) for more details) and is simply the cross-entropy between the empirical distribution and the model distribution

$$\mathrm{KL}\left(\hat{\mathrm{P}}|\mathrm{P}_{\boldsymbol{\theta}}\right) = \mathbb{E}_{\mathbf{x}\sim\hat{\mathrm{P}}} \log \hat{\mathrm{P}}(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim\hat{\mathrm{P}}} \log \mathrm{P}(\mathbf{x}|\boldsymbol{\theta}) \,.$$

As the first term does not depend on the model, the MLE can be thought as minimizing the discrepancy between the empirical data and model distribution, with ideal objective to match the true data-generating distribution $\mathrm{P}(\mathbf{x})$.

Conditional likelihood In a supervised learning perspective, where models are trained *end-to-end*, the dataset is in fact made of inputs and outputs $\mathbb{D} = \{\mathbf{X}, \mathbf{y}\}$ and the likelihood shall be replaced by the conditional likelihood $L(\boldsymbol{\theta}|\mathbb{D}) = \boldsymbol{\theta} \mapsto \log \mathrm{P}(\mathbf{y}|\boldsymbol{\theta},\mathbf{X})$. The corresponding maximum likelihood estimator readily generalizes to

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{mle}}(\mathbb{D}) &\equiv \mathrm{argmax}_{\boldsymbol{\theta}}\left\{L(\boldsymbol{\theta}|\mathbb{D})\right\} \\ &= \mathrm{argmax}_{\boldsymbol{\theta}}\left\{\mathbb{E}_{(\mathbf{x},y)\sim\hat{\mathrm{P}}(\mathbf{x},y)} \log \mathrm{P}(y|\boldsymbol{\theta},\mathbf{x})\right\},\end{aligned}$$

and is a central estimator in most supervised learning settings. It turns out in particular that under the assumption that the ground truth $\mathrm{P}^\star(\mathbf{x},y)$ lies within the probability density family $\mathbb{K}_{\boldsymbol{\theta}} = \left\{\mathrm{P}_{\boldsymbol{\theta}}(y|\boldsymbol{\theta},\mathbf{x}), \boldsymbol{\theta} \in \mathbb{R}^{d_\theta}\right\}$, the MLE estimator becomes *optimal* in the asymptotic infinite number of samples $n \to \infty$. As it converges the fastest towards the true parameters $\boldsymbol{\theta}^\star$, the estimator is qualified of *consistent* and also *efficient* as moreover its generalization error decreases in this limit. Indeed, the Cramer-Rao bound states that any unbiased estimator has a variance bounded by the inverse of the Fisher information:

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) \geq \left(\mathbb{E}\left[\left(\partial_{\boldsymbol{\theta}} \log \mathrm{P}(y|\boldsymbol{\theta},\mathbf{x})\right)^2\right]\right)^{-1},$$

and this lower bound is attained by the MLE in the large number of sample regime $n \to \infty$. This means that in this regime of large number of data, maximimzing the likelihood should be preferred to any other statistical estimator. Unfortunately when the number of data is limited, the MLE is not optimal and leads to overfitting that can be avoided by adding a regularization term. In fact in this regime, more prior-knowledge information is required to perform optimal reconstruction, which can be achieved with the Bayesian approach presented in the next section.

1.2.8.B BAYESIAN APPROACH

In contrast with the frequentist approach, which relies on a *worst-case analysis*, Bayesian statistics makes use of *prior information* or *knowledge beliefs* and relies on a *typical case analysis*. The Bayesian approach considers all possibles values of the estimators to make a prediction and it follows essentially Bayes (Bayes, 1763) and Laplace works.

While the frequentist perspective assumes that the ground truth parameter $\boldsymbol{\theta}^\star$ is unknown but fixed, the Bayesian approach uses probabilities to reflect prior knowledges, so that $\boldsymbol{\theta}^\star$ is considered as an uncertain random variable with *prior* distribution $\mathrm{P}(\boldsymbol{\theta}^\star)$. Also while MLE makes predictions using a point-wise estimate, the Bayesian approach makes a predictions using the full distribution over $\boldsymbol{\theta}$. Therefore, observing a supervised learning dataset $\mathbb{D} = \{\mathbf{X},\mathbf{y}\}$, we can make use of the observations of the data to model the probability of the parameter $\boldsymbol{\theta}$ essentially by means of the Bayes formula

$$\mathrm{P}(\boldsymbol{\theta}|\mathbf{y};\mathbf{X}) = \frac{\mathrm{P}(\mathbf{y}|\boldsymbol{\theta};\mathbf{X})\,\mathrm{P}(\boldsymbol{\theta})}{\mathrm{P}(\mathbf{y};\mathbf{X})} \tag{10}$$

where $\mathrm{P}(\mathbf{y}|\boldsymbol{\theta};\mathbf{X})$ denotes the *conditional likelihood*, $\mathrm{P}(\mathbf{y};\mathbf{X})$ the *evidence*, denoted later in the manuscript $\mathcal{Z}(\mathbf{y};\mathbf{X})$ also called the *partition function*, and finally $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y};\mathbf{X})$ is called the *a posteriori* or *posterior* distribution. With this prior informations, which model the external world, Bayesian methods generalize typically much better when the training set is small and does not contain enough information. However, we can already notice that computing the average over the posterior $\mathrm{P}(\mathbf{y};\mathbf{X}) = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}\ \mathrm{p}(\boldsymbol{\theta}|\mathbf{y};\mathbf{X})$ will strongly suffer in the high-dimensional regime, where $d, n \to \infty$, and is in fact very often intractable.

Notice that taking a Gaussian prior $\mathrm{P}(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\theta}}(0,1)$ is equivalent to add a ℓ_2 regularization term $-\log \mathrm{P}(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2$ to the log-likelihood.

How to choose the prior? Bayesian methods make deep use of the prior information $\mathrm{P}(\boldsymbol{\theta})$ which is unknown in general. The prior information is useful in the sense it shifts the probability density towards more probable regions of parameters. In particular, it might be used to promote models that are simpler or more smooth, and can be already understood as a *regularization* factor. As frequentists blame Bayesian to bias estimation by injecting prior information that may be wrong, we should decide how to select *correctly* the prior information $\mathrm{P}(\boldsymbol{\theta})$. This question was addressed and answered in (Jaynes, 1957; Jaynes, 2003) who advocated that in order to bias as few as possible the estimation, we should select priors according to the *maximum entropy principle* presented in more details in Sec. 4.2.2.b. In practice we often start with a Gaussian distribution, which is known to maximize the entropy under certain constraints, with wide variance to reflect the high degree of uncertainty in $\boldsymbol{\theta}$ and then decrease it along the training.

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate. It is least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information. ET. Jaynes, 1957

In practice computing the full posterior distribution in high-dimensions is often intractable. For simplicity it is therefore of practical interest to reduce the problem to simple point-wise estimates such as the mean and the maximum of the posterior distribution $\mathrm{P}(\boldsymbol{\theta}|\mathbf{y};\mathbf{X})$ corresponding to the so-called Minimum Mean Squared Error (MMSE) and Maximum A Posteriori (MAP) estimators.

Minimum Mean Squared Error The MMSE estimator is simply defined as the mean of the posterior distribution

$$\hat{\boldsymbol{\theta}}_{\mathrm{mmse}} = \mathbb{E}_{\mathrm{P}(\boldsymbol{\theta}|\mathbf{y};\mathbf{X})}\left[\boldsymbol{\theta}\right], \tag{11}$$

and will be of central interest in the rest of the manuscript. Indeed, ideally we hope to minimize the reconstruction error with the ground truth parameter $\boldsymbol{\theta}^\star$, i. e. the Squared Error (SE)

$$\mathrm{SE}\left(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}\right) = \frac{1}{d}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2. \tag{12}$$

However, as very often the ground truth parameter $\boldsymbol{\theta}^\star$ is not accessible, we would simply require to minimize the error in average, i. e. the MSE defined by

$$\mathrm{MSE}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{d}\int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}\ \mathrm{p}\left(\boldsymbol{\theta}|\mathbf{y};\mathbf{X}\right) \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2 . \tag{13}$$

Taking the derivative with respect to $\boldsymbol{\theta}$ directly yields the definition of the MMSE estimator $\hat{\boldsymbol{\theta}}_{\mathrm{mmse}}$ that therefore has the nice property to minimize the MSE reconstruction error. This estimator is very powerful but unfortunately very rarely tractable in practice as it requires to average over the high-dimensional posterior distribution $\mathrm{P}\left(\boldsymbol{\theta}|\mathbf{y};\mathbf{X}\right)$. An approach to compute this estimator would be to make use of Markov-Chain Monte-Carlo (MCMC) algorithms to sample the posterior distribution. But in high-dimensions, sampling methods are very inefficient and require a huge number of samples. As a spoiler, a main part of this work is concerned with computing this high-dimensional object with *heuristic methods* from statistical physics.

Maximum A Posteriori The other simple point-wise estimate is taking the maximum of the posterior distribution and not the mean as for the MMSE. The MAP estimator is naturally defined as

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{map}} &\equiv \mathrm{argmax}_{\boldsymbol{\theta}} \log \mathrm{P}\left(\boldsymbol{\theta}|\mathbf{y};\mathbf{X}\right)\\ &= \mathrm{argmin}_{\boldsymbol{\theta}}\left\{-\log \mathrm{P}\left(\mathbf{y}|\boldsymbol{\theta},\mathbf{X}\right) - \log \mathrm{P}\left(\boldsymbol{\theta}\right)\right\}\end{aligned}$$

and can be turned into a minimization problem such as in ERM. Under this formulation, we notice easily that MAP Bayesian estimation with $\mathrm{P}(\boldsymbol{\theta})$ a priori information is strictly equivalent to MLE estimation in the presence of a regularizer $-\log \mathrm{P}(\boldsymbol{\theta})$ and has the advantage to provide a way to design complicated yet interpretable regularization terms. In comparison with the MLE, it has the advantage to leverage prior information not contained in the training data at the price to increase the bias.

1.2.9 CLASSICAL MODELS

In this section, we briefly present the main models and architectures mostly used in modern supervised ML, ranging from linear models to deep neural networks.

1.2.9.A GENERALIZED LINEAR MODELS

The simplest and wildest class of models used in many ML applications is *linear models*. To perform classification or regression linear models are very popular because of their simplicity. However, to produce discrete outputs for instance, one often considers a wider hypothesis class known as the Generalized Linear Model (GLM) hypothesis class

$$\mathbb{H}_{\mathrm{glm}} = \left\{ f_{\mathbf{w}} : \begin{cases} \mathbb{R}^d \mapsto \mathbb{R} \\ \mathbf{x} \mapsto \varphi_{\mathrm{out}}\left(\mathbf{w}^\intercal \mathbf{x} + w_0\right) , \end{cases} \quad (\mathbf{w}, w_0) \in \mathbb{R}^{d+1} \right\}$$

It contains affine functions parametrized by a vector $\mathbf{w} \in \mathbb{R}^d$ applied as a scalar product with the features $\mathbf{x}$, and a *bias* or *intercept* w_0. In addition, φ_{out} represents a deterministic or stochastic element-wise activation function, potentially non-linear, added on top of the linear operation. In other words, GLM are simple models based on a linear weighted sum of the features components shifted by a bias w_0. The parameter $\mathbf{w} = \{w_i\}_{i=1}^d$ can be thought as the *weights* associated at each sample features $\mathbf{x}_\mu = \{x_{i\mu}\}_{i=1}^d$. This affine operation is called a *formal neuron*. Even though very simple, it is the elementary brick at the origin of more complex modern feed-forward DNN. The decision boundary of linear models is essentially a high-dimensional *hyperplane* that splits *linearly* the input space. For classification tasks, considering a sign output function $\varphi_{\mathrm{out}}(z) = \mathrm{sign}(z - K)$ or an Heaviside step function $\varphi_{\mathrm{out}}(z) = \Theta(K - z)$ refers to the historical *perceptrons* with a stability threshold K. In particular, we will illustrate our statistical physics approach on this simple model class notably in Sec. 4.1.5, 4.3.4, 4.4.1 and 4.4.3.

Linear regression: pseudo inverse, ridge & lasso Consider we want to *predict the output* $y \in \mathbb{R}$ of input vector $\mathbf{x} \in \mathbb{R}^d$, we first consider a linear predictor that outputs $\hat{y} = \mathbf{w}^\intercal \mathbf{x} + w_0$. Taking the MSE as our performance measure, we would like to minimize the generalization error on the test set $\mathbb{D}_{\mathrm{test}}$

$$\mathrm{MSE}_{\mathrm{test}}(\hat{\mathbf{w}}) = \mathbb{E}_{(y,\mathbf{x}) \sim \mathbb{D}_{\mathrm{test}}} \left(y - \hat{y}(\hat{\mathbf{w}})\right)^2 .$$

As as a surrogate, we minimize instead the empirical risk on the training set $\mathbb{D}_{\mathrm{train}}$. The goal is therefore to find an hyperplane that minimizes the sum of the squared errors between the observations y and predictions $\hat{y}$. In this simple case, we can derive an explicit expression of the parameters $\hat{\mathbf{w}}$ that minimize the MSE on the training set. Taking the gradient of the empirical risk to $\mathbf{0}$, we easily obtain the *pseudo-inverse* estimator, also called the *normal equations*:

$$\nabla_{\mathbf{w}} \|\mathbf{y}_{\mathrm{train}} - \mathbf{X}_{\mathrm{train}} \hat{\mathbf{w}}\|_2^2 = \mathbf{0} \Rightarrow \hat{\mathbf{w}}^{\mathrm{pseudo}} = \left(\mathbf{X}_{\mathrm{train}}^\intercal \mathbf{X}_{\mathrm{train}}\right)^{-1} \mathbf{X}_{\mathrm{train}}^\intercal \mathbf{y}$$

However, in practical applications of linear regression, the number of features d is often very large, and even larger than the number of samples d, so that in this case the problem has an infinite number of solutions.

To obtain a finite number of solutions, we often try to reduce it by selecting an appropriate set of *features* that describe correctly the underlying distribution. A *feature selection* method consists in projecting the data in a basis where the data are *sparse*, see (Hastie et al., 2015) for a comprehensive discussion. Nonetheless, the modern *feature selection* approach is to use regularization that slowly pushes the effects of irrelevant features towards zero while keeping only interesting features, see Sec. 1.2.6.d. Regularized regression coincides equivalently to penalized models or shrinkage methods. Minimizing the regularized empirical risk (6), that can be thought as the trade-off between minimizing the squared error and having *small* coefficients, constrains the initial hypothesis class $\mathbb{H}$ to particular solutions with small magnitude and fluctuations of the parameters.

In the case of linear regression with the squared loss, three main cases are widely considered: *LASSO* (Tibshirani, 1996) with $\Omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$, *ridge regression* (Hoerl et al., 1970) with $\Omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$ and a combination of them called *elastic net* (Zou et al., 2005).

Binary classification: perceptron & logistic For a binary classification task such as represented in Fig. 12 **(Left)** , the decision boundary can be estimated by a linear hyperplane such that on each side of the decision boundary the labels are positive or negative. This setup is known as the classical *perceptron* $y = \text{sign}\left(\mathbf{w} \cdot \mathbf{x} + w_0\right)$. To train this model and estimate the parameters $\{\mathbf{w}, w_0\}$, the original perceptron algorithm (Rosenblatt, 1958) and many variant rules have been proposed. The *perceptron* model has been the subject of a rich statistical physics literature, see (Engel et al., 1993) for a comprehensive review, and it will be discussed in Sec. 3. Modern ML tasks are very often formulated as minimization problems of the empirical risk (5). Keeping our generalized linear model hypothesis class, we still have the choice of the loss function ℓ. In the case of binary classification, let us mention the widely used *logistic regression* with $\ell^{\text{logistic}}(y, \hat{y}) = \log(1 + \exp(-y\hat{y}))$, which is equivalent to the binary cross entropy loss $\ell^{\text{bce}}(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$ with a sigmoid activation $\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{x} + w_0)$.

Support Vector Machines and hinge loss In the case where the training examples are *linearly-separable*, the *perceptron*'s solution is ill-defined as there exists an infinite number of hyperplanes that classify correctly the training set. To select a robust solution, the idea of the influential SVM is to select the perceptron with the widest margin (Boser et al., 1992; Vapnik, 2013). In the context of a binary classification task, in order to generalize as well as possible to variations of the dataset we should select the hyperplane that maximizes the distance to the nearest examples in the two classes, as illustrated in Fig. 12 **(Left)** . In more details, for $(\mathbf{w}, w_0) \in \mathbb{R}^{d+1}$, we require that on the margins $y = \text{sign}\left(\mathbf{w} \cdot \mathbf{x} + w_0\right) \Leftrightarrow 1 = y\left(\mathbf{w} \cdot \mathbf{x} + w_0\right)$, so that the distance of the decision boundary to the margins $\mathbf{x}_{\pm}$ is $\mathbf{w} \cdot \mathbf{x}_{\pm} = 1 \mp w_0$ and the width of the margin equals $\gamma = \frac{2}{\|\mathbf{w}\|_2}$. As a consequence, to maximize the margin γ we may equivalently minimize a ℓ_2 reguralization term $\frac{1}{2}\|\mathbf{w}\|_2^2$. To be more precise, the primal form reads

$$\text{Minimize } \frac{1}{2}\|\mathbf{w}\|_2^2, \text{ under the constraints } y_\mu \left(\mathbf{w} \cdot \mathbf{x}_\mu + w_0\right) \geq 1.$$

The Karush-Kuhn Tucker conditions on the associated dual formulation lead to a well-defined and unique solution and finally reduces the hypothesis class. Indeed while the VC dimension of the GLM hypothesis class $\mathbb{H}_{\text{glm}}$ is $d+1$, for the SVM the margin constraint γ shrinks it to $d_{\text{vc}} = \min\left(\frac{2R^2}{\gamma}, d\right) + 1$ that can be be much smaller than $d+1$, with R the radius of the smallest sphere comprising the training samples. Moreover, the primal problem may be formulated in a practical regularized version

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \sum_{\mu=1}^{n} |1 - y_\mu f_{\mathbf{w}}(\mathbf{x}_\mu)| + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 ,$$

by minimizing the *hinge loss* $\ell^{\text{hinge}}(y,\hat{y}) = \max(0, 1 - y\hat{y})$ which is another common choice to perform binary classification.

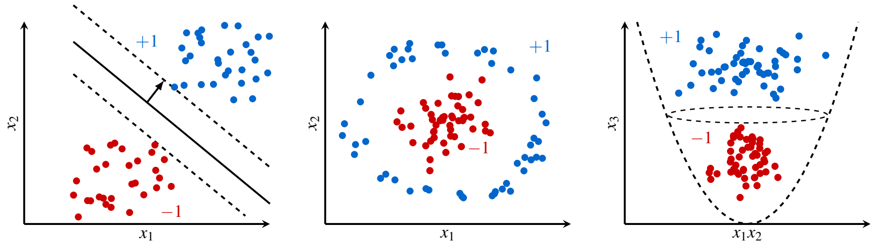


Figure 12: Illustration of a classification task for (**Left**) a linearly separable dataset that can be classified with a large margin SVM, (**Center**) and a non-linearly separable dataset that a generalized linear model cannot fit. (**Right**) Projection of the non-linearly separable dataset into a higher dimensional space $(x_1, x_2) \mapsto (x_1, x_2, x_3)$.

Limitations Linear models such as linear regression or binary classification with perceptrons or SVM illustrated in Fig. 4 are very simple from a practical viewpoint. Unfortunately the low-capacity hypothesis class $\mathbb{H}_{\text{glm}}$ is extremely limited to simple tasks and dataset and cannot fit correctly more complex tasks. In particular, the XOR function or a more complex donut-like set of points such as in Fig. 12 (**Center**) . However, linear models do not allow to classify non-linearly separable points.

Kernel methods To circumvent this issue, the very elegant idea of *kernels methods* is to change the representation of the input features $\mathbf{X} = \{\mathbf{x}_\mu\}_{\mu=1}^{n}$ by projecting them in a higher-dimensional latent space, in which data become evenetually linearly separable. For more details on kernel methods, see (Williams et al., 1996; Scholkopf et al., 1999). Kernel methods rely on the *kernel trick* (Aizerman, 1964) based on the Mercer's theorem. It follows from the observation that the dot product between the parameters $\mathbf{w}$ and a feature vector $\mathbf{x}$ can be written as a linear decomposition with some coefficients $\{\theta_\mu\}_{\mu=1}^n$ so that

$$\mathbf{w} \cdot \mathbf{x} + w_0 = \theta_0 + \sum_{\mu=1}^{n} \theta_\mu \mathbf{x}^\intercal \mathbf{x}_\mu \,.$$

Having this trick in mind, it has been extended to more complex *kernels* $k : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}$, where $\mathbf{x}^\intercal \mathbf{x}_\mu$ is replaced by a dot product in a high-dimensional space $k\left(\mathbf{x}, \mathbf{x}_\mu\right) = \phi\left(\mathbf{x}\right)^\intercal \phi\left(\mathbf{x}_\mu\right)$. By projecting the features in a new, possibly higher dimensional, space through the mapping ϕ, we eventually transform the dataset in a linearly separable representation. Indeed, the main interest of kernel methods is that the new estimator parametrized by $\boldsymbol{\theta}$

$$f_{\boldsymbol{\theta}}\left(\mathbf{x}\right) = \theta_0 + \sum_{\mu=1}^{n} \theta_i k\left(\mathbf{x}, \mathbf{x}_\mu\right)$$

is non-linear with respect to the examples $\mathbf{x}_\mu$, yet it is linear in the new features $\phi\left(\mathbf{x}_\mu\right)$. In other words, a kernel method is simply a linear model on pre-processing data in the space $\phi\left(\mathbf{X}\right)$. By considering ϕ fixed, we only need to optimize over $\boldsymbol{\theta}$, similarly to linear regression except that the model is now non-linear and more expressive. In particular, SVM may be used in parallel of the kernel trick and are called *kernel*-SVM in this context. We need to construct the $n \times n$ Gramm matrix $k_{\mu,\nu} = k\left(\mathbf{x}_\mu, \mathbf{x}_\nu\right)$ from the dot product of $\{\phi\left(\mathbf{x}_\mu\right)\}_{\mu=1}^n$. This operation is computationally inefficient as $\Theta(n^2)$ and certainly hopeless for training sets containing millions of examples. In practice the kernel k is not computed but commonly taken among simple tractable forms such as the Gaussian, also called Radial Basis Function (RBF) kernel $k(\mathbf{a}, \mathbf{b}) = \mathcal{N}\left(\mathbf{a} - \mathbf{b}, \sigma^2 \mathbf{I}\right)$, or even polynomial, Laplace, or sigmoid kernels. Recently, kernel methods started experience a decline in popularity with the advent of DL and DNN and especially when for the first time a neural network outperformed a Gaussian kernel SVM on MNIST (Hinton et al., 2006).

1.2.9.B DEEP FEED-FORWARD NEURAL NETWORKS

In the recent years, the wide class of DNN models entered the scene (LeCun et al., 2015). Just as the wings of plane are inspired by the wings of birds or many other biomimetics systems, DNN have been inspired by the brain mechanism to simulate AI. Yet the corresponding DL branch of research became far apart of the initial neuroscience field. ANN and DNN are henceforth a class of models made of a cascade connection of simple elementary bricks based on the *perceptron*, as illustrated in Fig. 13. Connecting several *formal neurons* into complex networks define a richer hypothesis class with higher capacity. For instance a *feed-forward* DNN of *depth* L is made of L *hidden layers* $\{\mathbf{h}^{(l)}\}_{l=1}^L$. Each hidden layer $\mathbf{h}^{(l)} = \{h_i^{(l)}\}_{i=1}^{n_l}$ of *width* n_l is composed of n_l *hidden units*. This high-expressivity model is parametrized by a set of weights matrices and

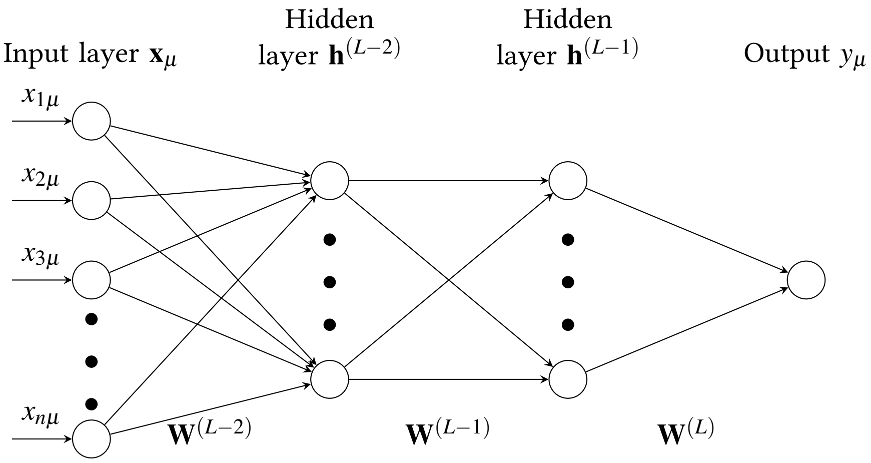


Figure 13: Representation of a deep feed-forward neural network with depth L. Each arrow represents a learnable scalar value and each hidden unit $h_i^{(l)}$ is the result of a formal neuron operation.

bias vectors $\boldsymbol{\theta} = \left\{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}, \forall l \in [\![L]\!]\right\}$. The architecture can be expressed mathematically as the following input to output mapping $\mathbf{X} \mapsto \mathbf{y}$:

$$\mathbf{y} = \sigma^{(L)}\left(\mathbf{W}^{(L)}\sigma^{(L-1)}\left(\cdots\sigma^{(1)}\left(\mathbf{W}^{(1)}\mathbf{X} + \mathbf{b}^{(1)}\right) + \cdots\right) + \mathbf{b}^{(L)}\right).$$

A given layer $\mathbf{h}^{(l+1)}$ is the result of a linear product of a *matrix of weights* $\mathbf{W}^{(l)}$ with the result of the previous layer $\mathbf{h}^{(l)}$ and adding a potential *bias* $\mathbf{b}^{(l)}$. This is followed by a non-linear operation continuous *activation function* σ acting component-wise: $\mathbf{h}^{(l+1)} = \sigma^{(l)}\left(\mathbf{W}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right)$ with $\mathbf{h}^{(L+1)} \equiv y$ and $\mathbf{h}^{(0)} \equiv \mathbf{x}$. The corresponding very expressive hypothesis class $\mathbb{H}$ is largely modular through the architecture of the network, as we can easily tune the depth, width, and activation choice, and it is one of the reasons for its success. Especially the *universal approximation theorem* (Cybenko, 1989; Hornik, 1991) showed that a two-layer neural network with $L = 2$ can approximate any smooth function. Yet, state-of-the-art DNN used nowadays are not limited to two layers and we observe an explosion of the numbers of layers to apply to various and more complex tasks. To illustrate, famous networks such as AlexNET contains 100 layers (Krizhevsky et al., 2012) so does a typical ResNET (He et al., 2016). Finally, to avoid the *vanishing gradient problem* during the training, typically with a gradient-descent based algorithm, it is preferable to choose smooth activations functions with non-vanishing gradients such as the popular Recitfied Linear Unit (ReLU) $\sigma(x) = \max(0, x)$ or, to a lesser extent, the hyperbolic tangent $\sigma(x) = \tanh(x)$.

To conclude, the main advantages of DNN with respect to kernel methods are their expressivity and scalability to be trained on larger and larger datasets.

A wide zoology of networks Nevertheless DNN are not restricted to *feed-forward* neural networks which are particularly suited to regression and classification. Depending on the task $\mathscr{T}$ and the kind of data, we observed emergence of various kind of networks. Notably, CNN are originally inspired from the biology and the visual cortex. By replacing the matrix product by a convolution product, they are particularly suited to processing arrays of numbers such as images in vision and pattern recognition (LeCun et al., 1998; LeCun et al., 1999; Krizhevsky et al., 2012). In contrast with the classical knowledge-based methods where filters to process images are smartly designed by hand, the power of CNN lies on the fact that these filters are directly learned from data. In the context of speech recognition and NLP, to take into account the global meaning of the sentences and correlations between words, *recurrent networks* (Rumelhart et al., 1986a) such as LSTM are quite popular since their high connectivity allows to simulate *memory*.

1.2.10 PRACTICAL ALGORITHMS

To conclude the global overview of the ML machinery, it remains to address algorithmic questions to perform statistical estimation of the model parameters $\boldsymbol{\theta}$. In this manuscript, we essentially focus on two classes of algorithms depending if the ML estimator is formulated as an *optimization* or an *averaging* problem. In one hand, many problems are formulated as minimizing an objective function or maximizing the likelihood with the data distribution, that can be handled by *gradient-based* algorithms. In the other hand, estimators based on the average over certain high-dimensional distributions require either to *sample* or *approximate* it. Notice that there exists other techniques such as *constrained optimization* with Frank–Wolfe algorithms, that we will not discuss in this manuscript.

1.2.10.A GRADIENT-BASED ALGORITHMS

Most of ML algorithms involve the minimization of a certain *smooth* and *differentiable objective* function with respect to the model parameters $\boldsymbol{\theta}$. The most popular objective function is the *negative log-likelihood* with respect to the training set $\mathbb{D}_{\text{train}}$, namely the empirical risk $\hat{\mathscr{R}}(\boldsymbol{\theta};\mathbb{D}_{\text{train}}) = -\mathbb{E}_{(\mathbf{x},y)\sim\hat{\mathrm{P}}(\mathbf{x},y)} \log \mathrm{P}(y|\boldsymbol{\theta},\mathbf{x})$ so that common estimators such as MLE and MAP can be formulated as

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \hat{\mathscr{R}}(\boldsymbol{\theta};\mathbb{D}_{\text{train}}) + \lambda\Omega(\boldsymbol{\theta}), \tag{14}$$

where the additional term $\lambda\Omega(\boldsymbol{\theta})$ may be added for regularization, see Sec. 1.2.6.d. The common strategy to train such parametric estimators is to consider simple first-order *gradient-based* algorithms (Cauchy, 1847) widely popularized with practical applications in (LeCun et al., 1998). Starting with some initial model parameters drawn randomly $\boldsymbol{\theta}^0 \sim \mathrm{P}(\boldsymbol{\theta}^0)$, the underlying idea of a majority of training algorithms consists in performing a *gradient-descent* on the empirical risk. Following the gradient-descent, the algorithm will certainly end up in a local minima, and eventually in a global one with good *generalization properties*. Recall this is the main difference between

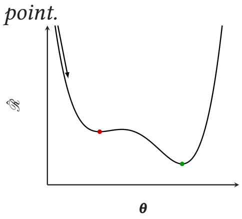
ML and optimization: while the later allows any global minima, the first requires at least a local minima that predicts correctly unseen data, in other words: that generalizes. This simple strategy is commonly known as Gradient-Descent (GD) defined by an update rule that computes in which way the weights $\boldsymbol{\theta}$ should be altered so that the proxy objective $\hat{\mathscr{R}}$ function can reach a minima:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \gamma^{t} \nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}} \left(\boldsymbol{\theta}; \mathbb{D}_{\text{train}} \right) = \boldsymbol{\theta}^{t} - \frac{\gamma^{t}}{n} \sum_{\mu=1}^{n} \nabla_{\boldsymbol{\theta}} \ell \left(y_{\mu}, f_{\boldsymbol{\theta}^{t}} \left(\mathbf{x}_{\mu} \right) \right) , \tag{15}$$

where the hyper-parameter γ^{t}, called the *learning rate*, controls the size of each decreasing gradient step and is usually fixed by performing line search on a validation set. Notice that this idea may be generalized to second-order methods such as the Newton's method that makes use of the second derivative. However, they are very rarely used in practical applications since computing the Hessian matrix, remains inefficient and costly to compute for a large amount of high-dimensional data. GD parameters update rule (15) has the advantage to be easy to implement and to understand, and its trajectory can be analyzed rigorously as soon the objective function is *convex*. Indeed in convex optimization (Boyd et al., 2004), most of algorithms have convergence guarantees by making strong the assumption that the Hessian of the objective function is always positive semi-definite to ensure there is no saddle points and local minima.

Convergence In most practical DL applications, data distribution $\mathrm{P}(\mathbf{x}, y)$ is very complex and the high-dimensional model may contain millions of parameters. The corresponding optimization problem (14) is very often *non-convex* and thus GD algorithm lacks convergence guarantees. In other words, the optimization problem is not guaranteed to converge even to a local minima in a finite time, despite this fact, in practice it often delivers quickly parameters with low values of the objective and good generalization properties.

In the presence of local minima (red), GD is not guaranteed to converge to the global minima (green), as it depends on the initialization point.



Variants and tricks Even though the GD algorithm is easy to understand and implement, it has the disadvantage to be possibly trapped in local minima and to be computationally inefficient on large datasets, since the full gradient has to be computed. Many variants of this simple gradient algorithms, such as Stochastic Gradient-Descent (SGD), have been introduced (Robbins, 1951), where the sum over the gradient of the full training set is replaced by the gradient over a single training example at a time. Thus *stochastic* refers to the randomness in the examples selection at each time step. Very interestingly, it turned out empirically that this variant was able to find other regions of parameters than simple GD, with low test error and therefore good generalization properties. Hence even though convergence is not guaranteed, these algorithms are strongly used in practice as moreover it solves the computational issue of storing in memory the gradient of the full dataset. In between, *mini-batch* GD is a good compromise and computes the gradient over small *batches*, of size n' with $1 \ll n' \ll n$, drawn uniformly from the training set and thus provides a more accurate estimate of the full gradient with some randomness. In the case where the dataset is redundant, this *mini-batch* version has also the advantage to converge faster than GD, since it does not require to explore the whole dataset to capture the underlying distribution. In particular, the size n' of the batch becomes another hyper-parameter we should tune on the validation set. In practice the mini-batch size is typically around hundred while the full batch contains millions of examples. Thus *full-batch* GD corresponds to the classical GD while *1-mini-batch* GD refers to SGD. In particular, these algorithms are widely used because even for infinitely large training set $n \gg 1$, the *complexity* of mini-batch GD remains $\Theta(1)$.

As convergence is still not guaranteed, other tricks have been developed to help and accelerate finding minima and avoid oscillations such as adding *momentum* (Sutskever et al., 2013) and *Nesterov accelerated gradient*. See (Goodfellow et al., 2016) for a detailed review. Also as fixing the learning rate may be tricky, new optimization variants with smart update learning-rate rules came to light, such as *Adagrad* (Duchi et al., 2011), *AdaDelta* (Zeiler, 2012) or *Adam* (Kingma et al., 2014). To conclude, many tricks and techniques on how training efficiently DNN are comprehensively described in (Bottou, 2010). In particular, we may briefly mention that the initialization scheme $\mathrm{P}(\boldsymbol{\theta}^0)$ seems to play a significant role as well as the batch normalization (Ioffe et al., 2015), since these tricks suggest to serve as an *inductive bias* and reduce adequately the effective hypothesis class of DNN.

Back-propagation In contrast with kernel methods which training suffers datasets of large size, DNN became very popular because of their scalability made possible thanks to a simple and robust training algorithm. The main difficulty in training a gradient-based algorithm according to the update (15) lies in *computing the gradient* of this loss with respect to the parameters $\boldsymbol{\theta}$. This has been made possible by the crucial observation that the gradient of the objective (4) with respect to the parameter $\boldsymbol{\theta}$ can be computed by the chain rule using simple algorithmic differentiation (Griewank, 1992). DNN can be trained efficiently, namely in linear time with the size of the network, by applying a simple chain-rule derivative, known as the *back-propagation* algorithm (Rumelhart et al., 1986b).

In more details, by matrix multiplication, adding biases and applying non-linearities across the different layers, the *forward-propagation* of the input $\mathbf{x}_\mu$ gives access to the predicted output $\hat{y}_\mu^t$ of the model and the loss $l(y_\mu, \hat{y}_\mu^t)$ at time t. To fix ideas, consider a two-layer neural network, without bias, of the form

$$\mathbf{z}_1^t = \mathbf{W}_t^{(1)} \mathbf{x}_\mu \,, \qquad \mathbf{z}_2^t = \mathbf{W}_t^{(2)} \sigma^{(1)}\left(\mathbf{z}_1^t\right) , \qquad \hat{y}_\mu^t = \sigma^{(2)}\left(\mathbf{z}_2^t\right)$$

with parameters at time t, $\boldsymbol{\theta}^t = \{\mathbf{W}_t^{(2)}, \mathbf{W}_t^{(1)}\}$. Computing the gradients of the empirical risk (5) with respect to the parameters $\mathbf{W}^{(2)}$, $\mathbf{W}^{(1)}$, for the squared loss $\ell(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2$ is simply performed as a succession of linear operations

$$\begin{aligned}
\partial_{\mathbf{W}^{(2)}} \ell(y_\mu, \hat{y}^t_\mu) &= -(y_\mu - \hat{y}^t_\mu) \cdot \partial_{\mathbf{z}_2} \sigma^{(2)}(\mathbf{z}^t_2) \cdot \sigma^{(1)}(\mathbf{z}^t_1)\,, \\
\partial_{\mathbf{W}^{(1)}} \ell(y_\mu, \hat{y}^t_\mu) &= -(y_\mu - \hat{y}^t_\mu) \cdot \partial_{\mathbf{z}_2} \sigma^{(2)}(\mathbf{z}^t_2) \cdot \mathbf{W}^{(2)}_t \partial_{\mathbf{z}_1} \sigma^{(1)}(\mathbf{z}^t_1) \cdot \mathbf{x}_\mu\,,
\end{aligned}$$

which intermediate results are stored in a computational graph for numerical efficiency. *Back-propagating* the derivatives over the whole DNN up to the input layer gives access to all the parameter updates (15) at time t. Linear in the size of the network and the number of data, it allows to scale the training procedure to very large networks.

To conclude this section, performing GD at each time step, we often monitor the training error e_{train} until convergence. Then we compute the validation error at the end of the training to tune hyper-parameters such as γ, n'. Once the model and hyper-parameters are properly selected, we can finally compute the error on the test set as a surrogate of the generalization performances, see Sec. 1.2.5.b.

1.2.10.B SAMPLING AND APPROXIMATING

As discussed in Sec. 1.2.8, Bayesian estimators such as the MMSE may be formulated instead as an *average* over the posterior distribution $\hat{\boldsymbol{\theta}} = \mathbb{E}_{\mathrm{P}(\boldsymbol{\theta}|\mathbf{y};\mathbf{X})}\left[\boldsymbol{\theta}\right]$. The average can be done explicitly only in cases where the posterior distribution $\mathrm{P}\left(\boldsymbol{\theta}|\mathbf{y};\mathbf{X}\right)$ is explicit and tractable. Unfortunately, in high-dimensions computing it is very often intractable and we shall investigate alternative strategies such as *sampling* or *approximations*. Approximating high-dimensional joint probability distribution such as $\mathrm{P}\left(\boldsymbol{\theta}|\mathbf{y};\mathbf{X}\right)$ is the goal of the *mean field methods* presented later on in Sec. 4.2. Alternatively, among sampling methods, we shall briefly mention *Gibbs sampling* performed with classical MC methods or more performant MCMC variants using notably *importance sampling*. Their simple idea relies on the Central Limit Theorem (CLT) that insures that the integral over the posterior can be approximated as

$$\hat{\boldsymbol{\theta}} = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}\ \mathrm{p}\left(\boldsymbol{\theta}|\mathbf{y};\mathbf{X}\right) \simeq \frac{1}{n_{\text{mc}}} \sum_{\mu=1}^{n_{\text{mc}}} \boldsymbol{\theta}_\mu \text{ where } \boldsymbol{\theta}_\mu \sim \mathrm{P}\left(\boldsymbol{\theta}|\mathbf{y};\mathbf{X}\right)$$

This kind of sampling methods is reviewed in details in (Andrieu et al., 2003; Allison et al., 2013; Craiu et al., 2014) and will not be at the heart of this manuscript, especially because they suffer slow convergence rate in very large dimensions and require *acceleration* and *variance reduction* to sample only useful regions of the high-dimensional probability distribution.

1.3 CHALLENGES AND OPEN QUESTIONS IN DEEP LEARNING

DL was designed to overcome the insufficiency of traditional ML to learn complex high-dimensional functions or probability distributions. It impressively brought unprecedented empirical progresses (LeCun et al., 2015) into various ML applications such as in image, text and speech processing. These recent successes were made possible thanks to the availability of much larger datasets and greater computational resources. Yet, as it relies essentially on ingenious engineering tricks, it brought as well many unanswered fundamental questions that still remain open. Strikingly, the early questions raised in (Breiman, 1995) are still of actuality. Because of this lack of theoretical foundations and guarantees, current practical DL is potentially sub-optimal in the sense the current *brut force* approach takes advantage of the computational efficiency of oversized DNN trained on large dataset.

1.3.1 CURSE OF DIMENSIONALITY AND OPTIMIZATION

The empirical loss minimization problem (14) becomes extremely difficult. With the explosion of the features size, this modern optimization problem lies in a high-dimensional space and was shown to be NP-complete (Blum et al., 1992). Indeed, as the problem dimensions increase, the number of configurations — i. e. the number of possible combinations of the different parameters — is exponential and therefore much larger than the number of training examples in the training set. The *curse of dimensionality*, faced by many computer science tasks in high-dimensions, refers to the statistical challenge to provide accurate predictions on large regions of parameters potentially not explored during the training.

On the other hand, our geometric intuition of this high-dimensional space seems to hit a paradox and trivially concludes that a gradient-based algorithm will naturally fall and remain stuck in one of the many existing local minima, if no additional help is given. In most of the practical cases, this task is doable for a large, but finite number of samples. Yet this remains very inefficient since a human baby that would recognize images of dog and cat with a few pictures whereas current state-of-the-art ML models require millions of images. Moreover, while convergence of gradient-based is guaranteed for quasi-convex loss functions, understanding why GD algorithms do not hit poor generalization local minima remains a burning open question. Indeed, even though minimizing highly-non convex losses is NP-hard, gradient-based algorithms such as GD, SGD or many other variants (Ruder, 2016) strikingly converge to regions of parameters with low generalization error and do not systematically lead to overfitting. Moreover even though many tricks are prescribed to help gradient-based algorithms to converge (Bottou, 2010; Bottou, 2012), building theoretical prescriptions is an active line of ML research.

1.3.2 GENERALIZATION PROBLEM

Classical statistical generalization bounds such as the VC dimensions or the Rademacher complexity are in theory used to justify the learning ability of some ML models. However, nowadays DNN contain millions of parameters and are so large that these classical worst-case statistical bounds became over-pessimistic and fail to predict DNN neural networks behavior. Indeed, as the number of parameters is larger than the number of examples PAC generalization bounds (Vapnik, 2013; Bartlett et al., 2002) predict they should largely overfit and therefore cannot explain their good generalization behavior observed in practice. Moreover, recent works showed that such traditional PAC bounds do not hold in DL, and should be refined. In particular the experimental work of (Zhang et al., 2016) showed that DNN were able to simultaneously learn complex rules as well as fitting random labels. Additionally, the traditional bias-variance trade-off to explain generalization performances is therefore obsolete and it is of actuality to understand why heavily parametrized high-capacity neural networks do not overfit the data (Neyshabur et al., 2017; Arora et al., 2018a). In fact empirical observations suggest that the optimization procedure induce a bias that reduces the effective dimension of neural networks, that can be captured by only a few order parameters. Highlighting them analytically is of course an intense line of research in the statistical learning community.

1.3.3 EXPRESSIVE POWER, UNIVERSALITY AND ARCHITECTURE

Modern ML relying essentially on the ANN and DNN provide a powerful hypothesis class $\mathbb{H}$ with large representation ability as stated by the strong universal approximation theorem (Cybenko, 1989; Hornik, 1991). However, this result for a two-layer *shallow* network is not *constructive* as it does not prescribe the *width*, i. e. the number of hidden units, or the *sample complexity* $\alpha = \frac{n}{d}$, with n the number of training examples d the input dimension, to correctly approximate a given target function $f^\star$, neither the estimator or training algorithm $\mathscr{A}$ to obtain model parameters $\boldsymbol{\theta}$ with good generalization properties. Also, increasing the depth was known for a long time (Minsky et al., 1969) as a solution to overcome simple perceptron limitations, the intuition that depth provides a natural hierarchal framework to learn different scales and representations across layers was recently advocated (Bengio et al., 2013) as well as the analogy with physics renormalization group (Mehta et al., 2014). Such intuition as well as theoretical principles on how to choose model-parameters such as the loss, activations, number of layers, sample complexity or hyper-parameters are fragile and the current understanding remains mainly empirical. On the unsupervised learning counterpart, even though VAE and GAN showed their impressive ability to produce realistic images, measuring the performance of the generative models by knowing in particular if they provide correct approximations of the true data distribution is an important ongoing line of research (Arora et al., 2018b). Mostly based on DNN, they naturally inherit of the theoretical challenges concerning their architecture, computational cost and training procedure in the supervised setting.

1.3.4 OPENING TOWARDS STATISTICAL PHYSICS

To conclude, the successes of DL rely essentially on both the type of structured data and substantial biases of gradient-based algorithms, that allow to reduce the hypothesis class and select an estimator with good generalization abilities. As presented in the next sections, statistical physics has a long history with the theory of ML and we believe that powerful statistical physics tools have a role to play in disentangling the joint roles of the data structure, training algorithm and the network architecture. Moreover, since the classical, overly pessimistic, *worst-case* analysis fails to capture high-dimensional generalization behavior of DNN, the *typical analysis* handled by *statistical mechanics* seems to be a fruitfully alternative approach. Indeed, as usual in physics, by dealing with simple architectures and synthetic data, statistical physics tries to highlight universal properties that will potentially hold in general and moreover for a practical usage. In this perspective, we will consider the simplest theoretical case of *supervised learning* with *feed-forward shallow networks*. This much simplified set-up for deep learning with a few hidden-layer, without convolutions, pooling, batch-normalization, etc., is already complex to understand and is believed to already capture some of the core difficulties. However, so far the simplicity of the models under consideration by the *statistical physics* community is still far away of being realistic to provide direct and practical guidances of the size, architecture, optimization procedure or sample complexity.

In the perspective of handling theoretically simple ML models within this statistical physics framework, we present a general introduction to it in Chap. 2 and see how it may help building theoretical foundations of DL in Chap. 3.

2

AN OVERVIEW OF STATISTICAL PHYSICS AND PHASE TRANSITIONS

In this chapter, we introduce the basic tools and concepts of statistical physics that we will use all along this manuscript. In particular, we advocate that statistical physics is a very powerful framework to describe phase transitions appearing in systems composed of a large number of interacting particles. In Chap. 2.1, we introduce the unfamiliar reader to the fundamental concepts of statistical mechanics. Chap. 2.2 and Chap. 2.3 are respectively devoted to describe the set of mathematical tools of statistical mechanics applied to *ordered* and *disordered* systems.

2.1 WHY STATISTICAL PHYSICS MATTERS?

This manuscript aims to analyze simple *machine learning models* presented in Sec. 1.2 through the singular lens of statistical physics. Even though at the first glance it appears unnatural, in this section we advocate that *statistical physics* is a generic framework that applies to various fields outside of pure physics such as computer science or mathematical problems.

Statistical physics is a branch of physics introduced in the 19th century by Maxwell, Boltzmann and Gibbs, whose objective is to understand the collective behavior that emerges from a system built of *many particles* in interaction. Very powerfully, statistical physics directly applies to various fields, starting with phenomenon observed in everyday life. For instance, without it we could not understand the description of the phase transition between solids, liquids and gas, or the difference of behaviors between metals and insulators, nor even we could not understand supraconductivity, or fermionic and bosons quantum systems (Balian et al., 1986; Georges et al., 2004). But its application range is much wider and goes beyond natural fields. In particular, statistical physics has been successfully applied to social sciences with the Schelling's model (Gauvin et al., 2009), information theory with error correcting codes (Mézard et al., 2009), percolation, combinatorial optimization problem (Krzakała et al., 2007), avalanches in financial and economy modelization (Mantegna et al., 2000; Bouchaud et al., 2003; Voit, 2013), as well as simple machine learning models such as perceptrons (Opper et al., 1991b; Engel et al., 1993), and many other systems. For a more detailed introduction to statistical physics, please refer to (Diu et al., 1989; Sethna et al., 2006; Kardar, 2007; Ma, 2018).

2.1.1 FROM MICROSCOPIC TO MACROSCOPIC SCALES

While historically physics strategy focused on describing macroscopic systems by ignoring the precise microscopic details, statistical mechanics is a *reductionist* and *statistical* description that deduces the macroscopic properties of a system from the microscopic interactions and laws which govern the behavior of its elementary constituents at smaller scales. At the heart of statistical mechanics, the transition from microscopic to macroscopic scales does not offer a completely new description of the nature, but instead adapt existing tools to describe the macroscopic behavior of systems composed of an extensive number of particles. For instance the atmospheric pressure results from collisions between microscopic molecules and statistical mechanics provides a microscopic justification of the laws of thermodynamics that govern macroscopic quantities such as the pressure, temperature and volume.

2.1.2 LAGRANGIAN MECHANICS VERSUS PROBABILITIES

For the sake of illustration, let us consider a simple glass of water that contains typically 10^{23} molecules of water. The large number of particles and degrees of freedom makes the corresponding configuration space so large that tracking over time the positions and speeds of each molecule, which undergoes potentially a huge number of events and interactions, is intractable in practice because of memory usage and precision. Thus the classical Lagrangian mechanics cannot be applied directly to properly describe the behavior of such a simple yet large system. Instead of describing in full details the microscopic states of the system at a given time, statistical mechanics takes advantage of a *probabilistic approach* to only quantify the probability of observing the system in a given microscopic configuration during its evolution.

2.1.3 INTERACTIONS AND COLLECTIVE BEHAVIOR

The emergence of unexpected spectacular *collective behaviors* arises in fact from the interaction of a *very large number d* of particles, that cannot be imagined from the microscopic laws o just a few particles. As an illustration increasing the pressure of our glass of water, at a certain *critical* threshold it will undergo a solidification *phase transition* and becomes solid. This common phenomena cannot be explained theoretically without invoking a *sudden* *change in microscopic interactions* between molecules. These observations are well summarized by the famous formulation from P. W. Anderson (Anderson, 1972) "*More is different*", that stresses the idea that macroscopic behavior cannot be fully described as the sum of non-interacting agents. In other words, the whole system cannot be thought as the simple sum of its components, and interactions play a fundamental role in the macroscopic behavior. Statistical physics aims to analyze the behavior of such macroscopic system and predict the arising critical *phase transition*, such as the classical liquid-gas-solid, para-ferromagnetic or metal-insulator phase transitions. Other spectacular collective behaviors can be observed beyond classical physics systems such as in finance, economy and social sciences in which strongly interacting agents may lead to collective phenomena and rare events such as crashes, reactions of panic and stampedes.

2.1.4 THERMODYNAMIC LIMIT AND CONCENTRATION

As in analytical mechanics, analyzing a very large number d of particles is often intractable. Statistical mechanics makes use of this large size system to describe it in the theoretical infinite size limit, the so-called *thermodynamic limit $d \to \infty$*. This *thermodynamic* limit is a favorable and powerful tool as the behavior of the system becomes asymptotically and surprisingly deterministic! Indeed if we assume that particles are i.i.d, the CLT ensures that the *equilibrium probability distribution* of the system concentrates around its most probable value with fluctuations that *decrease* with the size of the system in $\Theta(d^{-1/2})$. Finally even though in practice particles are very often not i.i.d the behavior of real systems will still be typically given by the thermodynamic limit with some finite-size fluctuations, so that describing the behavior in the thermodynamic limit plays a role of the utmost importance in statistical mechanics.

2.2 DESCRIBING THE SYSTEM BEHAVIOR

Throughout the manuscript, we consider a set of d interacting particles denoted by a vector $\boldsymbol{\sigma} = (\sigma_i)_{i=1}^d \in \chi_d$ which lies in the *configuration space* χ_d. $\forall i \in [\![d]\!], \sigma_i$ belongs to an alphabet χ that represents the degrees of freedom of each spin, and it can be discrete (e. g. $\chi_d = \{\pm 1\}^d$) or continuous (e. g. $\chi_d = \mathbb{R}^d$). The vector $\boldsymbol{\sigma}$ may represent different physical systems such as particles, magnetic spins, pixels, model parameters, etc., depending on the scope of application among image processing, information theory, computer science, physics, biology, error correcting codes or ML. As physicists we will generally call $\boldsymbol{\sigma}$ a vector of *spins* for historical reasons and its values a *configuration* that refers to a given realization. For the purpose of the illustration, we may imagine d magnetic moments, also called spins, that can precess around a

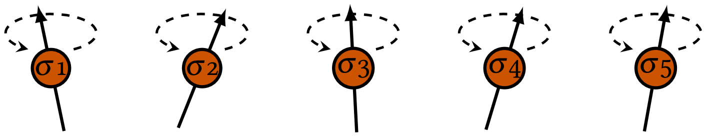


Figure 14: System of magnetic moments that can interact with their neighbors.

vertical axis so that its extremity evolves on a sphere $\mathbb{S}^2 \subseteq \mathbb{R}^3$ and interact with the neighboring spins by magnetic interactions. Each spin configuration $\sigma_i \in \chi = \mathbb{S}^2$ represents the precession degrees of freedom, as illustrated in Fig. 14.

2.2.1 GRAPHICAL MODELS AND FREE ENTROPY

2.2.1.A JOINT PROBABILITY DISTRIBUTION

As stressed in the introduction Sec. 2.1, the probabilistic description of the system is inevitable to analyze collective behaviors of systems with many particles. The behavior of the interacting Random Variable (RV) $\boldsymbol{\sigma} \in \chi_d$ is therefore modelled by a Joint Probability Distribution (JPD)

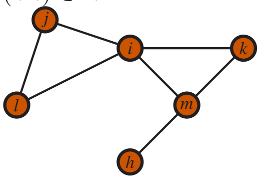
$$\mathrm{P}_d\left(\boldsymbol{\sigma}\right) \equiv \mathrm{P}_d\left(\sigma_1, \cdots, \sigma_d\right) , \tag{16}$$

that may hardly be tractable in large dimensions. The ultimate goal is to compute the marginals distributions $\mathrm{P}(\sigma_i) = \int_{\chi_{d-1}} \mathrm{d}\boldsymbol{\sigma}_{\setminus i} \mathrm{P}_d\left(\boldsymbol{\sigma}\right)$, i. e. the behavior of a single spin variable, by integration over the configurations of the other $d-1$ spins, denoted $\boldsymbol{\sigma}_{\setminus i}$. Analyzing the JPD (16) is a complex task at the heart of this manuscript. Yet very often computationally hard, the computation of the marginals rely in most of the cases on mean-field approximations presented in details in Sec. 4. However, in the simplest case of *non-interacting* spins, the JPD is degenerated and trivially given by the product of the marginal probabilities $\mathrm{P}_d\left(\boldsymbol{\sigma}\right) \equiv \prod_{i=1}^d \mathrm{P}\left(\sigma_i\right)$. Even though the latter factorized decomposition is very useful for approximations, see Sec. 4.2.3.a, the existence of such non-interacting systems is idealist and essentially pedagogical. Therefore the study of non-interacting variables is instructive but very limited in practice. Indeed, interesting and more realistic behaviors mostly appear with the existence of complex interactions between the particles encoded in the JPD.

2.2.1.B GRAPHICAL MODELS

However, representing such complex interacting system is difficult, especially in the high-dimensional regime $d \to \infty$, called the *thermodynamic limit*. Indeed in this limit, the main difficulty lies in describing all the interactions between each spin $\sigma_i, \forall i \in [\![d]\!]$. Therefore, we need a practical way of representing the joint distribution $\mathrm{P}_d(\boldsymbol{\sigma})$ eq. (16). To this extent, we introduce *graphical models* that give a very generic, intuitive and powerful way to think of probabilistic models for finite or infinite size systems. Indeed, interacting spins can be represented conveniently by a *graph* $\mathscr{G}\left(\mathrm{V},\mathrm{E}\right)$, either *directed* or *undirected*, composed of

- a set of nodes V, also called vertices, that represent the spin configuration $\boldsymbol{\sigma} = \{\boldsymbol{\sigma}_i\}_{i=1}^{d}$, so that $|\mathrm{V}| = d$,
- a set of edges E that connects the nodes with $|E| = n$. The edges represent the statistical dependencies, i. e. the interactions, between the random variables $\boldsymbol{\sigma}$. Directed graphs refers to directed interactions of the form $(i \to j)$, while undirected graphs deal with undirected pairs of vertices (ij).



The graph $\mathscr{G}\left(\mathrm{V},\mathrm{E}\right)$ *contains 6 vertices* $h,i,j,k,l,m \in \mathrm{V}$ *and 7 edges* (ij), (ik), (il), (im), (jl), (km), $(hm) \in \mathrm{E}$.

We distinguish *directed* graphical models called Directed Acyclic Graphs (DAG) from *undirected* graphical models known as Markov Random Field (MRF). We will mainly focus on the latter ones in the following and especially in Sec. 2.2.1.c. To describe more formally the geometry of undirected graphs, we often introduce the *adjacency matrix* $\mathbf{A}$ of size $d \times n$ with binary entries such that $a_{ij} = \mathbb{1}\left[(ij) \in \mathrm{E}\right]$. It is in particular useful to compute the *connection degree*, i. e. the size of set of neighbors of a node i, denoted ∂_i

$$|\partial_i| = \sum_j a_{ij} = \sum_j \mathbb{1}\left[(ij) \in \mathrm{E}\right] .$$

For the sake of the illustration, in the case of independent RV, the JPD factorizes and the set of edges of the corresponding graph $\mathscr{G}$ reduces to an empty set $\mathrm{E} = \emptyset$, so that $\forall i \in \mathrm{V}, |\partial_i| = 0$.

This simple graphical formulation gives a convenient and geometrical representation to encode the conditional dependencies of a large number of interacting variables $\boldsymbol{\sigma}$. Interestingly, it will naturally lead to the design of powerful dynamical equations such as the *cavity method* and *belief propagation* discussed in Sec. 4. Notice that even though graphical models may be used for finite size systems, their crucial power lies in their ability to represent as well high-dimensional probability distributions. The interested reader may find a comprehensive introduction with more details about graphical models in (Yedidia et al., 2001a; MacKay et al., 2003; Jordan et al., 2004; Wainwright et al., 2008; Koller et al., 2009)

2.2.1.c GENERAL MARKOV RANDOM FIELDS

On non-regular graphs $\mathscr{G}$ with arbitrary connectivity, counting and describing properly the graph $\mathscr{G}\left(\mathrm{V},\mathrm{E}\right)$ associated with the JPD might be tricky. Fortunately, it is often the case that the RV present a certain structure and independence properties. For this reason, we introduce the notion of *clique*, defined as a subset $\mathrm{C} \subseteq \mathrm{V}$ of fully connected nodes. Indeed the Hammersley and Clifford theorem (Hammersley et al., 1971) insures that if the global independency Markov property

$$\forall\, \mathrm{V}_1, \mathrm{V}_2, \mathrm{V}_3 \subset \mathrm{V}, \;\; \mathbb{P}\left(\boldsymbol{\sigma}_{\mathrm{V}_1 \cup \mathrm{V}_2} | \boldsymbol{\sigma}_{\mathrm{V}_3}\right) = \mathbb{P}\left(\boldsymbol{\sigma}_{\mathrm{V}_1} | \boldsymbol{\sigma}_{\mathrm{V}_3}\right) \mathbb{P}\left(\boldsymbol{\sigma}_{\mathrm{V}_2} | \boldsymbol{\sigma}_{\mathrm{V}_3}\right) ,$$

is verified, the JPD may be decomposed as a general compact MRF

$$\mathrm{P}_d(\boldsymbol{\sigma}) = \frac{1}{\mathcal{Z}_d}\prod_{c\in\mathrm{C}}\Psi_c(\boldsymbol{\sigma}_c) = \frac{1}{\mathcal{Z}_d}\prod_{i=1}^{d}\phi_i(\sigma_i)\prod_{(ij)}\Psi_{ij}(\sigma_i,\sigma_j)\prod_{(ijk)}\cdots, \tag{17}$$

where we introduced some *potential functions* corresponding to cliques with different sizes $\{\Psi_c\}_{c\in\mathrm{C}}$, see (Yedidia et al., 2001a; Jordan et al., 2004). Partitioning over the sizes reveals successive contributions of many-body interactions. For instance, $\{\phi_i\}_i$ represent the one-spin interactions and $\{\Psi_{ij}\}_{i\neq j}$ the two-spin interactions, etc.

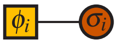
2.2.1.D FACTOR GRAPH REPRESENTATION

The general MRF formulation (17) remains quite cumbersome as the size of the cliques may be very large and involve a large number of spins. Therefore, to obtain a more compact representation of the JPD P_d that highlights the conditional dependencies between RV, it is helpful to replace them with n *factors* or *constraints* $\{\Psi_\mu(\boldsymbol{\sigma}_{\partial_\mu}) : \mu \in [\![n]\!]\}$ that are already factorized, where ∂_μ denotes the subset of neighboring nodes of the factor μ. The JPD can therefore be written in full generality as the product over all possible factors

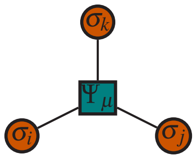
$$\mathrm{P}_d(\boldsymbol{\sigma}) = \frac{1}{\mathcal{Z}_d}\prod_{i=1}^{d}\phi_i(\sigma_i)\prod_{\mu=1}^{n}\Psi_\mu\left(\boldsymbol{\sigma}_{\partial_\mu}\right), \tag{18}$$

where $\mathcal{Z}_d = \sum_{\boldsymbol{\sigma}}\prod_{i=1}^{d}\phi_i(\sigma_i)\prod_{\mu=1}^{n}\Psi_\mu(\boldsymbol{\sigma}_{\partial_\mu})$ represents a normalizing constant. The above factorization ends up with a simple *bipartite factor graph representation* $\mathcal{G} = (\mathrm{V},\mathrm{F},\mathrm{E})$ of the JPD composed of *variable nodes* $\boldsymbol{\sigma} \in \mathrm{V}$ represented by circles and *factor nodes* $\phi,\Psi \in \mathrm{F}$ represented with squares, connected with edges E, as illustrated in Fig. 15. In particular, each *non-negative* factor Ψ_μ is connected to neighboring variables $\boldsymbol{\sigma}_{\partial_\mu} = \{\sigma : \sigma \in \partial_\mu\}$.

For clarity, local fields or equivalently one-body interaction are described by leaf factors and drawn in yellow



and many-body interactions are represented with green constraint factors



The factor graph formalism provides a powerful and very convenient representation of the JPD that gained a lot of interest in various fields such as constraint satisfaction and combinatorial optimization, error-correcting codes, bioinformatics, language and speech processing, image processing and spatial statistics. See a review of a wide range of applications in (Wainwright et al., 2008; Koller et al., 2009).

2.2.1.E CONNECTION WITH PHYSICS: HAMILTONIAN AND THE GIBBS MEASURE

The connection between physics and the factor graph formalism can be made explicit if we exponentiate the above formulation (18) to introduce the fundamental *Hamiltonian* energy $\mathcal{H}_d$ that measures the interaction energies of the local fields $\log \phi_i\left(\sigma_i\right)$ applied point-wise on each spin $\sigma_i, \forall i \in \llbracket d \rrbracket$ and the n constraints $\log \Psi_\mu\left(\boldsymbol{\sigma}_{\partial_\mu}\right)$ that the spin configuration $\boldsymbol{\sigma}$ shall satisfy

$$\mathscr{H}_d\left(\boldsymbol{\sigma}\right) = -\sum_{i=1}^{d} \log \phi_i\left(\sigma_i\right) - \sum_{\mu=1}^{n} \log \Psi_\mu\left(\boldsymbol{\sigma}_{\partial_\mu}\right) . \tag{19}$$

The Hamiltonian of the system describes the microscopic interactions between spin variables $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{d}$ so that the corresponding energy $\mathscr{H}_d\left(\boldsymbol{\sigma}\right)$ measures the probability of each configuration $\boldsymbol{\sigma}$ according to the *Gibbs distribution*, also called the *Boltzmann* distribution

$$\mathrm{P}_d(\boldsymbol{\sigma};\beta) \equiv \frac{e^{-\beta\mathscr{H}_d(\boldsymbol{\sigma})}}{\mathscr{Z}_d(\beta)} , \tag{20}$$

where we introduced a parameter β, called the *inverse temperature*, that allows to explore all energy levels above the ground state energy. The inverse temperature will be taken to $\beta = 1$ in the most considered cases unless mentioned otherwise. In this case, the Gibbs distribution (20) is equivalent to the JPD formulation in (18). The Gibbs distribution is the central equilibrium measure in statistical physics and its exponential form can be justified by the *maximum entropy principle* detailed in Sec. 4.2.2.b. Notice that by construction, the most probable configuration is the one that achieves the smallest Hamiltonian energy (19). It is called the *ground state* configuration and is associated to a *ground state energy*. Moreover, notice that we introduced the normalizing constant at inverse temperature β of the random measure dP_d, called the *partition function*. Indeed imposing the normalization $\int_{\chi_d} \mathrm{dP}_d(\boldsymbol{\sigma};\beta) = 1$, the partition function is naturally given by the *sum over all the possible configurations* weighted by their Gibbs weights probability $e^{-\beta\mathscr{H}_d(\boldsymbol{\sigma})}$:

$$\mathscr{Z}_d\left(\beta\right) \equiv \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\ e^{-\beta\mathscr{H}_d(\boldsymbol{\sigma})} . \tag{21}$$

The partition function is a crucial quantity in statistical mechanics as it contains the important informations on the equilibrium distribution of all possible spin configurations of the system. Indeed $\mathscr{Z}_d\left(\beta\right)$ is known to be the *moment generating function*, because successive derivatives give access to the moments of the Gibbs measure. The *Gibbs average* over the Gibbs measure (20) is traditionaly denoted $\langle . \rangle_\beta$, and we may also use the notation $\mathbb{E}_{\boldsymbol{\sigma}\sim\mathrm{P}_d}$.

2.2.1.F THE FREE ENTROPY AS A CUMULANT GENERATING FUNCTION

Because the JPD P_d of the spin $\boldsymbol{\sigma}$ becomes exponentially peaked in regions of most probable configurations that dominate the whole distribution, we are only interested in its *large deviation* behavior. Thus, taking the logarithm of the partition function refers and defines the *free entropy* in information theory and statistical physics.

We may choose as well the free energy as historically in statistical physics. Notice that the literature is sometimes confusing and clumsy on the naming of these quantities.

Free entropy and energy We define respectively the *free entropy* Φ_d and free energy φ_d of a system of size d at inverse temperature β by:

$$\Phi_d(\beta) \equiv \frac{1}{d} \log \mathscr{Z}_d(\beta) , \qquad \varphi_d(\beta) \equiv -\frac{1}{d\beta} \log \mathscr{Z}_d(\beta) . \tag{22}$$

In order to avoid confusion with sign conventions and temperature prefactors, we mainly consider the free entropy as our central object of study. As stressed in the introduction Sec. 2.1, in the *thermodynamic limit $d \to \infty$* the free entropy of many systems *concentrate* around an asymptotic and deterministic value given (when it exists) by

$$\Phi(\beta) \equiv \lim_{d\to\infty} \Phi_d(\beta) , \qquad \varphi(\beta) \equiv \lim_{d\to\infty} \varphi_d(\beta) . \tag{23}$$

As the finite size behavior generally fluctuates around these asymptotic quantities, their computation is of crucial interest to understand the collective behavior of the system. In particular their study reveals potential phase transitions, as illustrated in the analysis of phase transitions in Sec. 2.2.2 and the presentation of simple examples in Sec. 2.2.3.

Large deviation principle The large deviation theory deals with the exponential decay of the JPD of random systems. In this paragraph, we provide justifications of the fact that computing the free entropy Φ in the study of equilibrium properties of systems with many-particles in interaction is equivalent to a large deviation theory. See (Oono, 1989; Varadhan, 2008; Touchette, 2008) for an extended review. Let $\boldsymbol{\sigma} \in \chi_d$, we say that the JPD $\mathrm{P}_d\left(\boldsymbol{\sigma}\right)$ satisfies a large deviation principle with rate $\mathscr{S}$ if

$$-\log \mathrm{P}_d\left(\boldsymbol{\sigma}\right) = d\mathscr{S} + o\left(d\right) \Rightarrow -\lim_{d\to\infty} \frac{1}{d} \log \mathrm{P}_d\left(\boldsymbol{\sigma}\right) \equiv \mathscr{S} ,$$

which is equivalent to say that the *dominant* behavior of P_d is decaying exponentially with the size of the system and is controlled by the rate function $\mathscr{S}$, called the *entropy* in physics. Indeed, the Gartner-Ellis theorem (Gärtner, 1977; Ellis et al., 1984) draws an explicit connection between the large deviation principle, the entropy and free entropy. Assuming the latter exists and is differentiable for any temperature $\beta \in \mathbb{R}$,

$$\Phi(\beta) = \lim_{d\to\infty} \frac{1}{d} \log \int_{\chi_d} \exp\left(-d\beta \mathscr{H}_d(\boldsymbol{\sigma})\right) \mathrm{d}\boldsymbol{\sigma} ,$$

it states that the JPD verifies a large deviation principle

$$\lim_{d\to\infty} -\frac{1}{d} \log \mathrm{P}_d\left(\mathscr{H}_d(\boldsymbol{\sigma}) = \mathscr{E}\right) = \mathscr{S}\left(\mathscr{E}\right) ,$$

where the rate function is given by the *entropy* $\mathscr{S}(\mathscr{E}) = \max_\beta \left(\Phi(\beta) + \beta\mathscr{E}\right)$ obtained by a Legendre transform detailed in Sec. 2.2.2.b. Back to statistical physics, proving the existence of a *thermodynamic limit* of the free entropy

$\Phi(\beta)$ in (23) is therefore equivalent to prove a large deviation principle of the Gibbs measure.

Cumulant generative function Finally, similarly to the partition function (21), the free entropy has the advantage to encode for all the useful informations of the system. It can be seen as a *cumulant generative function*. Namely successive cumulants of the Gibbs distribution can be obtained by taking higher order derivatives. In particular, the free entropy gives access to important quantities such as the *magnetization*, the corresponding *average energy* or the *ground state energy* associated to ground state configuration. For the sake of illustration, we assume that the local one-body interaction simply reads, as very often in physics, $\forall i \in [\![d]\!]$, $\log \phi_i(\sigma_i) = h_i \sigma_i$, where $\mathbf{h} = \{h_i\}_{i=1}^d$ is called the *external field*.

- *Magnetization*

The *magnetization* m_d at zero external field $\mathbf{h} = \mathbf{0}$ is defined as the averaged value of the spin configuration over the Gibbs distribution. It is simply obtained by taking the derivative of the free energy (22) with respect to the vanishing external field $\mathbf{h} \to \mathbf{0}$

$$\begin{aligned} m_d &\equiv \left\langle \frac{1}{d} \sum_{i=1}^d \sigma_i \right\rangle_\beta = \frac{1}{d \mathcal{Z}_d(\beta)} \int_{\chi_d} \left(\sum_{i=1}^d \sigma_i \right) e^{-\beta \mathcal{H}_d(\boldsymbol{\sigma})} \mathrm{d}\boldsymbol{\sigma} \\ &= - \lim_{\mathbf{h} \to \mathbf{0}} \partial_{\mathbf{h}} \varphi_d(\beta) \,. \end{aligned} \tag{24}$$

- *Average energy and variance*

The average energy at zero external field $\mathbf{h} = \mathbf{0}$ is simply the Gibbs average of the Hamiltonian energy given by

$$e(\beta) \equiv \left\langle \frac{\mathcal{H}_d(\boldsymbol{\sigma})}{d} \right\rangle_\beta = - \lim_{\mathbf{h} \to \mathbf{0}} \partial_\beta \Phi_d(\beta) \,, \tag{25}$$

while the second cumulant, the variance, is naturally given by the second derivative of the free entropy Φ_d

$$\frac{1}{d} \left(\langle \mathcal{H}_d(\boldsymbol{\sigma})^2 \rangle_\beta - \langle \mathcal{H}_d(\boldsymbol{\sigma}) \rangle_\beta^2 \right) = \lim_{\mathbf{h} \to \mathbf{0}} \partial^2_{\beta^2} \Phi_d(\beta) \,. \tag{26}$$

- *Ground state energy*

The ground state energy $e_{\text{gs},d}$ is the minimum energy that can be reached by at least one configuration $\boldsymbol{\sigma}$. It can be computed by taking the *zero-temperature* limit $\beta \to \infty$ of the Gibbs random measure

$$e_{\text{gs},d} \equiv \min_{\boldsymbol{\sigma} \in \chi_d} \left\{ \frac{\mathcal{H}_d(\boldsymbol{\sigma})}{d} \right\} = \lim_{\beta \to \infty} \left\langle \frac{\mathcal{H}_d(\boldsymbol{\sigma})}{d} \right\rangle_\beta \,. \tag{27}$$

2.2.1.G ILLUSTRATION OF SIMPLE GRAPHICAL MODELS

For the sake of clarification, in this section we briefly present some simple and common models and their graphical representation such as the *k-SAT* problem, general tree factor graphs and regular pairwise MRF.

K-SAT problem The k-SAT problem is a Constraints Satisfaction Problem (CSP) at the interface between information theory and error correcting codes. It is specified by d boolean variables $\boldsymbol{\sigma} \in \chi_d = \{0,1\}^d$ that must verify simultaneously the AND logical operator of n constraints with k-body interactions, also called *clauses*, that depend on a subset of k boolean variables. Random CSP are a variant in which the clauses are drawn from a random ensemble. As generally in CSP, such as the graph-coloring problem, the traveling salesman problem and many others, the problem can be easily described by a factor graph eq. (18) (Dechter et al., 1988). Namely for the k-SAT problem, each factor denotes a *hard constraint* represented by the indicator function $\log \Psi_\mu = \mathbb{1}\left(\boldsymbol{\sigma}_{\partial\mu}\right)$ so that the Hamiltonian energy counts the number of satisfied clauses

$$\mathscr{H}_d(\boldsymbol{\sigma}) = -\sum_{\mu=1}^{n} \mathbb{1}\left(\boldsymbol{\sigma}_{\partial\mu}\right) . \tag{28}$$

The ground state configuration is reached if the n clauses are verified such that the Hamiltonian energy $\mathscr{H}_d$ is minimal. As an illustration we give the JPD of a k-SAT problem realization at zero temperature for $k=3$, $d=4$ and $n=3$

$$\mathrm{P}_d\left(\boldsymbol{\sigma}\right) \propto \underbrace{\mathbb{1}\left(\bar{\sigma}_1 \vee \bar{\sigma}_2 \vee \bar{\sigma}_4\right)}_{\Psi_1(\boldsymbol{\sigma}_{\partial 1})} \underbrace{\mathbb{1}\left(\sigma_2 \vee \sigma_3 \vee \sigma_4\right)}_{\Psi_2(\boldsymbol{\sigma}_{\partial 2})} \underbrace{\mathbb{1}\left(\sigma_1 \vee \sigma_2 \vee \sigma_3\right)}_{\Psi_3(\boldsymbol{\sigma}_{\partial 3})} , \tag{29}$$

where $\vee$ denotes the OR operator and $\bar{\sigma}$ the negation of a boolean variable. The corresponding factor graph is represented in Fig. 15 (**Left**).

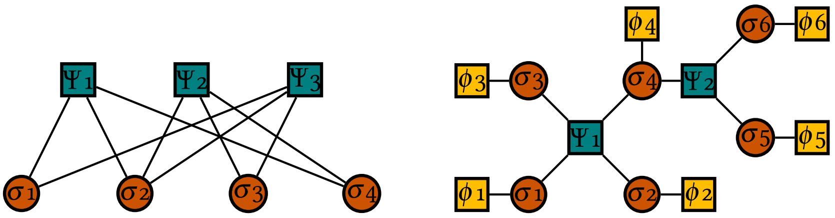


Figure 15: Factor graph representations: the red circles represent the spin variables, the green squares represent the factors that account for statistical dependencies between the variables and the yellow squares represent the single variable factors. (**Left**) A 3-SAT problem in (29). (**Right**) A tree factor graph in (30).

Notably one important challenge in CSP is to compute the maximum constraints density that can satisfy the spin variables $\boldsymbol{\sigma}$, called the SAT-*threshold* $\alpha_c = n_c/d$. The SAT-threshold α_c can be computed from the analysis of the JPD or the free entropy Φ_d in (22), by probing the existence of at least one configuration with strictly positive probability in the thermodynamic limit. See (Mézard et al., 2009) and Sec. 3.2.3 for more details.

Tree factor graphs Tree-like factor graphs are a class of graphical model with the advantageous property of not presenting any loops, meaning that interactions are local and involve only the nearest neighboring spins. Tree factor graphs play an important practical and theoretical role because the full graph can be scanned with linear time complexity $\Theta(d)$ and inference can be performed exactly. As an illustration, we give an example of a tree-like JPD represented in Fig. 15 (**Right**),

$$\mathrm{P}_d(\boldsymbol{\sigma}) = \frac{1}{\mathcal{Z}_d}\prod_{i=1}^{6}\phi_i(\boldsymbol{\sigma}_i)\Psi_1(\sigma_1,\sigma_2,\sigma_3,\sigma_4)\Psi_2(\sigma_4,\sigma_5,\sigma_6)\,. \tag{30}$$

Pairwise Markov random fields Among general MRF models, a large class of common models focuses on regular factor graphs with at most *pairwise interactions* known as *pairwise Markov random fields*. Therefore, we consider the JPD in eq. (18) by absorbing all potentials Ψ_μ with strictly more than two-body interactions, such that the JPD simply reads

$$\mathrm{P}_d(\boldsymbol{\sigma}) = \frac{1}{\mathcal{Z}_d}\prod_{i=1}^{d}\phi_i(\sigma_i)\prod_{(ij)}\Psi_{ij}(\sigma_i,\sigma_j)\,, \tag{31}$$

where we have introduced pairwise symmetric potentials $\Psi_{ij} = \Psi_{ji} = \Psi_\mu$ by re-indexing all interacting pairs $\mu = (ij) = (ji)$, with $\mu \in [\![n]\!]$. The corresponding Hamiltonian (19) simplifies to

$$\mathcal{H}_d(\boldsymbol{\sigma}) = -\sum_{i=1}^{d}\log\phi_i(\sigma_i) - \sum_{(ij)}^{n}\log\Psi_{ij}(\sigma_i,\sigma_j)\,. \tag{32}$$

A large part of the statistical physics literature focuses on such theoretical pairwise MRF on regular lattices, called alternatively *Ising-like models*. For the sake of illustration, we consider such a system of magnetic spins, illustrated in Fig. 14, immersed in an uniform external magnetic field $\mathbf{h}$ and local neighboring interactions. This system can be represented by a spin model $\boldsymbol{\sigma} \in \chi_d$ associated to a graph $\mathcal{G}(\mathrm{V},\mathrm{E})$ with pairwise exchange interactions described by a matrix $\mathbf{J} \in \mathbb{R}^{d\times d}$ and local external fields $\mathbf{h} \in \mathbb{R}^d$, so that

$$\log\Psi_{ij}(\sigma_i,\sigma_j) = J_{ij}\sigma_i\sigma_j\,, \qquad\qquad \log\phi_i(\sigma_i) = h\sigma_i\,.$$

The corresponding factor graph is represented in Fig. 16. Notice that the energy term $-\log\Psi_{ij}(\sigma_i,\sigma_j) = -J_{ij}\sigma_i\sigma_j$ is a convention that insures that for ferromagnetic interactions $J_{ij} > 0$, the energy term decreases the total Hamiltonian energy if the spins σ_i and σ_j are aligned. The one-body interaction term $-\log\phi_i(\sigma_i) = -h\sigma_i$ represents the interaction of each spin with a uniform external field h that tends to align all the spins in its direction. The coupling constants J_{ij} represent the strength of the *pairwise interaction* between the spin σ_i and σ_j. In particular the interactions may be positive $J_{ij} > 0$ or negative $J_{ij} < 0$ and the corresponding models are respectively qualified of *ferromagnetic* and *antiferromagnetic*. The many variants of this general model depend mainly on the geometry and connectivity of the interactions E, the distribution $\mathrm{P}(\mathbf{J})$ of the interaction matrix $\mathbf{J}$ and the configuration space χ_d, that lead to a profusion of theoretical models. We briefly recall the different well-know pairwise MRF models with *continuous* or *discrete* variables and pairwise interactions such as the Ising, Potts, XY, Heisenberg and $\Theta(N)$ models.

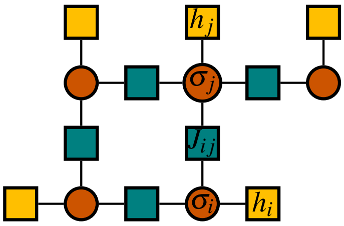


Figure 16: Factor graph of an Ising-like model on a regular lattice.

- *Discrete models: Ising and Potts*

Discrete models such as the Potts (Wu, 1982) model assumes that each spin lies in a discrete alphabet $\chi = \mathbb{Z}^q$ with q characters. For instance, it has been notably considered for *hyper-graph coloring* problems, where each alphabet value represent a color, so that positive interaction happen only if interacting spins have the same color $\Psi_{ij}(\sigma_i, \sigma_j) = J_{ij}\delta\left(\sigma_i, \sigma_j\right)$. In particular, the model reduces to the classical Ising model with two colors for $q = 2$ with $\chi_d = \{\pm 1\}^d$.

- *Continuous models: continuous Ising, Heisenberg, XY, $\Theta(N)$*

The $\Theta(N)$ model (Stanley, 1968; Gennes, 1972; Gaspari et al., 1986) considers instead d continuous variable in N dimensions so that the interaction term depends on the scalar product between vectorial spins $\Psi_{ij}(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j) = J_{ij}\boldsymbol{\sigma}_i \cdot \boldsymbol{\sigma}_j$. For $N = 1$, we recover the continuous Ising model ($\sigma \in \mathbb{R}$). The cases $N = 2$ and $N = 3$ correspond respectively to the XY model ($\boldsymbol{\sigma} \in \mathbb{R}^2$) and the Heisenberg model ($\boldsymbol{\sigma} \in \mathbb{R}^3$).

2.2.2 PHASE TRANSITIONS TYPOLOGY

One of the most spectacular consequences of interactions among particles is the emergence of collective behaviors that would not have been observed in the presence of only a few particles. Indeed in nature, many physical compounds exist under different forms, also called *phases* or *states*. As you change the macroscopic variables of a large system, called *order parameters*, sometimes the system will abruptly change and move to another *phase*. As these *phase transitions* affect dramatically the macroscopic behavior and properties of the system, they shall correspond to singularities in the free energy. Therefore, studying the free energy, that explicitly describes the interplay between energy and entropy contributions, is crucial to detect phase transitions.

2.2.2.a A FIRST PHASE TRANSITION: THE SOLID-LIQUID-GAS PHASE TRANSITIONS

For the sake of clarification, let us consider the simplest example observable in everyday life: the phase transitions of water. Consider a *solid* ice cube at low temperature T and constant pressure P. Increasing the temperature (or decreasing the pressure), we may observe the ice transforming in the *liquid* water at $T_{\text{fus}} = 0°C$ before vaporizing in the water *vapor* at $T_{\text{vap}} = 100°C$. The successive transformations are represented in the *phase diagram* in Fig. 17 and the different states are delimited by some solid lines which represent where the phase transitions occur. Notice that during the phase transitions, the system is still composed of the same number of particles of water. The only difference is the trade-off between the *energetic* and *entropic* contributions in the free energy that are modified so that the system adapts to the most stable collective configuration with the *lowest free energy*. We could mention as well the *ferromagnetic–paramagnetic* phase transition of a metal, that will be discussed with the Ising and Currie-Weiss models in Sec. 2.2.3.b . The analysis of this kind of phase transition in physical systems is particularly suited to statistical physics, whose phenomenology is general and applies to other fields such as information theory, optimization, computer science, biology, social sciences, economy, etc.

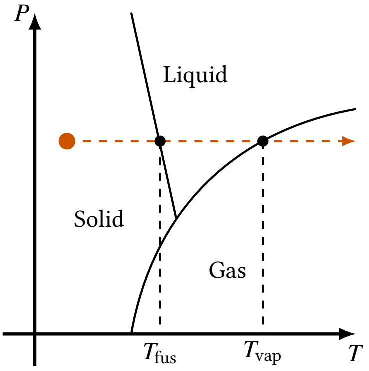


Figure 17: Phase diagram (T,P) of water. Increasing the temperature T at constant pressure P, the ice will melts at T_{fus} and vaporizes at T_{vap}.

2.2.2.b ENERGY-ENTROPY DECOMPOSITION: LEGENDRE TRANSFORM

We stressed that the Gibbs measure often verifies a large deviation principle and is peaked in the most probable regions. However, we did not discuss *which* or *how many configurations* contribute to this dominant equilibrium configurations. Indeed counting the number of configurations at a certain energy level that participates to the partition sum and the free entropy Φ_d is very instructive to introduce the energy-entropy decomposition. Let us denote Ω the number of configurations or equivalently the volume of phase space that achieve a given *energy* $\mathscr{E}$.

Legendre transform The partitioning sum $\mathscr{Z}_d(\beta)$ in (21) can be partitioned instead over configurations with a particular level of energy $\mathscr{E} \equiv \frac{1}{d}\mathscr{H}_d(\boldsymbol{\sigma})$. Introducing $\Omega(\mathscr{E}) = \int_{\mathbb{R}} \mathrm{d}\boldsymbol{\sigma}\, \delta\left(\mathscr{E} - \mathscr{H}_d(\boldsymbol{\sigma})\right)$ the number of such configurations, it is linked to the entropy density $\mathscr{S}(\mathscr{E}) \equiv \frac{1}{d}\log\Omega(\mathscr{E})$. With this decomposition, the partition function writes

$$\mathscr{Z}_d(\beta) \equiv \exp\left(d\Phi_d(\beta)\right) \equiv \int_{\chi_d} e^{-\beta\mathscr{H}_d(\boldsymbol{\sigma})}\mathrm{d}\boldsymbol{\sigma}$$
$$= \int_{\mathbb{R}} e^{-\beta d\mathscr{E}}\Omega(\mathscr{E})\,\mathrm{d}\mathscr{E} \equiv \int_{\mathbb{R}} e^{d(\mathscr{S}(\mathscr{E})-\beta\mathscr{E})}\mathrm{d}\mathscr{E}$$

so that using a Laplace method (Wong, 1989) in the thermodynamic limit $d \to \infty$, we obtain

$$\Phi(\beta) = \max_{\mathscr{E}}\left(\mathscr{S}(\mathscr{E}) - \beta\mathscr{E}\right) = \mathscr{S}(\mathscr{E}^\star) - \beta\mathscr{E}^\star \tag{33}$$

where the equilibrium energy $\mathscr{E}^\star$ verifies $\partial_{\mathscr{E}}\mathscr{S}|_{\mathscr{E}=\mathscr{E}^\star} = \beta$. This last formulation shows that the *free entropy* $\Phi(\beta)$ is the Legendre transformation (Zia et al., 2009) of the *entropy* $\mathscr{S}(\mathscr{E})$. Notice that the free entropy $\Phi(\beta)$ is a function of the inverse temperature β, which plays the role of a *control parameter*, while the entropy $\mathscr{S}(\mathscr{E})$ is a function of the *response parameter* the energy $\mathscr{E}$. Indeed the main advantage of the Legendre transformation is to exchange the role of the variables associated with control and response.

Inverse Legendre transform Similarly, with the definitions of $\mathscr{Z}_d$ and Ω, we can introduce the inverse Laplace transform that leads to the inverse Legendre transform

$$\Omega(\mathscr{E}) \equiv e^{d\mathscr{S}(\mathscr{E})} = \int_{\mathbb{R}} \mathscr{Z}_d(\beta)e^{\beta d\mathscr{E}}\mathrm{d}\beta = \int_{\mathbb{R}} \exp\left(d\left(\Phi_d(\beta) + \beta\mathscr{E}\right)\right)\mathrm{d}\beta\,,$$

and using again a Laplace method in the thermodynamic limit $d \to \infty$, we obtain that the entropy $\mathscr{S}(\mathscr{E})$ is the Legendre transform of the free entropy $\Phi(\beta)$:

$$\mathscr{S}(\mathscr{E}) = \max_{\beta}\Phi(\beta) + \beta\mathscr{E} \Leftrightarrow \mathscr{S}(\mathscr{E}) = \Phi(\beta^\star) + \beta^\star\mathscr{E}\,, \tag{34}$$

where the critical temperature $\beta^\star$ is such that the slope of the free entropy verifies $\partial_\beta\Phi|_{\beta=\beta^\star} = -\mathscr{E}$.

Free entropy decomposition and collective behaviours The Legendre transform (33) reveals that the free energy, or respectively the free entropy, decomposes in two contributions: the energy $\mathscr{E}^\star$ and the entropy $\mathscr{S}\,(\mathscr{E}^\star)$:

$$\varphi(\beta) = \mathscr{E}^\star - \frac{1}{\beta}\mathscr{S}(\mathscr{E}^\star)\,.$$

This decomposition is crucial to understand the emergence of collective behaviors in large systems. Indeed without the entropic term, the free energy would simply be given by the energetic term $\mathscr{E}^\star$ that measures the *cost* of a typical configuration. It will not change when the inverse temperature β or any other control parameter is modified. The appearance of macroscopic collective behavior happens therefore as soon as the number of equilibrium configurations $\Omega^\star$ scales exponentially with the size of the system so that the entropic term $\mathscr{S}(\mathscr{E}^\star)$ becomes comparable to the energetic term $\mathscr{E}^\star$. The inverse temperature is a free parameter that plays the role of a tension between the energy and the entropy, and controls the trade-off between being in a disordered phase with a high entropy or in an ordered phase with a low energy.

Phase transition: first and second order More generally, the free entropy (or the free energy) is the central object of study to analyze the behavior of large systems because very interestingly it captures the behavior of the different phases according to some carefully chosen *order parameters*, denoted $\mathbf{q}$. Indeed, non trivial behaviors and singularities in the free entropy reveal the *phase transitions*, which are very often abrupt and happen at precise values of the order parameters. When it is possible, the free energy of the large size interacting system is computed and mapped to an extremization problem over a set of order parameters:

$$\varphi\,(\beta) = \mathbf{extr}_{\mathbf{q}}\left\{\Psi(\mathbf{q},\beta)\right\}\,, \tag{35}$$

where we introduced a variational free energy $\Psi(\mathbf{q},\beta)$ that depends on the inverse temperature β and the order parameters $\mathbf{q}$. The equilibrium behavior of the system can be analyzed by solving this simple optimization problem.

Interestingly, depending on the shape of the variational free energy $\Psi(\mathbf{q})$, we can observe multiple local and global minima that lead to different kind of phase transitions. In general, as well as in this manuscript, phase transitions can be classified into two main classes: *discontinuous* first order and *continuous* second order transitions. This has been formalized by the Landau theory reviewed in (Toledano et al., 1987). It allows to describe the phase transition phenomenology with a simple formalism by assuming the variational free energy may be written as a polynomial in a scalar order parameter q: $\Psi(q,\{\alpha_i\}_i) = \sum_{\{\alpha_i\}}\alpha_i q^i$. Depending on how the shape of the variational free entropy evolves with the control parameter, we may observe continuous or discontinuous phase transitions of the stable phase, which is the one with the lowest free energy. A *second order* phase transition happens when the order parameter of the most stable phase *evolves continuously*. In contrast, a *first order* transition is observed when a local minima becomes at some point the global minima, so that the order parameter *jumps from one phase to the other*. For instance, it is the case of the liquid-solid phase transition. First and second order phase transitions are illustrated in Fig. 18. Interestingly with the Landau approach, distinct systems can be gathered in the same universality class, characterized by the same non-zero coefficients in their variational free energy, such that their phase transition description is identical. Indeed, the *critical exponents*, that describe the behavior of the order and control parameters close to the phase transitions, are believed to be *universal* and can be computed with *renormalization group* (Wilson, 1983) technics.

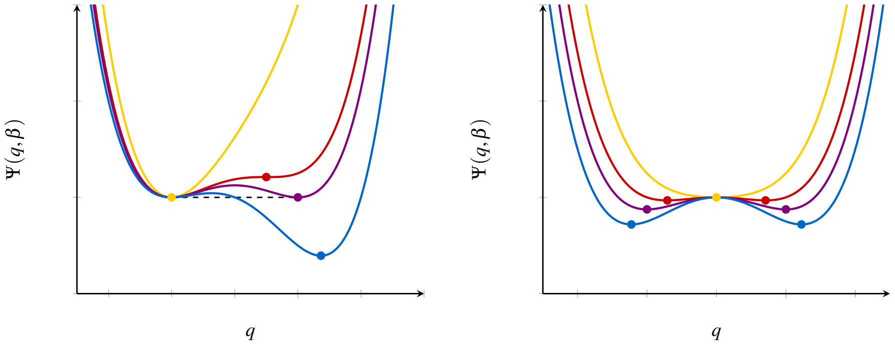


Figure 18: (**Left**) First order phase transition for $\Psi(q) = \beta q^2/2 - q^3 + q^4/4$. For $\beta \ll 1$, the free energy has a single minimum $q = 0$ (yellow). At $\beta = \beta_{\rm sp}$, a second local minimum with higher free energy appears (red) and the most stable phase remains in $q = 0$. At $\beta = \beta^\star$, there exists two global minima (violet). For $\beta > \beta^\star$, the order parameter jumps from 0 to $q > 0$ (blue). (**Right**) Second order phase transition for $\Psi(q) = \beta q^2/2 + q^4/4$. For $\beta \ll 1$, the free energy has a single minimum (yellow) in $q = 0$. At $\beta = \beta^\star$, this minimum becomes unstable and two global minima appear continuously (red) and becomes more and more stable (violet-blue).

Metastable phases and ergodicity breaking The variational free energy landscape can be complex with the presence of various local minimum. As a consequence, *initializing* the system in a configuration close to such a locally stable *state*, if the system is not perturbed, it will remain in this phase. However large fluctuations can destroy this local stability and in this case the system should adapt and move to another phase that corresponds to the global minima of the free energy. This kind of locally stable minima is called a *metastable state*, i. e. a state that remains stable if the system is not perturbed too much. Such systems undergo a harmful *ergodicity breaking* of the phase space, which means that the *ensemble average* and the *time average* are no longer equal and breaks the fundamental hypothesis of statistical mechanics. Indeed, by initializing the system in any metastable state, the system should visit all other possible states, eventually after an infinite time. Yet, on finite time scales, we could only observe the system in this state even though it is not the global minima of the free energy. After a finite amount of time, since the JPD remains unchanged, we could conclude that the system reached equilibrium. But in the case of non-ergodic systems, such as structural glasses, they remain stuck in a small portion of the configuration space and do not reach the globally stable equilibrium configuration.

To conclude this introduction, we illustrate and apply these notions of metastability with the analysis of the para/ferro-magnetic phase transition on the celebrated Ising model.

2.2.3 A CLASSICAL EXAMPLE OF LATTICE MODEL

2.2.3.A THE ISING MODEL

To describe ferromagnetism observed in metals, the battle-horse model of standard statistical physics is certainly the *Ising model.* Indeed, it is the simplest regular pairwise MRF, that describes the collective behavior of magnetic *spins* $\boldsymbol{\sigma} \in \chi_d = \{\pm 1\}^d$. In general an Ising-like model is defined on a graph $\mathscr{G}(\mathrm{V},\mathrm{E})$, so that spins lie on the vertices of the graph V and interacts with neighbors, defined by the edges E, through exchange interactions $\mathbf{J} \in \mathbb{R}^{d\times d}$ and a potential external field $\mathbf{h} = h \cdot \mathbf{1} \in \mathbb{R}^d$, whose Hamiltonian is given in (32). As already stressed, there exists many variants to the Ising model depending on the geometry of the structure of the adjacency matrix. We focus on the simple Ising model defined on a N-dimensional regular lattice illustrated in Fig. 16, such that each spin has $2N$ interacting nearest-neighbors. It is formalized by the following Hamiltonian

$$\mathscr{H}_d(\boldsymbol{\sigma};\mathbf{J},\mathbf{h}) = -\frac{1}{2}\sum_{\langle ij\rangle} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i\,, \tag{36}$$

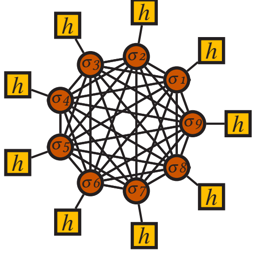
where $\langle ij\rangle$ denotes all possible pairs of neighboring spins on the regular lattice V. Solving the Ising model at finite dimensions for $N=1$ is a simple exercise and easy to solve with the transfer matrix technic. Unfortunately the model does not show any phase transition as the *lower critical dimension* of the Ising model is $N_- = 1$, under which there does not exist any collective behavior and ordered phase. For $N=1$, the fluctuations are so large that they kill the potential ordered phase. The *up-down* symmetry is therefore preserved in a *disordered paramagnetic phase* with zero macroscopic magnetization $m_d = 0$. Nonetheless, above this lower critical dimension the model exhibits a spontaneous symmetry breaking though and an interesting phase transition. Indeed, for $N=2$, the more cumbersome Ising model has been solved exactly in (Onsager, 1944), while the much harder case $N=3$ still witnesses important research works. In the other hand, above the *upper critical dimension* $N_+ = 4$, it turns out that the *mean-field approximation* $N \to \infty$ of the Ising model is exact, see (Kardar, 2007) for more details.

As a pedagogical illustration, we present this latter mean-field approximation of the Ising model, called in this context the Curie-Weiss model, which is much easier to solve analytically.

2.2.3.B THE CURIE-WEISS MODEL

The Curie-Weiss model (Curie, 1895) is the mean-field approximation of the Ising model with *fully-connected* interactions in the limit of a high-dimensional lattice. As very often, mean-field or fully connected approximation have the advantage to make the model much easier to solve analytically. See Sec. 4.2 for a different approach to mean-field approximations. In contrast with the Ising model, the interactions of the *mean-field* Curie-Weiss model are fully-connected and long-range such that each spin is connected to all other spins $\boldsymbol{\sigma} \in \chi_d = \{\pm 1\}^d$ including itself, with a weak homogeneous coupling constant $J_{ij} = \frac{J}{2d}$ scaling with the total number of spins to ensure the existence of the thermodynamic limit. Each spin can also interact with a homogeneous external field h so that the Hamiltonian of the Curie-Weiss model trades the summation over neighboring pairs with all possible long range pairs, such that it reduces to

$$\mathscr{H}_d(\boldsymbol{\sigma};J,h) = -\frac{J}{2d}\sum_{i,j=1}^{d}\sigma_i\sigma_j - h\sum_{i=1}^{d}\sigma_i\,. \tag{37}$$



The factor graph of the Curie-Weiss model is completely symmetric as every spin is connected to every others. For clarity we do not draw the interaction factors $J/(2d)$

Because of this absence of geometric structure the JPD at inverse temperature β is simply given by

$$\mathrm{P}_d(\boldsymbol{\sigma};\beta,J,h) = \frac{1}{\mathscr{Z}(\beta,J,h)}\prod_{i,j}^{d} e^{\frac{\beta J}{2d}\sigma_i\sigma_j}\prod_{i} e^{\beta h\sigma_i}$$

and is represented by a fully-connected symmetric factor graph. Let us introduce a simple order parameter: the *magnetization* in eq. (24). Defined as the macroscopic averaged magnetic moment $m_d \equiv \frac{1}{d}\sum_i \mathbb{E}_{\boldsymbol{\sigma}\sim \mathrm{P}_d}\sigma_i$, it takes $2d+1$ possible values $m_d \in \mathbb{M}_d = \{-1+\frac{k}{d}, k \in [\![0:2d]\!]\}$. At high temperatures, each spin is *free to flip upside down* and is not affected by the interactions with its neighbors. As a consequence, by symmetry the magnetization is, in average, basically zero $m_d = 0$ and the system lies in the *paramagnetic phase*. The specificity of a ferromagnet is that below a certain *critical temperature* the influence of neighbors increase such that a spontaneous magnetization $m_d \neq 0$ appears in the absence of any external magnetic field $h \to 0$. This transition is the so-called *paramagnetic-ferromagnetic* transition. To describe quantitatively this para-ferro phase transition, let us derive the free entropy $\Phi_d(\beta,J,h) = \frac{1}{d}\log \mathscr{Z}_d(\beta,J,h)$ with two different methods: a direct combinatorial one that makes use of the finite set $\mathbb{M}_d$ of possible values taken by the magnetization, which is specific to this case, and the general Fourier method that we intensively use in the rest of the manuscript.

By first introducing the order parameter with a Dirac-delta integral $1 = \int_{\mathbb{M}_d} \mathrm{d}m_d \delta\left(m_d - \frac{1}{d}\sum_{i=1}^d \sigma_i\right)$ the partition function can be expressed as

$$\mathcal{Z}_d(\beta, J, h) = \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma} \int_{\mathbb{M}_d} \mathrm{d}m_d \delta\left(m_d - \frac{1}{d}\sum_{i=1}^d \sigma_i\right) \\ \times \exp\left(\frac{\beta J}{2d}\left(\sum_{i=1}^d \sigma_i\right)^2 + \beta h \sum_{i=1}^d \sigma_i\right) \tag{38}$$

Combinatorial method The partition function can be directly computed by a combinatorial argument. Indeed, fixing the total magnetization m_d and denoting d_+, d_- the number of positive and negative spins, we therefore have $d \cdot m_d = d_+ + d_-$ and $d = d_+ - d_-$, so that $d_+ = \frac{d(1+m_d)}{2}$ and $d_- = d - n_+ = \frac{d(1-m_d)}{2}$. Defining $\Omega_d(m_d) = \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma} \delta\left(m_d - \frac{1}{d}\sum_{i=1}^d \sigma_i\right)$ the number of configurations that give the same magnetization m_d, it is simply given by the number of possibility to choose d_+ positive spins:

$$\Omega_d(m_d) = \binom{d}{\frac{d(1+m_d)}{2}} = \frac{d!}{\left(d\frac{1-m_d}{2}\right)! \left(d\frac{1+m_d}{2}\right)!} \simeq e^{d\mathrm{H}_{\text{binary}}\left(\frac{1+m_d}{2}\right)}.$$

Up to negligible terms, that do not scale exponentially with the system size, it can be expressed as a function of the Shanon binary entropy $\mathrm{H}_{\text{binary}}(x) = -x\log(x) - (1-x)\log(1-x)$, see Sec. 4.2.1.a, by using the Stirling approximation of $d! \sim \sqrt{2\pi d}\,(d/e)^d$, in the large size limit $d \to \infty$. Finally, the partition function can be transformed as

$$\begin{aligned} \mathcal{Z}_d(\beta, J, h) &= \int_{\mathbb{M}_d} \mathrm{d}m_d \Omega_d(m_d) \exp\left(d\left(\frac{\beta J}{2} m_d^2 + \beta h m_d\right)\right) \\ &\simeq \int_{\mathbb{M}_d} \mathrm{d}m_d \exp\left(d\left(\mathrm{H}_{\text{binary}}\left(\frac{1+m_d}{2}\right) + \frac{\beta J}{2} m_d^2 + \beta h m_d\right)\right) \\ &\equiv \int_{\mathbb{M}_d} \mathrm{d}m_d \exp\left(d\Psi\left(m_d; \beta, J, h\right)\right), \end{aligned}$$

where we introduced the free entropy potential

$$\Psi\left(m; \beta, J, h\right) = \mathrm{H}_{\text{binary}}\left(\frac{1+m}{2}\right) + \frac{\beta J}{2} m^2 + \beta h m. \tag{39}$$

Notice that this mean-field approximation can also be obtained from a more elegant variational principle based on the Gibbs inequality presented in Sec. 4.2. In the thermodynamic limit $d \to \infty$, since the integral is dominated by its maximum, the partition function can be evaluated with a Laplace method, also called a *saddle point* method (Wong, 1989), so that the free entropy yields

$$\Phi(\beta, J, h) = \lim_{d\to\infty} \frac{1}{d} \log \mathcal{Z}_d(\beta, J, h) = \max_{m \in [-1;1]} \Psi(m; \beta, J, h) \tag{40}$$

where we used the fact that $\mathbb{M}_d \underset{d\to\infty}{\longrightarrow} [-1;1]$. For a pedagogical purpose, we present as well the equivalent computation of the mean-field free entropy with the Fourier transform method that can be generalized

Fourier transform method The general method consists in introducing the Fourier representation of the Dirac-delta distribution according to

$$\delta\left(x\right) = \frac{1}{2\pi i}\int_{i\mathbb{R}} \mathrm{d}\hat{x} e^{\hat{x}x} ,$$

in (38) so that

$$\begin{aligned}\mathscr{Z}_d(\beta,J,h) &\propto \int_{\mathbb{M}_d} \mathrm{d}m_d \int_{i\mathbb{R}} \mathrm{d}\hat{m} e^{d\left(\hat{m}m_d + \frac{\beta J}{2}m_d^2 + \beta h m_d\right)} \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma} e^{-\hat{m}\Sigma_{i=1}^d \sigma_i} \\ &\propto \int_{\mathbb{M}_d} \mathrm{d}m_d \int_{i\mathbb{R}} \mathrm{d}\hat{m} e^{d\left(\hat{m}m_d + \frac{\beta J}{2}m_d^2 + \beta h m_d\right)} \left(2\cosh\hat{m}\right)^d ,\end{aligned}$$

The Cauchy theorem states that the integral of a holomorphic function $f : \Gamma \mapsto \mathbb{C}$ on a simply connected Γ open subset is null $\int_\Gamma f(z)dz = 0$

where we omitted the negligible pre-factors in the thermodynamic limit. Deforming the integration contour with the Cauchy theorem, the free entropy can be formulated as a *saddle point* and evaluated by

$$\Phi \equiv \lim_{d\to\infty} \frac{1}{d} \log \mathscr{Z}_d(\beta,J,h) = \mathbf{extr}_{m,\hat{m}} \tilde{\Psi}(m,\hat{m};\beta,J,h) ,$$

with $m = \lim_{d\to\infty} m_d \in [-1;1]$ and $\tilde{\Psi}(m,\hat{m};\beta,J,h) \equiv \hat{m}m + \frac{\beta J}{2}m^2 + \beta h m + \log\cosh\hat{m}$. Taking the saddle point condition over $\hat{m}$ and using the fact that $\mathrm{H}_{\text{binary}}\left(\frac{1+m}{2}\right) = -m\,\mathrm{atanh}(m) + \log\cosh\mathrm{atanh}(m)$, we finally recover the same free entropy potential (39)

$$\begin{aligned}\tilde{\Psi}(m,\hat{m}^\star;\beta,J,h) &= \frac{\beta J}{2}m^2 + \beta h m - m\,\mathrm{atanh}(m) + \log\cosh\mathrm{atanh}(m) \\ &= \frac{\beta J}{2}m^2 + \beta h m + \mathrm{H}_{\text{binary}}\left(\frac{1+m}{2}\right) = \Psi(m;\beta,J,h) .\end{aligned}$$

Paramagnetic-ferromagnetic phase transition To conclude, the computation of the free entropy of $d\to\infty$ interacting spins reduces to a one-dimensional optimization problem over the magnetization order parameter. This extremization (40) can therefore be analyzed easily to finally describe the phase transition occurring in the Currie-Weiss model, which is nothing but the mean-field approximation of the Ising model. Taking the extremization over $m \in [-1;1]$, $\partial_m \Psi(m) = 0$, we obtain that the maximum verifies $m = \tanh\left(\beta\left(Jm + h\right)\right)$, it can be solved numerically as illustrated in Fig. 19. In the limit of a vanishing external field $h = 0^+$, fixing the coupling constant to $J = 1$, we can explore the free entropy behavior as a function of the inverse temperature control parameter β. For high temperature, i. e. $\beta \to 0$, the global maxima of the variational free entropy is given by a paramagnetic phase with $m = 0$. At the critical inverse temperature $\beta = 1$, we observe the continuous apparition of two global maxima with $m \neq 0$ that correspond to

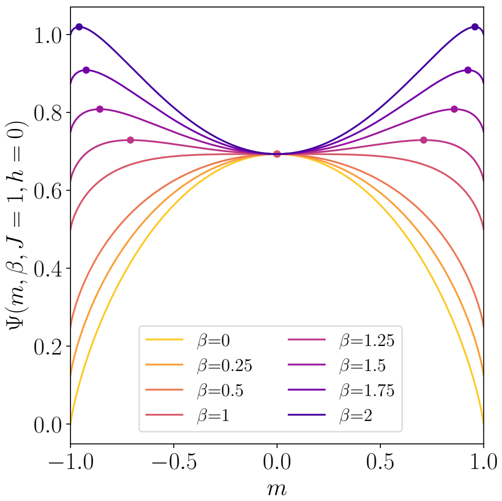


Figure 19: Variational free entropy $\Psi(m;\beta,J=1,h=0)$ of the Curie-Weiss model. Above the critical temperature $\beta < \beta^\star = 1$ (yellow-red), the global maxima of the potential is achieved for $m = 0$ that corresponds to a paramagnetic phase and no ordered phase exists. For smaller temperature $\beta > \beta^\star$ there exists an ordered phase with strictly positive or strictly negative magnetization that corresponds to ferromagnetic phases (pink-blue).

ferromagnetic phases with either a majority of spins *up* $m > 0$ or *down* $m < 0$, while the paramagnetic local maxima $m = 0$ becomes a metastable state.

2.3 EXTENSION TO DISORDERED SYSTEMS AND SPIN GLASSES

In the previous section, we focused on systems and models qualified as *ordered* in the sense that all parameters such as the exchange coupling $\mathbf{J}$ and the external field $\mathbf{h}$ are deterministic. However in nature, no material is perfectly homogeneous and deterministic. They usually present some sources of randomness like interstitial impurities or random external/local environment. For instance, consider a real magnetic material: the description that the local magnetic moments interact in a simple homogeneous way, as illustrated in the Ising and Curie-Weiss models in Sec. 2.2.3.a, is over-idealistic. In reality there exists some impurities that modify the interactions and make their behavior more complex. In more details, a small fraction of a transition metal may be diluted into a noble metal, to obtain an alloy with magnetic moment randomly localized: this is called a *spin glass* (Mézard et al., 1987). Such a system belongs to the large class of *disordered systems* in which some source of randomness emerges in the spin interactions. Statistical physics started to study this new class of models in the 60-70's and was the source of a rich literature since then. Indeed, incorporating randomness in the classical statistical physics tools, presented in Sec. 2.2, allowed to democratize the approach to various fields and to highlight the existence of new interesting phenomena and phase transitions. In particular, under a *weak disorder* assumption, the description of the ordered phase transitions and critical phenomenon presented in Sec. 2.2.2 may be either conserved or smoothed so that first oder become second order phase transitions. However, in the presence of a *strong disorder* it *strongly affects* and changes the nature of the phase transitions especially at low temperature where we observe the appearance of a singular *glassy phase* with many local metastable states.

In this section we discuss several kinds of disorder and models that account for it, and we focus the discussion on spin glasses that are more relevant according to the rest of the manuscript. We present a brief overview of the wide literature of spin glasses mainly based on (Mézard et al., 1987; Castellani et al., 2005; De Dominicis et al., 2006) to illustrate the basic ideas required to understand the rest of the manuscript. The discssion can be extended with more specific and influent contributions (Franz et al., 1997; Bouchaud et al., 1998; Biroli et al., 2001; Cugliandolo, 2002; Franz et al., 2011; Berthier et al., 2011).

2.3.1 QUENCHED AND ANNEALED DISORDER

There exists two main types of disordered systems: the ones with *explicit disorder* in the Hamiltonian of the model, and the ones such that the disorder is self-generated. The latter class can be simply illustrated with *structural glasses* in which many interacting particles are moving so that each particle feels a local *random disordered environment*. However, in this manuscript, we consider only systems with explicit disorder and we refer the reader to (Kirkpatrick et al., 1987a; Mézard et al., 2000; Lubchenko et al., 2007; Charbonneau et al., 2014) for more details on amorphous solids and structural glasses.

We therefore consider systems of spins $\boldsymbol{\sigma}$ with *explicit* disorder in the Hamiltonian (19), for example through the influence of random parameters such as the coupling constant or the external field $(\mathbf{J}, \mathbf{h})$, that we call for historical reasons *impurities*. We assume that the latter impurities are some RV that evolve at a typical time scale τ_q, while the system of spins $\boldsymbol{\sigma}$ evolve a time scale τ. Depending on how these time scales compare, we shall distinguish *quenched* and *annealed* disorders. *Annealed* disorder refers to systems such that $\tau_q \simeq \tau$. In other words, the random impurities $(\mathbf{J}, \mathbf{h})$ and the spins variables $\boldsymbol{\sigma}$ evolve and fluctuate on a similar time scale (Palmer, 1982). Therefore, they play the same role and should be considered on an equal footing. As a consequence, in the presence of an annealed disorder, the impurities are in thermal equilibrium and can simply be included in the statistical description of the system. In contrast, *quenched* disorder refers to systems such that $\tau_q \gg \tau$: the impurities are *static* and remain fixed while the spin variables $\boldsymbol{\sigma}$ fluctuate. Each realization of the quenched disorder thus corresponds to a *new experiment* with new sampled parameters. Therefore distinguishing the *slow-evolving* quenched impurities $\mathbf{J}$ from the thermal spins $\boldsymbol{\sigma}$ time scales is crucial. In particular, the equilibrium properties and the corresponding thermodynamics cannot be computed in the same way than for systems with annealed disorder. In order to take into account the random impurities and not compute properties of the system which depend on a single realization of the randomness, we would like to average over the randomness. However, the specific disorder time scale makes the average over the Gibbs random measure harder. The different time scales and the effect of the randomness in the Hamiltonian require therefore specific analytical treatments that we describe and develop in Chap. 4. Notice, nonetheless, that in some specific cases, quenched disordered systems behave as annealed systems and are easier to tackle analytically.

Even though other disordered systems such as Random Field Ising Model (RFIM) (Belanger et al., 1991; Mézard et al., 1992) have been considered in the literature (Imbrie, 1984; Belanger et al., 1991), in the rest of the manuscript we mainly focus the discussion on *spin glasses* with *quenched disorder*.

2.3.2 SPIN GLASSES WITH QUENCHED DISORDER

Spin glasses refer historically to metallic alloys in which during the chemical preparation of the sample magnetic impurities substitute to the original atoms in randomly selected positions (Binder et al., 1986; Fischer et al., 1993; Mézard et al., 1987). In order to theoretically understand the properties of these materials, various models have been proposed based on a spin model with a *quenched* disorder through the random exchange interaction $\mathbf{J}$ drawn from a distribution $\mathrm{P}(\mathbf{J})$. Unlike the simple RFIM case where the randomness only affects the one-body interaction term $\log \phi_i$, the disordered interactions $\mathbf{J}$ dramatically affect the two-body interactions and the thermodynamics of mean-field models spin glasses.

In this manuscript, we consider essentially models that can be formulated as *spin glass* models with *quenched disorder*. More precisely, we consider a system of d spins with $\boldsymbol{\sigma} \in \chi_d$ with an Hamiltonian $\mathscr{H}_d(\boldsymbol{\sigma};\mathbf{J},\mathbf{h})$ that explicitly depends on the quenched RV, e. g. the coupling constant $\mathbf{J}$ completely specified by its probability distribution $\mathrm{dP}(\mathbf{J}) = \mathrm{p}(\mathbf{J})\mathrm{d}\mathbf{J}$, and a fixed external field $\mathbf{h} \in \mathbb{R}^d$. For the sake of illustration, let us introduce the Ising-like spin glass model, historically considered in (Toulouse et al., 1987), which became the battle horse of the spin glass literature. Consider a graphical model $\mathscr{G}(\mathrm{V},\mathrm{E})$ with spins at the vertices V and pairwise interactions between spins on edges E:

$$\mathscr{H}_d(\boldsymbol{\sigma};\mathbf{J},\mathbf{h}) = -\frac{1}{2}\sum_{(ij)\in\mathrm{E}} J_{ij}\sigma_i\sigma_j - \sum_{i\in\mathrm{V}} h_i\sigma_i \,,$$

associated to the Gibbs thermal average and the partition sum

$$\mathscr{Z}_d(\beta,\mathbf{J},\mathbf{h}) = \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\; e^{-\beta\mathscr{H}_d(\boldsymbol{\sigma};\mathbf{J},\mathbf{h})} \,.$$

Unlike the case of the Ising ferromagnet 2.2.3.a (respectively anti-ferromagnet) with $\forall (ij) \in \mathrm{E}, J_{ij} > 0$ (respectively $J_{ij} < 0$), in spin glass models, the exchange interaction matrix is random so that J_{ij} associated to the edge (ij) has a random sign and can be either positive or negative. The local interaction is called *ferromagnetic* (respectively anti-ferromagnet) if $J_{ij} > 0$ (respectively $J_{ij} < 0$). The coupling being random, in *average*, the model is called ferromagnetic (respectively anti-ferromagnetic) if there exists a *bias* such that $\mathbb{E}_{\mathbf{J}} J_{ij} > 0$ (respectively $\mathbb{E}_{\mathbf{J}} J_{ij} < 0$). For conciseness, we leave aside the external field in the following.

2.3.3 FRUSTRATION

Understanding spin glasses is more involved that classical ferromagnetic models. Indeed, the quenched disorder may be the source of *frustration* between the spins of the system, so that finding an optimal configuration becomes harder and takes much longer time. In particular, the randomness of the coupling interactions $\mathbf{J}$ signs breaks down the spatial homogeneity and creates heterogeneity, called *frustration*. This collective behavior appears when the best possible spins configuration cannot satisfy all the local local two-body *constraints* and minimize all interactions terms in the Hamiltonian, as illustrated in Fig. 20. As a serious consequence, many distinct configurations may achieve the same energy level, so that one expects the existence of many *local minima* in the free energy landscape leading to a *glassy behavior*. Such frustrated systems show non-trivial properties richer than systems without disorder considered in Sec. 2.2 and the quenched random disorder dramatically affects their thermodynamic behavior.

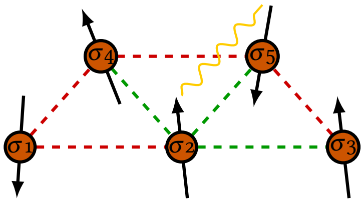


Figure 20: Illustration of frustrated spins on a regular "plaquette": frustration appears when a spin undergoes a positive (green) and a negative (orange) interaction at the same time. Spins σ_2, σ_5 are frustrated as the coupling constant is positive but the corresponding spins are anti-aligned.

2.3.4 AVERAGING AND SELF-AVERAGING

As discussed in the previous section, the quenched disorder $\mathbf{J}$ plays a singular role with respect to the thermal fluctuations of the spins $\boldsymbol{\sigma}$. Since it evolves at a much slower time scale $\tau_q \gg \tau$, for a given experiment the quenched disorder is considered as *static*. It means that at each experiment we draw a new realization of the disorder $\mathbf{J}$, from the distribution $\mathrm{P}(\mathbf{J})$, that is considered to be fixed all along the spin dynamics. As a consequence the free entropy explicitly depends on the realization of the disorder

$$\Phi_d(\beta,\mathbf{J}) \equiv \frac{1}{d}\log \mathscr{Z}_d(\beta,\mathbf{J}), \qquad \varphi_d(\beta,\mathbf{J}) \equiv -\frac{1}{d\beta}\log \mathscr{Z}_d(\beta,\mathbf{J}). \tag{41}$$

However, we do not want the description of the system to depend on the realization of the disorder $\mathbf{J}$. Instead, and specifically to quenched systems, we introduce the *averaged free entropy* and *free energy* by adding the *quenched average* over the disorder, denoted denote $\mathbb{E}_{\mathbf{J}}$, on top of (41), and crucially after the Gibbs thermal average contained in the partition sum

$$\Phi_d(\beta) \equiv \mathbb{E}_{\mathbf{J}}\Phi_d(\beta,\mathbf{J}), \qquad \varphi_d(\beta) \equiv \mathbb{E}_{\mathbf{J}}\varphi_d(\beta,\mathbf{J}). \tag{42}$$

Averaging over all possible disorder *realizations* refers to the so-called *typical scenario*. This is to oppose to the *worst case* scenario, that deals with the worst possible disorder instance to obtain a strong upper-bound of the system description. In the perspective of averaging over many experiments, we restrict ourselves to the *typical case* that suits much better our purpose, so that all quantities shall be averaged over the disorder. However, it is important to remark that for the moment there is no reason why the *averaged* free entropy would correctly describe the system for a single realization of the disorder.

Self-averaging In fact the most remarkable property of spin glasses is that some extensive observables become *self-averaging* in the thermodynamic limit, meaning that they are correctly described by their averaged behavior. In other words, we say that a RV $\mathrm{X}_d \in \mathbb{R}^d$ is self-averaging if it concentrates around its mean, namely if

$$\forall \varepsilon > 0, \lim_{d\to\infty} \mathbb{P}\left(\|\mathrm{X}_d - \mathbb{E}\left[\mathrm{X}_d\right]\| > \varepsilon\right) = 0. \tag{43}$$

Very importantly the free entropy is such a self-averaging quantity. Thus at fixed disorder realization $\mathbf{J}$, $\Phi_d(\beta,\mathbf{J})$ in (41) *concentrates* in the thermodynamic limit around a deterministic free entropy given by the average over the disorder $\Phi(\beta)$ in (42). Therefore, in the thermodynamic limit $d \to \infty$ such observables have the same value for each realization of the disorder $\mathbf{J}$ and this legitimates the fact of considering the *typical* scenario. In other words, the free entropy description no longer depends on the specific realization of the disorder and the sample fluctuations are vanishing in the large system size limit, so that its typical value coincides with the deterministic average value. As a final remark, notice that in general observables that involve the sum of an infinite number of particles are expected to self-average. On the contrary, non-self averaging quantities, such as correlation functions, can fluctuate significantly and computing them remains a harsh difficulty.

Finally, the self-averageness of the free entropy Φ can be described for both short and long range systems as explained in the following.

Short-range argument For short-range interactions mode, we recall a simple argument based on the CLT that shows that the free entropy is self-averaging (Mézard et al., 1987; Thouless et al., 1977; Orlandini et al., 2002). In the case the interactions are short-range, we can split the whole system of total volume $V = L^d$, where L is the typical size of the system, in N macroscopic sub-systems of volume $v = l^d$ with $V = Nv$ and $N = \left(\frac{L}{l}\right)^d$ defined in such way that they weakly interact with each other. If we assume that the interactions have a typical range $\lambda \ll l$, the free entropy decomposes in bulk and surface contributions

$$\begin{aligned}\Phi_d\left(\beta,\mathbf{J}\right) &= \frac{1}{d}\log \mathscr{Z}_d\left(\beta,\mathbf{J}\right) = \frac{1}{d}\log\sum_{\boldsymbol{\sigma}} e^{-\beta\mathscr{H}_{\text{bulk}}(\boldsymbol{\sigma};\mathbf{J})-\beta\mathscr{H}_{\text{surface}}(\boldsymbol{\sigma};\mathbf{J})} \\ &\simeq \frac{1}{d}\log\sum_{\boldsymbol{\sigma}} e^{-\beta\mathscr{H}_{\text{bulk}}(\boldsymbol{\sigma};\mathbf{J})} + \frac{1}{d}\log\sum_{\boldsymbol{\sigma}} e^{-\beta\mathscr{H}_{\text{surface}}(\boldsymbol{\sigma};\mathbf{J})} \\ &= \Phi_{\text{bulk}}\left(\beta,\mathbf{J}\right) + \Phi_{\text{surface}}\left(\beta,\mathbf{J}\right) \simeq \Phi_{\text{bulk}}\left(\beta,\mathbf{J}\right) \\ &\simeq \sum_{k=1}^{N}\log\sum_{\boldsymbol{\sigma}_k} e^{-\beta\mathscr{H}_{\text{bulk}}(\boldsymbol{\sigma}_k;\mathbf{J})}\,,\end{aligned}$$

where we first assumed that interactions between the bulk and the surface are negligible, and the last equality generally holds if the surface interaction is negligible with respect to the N blocks of the bulk contribution. In the thermodynamic limit as $l \ll L, N \to \infty$ and the CLT applies, the free entropy is therefore the sum of independent variables and becomes a Gaussian variable centered around its average $\Phi_d\left(\beta\right) \equiv \mathbb{E}_{\mathbf{J}}\Phi_d\left(\beta,\mathbf{J}\right)$ with fluctuations of order

$$\frac{\mathbb{E}_{\mathbf{J}}\left[\Phi_d\left(\beta,\mathbf{J}\right)^2\right] - \Phi_d\left(\beta\right)^2}{\Phi_d\left(\beta\right)} = \Theta\left(d^{-1/2}\right) \underset{d\to\infty}{\longrightarrow} 0\,.$$

Long-range systems Yet this simple general argument does not apply to long-range interactions. Anyway the self-averaging property is in fact very often assumed and may be proven in some long-range interactions model such as the Sherrington-Kirkpatrick (SK) model (Guerra et al., 2002a; Guerra et al., 2002b) or in more recent works for various models (Barbier et al., 2016; Barbier et al., 2018a; Barbier et al., 2019b; Barbier et al., 2019a).

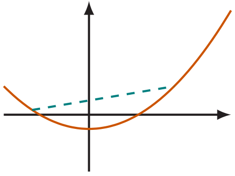
2.3.5 ANNEALED AVERAGES

The quenched nature of the disorder imposes to *average the free entropy itself* and *not the partition function*. Taking the quenched disorder average of the log-partition function in (42) is the main challenge and is unfortunately rarely analytically tractable. In order to circumvent this difficulty, the cumbersome replica method presented in Sec. 4.1 is a very powerful tool that we use intensively in many applications of this manuscript. Another option to avoid computing this average is to exchange instead the order of the expectation and the logarithm. This refers to the *annealed* average that leads to the much simpler computation of the *annealed free-entropy* defined by

$$\Phi_d^{\mathrm{a}}(\beta) \equiv \frac{1}{d\beta} \log \mathbb{E}_{\mathbf{J}} \mathscr{Z}_d(\beta, \mathbf{J}) . \tag{44}$$

As discussed in Sec. 2.3.1, this simplification is justified only for systems with *annealed disorder* so that the spins and the disorder fluctuate with the same time scales $\tau \sim \tau_q$ and appear on the same footing. Consequently, in this case it is necessary to take the thermal Gibbs average $\mathbb{E}_{\boldsymbol{\sigma}}$ and disorder average $\mathbb{E}_{\mathbf{J}}$ simultaneously before taking the logarithm, which keeps only the large deviation behavior of the system. Even though this simplification is justified for annealed disorder, in the presence of a quenched disorder this abusive annealed simplification for quenched disorder provides in fact an *approximation* of the cumbersome quenched average. More precisely, because of the concavity of the logarithm and using the Jensen inequality, we observe that the annealed free entropy is an *upper bound* of the quenched free entropy

Jensen inequality states that for a convex function f, the secant line of a convex function lies above the graph of the function



$$\Phi_d(\beta) \equiv \frac{1}{d} \mathbb{E}_{\mathbf{J}} \log \mathscr{Z}_d(\beta, \mathbf{J}) \leq \frac{1}{d} \log \mathbb{E}_{\mathbf{J}} \mathscr{Z}_d(\beta, \mathbf{J}) \equiv \Phi_d^{\mathrm{a}}(\beta) .$$

Finally in the cases where the quenched average is intractable, we can still hope that the simpler annealed average provides a good approximation.

2.3.6 ON THE SPIN GLASS PHASE

The thermodynamic behavior of spin glass systems are drastically affected by the appearance of the quenched disorder in the Hamiltonian that is responsible for the emergence of a new collective behavior: the *spin glass* phase. Remarkably, as a consequence of the frustration many local constraints may not be satisfied at the same time and thus there eventually exists many distinct ground state configurations with the same, strictly positive, energy level. As a result, in contrary with the classical Ising model in Sec. 2.2.3.a where only two phases with positive or negative macroscopic magnetizations emerged, in the spin glass phase we observe a highly non-trivial *ergodicity breaking* of the configuration space such that the Gibbs distribution exhibits many metastable states. In other words, the system is stuck in some sub-regions of the configuration space χ_d and can take exponential time (in the size of the system d) to explore the whole configuration space.

To illustrate the remarkable properties observed in experimental spin glasses, we recall the seminal Edwards-Anderson (EA) model (Edwards et al., 1975) that tries to capture the main features of the spin glass phase.

The Edwards-Anderson model Since the interactions in metal alloys are short-range, in order to replace the randomness induced by the impurities positions, it was proposed to consider an Hamiltonian defined on a regular graph $\mathscr{G}(\mathrm{V},\mathrm{E})$ with *nearest-neighbor* interactions:

$$\mathscr{H}_d(\boldsymbol{\sigma};\mathbf{J},\mathbf{h}) = -\sum_{<ij>\in\mathrm{E}} J_{ij}\sigma_i\sigma_j - \sum_{i\in\mathrm{V}} h_i\sigma_i\,. \tag{45}$$

The exchange couplings can be chosen either Gaussian $\mathrm{P}(J_{ij}) = \mathscr{N}_{J_{ij}}\left(0,\frac{J_0}{d}\right)$ or binary $\mathrm{P}(J_{ij}) = \frac{1}{2}\left(\delta(J_{ij}-J_0/d)+\delta(J_{ij}+J_0/d)\right)$ such that $\mathbb{E}_{\mathbf{J}}J_{ij}=0$ and $\mathbb{E}_{\mathbf{J}}J_{ij}^2 = \frac{J_0}{d}$, for some $J_0>0$. Let us introduce the averaged total magnetization m_d and the celebrated EA order parameter specifically designed to reveal the spin glass phase:

$$m_d = \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}_{\mathbf{J},\boldsymbol{\sigma}}\left[\sigma_i\right], \qquad q_{\mathrm{ea}} = \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}_{\mathbf{J},\boldsymbol{\sigma}}\left[\sigma_i\right]^2,$$

where the average are first taken with respect to the Gibbs distribution before taking the average with respect to the disorder $\mathbf{J}$. The particularity of the EA model is to present no ferromagnetic nor anti-ferromagnetic phase. In fact, as expected at high-temperature $\beta\to 0$, we observe a paramagnetic phase with a global magnetization $m=0$ and $q_{\mathrm{ea}}=0$. At the critical *glass transition* $\beta\geq\beta_g$ the system enters the so-called *spin glass phase* characterized by zero global magnetization $m=0$ but a non-zero EA order parameter $q_{\mathrm{ea}}\neq 0$. In other words, even though there is no global ordering of the system as the global magnetization stays zero in average, each individual spin dynamics is still frozen in a preferred orientation. Indeed, the time auto-correlation function is non-zero and given by the EA order parameter $C(t) = \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}_{\boldsymbol{\sigma}}\sigma_i(t)\sigma_i(0) \underset{t\to\infty}{\longrightarrow} q_{\mathrm{ea}}\neq 0$. This means that the system at time t $\boldsymbol{\sigma}(t)$ is strongly correlated to the initial configuration of the system $\boldsymbol{\sigma}(t=0)$. In other words, the system has a *strong memory* of the *initial preparation* of the system. This *aging* phenomenon (Sompolinsky et al., 1982), measured by the new EA order parameter, has been experimentally observed, for instance, by measuring the magnetic susceptibility with different system initialization (Vincent, 2007).

2.3.7 SPIN GLASS MODELS AND COMPUTER SCIENCE

After the EA model breakthrough, various disordered models came up to light and boosted the spin glass literature. Let us mention the celebrated SK model (Sherrington et al., 1975), whose dynamics have been studied in (Cugliandolo et al., 1994) and rigorously proven in (Talagrand, 1998; Panchenko, 2013), the p-spin for $p\geq 3$ interactions with binary or continuous variables for structural glass theory (Gardner, 1985; Kirkpatrick et al., 1987c; Kirkpatrick et al., 1987b; Crisanti et al., 1995; Crisanti et al., 1992). We shall mention as well the *Random Energy Model* (Derrida, 1981; Mézard et al., 2009), which is one of the simplest toy model that exhibits a glassy phase; the *KPZ equation*

(Kardar et al., 1986) that describe the behavior of particles in a rough random landscape whose theory has been recently confirmed numerically by precise simulations (Hartman, 1982); or the *Stochastic Block Model* (Decelle et al., 2011) for community detection on random graphs.

This kind of glassy dynamics is believed to be present in many systems such as in computer science problems that we will be interested in the main contributions of this manuscript. Notably, *combinatorial optimization* and *random constraints satisfaction problems* gained in importance with especially error correcting code such as LDPC in noisy communication channels (Shannon, 1948; MacKay et al., 1996), *the minimum spanning tree*, *Eulerian circuits*, *Hamiltonian cycles*, the *Travelling salesman problem* or *partitioning* problems (Mézard et al., 2009). These random optimization problems such as random k-satisfiability problems (Ricci-Tersenghi et al., 2001; Mézard et al., 2002; Mézard et al., 2005) or graph coloring (Jensen et al., 2011; Mulet et al., 2002; Zdeborová et al., 2007) can be formulated as generic CSP whose Gibbs distribution was studied in details in (Krzakała et al., 2007). Similarly to a physical system being frozen in a sub-region of the configuration space, a similar ergodicity breaking might dramatically impact the algorithmic performances of sampling and optimization algorithms in computer science problems. Indeed it is believed that the existence of exponential metastable states may drastically harm the computational performances and explain the computational hardness in random problems such as CSP (Mézard et al., 2009; Zdeborová et al., 2016a) and other optimization problems (Moore et al., 2011).

In the recent years, due to the accession of machine learning and neural networks, we observed a renewed interest of statistical physics in these computer science problems. Interestingly, very often they can be formulated as spin glass models and treated with the corresponding set of powerful tools. In the next section, we propose a brief historical review of the exchange of ideas between statistical physics and computer science.

3 STATISTICAL PHYSICS AND MACHINE LEARNING BACK TOGETHER

Analyzing machine learning problems with statistical physics tools may be unusual to most of the computer science community. Yet, there exists a rich literature with influential connections between these two fields, that we briefly review in this chapter. On one hand, statistical physics aims to understand collective behaviors of matter and phase transitions as illustrated in Chap. 2. However, its powerful formalism readily applies to various fields such as *statistical inference*, whose goal is to *detect* and *recover* a *hidden signal* from observations. In particular, the *high-dimensional statistical regime* in which the number of data and parameters diverge fits perfectly the underlying fundamental large-size hypothesis of the statistical physics framework. Thus, approaching high-dimensional inference and other machine learning problems with statistical physics has a long tradition and an intimate connection which is widely depicted in the literature of *statistical physics of machine learning* (Nishimori, 2001; MacKay et al., 2003; Mézard et al., 2009; Grassberger et al., 2012; Zdeborová et al., 2016a; Advani et al., 2016b; Zdeborová, 2017; Biehl et al., 2019; Zdeborová, 2020).

In this section, we first recall the main interactions between these two fields by presenting a short historical overview in Sec. 3.1. Then we focus on the main contributions that influenced the current statistical physics approach in Sec. 3.1.2, before presenting our global approach in Sec. 3.2 that we deeply use in the main contributions Part II. In particular, we depict how mean-field methods such as the replica method or message passing algorithms, presented in details in Sec. 4.1-4.3, became central to analyze the phase transitions of inference problems and can lead to the design of new algorithms. All along this work, we try to especially highlight and compare the algorithmic phase transitions to optimal statistical thresholds in light of glassiness behaviors and computational hardness.

3.1 A COMMON HISTORY OF MACHINE LEARNING AND STATISTICAL PHYSICS

The intimate connection between machine learning and statistical physics basically started in the 80's and was recently renewed with the democratization and accession of ANN. After the recent successes of DL in many applications, the scientific communities from various fields try to address many of the theoretical challenges raised by their empirical successes. In particular, statistical physics experienced a renewed interest in ANN research with in particular the emergence of rigorous justifications of former heuristic statistical physics methods. The goal of this short section is to briefly recount the main influential works of the statistical physics approach on open ML questions as well as the intricate story between physics and neural networks.

3.1.1 FROM SPIN GLASS THEORY TO RIGOROUS MACHINE LEARNING

The connection of Information Theory (IT) with physics dates probably back to the end of the 1900's early 2000's with Maxwell, Boltzmann, Szilard that study the entropy in thermodynamical systems. It opened a breach for IT whose Shannon became the pioneer followed later on in the 60's by the Gibbs-Bogoluibov-Feynman variational principle that became a central tool in approximate statistical inference, as detailed in Sec. 4.2.

3.1.1.A EMERGENCE OF THE SPIN GLASS COMMUNITY

Later on, during the second AI winter, the physicists Hopfield (Hopfield, 1982) revived the ANN-oriented research by proposing the celebrated eponym *Hopfield model* to explain *associative memory* as a variant of the Ising model with pairwise interactions generated from a set of n patterns $\{\boldsymbol{\xi}_\mu\}_{\mu=1}^n$ such that $J_{ij} = \frac{1}{n}\sum_{\mu=1}^n \xi_{i\mu}\xi_{j\mu}$. This energy based model crystallized particularly the interest of physicists and is certainly responsible of the emergence of an entire branch of the statistical physics dedicated to ML models. By taking advantage of the heuristic tools of the spin glass community, mainly developed with the previous EA and SK models (Edwards et al., 1975; Sherrington et al., 1975; Thouless et al., 1977), such that the Thouless-Anderson-Palmer (TAP) approach, the Hopfield model is analyzed heuristically (Amit et al., 1985b; Amit et al., 1987) and opened the door to more complex ANN models. The same heuristic mean-field methods are then used to analyze ANN and started being popularized in (Amit et al., 1985a). For instance Boltzmann machines, which are nothing more than a stochastic version of the SK spin glass model, have been brought to light in the computer science community (Ackley et al., 1985), making the connection of statistical physics and machine learning even closer. In parallel, computer scientists developed the PAC theory to analyze the generalization property of neural networks (Valiant, 1984), which turns out to be completely orthogonal to the physicist approach used to analyze the Hopfield model. Hence, even though the physics community deeply contributed in the early analysis of ANN models, the computer science community largely ignored the corresponding approach based on heuristic techniques.

In the late 90's, E. Gardner introduced the *replica method* to analyze the *maximum storage capacity* (Gardner, 1987; Gardner, 1988) that is known to be closely related to the VC dimension mostly considered in the ML community. These very influential works introduced a powerful technic used to compute the typical configurations space volumes in order to count how many networks achieve a certain level of error. Many heuristic papers followed (Derrida et al., 1987; Gardner et al., 1988; Krauth et al., 1989) and readily apply the *Gardner approach* to supervised learning with *randomly-quenched disorder* for which the random labels are not correlated with the inputs. In parallel, the training method of SVM (Boser et al., 1992) was inspired by a physics intuition (Krauth et al., 1987). During the same years, a deeply influential review (Mézard et al., 1987) gathered the main mean-field treatments from the fruitful research on spin glasses whose publication accelerated and democratized their use.

3.1.1.B FROM RANDOM LABELS TO LEARNING A RULE

After having widely studied the storage capacity problem with random quenched disorder and random labels, the research shifts towards the *statistical inference* of a hidden *signal*, that a supervised model shall recover from observations. The idea that the training set contains a hidden *planted configuration*, representing a *crystal* configuration in the physicist language, also called a *rule*, refers to the so-called Teacher-Student (T-S) scenario. Both these *random* and *structured* settings were in fact already introduced in the seminal work (Gardner et al., 1989). The replica method and the TAP approach started being applied to more general inference problems such as this T-S for the simplest ANN, the perceptron. The first learning curves and physics-like phase transitions (Györgyi, 1990) are derived and exhibit interesting physics: first and second order phase transitions with the existence of metastable states are observed (Sompolinsky et al., 1990; Opper et al., 1990; Hansel et al., 1990; Opper et al., 1996a) keeping the interest of physicists at the highest level. This simple model architecture is then pushed forward with a second untrained layer: the *committee machine* (Schwarze et al., 1992; Schwarze, 1993; Schwarze et al., 1993). In another direction, the usage of gradient-descent-like algorithms is studied in an online setting (Saad et al., 1995b), where a single example, from an unrealistic infinite reservoir of examples, is observed at each time step.

Unfortunately, physics contributions had almost no impact in the ML community that largely frustrated the physics community. Even though the approach was very elegant for the physicist oriented mind, it was not taken fully seriously mainly because of its lack of rigor. Moreover with the decline of AI attraction, in the late 90's the physics research globally stopped in this direction.

3.1.1.c A RENEWED INTEREST OF PHYSICISTS AND RIGOROUS JUSTIFICATIONS

With the recent flourishing numerical successes of ML and DL, the theoretical research activity around these disciplines grew up again in the recent years. Especially because traditional ML theory, based on data-independent PAC generalization bounds (Vapnik et al., 1994), predicted that models such as DNN with a number of parameters similar to a number of data should overfit. Thus it failed explaining the empirical and striking *generalization problem* of DNN that does not experienced overfitting. Therefore, statistical physics community stroke back and started to work in this direction as they believe that their singular typical case approach, yet on unreasonable simple models, may contribute to understand this challenge and answer fundamental ML questions. In order to finally bring impact of the physicists heuristic methods to the ML community, the mathematical-physics research started to prove rigorously results previously derived in the spin glass literature with the so-called *replica method*, see Sec. 4.1. This stage starts with a first rigorous tentative (Haussler et al., 1996) where they rigorously showed the existence of phase transitions in the learning curve behavior. The analogy between spin glasses of dynamical systems and machine learning seem very interesting but again failed to fully break through. Finally, with the works of (Guerra et al., 2002b; Talagrand, 2003; Panchenko, 2013), which rigorously proved heuristic results of the 80's in the context of the SK model, IT started slowly to consider the statistical mechanics approach. This renewed approach of statistical mechanics techniques are currently gaining in popularity as well in ML because of their recent rigorous justifications for instance in the case of the GLM (Barbier et al., 2016; Reeves et al., 2016b) and the committee machine (Aubin et al., 2018b), whose models have been both studied heuristically in the 90's, or deeper architectures (Gabrié et al., 2017).

3.1.1.d ALGORITHMS AND COMPUTATIONAL COMPLEXITY

The influence of statistical physics is even more obvious in combinatorial optimization and CSP. Indeed, early 2000 many graphical model algorithms such as Belief Propagation (BP) (Pearl, 1982; Yedidia et al., 2001b; Yedidia et al., 2001a) are popularized. These algorithms derived in different fields under different names such as the Viterbi algorithm, Pearl's BP, Gallager codes, Kalman filter, transfer-matrix approach (Yedidia et al., 2001a) are closely related to the physics *cavity method* (Mézard et al., 2009). The simplification of the BP equations under a set of assumptions, see Sec. 4.4, leads to Approximate Message Passing (AMP) algorithms, introduced in the context of Compressed Sensing (CS) in (Donoho, 2006; Maleki, 2011) and popularized in (Montanari, 2012; Rangan, 2011). These physics-inspired algorithms are applied to various CSP whose general Gibbs measure description was studied in (Krzakała et al., 2007). Very importantly, this renewed line of ANN research made a clear connection with the algorithmic computational complexity (Moore et al., 2011) that was never considered in the early statistical physics literature.

3.1.2 RECENT AND CURRENT LINE OF RESEARCH

Statistical physics is currently pursuing actively this line of research and attempting to answer fundamental questions raised by the increasing use of DNN. We present below a short and, inevitably, biased selection of important research directions from a physicist point of view.

3.1.2.A FROM AMP TO THE ANALYSIS OF GD ALGORITHMS

In many models the BP algorithm and variants such as AMP are of theoretical interest since they have been shown and believed to achieve the optimal statistical performances in large regions of parameters. Easily derived and implemented for finite sizes, their performances are nonetheless not guaranteed. But powerfully, the statistical physics approach turns out very useful as it allows to derive and prove the high-dimensional asymptotic performances of the corresponding AMP algorithms, called in this context the *state evolution*. While this requires in principle to compute a high-dimensional JPD, the physics mean-field methods reduce it to a simple optimization problem over a small set of order parameters. Yet the prevalence of gradient-based algorithms in DL recently shifted the current research towards understanding GD dynamics. Indeed, dynamics of GD is believed to be very important as it induces a bias that reduces the wide hypothesis class along training by diffusion and is responsible for good generalization. Even though the analysis of GD dynamics has been performed for linear models in (Baldi et al., 1991; Baldi et al., 1995; Dunmur et al., 1993; Krogh et al., 1992; Advani et al., 2017), generalizing it to non-linear models remain challenging. Yet, first steps in this direction have been recently performed. Following the early works of (Saad et al., 1995a), the dynamics of *online* SGD was studied in more details and generalized to more complex architectures (Goldt et al., 2019a). In the other hand, following the dynamical approach studied in the early works (Cugliandolo et al., 1993; Ben Arous et al., 2006) in the context of the p-spin model, it was recently extended to the perceptron (Agoritsas et al., 2018), the spiked matrix model (Mannelli et al., 2020) and a Gaussian mixture classification task (Mignacco et al., 2020b). Generalizing this dynamical approach to more complex architectures and data structures is certainly a fruitful direction of research.

3.1.2.B THE ROLE OF DATA: FROM IID TO A MANIFOLD

Another essential ingredient in understanding DL performances is definitely the essential role of data. Most theoretical statistics works commonly assume that data come from a i.i.d factorized probability distribution, without explicitly modelling the training dataset. As a consequence, these approaches lack capturing the deep correlations of real datasets and their fundamental impact on the training of DNN. A first step to overtake this i.i.d limitation was performed in (Kabashima, 2008) by generalizing it to rotationally invariant inputs in perceptrons and later on to the weights in DNN (Gabrié et al., 2018).

Moreover, the original T-S scenario fed with i.i.d samples is also gradually challenged as the dynamics of DNN on real-life tasks such as MNIST classification do not reveal the same dynamics than for a T-S synthetic dataset. To capture this particular learning dynamics on MNIST, (Goldt et al., 2019b) introduced the *Hidden manifold model* to represent the input data by a low-dimensional structure, that was studied later on in (Gerace et al., 2020) in the context of random features. This rich data modelling idea is another promising step to take into account the importance of real-data distributions in the learning dynamics.

3.1.2.c FROM A FEW HIDDEN UNITS TO DEEP/WIDE LAYERS

Multi-layer and over-parametrization Finally, the last ingredient responsible for the DNN success is certainly the wide and deep architectures of networks that form a large hypothesis class with a great expressivity. While early rigorous works in statistical physics focused on simple single-layer perceptron (Barbier et al., 2016), the current trend consists in analyzing models with increasing sizes, starting with a simple two-layers extension (Aubin et al., 2018b). In parallel, another mean-field scaling limit was recently proposed, where the number of hidden units K is much larger than the input size $d = o(K)$. In a recent line of research, (Jacot et al., 2018; Du et al., 2018; Allen-Zhu et al., 2019; Arora et al., 2019; Lee et al., 2019) observed that with a scaling of the weights as $\Theta(K^{-1/2})$, the dynamics enters a *lazy regime* governed by the Neural Tangent Kernel (NTK). As a consequence, it remains stuck close to the initialization and can be therefore trivially analysed. In contrast, in the same infinitely wide limit, but under a different scaling $\Theta(K^{-1})$, (Chizat et al., 2018; Mei et al., 2018; Rotskoff et al., 2018) observed another, yet more interesting, *feature learning* regime in which the NTK really learns. Closely related, *random features* was the subject of various works notably to understand the learning curve behavior and double descent generalization phenomena (Belkin et al., 2019a; Mei et al., 2019; d'Ascoli et al., 2020). It was also studied with *Random Matrix Theory* (RMT) applied to single-layer random neural networks by analyzing the Gram matrix of the hidden units (Louart et al., 2018; Couillet et al., 2011). In another direction, in order to evaluate the *information bottleneck* theory in DNN (Tishby et al., 2015) suggested a connection with representation compression whereas (Gabrié et al., 2018) developed a rigorous scalable formula for mutual information between layers of multi-layer neural networks. Finally, another approach (Mehta et al., 2014) consists in applying ideas of the physics renormalization group to DNN whose idea is to learn hierarchal representations across layers.

Beyond separable priors Recently, we observed a practical and intense use of deep neural-network-based generative priors for estimation problems (Bora et al., 2017; Tramel et al., 2016b). Whereas in classical statistics, we often assume that the hidden ground truth signal is drawn from a separable prior, this overly simple hypothesis may be replaced by a generative prior to model a more complex, non-separable JPD of the signal. Therefore the practical use of generative priors angled research towards understanding them in simple estimation problems such as compressed sensing, phase retrieval of spike matrix models (Aubin et al., 2019c; Aubin et al., 2020c) and to design multi-layer approximate message passing algorithms (Manoel et al., 2017; Fletcher et al., 2018).

3.2 STATISTICAL INFERENCE AND CSP AS A STATISTICAL PHYSICS PROBLEM

In this section, we present the general approach used in the main contributions of this manuscript to analyze various models. Among them, we will consider two large classes of problems already mentioned: Statistical Inference (SI) and CSP. Both kind of problems can be formulated as a statistical physics model, and classical tools of disordered systems, presented in Sec. 4, readily apply in certain scaling limits. The connection between statistical physics, SI and CSP is not relatively new and has in fact a long history as sketched in Sec. 3.1. Especially, influential and seminal works (Shannon, 1948; Jaynes, 1957) brought to light the link between IT, Bayesian inference, thermodynamics and statistical physics. The connection was more recently renewed during the second AI winter (Grassberger et al., 2012) and was celebrated during a recent summer school *Statistical Physics and Machine Learning back together*. All along this manuscript, we stress and make an intense use of the deep connection between Bayesian inference and spin glass techniques (Mézard et al., 1987) that were early applied to error correcting codes (Sourlas, 1989), perceptrons (Seung et al., 1992; Watkin et al., 1993) and sparse random graphs (Mézard et al., 2001; Mézard et al., 2003). See (Nishimori, 2001; Mézard et al., 2009) for an extended review. Moreover, as early heuristic works were not focussing on algorithmic considerations, in this manuscript we will put the accent on rigorous results and algorithmic thresholds. Before closing this chapter and presenting in details the mean-field methods, we provide a high-level perspective of the general approach of this work. The same approach will be used on different problems and it is therefore useful to summarize it once for all. Finally we present the generic phase diagram descriptions of CSP and SI models.

3.2.1 BAYESIAN INFERENCE IN THE HIGH-DIMENSIONAL REGIME

3.2.1.A HIGH DIMENSIONAL REGIME

Classical statistics traditionally considers models with a finite number of parameters d. Yet, the recent progresses of DNN drove the usage of modern ML models with increasing number of parameters. Additionally with the widely increasing availability of data, the classical statistical regime must be rethought. In contrast with classical statistics, the sizes of the dataset n and the number of parameter d are assumed to very large and even *infinite*. These limits are particularly suitable to the statistical physics approach that requires the *thermodynamic limit* $d \to \infty$ to proceed. Therefore, we mostly consider that both the number of parameters $d \to \infty$ as well as the number of observed data $n \to \infty$. To be able to tackle analytically these ill-defined behaviors we assume that they both go to infinity with a fixed and finite ratio $\alpha \equiv \frac{n}{d} = \Theta(1)$. This simplifying assumption is however not arbitrary and reflects quite correctly the practical dimensions. For instance the MNIST (LeCun et al., 2010) dataset contains 60000 images and can be learned correctly by a two-layers network with biases with $d = (784+1) \times 32 + (32+1) \times 10 = 25450$ parameters, so that the ratio α is indeed of order one.

3.2.1.B STATISTICAL INFERENCE

Let us consider a set of interacting variables $\boldsymbol{\sigma}$ defined on a graph $\mathcal{G}(\mathrm{V},\mathrm{E})$. The overall goal in CSP and SI problems is to compute the *marginal distributions* $\mathrm{P}_d(\sigma_i) = \int_{\chi_{d-1}} \mathrm{d}\boldsymbol{\sigma}_{\setminus i} \mathrm{P}_d(\boldsymbol{\sigma})$ accessible from the knowledge of the JPD $\mathrm{P}_d(\boldsymbol{\sigma})$. However in the high-dimensional regime this JPD is a high-dimensional object that is very often not tractable analytically. That is where statistical physics comes into play. Indeed, statistical physics with its long-history provides a suitable and powerful set of tools, see Sec. 2, to analyze and characterize the corresponding high-dimensional JPD $\mathrm{P}_d(\boldsymbol{\sigma})$. Moreover its extension to disordered systems and spin glasses, see Sec. 2.3, makes it singular and a very powerful approach to analyze a JPD of the form $\mathrm{P}(\boldsymbol{\sigma}|\mathbf{y})$ in presence of a *quenched disorder* $\mathbf{y}$, such as the randomness in the *observed data*. In the perspective that we will apply these tools to supervised ML applications, let us draw the correspondence between physics spin models and SI. While in physics spin models, the randomness steps in through the exchange interactions $\mathbf{J}$ that follow a particular distribution $\mathrm{P}(\mathbf{J})$, see Sec. 2.3.1, in most of ML applications, the randomness intervenes in the distribution of the input data $\mathbf{X}$ and the corresponding labels $\mathbf{y}$ in a supervised setting. Thus, the spin configuration $\boldsymbol{\sigma}$ will naturally denote the value of the *parameters* of the ML model. More details on the connection between statistical physics and Bayesian-inference can be found in (Engel et al., 1993; Nishimori, 2001; Mézard et al., 2009; Grassberger et al., 2012; Zdeborová et al., 2016a; Advani et al., 2016b).

3.2.1.C BAYESIAN INFERENCE AS A STATISTICAL PHYSICS MODEL

In order to compute the JPD, we use a particularly suitable Bayesian approach based on the *Bayes-formula* decomposition, see Sec. 1.2.8.b,

$$\mathrm{P}_d\left(\boldsymbol{\sigma}|\mathbf{y}\right) = \frac{\mathrm{P}\left(\mathbf{y}|\boldsymbol{\sigma}\right)\mathrm{P}\left(\boldsymbol{\sigma}\right)}{\mathrm{P}\left(\mathbf{y}\right)} = \frac{1}{\mathrm{P}\left(\mathbf{y}\right)} \prod_{\mu=1}^{n} \mathrm{P}\left(y_\mu|\boldsymbol{\sigma}_{\partial_\mu}\right) \prod_{i=1}^{d} \mathrm{P}(\sigma_i)\,, \tag{46}$$

where we used the fact that, in many examples, the joint *channel* and *prior* distributions $\mathrm{P}\left(\mathbf{y}|\boldsymbol{\sigma}\right)$ and $\mathrm{P}\left(\boldsymbol{\sigma}\right)$ respectively factorize over the n observations and d input dimensions. This decomposition is very interesting in the sense it explicitly shows the distributions used to model the observations: the prior distribution $\mathrm{P}\left(\boldsymbol{\sigma}\right)$ describes the prior knowledge we have on the variables $\boldsymbol{\sigma}$, e. g. discrete binary, Laplace, Gaussian, etc. whereas the distribution $\mathrm{P}\left(\mathbf{y}|\boldsymbol{\sigma}\right)$ models how the observations are related to the variables, for instance through a noisy Gaussian channel, a linear matrix multiplication, etc. To properly cast this problem into a statistical physics formalism, we shall introduce the Hamiltonian

$$\mathcal{H}_d\left(\boldsymbol{\sigma};\mathbf{y}\right) \equiv -\sum_{\mu=1}^{n} \log \mathrm{P}\left(y_\mu|\boldsymbol{\sigma}_{\partial_\mu}\right) - \sum_{i=1}^{d} \log \mathrm{P}(\sigma_i) \tag{47}$$

so that the JPD in eq. (46) can be formulated as a Gibbs distribution

$$\mathrm{P}_d\left(\boldsymbol{\sigma}|\mathbf{y}\right) = \frac{e^{-\beta\mathcal{H}_d(\boldsymbol{\sigma};\mathbf{y})}}{\mathcal{Z}_d(\mathbf{y})}\,, \text{ with } \mathcal{Z}_d(\mathbf{y}) \equiv \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\, \mathrm{p}\left(\mathbf{y}|\boldsymbol{\sigma}\right)\mathrm{p}\left(\boldsymbol{\sigma}\right), \tag{48}$$

as soon as the inverse temperature is set to $\beta = 1$. Yet, the temperature parameter may be freely chosen depending on the statistical estimator that we will consider. For instance to obtain the MAP behavior, we should take the zero temperature limit $\beta \to \infty$. Moreover, in a physics language, $\log \mathrm{P}(\sigma_i)$ is exactly analogous to the local external field interaction $h_i\sigma_i$ in spin systems, while the term $\log \mathrm{P}(y_\mu|\boldsymbol{\sigma}_{\partial_\mu})$ represents the interaction term between $|\partial_\mu|$ variables. In particular, this analogy allows to compute easily the marginal probability $\mathrm{P}(\sigma_i) = \int_{\chi_{d-1}} \mathrm{d}\boldsymbol{\sigma}_{\setminus i}\mathrm{P}(\boldsymbol{\sigma})$ as a simple local magnetization. Written under this general formulation with generic prior distributions, it has the deep advantage to encompass a large class of models: Ising, SK, p-spin models, GLM, committee machine,... with a various choice of prior distributions ranging from discrete to continuous variables.

3.2.1.D FREE ENTROPY AND REPLICA COMPUTATION

As stressed in Sec. 2.2.1.f, the averaged free entropy $\Phi_d = \frac{1}{d}\mathbb{E}_{\mathbf{y}} \log \mathcal{Z}_d(\mathbf{y})$ being effectively the cumulant generative function of many useful quantities is therefore a central object in statistical physics. In the high-dimensional regime, $d \to \infty, n \to \infty, \alpha = \Theta(1)$, we focus instead in the asymptotic averaged free entropy $\Phi = \lim_{d\to\infty} \Phi_d$ that can be computed with the *replica method* that plays a crucial role in this work and detailed in Sec. 4.1. To quickly summarize, while computing the high-dimensional JPD is often intractable, the replica method allows to reduce the high-dimensional inference problem to a simple *optimization problem* of a free entropy potential Ψ over a set of a few *order parameters*, e. g. $q, \hat{q}$,

$$\Phi = \textbf{extr}_{q,\hat{q}}\{\Psi(q,\hat{q})\}\,. \tag{49}$$

The free entropy behavior allows to detect statistical thresholds and exhibit potential information theoretical phase transitions. Indeed, these *order parameters* $q, \hat{q}$, called *overlaps*, have a deep interpretation as they directly provide knowledge on the solution space: either the correlation with the ground truth solution in the case of SI, or the typical distance between solutions in the context of CSP.

3.2.1.e Towards rigor

Yet, the replica method, that we use intensively, is unfortunately not a rigorous method in the mathematical sense: a few important steps are not justified and may even seem absurd. But it turns out that in many cases the result was either proven or believed to be *correct*. As stressed in the historical part in Sec. 3.1, progresses and results in statistical physics have very often not been taken seriously in the mathematics and computer science community because of this lack of rigor. That is one of the reasons why researchers at the interface between mathematics and physics recently undertook to rigorously prove these results, which may have been obtained 20-30 years earlier by physicists. Even though proving the heuristic results does not provide new essential understanding of the system behavior, it nonetheless has the profound benefit to bring greater impact and visibility outside of physics. To give a flavor, most of the proofs of this manuscript will be based on *Guerra-interpolation* (Guerra et al., 2002a; Guerra, 2003; Talagrand, 2006a) that require fundamentally to previously derive the heuristic replica result. Other proofs are simply based on moment bounds (Friedgut, 1999; Achlioptas et al., 2002; Brémaud, 2017) and Gordon's Convex Gaussian Min-max Theorem (CGMT) (Gordon, 1985).

3.2.2 Algorithmic perspectives

The free entropy computation gives access to the information theoretical phase transitions of the system, also called the *statistical thresholds*. However, for practical purposes, we are interested in algorithms that are able to reach these theoretical performances. Another interesting feature of statistical physics approach is that, very often, computations can be turned in very powerful polynomial algorithms, whose behaviors show new phase transitions. Among them, in this dissertation we will focus on *message passing* algorithms which have a long history with physics. On top of that, they have the advantageous property of being proven optimal in many applications (Maleki, 2011; Donoho et al., 2013a; Barbier et al., 2016) or believed so. To derive such algorithms, the first step is to represent the high-dimensional JPD with a factor graph, as illustrated in Fig. 21. From this factor graph, we may apply the *cavity method* (Mézard et al., 2001; Mézard et al., 1987) or equivalently the BP equations, which can be simplified in an AMP algorithm under a Gaussian assumption in the thermodynamic regime. Notice that for more complex models, a more general approach was developed known as *survey propagation* (Braunstein et al., 2005). This set of iterative equations

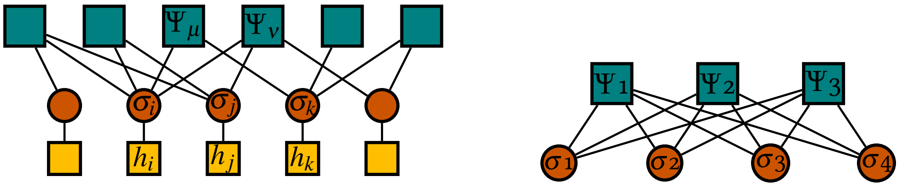


Figure 21: Factor graph representation of (**Left**) the joint probability distribution $\mathrm{P}_d(\boldsymbol{\sigma}|\mathbf{y})$ eq. (46), (**Right**) the linear system eq. (50).

of the form $\forall i \in [\![d]\!]$, $\hat{\sigma}_i^{t+1} = f_i(\hat{\boldsymbol{\sigma}}^t)$ can be iterated and gives an estimate of the marginal probabilities. In addition of being very often optimal, AMP algorithms have the exceptional particularity that in certain situations their average infinite behavior, called *state evolution*, is exactly characterize by the replica free entropy potential, that allows to compare their performances to the information theoretical statistical thresholds. Notably, if we define for instance the self-overlap parameter

$$q^t = \lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathbf{y}} \left[\boldsymbol{\sigma}^t \cdot \boldsymbol{\sigma}^t \right] ,$$

at convergence and under certain conditions, the performance measures such as the generalization error or the MSE are characterised by the asymptotic overlap $q^{t=\infty}$, which is, strikingly, equivalently the solution of the replica free entropy extremization problem in eq. (49). In particular, this means that at each iteration AMP follows the gradient of the replica free entropy, until convergence to a maxima. However, whereas the information theoretical performances are characterized by the global maxima of the replica free entropy, since AMP iterations start with non-informative initializations, the algorithm may converge to some local maximum and achieve sub-optimal performances. This key observation reveals in particular the existence of *hard* algorithmic regions as soon as the free entropy potential presents metastable states. To conclude, we already see that the replica free entropy and the AMP algorithm are two sides of the same coin, and this observation will follow in extended discussions in all the considered applications.

Polynomial refers to the space complexity, meaning that at each time iteration, the algorithm requires a polynomial (in the size of system) number of operations.

Hereafter, we present the two types of problems we will mainly describe in the application part of this manuscript: CSP and SI, which are very similar but do not show the same phase transitions typology because the quenched disorder is, crucially, not of the same nature. In fact historically physics was first interested in CSP (Mézard et al., 1986a; Mézard et al., 1986b; Gardner et al., 1988; Krauth et al., 1989; Mézard et al., 2002; Krzakała et al., 2007; Zdeborová et al., 2007) before recently shifting towards phase transitions in SI (Decelle et al., 2011; Krzakala et al., 2012b; Lesieur et al., 2017a; Barbier et al., 2019b). Even though the phase transitions are slightly different, the above general Bayesian approach readily apply to their analysis.

3.2.3 RANDOM CONSTRAINT SATISFACTION PROBLEMS

Let us introduce general combinatorial optimization problems, generally called CSP (Apt, 2003; Mézard et al., 2009; Tsang, 2014). This is a general definition that applies to various problems such as the k-SAT, sphere packing, Eulerian and Hamiltonian paths, travelling salesman, the q-coloring and the vertex-covering problems and many others. A CSP is specified by d variables $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{d} \in \chi_d$, where χ_d denotes an alphabet, that must satisfy a set of n constraints $\{\Psi_\mu(\boldsymbol{\sigma}_{\partial_\mu})\}_{\mu=1}^{n}$ within a given collection. We say that a constraint is satisfied by the variables (resp. non-satisfied) if $\Psi_\mu(\boldsymbol{\sigma}_{\partial_\mu}) = 1$ (resp. 0). The problem is satisfiable (SAT) if there exist at least one configuration that satisfies all the constraints, and UNSAT otherwise. As a generalization, random Constraints Satisfaction Problem (rCSP) are a particular case when the constraints are drawn randomly among the collection (Franco et al., 1983). The randomness may be based either on the graph geometry by a random connectivity or any other random source in the constraints for fully connected variables. Such models can be seen as spin glass models with *random quenched disorder* as the constraints are completely random. In order to maximize the number of satisfied constraints, we introduce an Hamiltonian defined as the number of violated constraints, that measures the energy of a configuration $\boldsymbol{\sigma}$

$$\mathscr{H}_d(\boldsymbol{\sigma}) \equiv \sum_{\mu=1}^{n} \left(1 - \Psi_\mu\left(\boldsymbol{\sigma}_{\partial_\mu}\right)\right),$$

and the Gibbs distribution at finite temperature $1/\beta$, that represents the level of exigence we require on the satisfiability of the constraints,

$$\mathrm{P}_d(\boldsymbol{\sigma}) = \frac{1}{\mathscr{Z}_d(\beta)} e^{-\beta\mathscr{H}_d(\boldsymbol{\sigma})} \underset{\beta\to\infty}{\longrightarrow} \frac{1}{\mathscr{Z}_d} \prod_{\mu=1}^{n} \Psi_\mu\left(\boldsymbol{\sigma}_{\partial_\mu}\right),$$

converges to a product of indicator functions at zero temperature $\beta \to \infty$. We introduce the rescaled number of constraints by the number of variables of the problem $\alpha = \frac{n}{d}$. To fix ideas, let us provide a simple example: solving an *affine system*. The interested reader may find various other examples in the literature starting with the reference book (Mézard et al., 2009).

A toy example: affine system of equations Consider you have a set of n *linear* equations, represented by the matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and a vector $\mathbf{b} \in \mathbb{R}^n$ with real coefficients, depending on d variables $\mathbf{x} \in \mathbb{R}^d$

$$\mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0} \quad \Leftrightarrow \quad \prod_{\mu=1}^{n} \Psi_\mu(\mathbf{x}) = 1\,. \tag{50}$$

This linear system can be rewritten as a fully-connected CSP involving the product of constraints $\{\Psi_\mu\}_{\mu=1}^{n}$ with $\Psi_\mu(\mathbf{x}) = \mathbb{1}\left[\mathbf{a}_\mu \cdot \mathbf{x} - b_\mu\right]$, where $\mathbf{a}_\mu$ represents the μ-th row vector of the matrix $\mathbf{A}$. The problem may be represented by a factor graph illustrated in Fig. 21 (**Right**). For a random or deterministic matrix $\mathbf{A}$, we would like to know when the linear system has at least one solution. We shall remember that if the number of constraints is smaller than the number of variables $n < d$, the linear system is undetermined and there exists many degenerated solutions. While if $n > d$ the system is over-constrained and there does not exist any solution, so that there exists an intermediate critical value $\alpha_{\rm sat} = 1$ such that solutions no longer exist, called the SAT-UNSAT phase transition.

On phase transitions of rCSP As illustrated with the above simple example, we understand intuitively that CSP and rCSP may undergo phase transitions as the constraints density α varies. In particular above a large number of constraints, if the system is heavily over-constrained it is intuitive that no configuration can be solution. In contrast, if the system is largely under-constrained, there will eventually exist many solutions. The SAT-UNSAT phase transition appears at the constraint density above which no solution exists. For instance in the case of the linear system, the SAT-UNSAT threshold is simply given by $n = d \Leftrightarrow \alpha_{\rm sat} = 1$. Yet, the phase diagram of rCSP is not limited to this SAT-UNSAT phase transition and reveals a richer description. We invite the reader to read more about it in (Krzakała et al., 2007; Mézard et al., 2009) where the whole phase transition phenomenology of rCSP is described with the *cavity* formalism. In more details, for a small constraint density $\alpha \ll 1$, we expect that many solutions exist in a large connected sub-region of parameters. As the density of constraints increases, a first remarkable *clustering* phase transition, also known as the dynamical phase transition in the context of structural glasses (Parisi et al., 2010; Charbonneau et al., 2017), is encountered at $\alpha_{\rm clust}$ and is characterized by the decomposition of the large set of solutions in an exponential number of disconnected sets, called clusters. Further increasing the density α, we observe that this exponential number of clusters reduces to a sub-exponential number of clusters at the *condensation* phase transition $\alpha_{\rm cond}$. If the alphabet χ_d is discrete, the system may as well undergo another phase transition: a *rigidity* or *freezing* phase transition (Semerjian, 2008) at $\alpha_{\rm freez}$ such that each cluster shrinks and contains only a finite number of solutions. Finally, above a certain constraint density $\alpha_{\rm sat}$, the problem becomes UNSAT, meaning that no configuration can satisfy all the hard constraints simultaneously such that at least one constraint is violated and the ground state energy is strictly positive. The description of the different phase transitions by increasing α is illustrated in Fig. 22, that can be completed by quantitative definitions (Krzakała et al., 2007; Gabrié et al., 2017).

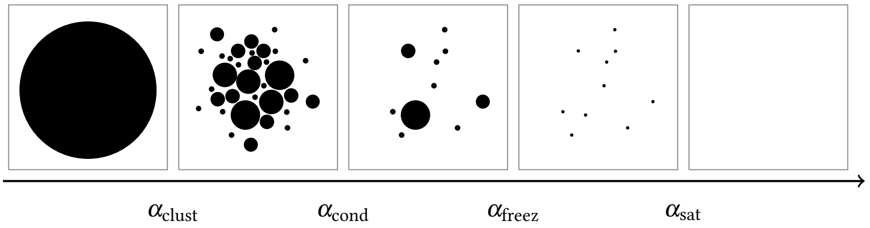


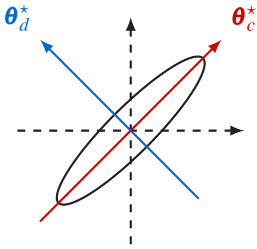
Figure 22: Illustration of the solution configuration space of a random CSP crossing clustering, condensation, freezing and SAT-UNSAT phase transitions as a function of the constraint density α, inspired from (Krzakała et al., 2007).

3.2.4 STATISTICAL INFERENCE AND SUPERVISED LEARNING

SI denotes the process of extracting useful informations and properties of an underlying high-dimensional joint probability distribution $\mathrm{P}_d(\boldsymbol{\sigma})$ from the observations of data. With the large amount of data nowadays available, statistical inference apply to many applications from computer science with ML, DL, signal processing and IT, to natural sciences with medicine, neuroscience, biology, social sciences or economy, etc. SI can be thought as the action of extracting informations from a large set of data or in other words recovering a hidden signal from a set of observations. The literature about IT and SI is very rich and more details can be found in (Barber, 2012; MacKay et al., 2003).

Ground truth representation The main difference between SI and CSP, depicted in the previous section, is that we assume there always exists a hidden solution to the problem, called the *ground truth* $\boldsymbol{\theta}^\star$ or *planted solution*, that we aim to recover. De facto, the SAT-UNSAT transition does not exist in inference problems and we expect the phase transition phenomenology to be different, yet even richer. The existence of the ground truth $\boldsymbol{\theta}^\star$ is ensured in many applications such as noisy communication channels, compressed sensing, phase retrieval, matrix factorization, and many others. In the case of supervised learning of real datasets, the ground truth is not explicit but we shall still assume it exists for our theoretical purposes. Indeed, even though the generative process of the data is hidden, we shall assume that the collected dataset contains a common hidden representation. For instance in a dataset containing images of cats and dogs, it seems natural to assume there exists some ground truth representations, yet unaccessible, $\boldsymbol{\theta}_c^\star, \boldsymbol{\theta}_d^\star$ that commonly characterize the images of cats and dogs.

We may imagine that the data projected in a particular space, potentially in higher dimensions, contains two principal components representing the images of cats and dogs



The teacher-student scenario and the planted ensemble In practice, the ground truth representation $\boldsymbol{\theta}^\star$ is not available for a direct comparison. Yet, for theoretical purposes, in order to measure the reconstruction performances of the hidden signal and depict the corresponding phase transitions, we naturally need to have access to the ground truth representation. To circumvent this difficulty, it gems from this idea the notion of hidden *rule* based on a signal $\boldsymbol{\theta}^\star$ that a *teacher* uses to generate a training set $\mathbb{X}_{\text{train}}$ (Patarnello et al., 1987; Gardner et al., 1989; Tishby et al., 1989; Sompolinsky et al., 1990; Seung et al., 1992; Watkin et al., 1993; Györgyi, 2001). This is called the T-S scenario: *student* aims to recover the hidden rule for the observations in the corresponding *synthetic* training set $\mathbb{X}_{\text{train}}$ generated by the *teacher*. Statistical inference and statistical physics show a narrow connection through the lens of this T-S and the *planted spin glass ensemble*. Indeed, in this context inferring the ground truth vector in statistical inference is similar to recovering a *crystal configuration* in planted spin glasses. One of the main advantage of this setting is that the Bayesian approach easily suits this framework and gives an optimal strategy that can be furthermore analyzed by statistical physics tools in the high-dimensional regime. Even though the T-S scenario and the randomly-quenched disorder in rCSP can be analyzed in a similar Bayesian approach, we shall keep in mind their striking difference that lead to very different phase diagrams typology. In particular, in the case of randomly-quenched disorder in rCSP, the observations correspond to independent random constraints, whereas in the case of the T-S scenario, the observations are not independent as they all depend on the ground truth hidden representation.

Statistical inference and estimators Let us define general inference problems we will focus on. One considers a $d-$dimensional hidden ground truth variable $\boldsymbol{\theta}^\star = \{\theta_i^\star\}_{i=1}^d \in \chi_d$, drawn from a probability distribution $\mathrm{P}_{\boldsymbol{\theta}^\star}$. The goal of SI is to infer it from n observations $\{\mathbf{X}, \mathbf{y}\} \in \mathbb{X}_{\text{train}}$ generated according to a *generative process*

$$\mathbf{y} = \varphi^\star_{\text{out}}\left(\mathbf{X}, \boldsymbol{\theta}^\star\right) \qquad \Leftrightarrow \qquad \mathbf{y} \sim \mathrm{P}^\star_{\text{out}}\left(.\right), \tag{51}$$

where $\varphi^\star_{\text{out}}$ represents a deterministic or stochastic function equivalently associated to a distribution $\mathrm{P}^\star_{\text{out}}$. Again, we introduce the parameter α as the ratio of the number of observations over the dimension of the problem, namely here $\alpha = \frac{n}{d}$. Inferring the above statistical model from observations $\{\mathbf{X}, \mathbf{y}\}$ can be tackled in several ways and consists in building an estimator $\hat{\boldsymbol{\theta}}$ that approaches the ground truth planted solution $\boldsymbol{\theta}^\star$. For instance we often use in this case the MSE $\ell(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|_2^2$, to measure the distance between the estimator $\hat{\boldsymbol{\theta}}$ and the hidden parameter $\boldsymbol{\theta}^\star$. Our Bayesian framework (46) is particularly suited to the analysis of two common estimators based on the high-dimensional, often intractable, posterior distribution (48). In one hand, the MMSE estimator for $\beta = 1$, consists in computing the mean of the of the posterior $\mathrm{P}\left(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}\right)$ according to

$$\hat{\boldsymbol{\theta}}_{\text{mmse}} = \mathbb{E}_{\mathrm{P}(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})}\left[\boldsymbol{\theta}\right],$$

which is well-known to minimize the MSE reconstruction error. In the other hand the MAP estimator consists in computing the maximum of the posterior distribution, that can be performed in the limit $\beta = \infty$. It can be formulated as a minimization problem according to

$$\hat{\boldsymbol{\theta}}_{\mathrm{map}} = \mathrm{argmax}_{\boldsymbol{\theta}} \mathrm{P}\left(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}\right) = \mathrm{argmin}_{\boldsymbol{\theta}} \left[\sum_{\mu=1}^{n} \ell\left(\boldsymbol{\theta}; y_\mu, \mathbf{x}_\mu\right) + r\left(\boldsymbol{\theta}\right) \right] ,$$

where the loss is simply mapped to $\ell\left(\boldsymbol{\theta};\mathbf{y},\mathbf{X}\right) = -\log \mathrm{P}\left(\mathbf{y}|\boldsymbol{\theta},\mathbf{X}\right)$ and the regularizer $r\left(\boldsymbol{\theta}\right) = -\log \mathrm{P}\left(\boldsymbol{\theta}\right)$, so that ERM can be analyzed in this framework. Thus both the study of MAP and MSE estimations can be casted in this general Bayesian approach and are simply reduced to the analysis of the posterior. Moreover, while the MMSE estimator is exactly the one performed by classical AMP algorithms, the MAP estimator can be thought as the ground state of the physical system and is closely related to ERM estimation performed by practical GD whose asymptotic behavior can be analyzed within this framework.

Bayes-optimal estimation and the Nishimori conditions In the idealistic case where the *student* knows all the correct prior distributions $\mathrm{P}_{\boldsymbol{\theta}} = \mathrm{P}_{\boldsymbol{\theta}^\star}$, $\mathrm{P}_{\mathrm{out}}\left(\mathbf{y}|\boldsymbol{\theta};\mathbf{X}\right) = \mathrm{P}_{\mathrm{out}^\star}\left(\mathbf{y}|\boldsymbol{\theta}^\star;\mathbf{X}\right)$, this scenario is called the *Bayes-optimal* setting. In the context of MSE reconstruction loss, performing the MMSE estimation in the Bayes-optimal case, yet unrealistic in practice, will be an important theoretical optimal baseline all along this manuscript. In this very specific and idealistic Bayes-optimal case, SI turns out to deeply simplify thanks to the Nishimori conditions (Opper et al., 1991a; Iba, 1999; Nishimori, 2001), presented in Appendix. A.3, and that will be intensely used in the following. These Nishimori conditions simply state that in average there is no statistical difference between the ground truth configuration and a configuration sampled uniformly at random from the posterior distribution, so that *overlaps* between the ground truth and the estimator is essentially the self-overlap of the estimator. As a consequence, under the Bayes-optimal assumption, the free entropy turns out to be exactly given by the *replica symmetric* ansatz. However, these powerful identities do not hold in the practical *mismatched setting* where the correct ground truth prior distributions are hidden during estimation.

Information theoretical phase transitions For the moment, without any algorithmic consideration, assuming we have access to exponential time and resources, we can already depict various *phase transitions* in the above SI problem, based on *information theoretical* predictions. They can be formalized from a quantitative analysis of the free entropy potential, but for conciseness we propose to only describe qualitatively the different phase transitions of the *optimal* estimator:

- With very few observations $\alpha \ll 1$, any algorithm is unlikely to infer correctly the hidden signal $\boldsymbol{\theta}^\star$. The estimator cannot *extract* any information correlated with the ground truth solution and the loss reaches its maximal value ρ: $\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^\star) = \rho$. This region is called the *undetectable phase* for $\alpha < \alpha_{\mathrm{weak}}$.

- From a certain number of samples α_{weak}, the estimator can *partially* reconstruct the signal such that the loss decreases but does not reach is minimal value $0 < \ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^\star) < \rho$, that corresponds to the *weak-recovery phase* $\alpha_{\mathrm{weak}} \leq \alpha < \alpha_{\mathrm{IT}}$.
- Above a critical observations density α_{IT}, it becomes theoretically possible to *perfectly* reconstruct the signal such that the loss $\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^\star) = 0$. This regime is called the *easy phase* for $\alpha \geq \alpha_{\mathrm{IT}}$.

Algorithmic phase transitions and computational efficiency With the recent success of ML applications, while statistics was often not concerned with algorithmic performances, the increasing number of parameters in the models raises the question of the computational efficiency. Hence, for practical purposes, we are interested in knowing if a particular algorithm can achieve the above *information theoretical* performances. Optimality of most algorithms is far to be theoretically guaranteed. Yet, in the case of MMSE estimation, the AMP algorithms under consideration are proven (or believed) to achieve *information theoretical* performances. However, very often, there exists some regions of parameters in which the optimal algorithmic reconstruction is not possible, while, theoretically, it is should be the case. Therefore, the *easy phase* shall be revised under this algorithmic perspective with *finite* resources. This region is called a *hard phase* that slots into the *weak recovery* phase and the *easy* phase: $\alpha_{\mathrm{IT}} < \alpha < \alpha_{\mathrm{alg}}$. It is related to the notions of computational complexity and notably to the distinctions between P, NP, and NP-complete classes. A more accurate discussion may be found in (Monasson et al., 1999; Percus et al., 2006; Arora et al., 2009; Moore et al., 2011). As a conclusion, the schematic phase diagram of SI is represented in Fig. 23. In the next sections, we will provide more details on these phase transitions and especially stress they have a clear and deep interpretation in the physics formalism.

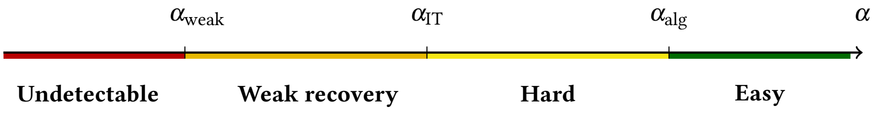


Figure 23: Illustration of the phase transitions happening in inference problems.

In the next chapter, we finally introduce the *mean-field* methods to analyze quantitatively the free entropy potential and depict quantitatively the phase diagrams of simple CSP and SI models.

4

FROM MEAN-FIELD METHODS TO ALGORITHMS

In this chapter, we propose a short review of the main mean-field methods used to study analytically spin glass models along this manuscript. Namely we start by presenting the *replica* method in Sec. 4.1, which is at the heart of this dissertation since it provides a powerful technique to compute the averaged quenched free entropy. It naturally gives access to the free entropy potential from which phase diagrams can be directly described. Next in Sec. 4.2, we discuss variational principles to derive various general mean-field methods. Finally, we present in Sec. 4.3 the BP equations that are a set of iterative equations, closely related to the cavity method (Mézard et al., 1987), and leading to a perfect inference algorithm on tree-like graphs. Under a Gaussian projection, the set of BP equations can be simplified to the AMP algorithm, highlighted in Sec. 4.4. The literature is quite extensive on the subject and the interested reader may find more details in (Mézard et al., 1987; Mézard et al., 2009; Zdeborová et al., 2016a; Advani et al., 2017; Gabrié, 2020).

4.1 THE REPLICA METHOD: A POWERFUL HEURISTIC MEAN-FIELD METHOD

This section is devoted to present the powerful *replica method* introduced in (Kac, 1968; Edwards et al., 1975) and reviewed in (Mézard et al., 1987; De Dominicis et al., 2006; Parisi et al., 2020). This method allows to tackle the logarithmic difficulty in the computation of the average over the quenched disorder $\mathbf{J}$ in the free entropy (42)

$$\Phi_d\left(\beta,\mathbf{J}\right) \equiv \frac{1}{d}\mathbb{E}_{\mathbf{J}} \log \mathscr{Z}_d\left(\beta,\mathbf{J}\right) . \qquad (52)$$

The method fundamentally relies on the so-called *replica trick*, which is a simple mathematical identity, that carries nonetheless profound physical consequences.

4.1.1 REPLICA TRICK

The replica trick is a simple identity that allows to exchange the expectation over the disorder and the logarithm, in exchange of computing the $r \in \mathbb{N}$ moments of the partition function $\mathcal{Z}_d^r$ according to

$$\mathbb{E}_{\mathbf{J}} \left[\log \mathcal{Z}_d \right] = \lim_{r \to 0} \frac{\partial \log \mathbb{E}_{\mathbf{J}} \left[\mathcal{Z}_d^r \right]}{\partial r} . \tag{53}$$

Proof. Suppose $r \in \mathbb{R}$ close to zero, then

$$\mathcal{Z}_d^r = e^{r \log \mathcal{Z}_d} = 1 + r \log \mathcal{Z}_d + \mathrm{o}(r) \quad \Rightarrow \quad \log \mathcal{Z}_d = \lim_{r \to 0} \frac{\mathcal{Z}_d^r - 1}{r} .$$

By exchanging the limit $r \to 0$ and the expectation, and assuming that $r \in \mathbb{N}$, we obtain

$$\mathbb{E}_{\mathbf{J}} \left[\log \mathcal{Z}_d \right] = \mathbb{E}_{\mathbf{J}} \left[\lim_{r \to 0} \frac{\mathcal{Z}_d^r - 1}{r} \right] = \lim_{r \to 0} \partial_r \log \left(\mathbb{E}_{\mathbf{J}} \left[\mathcal{Z}_d^r \right] \right) .$$

□

As a result, the replica trick reduces the quenched average of the logarithm to the average of the moments of the partition function $\mathcal{Z}_d^r$, that are more tractable. Moreover, as soon as $r \in \mathbb{N}$, the moment $\mathcal{Z}_d^r$ represents in fact the product of r identical partition functions, namely the partition function of a system containing r non-interacting copies, called *replicas*, of the original system

$$\mathcal{Z}_d^r (\beta, \mathbf{J}) = \prod_{a=1}^{r} \mathcal{Z}_d^a (\beta, \mathbf{J}) = \prod_{a=1}^{r} \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}^a \, e^{-\beta \mathcal{H}_d (\boldsymbol{\sigma}^a ; \mathbf{J})} , \tag{54}$$

where $a \in [\![r]\!]$ denotes the replica indices. However, under the disorder average, the initial r non-interacting replicas are transformed in a highly non-trivial interacting particles problem charaterized by a matrix order parameter $\mathbf{Q} \in \mathbb{R}^{r \times r}$

$$\mathbb{E}_{\mathbf{J}} \mathcal{Z}_d^r (\beta, \mathbf{J}) = \int_{\mathbb{R}^{r \times r}} \mathrm{d}\mathbf{Q} \, e^{\Phi^{(r)}(\mathbf{Q})} , \tag{55}$$

where $\Phi^{(r)}$ denotes the replica potential. This simple mathematical trick has profound consequences as non-trivial properties can emerge from the interactions between these *coupled replicas*. Additionally, notice that the average of the replicated partition function has substituted the initial exponentially large summation $\mathbb{E}_{\boldsymbol{\sigma}}$ by an analytical formula involving a new order parameter. In return, the difficulty is from now on to analyze the complex structure of the order parameter $\mathbf{Q}$. In particular, the initial invariance of the replicas can be conserved in certain situations. This solution is called the Replica Symmetry (RS) Ansatz, in contrast to Replica Symmetry Breaking (RSB) Ansätze in which the mean-field solution breaks the initial invariance of the replicas permutation. As soon as the symmetry is broken, choosing the correct structure for the matrix $\mathbf{Q}$ in the replica space is highly non-trivial. As a conclusion, the replica trick and the average of the replicated partition function substituted the complex analysis of interacting disordered models to finding the values of a matrix order parameter of finite size. In general, they can be found as the solution of a closed set of non-linear equations that require only a polynomial number of operations.

4.1.2 PURE STATES AND OVERLAP DISTRIBUTION

Analyzing the overlap matrix distribution becomes essential to understand the behavior of this new interacting problem. In this end, we introduce the probability distribution averaged over the quenched disorder $\mathbf{J}$

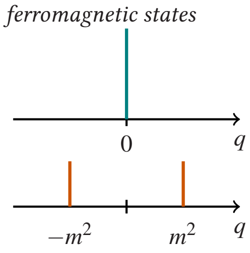
$$\mathrm{P}(q) = \mathbb{E}_{\mathbf{J}} \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\, e^{-\beta \mathscr{H}_d(\boldsymbol{\sigma},\mathbf{J})} \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}'\, e^{-\beta \mathscr{H}_d(\boldsymbol{\sigma}',\mathbf{J})} \frac{\delta\left(q - \frac{1}{d}\boldsymbol{\sigma}\cdot\boldsymbol{\sigma}'\right)}{\mathscr{Z}_d(\beta)^2} \,.$$

that two configurations $\boldsymbol{\sigma}, \boldsymbol{\sigma}'$ have a mutual overlap q at equilibrium. The overlap distribution $\mathrm{P}(q)$ reveals important knowledges about the thermodynamics of the model and especially the distance between typical equilibrium configurations. In particular, the Gibbs measure at equilibrium is carried by a few *pure states* (Mézard et al., 1987) that respectively describe distinct ergodic connected components of the configuration space. Indeed, denoting α these pure states, the Gibbs average can be decomposed as

$$\langle\cdots\rangle_\beta = \sum_\alpha \underbrace{\frac{\mathscr{Z}_\alpha(\beta)}{\mathscr{Z}(\beta)}}_{w_\alpha(\beta)} \underbrace{\int \mathrm{d}\boldsymbol{\sigma}_\alpha \cdots \frac{e^{-\beta\mathscr{H}_d(\boldsymbol{\sigma},\mathbf{J})}}{\mathscr{Z}_\alpha(\beta)}}_{\langle\cdots\rangle_\alpha} = \sum_\alpha w_\alpha(\beta)\langle\cdots\rangle_\alpha \,,$$

with $w_\alpha(\beta)$ the thermodynamic weight of the state α that contributes to the non-trivial structure of overlap distribution $\mathrm{P}(q)$. In the presence of different pure states α, β, we define the overlap between them $q_{\alpha\beta} = \frac{1}{d}\sum_{i=1}^d \langle\sigma_i\rangle_\alpha \langle\sigma_i\rangle_\beta$ so that the averaged overlap distribution reads $\mathrm{P}(q) = \sum_{\alpha,\beta} w_\alpha w_\beta \delta\left(q - q^{\alpha\beta}\right)$ because of the *clustering property* $\langle\sigma_i\sigma_j\rangle_\alpha = \langle\sigma_i\rangle_\alpha\langle\sigma_j\rangle_\alpha$. As an illustration, in the Curie-Weiss model presented in Sec. 2.2.3.b, we observed that below the critical inverse temperature $\beta^\star = 1$ there exists a single pure paramagnetic state $q = 0$, such that the distribution contains a single ergodic component $\mathrm{P}(q) = \delta(q)$. Above the critical temperature, we observed the emergence of two ferromagnetic states $q = -m^2$ and $q = m^2$ with $m > 0$, such that the distribution splits into two connected components $\mathrm{P}(q) = \frac{1}{2}\delta(q - m^2) + \frac{1}{2}\delta(q + m^2)$.

Overlap distribution for a single paramagnetic pure state and two ferromagnetic states



4.1.3 REPLICA ANSATZ

In general, the full replica computation boils down to a Lagrangian similar to (55) expressed in terms of a symmetric matrix order parameter $\mathbf{Q}$. The computation for unconstrained symmetric matrices is unfortunately intractable, and (Parisi, 1983) proposed an iterative scheme to approximate the corresponding overlap distribution $\mathrm{P}(q)$. We present the RS and RSB simple Ansätze that turn out to be stable in various models.

4.1.3.A REPLICA SYMMETRIC

The simplest RS Ansatz is particular as it assumes that the initial permutation invariance of the fictive replicas is conserved so that the overlap between two arbitrary replicas is identical and fixed to $q_0 = \frac{1}{d}\boldsymbol{\sigma}^{(a)} \cdot \boldsymbol{\sigma}^{(b)}$, $\forall (a,b) \in [\![r]\!]^2$. The overlap distribution is therefore given by $\mathrm{P}^{(\mathrm{rs})}(q) = \delta(q - q_0)$ such that the overlap matrix $\mathbf{Q}^{(\mathrm{rs})} = (Q - q_0)\mathbf{I}_r + q_0 \mathbf{J}_r \in \mathbb{R}^{r\times r}$, illustrated in Fig. 24, where $Q = \frac{1}{d}\|\boldsymbol{\sigma}\|_2^2$ denotes the self-overlap. This Ansatz turned out to be stable in many situations such as on the Nishimori line in the context of the SK model (Nishimori, 1980; Nishimori, 1981; Georges et al., 1985) or in the *Bayes-optimal* setting in SI, where the Nishimori conditions detailed in Appendix. A.3 allow to rigorously prove the validity of the RS Ansatz.

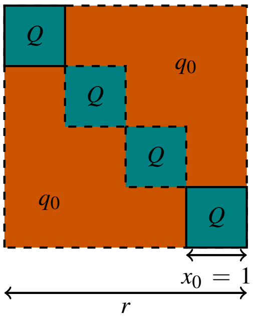


Figure 24: Illustration of the replica symmetric overlap matrix $\mathbf{Q}^{(\mathrm{rs})}$.

4.1.3.B ANSATZ STABILITY: DE ALMEIDA-THOULESS TRANSITION

Otherwise, the correctness of a given Ansatz can be highlighted by estimating its *stability* (Almeida et al., 1978; Thouless et al., 1977). For instance, in the case of the RS Ansatz, its stability is evaluated by expanding the replica potential $\Phi^{(r)}(\mathbf{Q})$ in eq. (55) around the RS fixed point

$$\Phi^{(r)}(\mathbf{Q}) = \Phi^{(r)}(\mathbf{Q}^{(\mathrm{rs})}) - \frac{1}{2} \sum_{a<b,c<d} \delta \mathbf{Q}^{ab} \mathscr{M}^{ab,cd} \delta \mathbf{Q}^{cd} \,. \tag{56}$$

By studying the fluctuations and eigenvalues of the Hessian matrix $\mathscr{M}^{ab,cd} = -\frac{\partial^2 \Phi^{(r)}}{\partial \mathbf{Q}^{ab} \partial \mathbf{Q}^{cd}}\big|_{\mathbf{Q}=\mathbf{Q}^{(\mathrm{rs})}}$, we can detect the so-called de Almeida Thouless (dAT) transition that occurs when the RS Ansatz becomes unstable, i. e. when the first negative eigenvalue appears, called in this context the replicon eigenvalue. Notice that the technique is not limited to this latter Ansatz and can be applied to more complex ones.

4.1.3.c REPLICA SYMMETRY BREAKING AND ERGODICITY BREAKING

In case the RS Ansatz is unstable and leads sometimes to unphysical results such as negative entropies (Gardner et al., 1988), more complex Ansätze should be investigated above the dAT line. Constructing such an Ansatz is not easy as it should respect physical constraints such as the positivity of the entropy and the overlap distribution $P(\mathbf{Q})$, and it should be stable with respect to Gaussian fluctuations. The first step forward was introduced in (Blandin, 1978; Blandin et al., 1980) where the idea of breaking the replicas symmetry into blocks emerged and it was further developed in (Sommers, 1978; Sommers, 1979; C. et al., 1979; Bray et al., 1980). Yet the permutation symmetry may be broken in many ways such that finding the correct Ansatz was the main focus of most theoretical research works in the spin glass literature (Sherrington et al., 1975; Derrida, 1981; Gardner et al., 1988; Crisanti et al., 1992). Finally, the general solution was delivered by Parisi in a series of works (Parisi, 1979; Parisi, 1980b; Parisi, 1980a; Parisi, 1983), in which he proposed a general scheme, which respect all the physical constraints, for progressively breaking the replica symmetry, called RSB, that eventually leads to the correct solution This scheme predicts that the stable Ansatz should perform a *infinite* and *continuous* hierarchy of symmetry breaking, the so-called Full Replica Symmetry Breaking (FRSB) Ansatz. However, very often the One-step Replica Symmetry Breaking (1RSB),Two-steps Replica Symmetry Breaking (2RSB) Ansätze give very accurate approximations that avoid to solve numerically the cumbersome FRSB equations. In the context of the SK model (Sherrington et al., 1975), this FRSB Ansatz turned out to be exact and was rigorously proven later on in (Guerra, 2003; Talagrand, 2006b).

RSB Parisi's scheme For the sake of illustration, let us illustrate the Parisi's scheme for breaking the replicas symmetry. The overlap matrices and distributions in the 1RSB and 2RSB Ansätze can be written as follows

$$\mathbf{Q}^{(1\mathrm{rsb})} = (Q-q_1)\mathbf{I}_r + (q_1-q_0)\mathbf{I}_{\frac{r}{x_0}} \otimes \mathbf{J}_{x_0} + q_0\mathbf{J}_r$$
$$\mathrm{P}^{(1\mathrm{rsb})}(q) \underset{r\to 0}{\longrightarrow} (1-x_0)\delta(q-q_1) + x_0\delta(q-q_0)$$

and

$$\mathbf{Q}^{(2\mathrm{rsb})} = (Q-q_2)\mathbf{I}_r + (q_2-q_1)\mathbf{I}_{\frac{r}{x_1}} \otimes \mathbf{J}_{x_1} + (q_1-q_0)\mathbf{I}_{\frac{r}{x_0}} \otimes \mathbf{J}_{x_0} + q_0\mathbf{J}_r$$
$$\mathrm{P}^{(2\mathrm{rsb})}(q) \underset{r\to 0}{\longrightarrow} (1-x_1)\delta(q-q_2) + (x_1-x_0)\delta(q-q_1) + x_0\delta(q-q_0)$$

and the corresponding overlap matrices are depicted in Fig. 25. Therefore, the RSB leads to consider that replicas play different roles and are clustered in different states with different inner and outer correlations, respectively q_1 and q_0 in the context of the 1RSB Ansatz with $q_1 > q_0$. In contrast the RS Ansatz, in the context of the RSB the ergodicity is broken in a nontrivial way and the phase space is organized into a hierarchical structure of pure states. In particular, given the multiplicity of ergodic components in the RSB Ansatz, that mainly appear at low temperature, the thermodynamic averages performed with the Gibbs measure are not equivalent to the average inside one state but it rather takes into account the presence of all the states. The overlap matrix $\mathbf{Q}$ is therefore hierarchically constant by *blocks* for a finite number k of RSB steps. Nonetheless, the Parisi scheme can be repeated for an infinite number of steps $k = \infty$, reaching a continuous limit and the so-called FRSB solution scheme as illustrated in Fig. 26. As an illustration, by iterating the Parisi's scheme, the 2RSB Ansatz can be obtained by simply imposing a similar fractal structure within the smallest blocks of the 1RSB Ansatz.

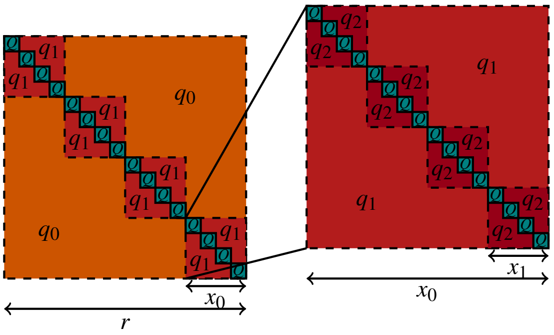


Figure 25: Illustration of the Parisi scheme: the 2-step RSB Ansatz $\mathbf{Q}^{(2\text{rsb})}$ is obtained by repeating the hierarchal structure inside the red block of the 1-step RSB Ansatz $\mathbf{Q}^{(1\text{rsb})}$ (**Left**).

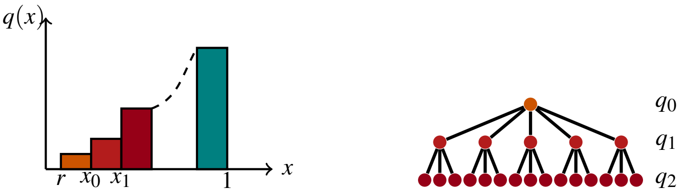


Figure 26: Illustration of the Parisi iterative scheme of the overlap distribution which reflects the multiplicity of ergodic components in RSB solutions. (**Left**) Evolution of the distribution of overlaps from a constant by parts to a continuous function $q(x)$ for an infinite number of RSB steps. (**Right**) Illustration of the hierarchical structure of the overlaps.

4.1.4 COMPLEXITY AND METASTABLE STATES

The replica method and the hierarchical of the FRSB scheme naturally reveal the ergodicity breaking and the existence of *metastable states* in spin glasses. A metastable state represents a region of configuration space separated from the rest of the space by a free energy barrier that diverge with the size of the system. Therefore, to escape this locally attracting valley, we shall go across higher free energy barriers. Equivalently, from the dynamic point of view, a metastable state is a region where the system will remain confined for finite times and could escape it only in a time scaling with the size of the system. The analysis of the *p-spin* model in particular (Thouless et al., 1977; Rieger, 1992; Crisanti et al., 1992) revealed that this spin glass model had a very large number of metastable states, i. e. locally stable solution with free energy higher than the ground state free energy. Moreover, this number of states $\mathcal{N}_\Phi$ turned out to scale exponentially with the size d of the system. As a result, to take into account the multiplicity of the metastable states, we define a new entropy measure, called the *complexity* or the *configurational entropy* in the glass community, defined as the logarithm of $\Omega(\Phi)$ the number of states at a given free entropy, i. e. $\Sigma(\Phi) = \frac{1}{d}\log\Omega(\Phi)$. The existence of such metastable states has potentially harmful consequences on dynamic systems such as structural glasses or optimization algorithms since they could eventually get stuck in metastable local minimum for an exponential time.

Complexity computation In order to quantify the existence of metastable states, (Monasson et al., 1995b) proposed a general method, comprehensively reviewed in (Zamponi, 2010), to compute the complexity $\Sigma(\Phi)$ as a function of the free entropy of the states. The idea consists in considering m *real replicas* of the original system that are coupled by a small interacting term that will push all copies in the same pure state. The total replicated free entropy of the m replicas can be well approximated by the sum of the contributions over all the states

$$\begin{aligned}\mathcal{Z}_m \equiv e^{d\Phi_m(m,\beta)} &= \sum_\alpha \exp\left(dm\Phi_\alpha\right) = \int \mathrm{d}\Phi \sum_\alpha \delta(\Phi-\Phi_\alpha)e^{dm\Phi} \\ &= \int \mathrm{d}\Phi\Omega(\Phi)e^{dm\Phi} = \int \mathrm{d}\Phi \ \exp\left(d\left(m\Phi+\Sigma\left(\Phi\right)\right)\right).\end{aligned}$$

In the thermodynamic limit $d\to\infty$, a Laplace method (Wong, 1989) allows to write the replicated free entropy as the Legendre transform of the complexity:

$$\Phi_m(m,\beta) = \max_\Phi\{m\Phi+\Sigma(\Phi)\} = m\Phi^\star(\beta)+\Sigma\left(\Phi^\star\right), \tag{57}$$

where the equilibrium free entropy $\Phi^\star$ can be computed as

$$\Phi^\star(\beta) = \frac{\mathrm{d}\Phi_m(m,\beta)}{\mathrm{d}m} \qquad \text{and} \qquad \Sigma\left(m,\beta\right) = \Phi_m(m,\beta) - m\Phi^\star(\beta).$$

Varying the Legendre parameter m at fixed temperature β, we can reconstruct the full complexity function $\Sigma(m, \beta)$ from the knowledge of the replicated partition function, which turns out to be closely related to the 1RSB free entropy $\Phi_m(m, \beta) = m\Phi^{(1\text{rsb})}(\beta)$.

4.1.5 APPLICATION - REPLICA COMPUTATION OF THE GLM

In this section, we finally illustrate how the replica method developed in the context of the spin glass theory can be readily applied to simple supervised ML models such as the GLM, defined in Sec. 1.2.9.a.

The generalized linear estimation problem consists to fit n observations $\mathbb{X}_{\text{train}} = \{\mathbf{X}, \mathbf{y}\}$ with a linear parametric model with weights $\mathbf{w} \in \mathbb{R}^d$ according to

$$\mathbf{y} = \varphi_{\text{out}}\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}\right) .$$

In other words, we try to fit the observation $\mathbf{y} \in \mathbb{R}^n$, which can be either discrete *labels* or continuous *outputs*, with a linear transformation of the *input* data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, up to a component-wise non-linear activation function $\varphi_{\text{out}^\star} : \mathbb{R} \mapsto \mathbb{R}$ which can be deterministic or stochastic. Moreover, we assume that the matrix of data inputs $\mathbf{X} \in \mathbb{R}^{n\times d}$ is drawn i.i.d with density $p_{\mathbf{X}}$. Specifically we will consider them to be Gaussian with zero mean and unit variance, namely $\forall \mu \in [\![n]\!]$, $\mathbf{x}_\mu \sim \mathcal{N}_{\mathbf{x}}(\mathbf{0}, \mathbf{I}_d)$.

On the data generative process As stressed in Chap. 3.2, different settings have been considered in the physics literature on how the ground truth observations $\mathbf{y}$ relate to the inputs $\mathbf{X}$. In particular (Gardner et al., 1989) in their influential paper introduced the two main generative processes constantly studied in the subsequent literature:

- The *random labels* setting: the labels $\mathbf{y}$ are uncorrelated from the input data $\mathbf{X}$. Namely,

$$\forall \mu \in [\![n]\!],\ y_\mu \sim \mathrm{P}_{\mathrm{y}}(.) \quad \text{and} \quad \mathbf{x}_\mu \sim \mathrm{P}_{\mathrm{x}}(.) \text{ with } y_\mu \perp \mathbf{x}_\mu . \tag{58}$$

  This setting has been studied in particular for perceptrons in (Gardner et al., 1988; Krauth et al., 1989) in the context of rCSP, see Sec. 3.2.3. Indeed the *randomly quenched disorder* over the input $\mathbf{X}$ and $\mathbf{y}$ are not correlated such that trying to fit this dataset can be equivalently seen as trying to satisfy random constraints.

- The *teacher-student* scenario or equivalently the *planted spin glass* model: the labels $\mathbf{y}$ are generated from a synthetic model designed by a *teacher*, from which, in the context of SI, the *student* aims to recover the teacher's parameters. Here we consider the ground truth as a linear model with weights $\mathbf{w}^\star$ according to the channel $\mathbf{y} = \varphi_{\text{out}^\star}\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}^\star\right)$ or equivalently

$$\mathbf{y} \sim \mathrm{P_y}(\mathbf{y}|\mathbf{X}) = \int_{\mathbb{R}^n} \mathrm{d}\mathbf{z}^\star \; \mathrm{p}_{\text{out}^\star}(\mathbf{y}|\mathbf{z}^\star) \\ \times \int_{\mathbb{R}^d} \mathrm{d}\mathbf{w}^\star \; \mathrm{p}_{\mathrm{w}^\star}(\mathbf{w}^\star)\delta\left(\mathbf{z} - \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}^\star\right), \tag{59}$$

with generic teacher densities $\mathrm{p}_{\mathrm{w}^\star}, \mathrm{p}_{\text{out}^\star}$. This T-S scenario perfectly fits in a supervised learning setting mentioned in Sec. 3.2.4. In this section, we assume that the student must infer the rule designed by the teacher, where both *teacher* and *student* belong to the same hypothesis class.

The full computation in the case of *random labels*, used in particular in Chap. 6-7, detailed in Appendix. B.2 is very similar and even simpler. In the T-S setting, the replica computation in the GLM for i.i.d data has been performed in many works such as (Schülke, 2016) and has been generalized to rotationally invariant matrices in (Kabashima, 2008). For the sake of illustration, in this section we show only the main steps of the replica computation for the GLM and we leave the cumbersome details in Appendix. B.1.1, presented in the context of the more general *committee machines* hypothesis class. Committee machines, that we investigate in Chap. 5, use instead K GLM estimators simultaneously, so that its parameters is a matrix $\mathbf{W} \in \mathbb{R}^{d\times K}$, to fit the training set according to

$$\mathbf{y} = \varphi_{\text{out}}\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}\right) = \varphi_{\text{out}}\left(\left\{\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}_k\right\}_{k=1}^{K}\right),$$

where $\varphi_{\text{out}} : \mathbb{R}^K \mapsto \mathbb{R}$. As a consequence, the classical GLM, we present in this section, is a particular case of committee machines for $K = 1$. Nonetheless, GLM are a wide class of linear models with various applications such as

- Compressed sensing: $\varphi_{\text{out}^\star}(y|z) = z + \sqrt{\Delta}\xi$,
- Phase retrieval: $\varphi_{\text{out}^\star}(y|z) = |z| + \sqrt{\Delta}\xi$,
- Perceptron: $\varphi_{\text{out}^\star}(y|z) = \text{sign}(z) + \sqrt{\Delta}\xi$,

where $\xi \sim \mathcal{N}(0,1)$ represents a potential Gaussian noise scaled by a variance $\Delta \geq 0$. Moreover the ground truth vector $\mathbf{w}^\star$ can be drawn according to common prior distributions such as

- Gaussian weights: $\mathrm{P}_{\mathrm{w}^\star}(\mathbf{w}^\star) = \mathcal{N}_{\mathbf{w}^\star}(\mathbf{0}, \rho_{\mathrm{w}^\star}\mathbf{I}_d)$,
- Spherical weights: $\mathrm{P}_{\mathrm{w}^\star}(\mathbf{w}^\star) = \delta\left(\|\mathbf{w}^\star\|_2^2 - d\right)$,
- Binary weights: $\mathrm{P}_{\mathrm{w}^\star}(\mathbf{w}^\star) = \prod_{i=1}^d \frac{1}{2}\left(\delta(w_i^\star - 1) + (\delta(w_i^\star + 1)\right.$.

On statistical estimation As stressed in Sec. 1.2.8.b, MMSE and MAP estimations boil down to the analysis of the joint distribution $\mathrm{P}_d\left(\mathbf{y},\mathbf{X}\right)$ involved in the high-dimensional posterior JPD according to the Bayes formula

$$\mathrm{P}_d\left(\mathbf{w}|\mathbf{y},\mathbf{X}\right)=\frac{\mathrm{P}\left(\mathbf{y}|\mathbf{w},\mathbf{X}\right)\mathrm{P}\left(\mathbf{w}\right)}{\mathrm{P}_d\left(\mathbf{y},\mathbf{X}\right)}=\frac{\mathrm{P}_{\mathrm{out}}\left(\mathbf{y}|\mathbf{w},\mathbf{X}\right)\mathrm{P}_{\mathrm{w}}\left(\mathbf{w}\right)}{\mathcal{Z}_d\left(\{\mathbf{y},\mathbf{X}\}\right)}\,. \tag{60}$$

To explicitly connect with the spin glass approach, the distribution $\mathrm{P}_d\left(\mathbf{y},\mathbf{X}\right)=\mathcal{Z}_d\left(\{\mathbf{y},\mathbf{X}\}\right)$ is also called the *partition function* and we define the corresponding Hamiltonian, for separable prior distributions $\mathrm{P}_{\mathrm{out}},\mathrm{P}_{\mathrm{w}}$ as

$$\begin{aligned}\mathcal{H}_d\left(\mathbf{w},\{\mathbf{y},\mathbf{X}\}\right)&=-\log\mathrm{P}_{\mathrm{out}}\left(\mathbf{y}|\mathbf{w},\mathbf{X}\right)-\log\mathrm{P}_{\mathrm{w}}\left(\mathbf{w}\right)\,,\\&=-\sum_{\mu=1}^{n}\log\mathrm{P}_{\mathrm{out}}\left(y_\mu|\mathbf{w},\mathbf{x}_\mu\right)-\sum_{i=1}^{d}\mathrm{P}_{\mathrm{w}}\left(w_i\right)\,.\end{aligned}$$

The spin variables denote the linear model weights $\mathbf{w}\in\mathbb{R}^d$ that interact through the quenched dataset $\{\mathbf{y},\mathbf{X}\}$, which plays the role of the exchange interaction. However, here the interactions are not *pairwise*, but instead *fully connected*, meaning that each variable w_i is connected to $\{w_j\}_{j\in\partial i\setminus i}$ through the factors $\mathrm{P}_{\mathrm{out}}\left(y_\mu|\mathbf{w},\mathbf{x}_\mu\right)$. The corresponding factor graph is represented in Fig. 27 and the partition function at temperature β is defined by

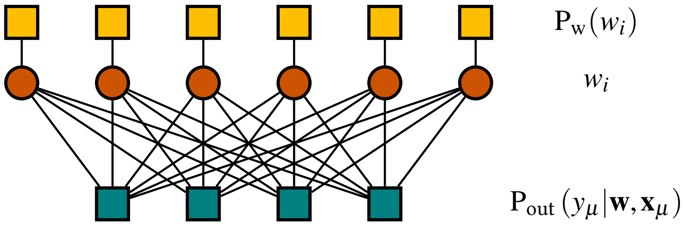

$\mathrm{P}_{\mathrm{w}}(w_i)$
w_i
$\mathrm{P}_{\mathrm{out}}\left(y_\mu|\mathbf{w},\mathbf{x}_\mu\right)$

Figure 27: Factor graph representing the GLM class. The variables w_i are fully connected through the factor $\mathrm{P}_{\mathrm{out}}\left(y_\mu|\mathbf{w},\mathbf{x}_\mu\right)$ that represent the constraint imposed by the μ-th example in the dataset. Each variable is connected to a one-body interaction with a separable prior distribution $\mathrm{P}_{\mathrm{w}}(w_i)$.

$$\mathcal{Z}_d\left(\{\mathbf{y},\mathbf{X}\};\beta\right)\equiv\mathrm{P}_d\left(\mathbf{y},\mathbf{X}\right)=\int_{\mathbb{R}^d}\mathrm{d}\mathbf{w}\;e^{-\beta\mathcal{H}_d(\mathbf{w},\{\mathbf{y},\mathbf{X}\})} \tag{61}$$

$$=\int_{\mathbb{R}^d}\mathrm{d}\mathbf{w}\;e^{\beta\left(\log\mathrm{p}_{\mathrm{out}}(\mathbf{y}|\mathbf{w},\mathbf{X})+\log\mathrm{p}_{\mathrm{w}}(\mathbf{w})\right)}=\int_{\mathbb{R}^d}\mathrm{d}\mathbf{w}\;\mathrm{p}_{\mathrm{out}}\left(\mathbf{y}|\mathbf{w},\mathbf{X}\right)\mathrm{p}_{\mathrm{w}}\left(\mathbf{w}\right)\,,$$

and can be mapped to Bayesian estimation for $\beta=1$. In the considered modern high-dimensional regime with $d\to\infty$, $n\to\infty$ with $\alpha=n/d=\Theta(1)$, we are interested to compute the *free entropy* Φ *averaged* over the input data $\mathbf{X}$ and teacher weights $\mathbf{w}^\star$ or equivalently over the output labels $\mathbf{y}$, defined as

$$\Phi(\alpha)\equiv\lim_{d\to\infty}\frac{1}{d}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\log\mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)\right]\,. \tag{62}$$

The replica method described in Sec. 4.1 allows to compute the above average over the dataset $\{\mathbf{X},\mathbf{y}\}$, that plays the role of the planted quenched disorder in usual spin glasses. The details of the computation can be found in Appendix. B.1.1 for committee machines in the case of a synthetic dataset $\mathrm{P_y}\left(\mathbf{y}|\mathbf{X}\right)$ in (59), whereas the similar computation for random labels (58) is derived in Appendix. B.2.

Replica computation We present here the replica computation of the averaged free entropy $\Phi(\alpha)$ in eq. (62) for general *student* prior and channel distributions $\mathrm{P_w}$ and $\mathrm{P_{out}}$. The average in eq. (62) is intractable in general, and the computation relies on the so called *replica trick*, see Sec. 4.1.1, that consists in applying the identity

$$\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\lim_{d\to\infty}\frac{1}{d}\log\mathscr{Z}_d\left(\mathbf{y},\mathbf{X}\right)\right]=\lim_{r\to 0}\left[\lim_{d\to\infty}\frac{1}{d}\frac{\partial\log\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathscr{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r\right]}{\partial r}\right]. \tag{63}$$

This is interesting in the sense that it reduces the intractable average to the computation of the moments of the averaged partition function, which are easier quantities to compute. Note that for $r\in\mathbb{N}$, $\mathscr{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r=\prod_{a=1}^r\mathscr{Z}_d\left(\mathbf{y},\mathbf{X}\right)$ represents the partition function of $r\in\mathbb{R}d$ identical non-interacting copies of the initial system, called *replicas*. Taking the quenched average over the disorder will correlate the replicas, before taking the number of replicas $r\to 0$. Therefore, we assume there exists an analytical continuation so that $r\in\mathbb{R}$ and the limit is well defined. Finally, notice we exchanged the order of the limits $r\to 0$ and $d\to\infty$. These technicalities are crucial points but are not rigorously justified and we will ignore them in the rest of the computation. Next, in order to decouple the contributions of the channel $\mathrm{P_{out}}$ and the prior $\mathrm{P_w}$, we introduce the variable $\mathbf{z}=\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}$ with a Dirac-delta integral

$$\mathscr{Z}_d\left(\{\mathbf{y},\mathbf{X}\}\right)=\int_{\mathbb{R}^n}\mathrm{d}\mathbf{z}\;\mathrm{p_{out}}\left(\mathbf{y}|\mathbf{z}\right)\int_{\mathbb{R}^d}\mathrm{d}\mathbf{w}\;\mathrm{p_w}\left(\mathbf{w}\right)\delta\left(\mathbf{z}-\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}\right),$$

so that the replicated partition function in eq. (63) can be written as

$$\begin{aligned}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathscr{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r\right]=\mathbb{E}_{\mathbf{X}}\int_{\mathbb{R}^n}\mathrm{d}\mathbf{y}\prod_{a=0}^{r}\int_{\mathbb{R}^n}\mathrm{d}\mathbf{z}^a\;\mathrm{p}_{\mathrm{out}^a}\left(\mathbf{y}|\mathbf{z}^a\right)\\ \times\int_{\mathbb{R}^d}\mathrm{d}\mathbf{w}^a\;\mathrm{p}_{\mathrm{w}^a}\left(\mathbf{w}^a\right)\delta\left(\mathbf{z}^a-\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}^a\right),\end{aligned} \tag{64}$$

with the decoupled channel $\mathrm{p_{out}}\left(\mathbf{y}|\mathbf{z}\right)=\prod_{\mu=1}^n\mathrm{p_{out}}\left(y_\mu|z_\mu\right)$ and prior $\mathrm{p_w}\left(\mathbf{w}\right)=\prod_{i=1}^d\mathrm{p_w}\left(w_i\right)$ densities. Interestingly the average over $\mathbf{y}$ is equivalent to the one over the ground truth vector $\mathbf{w}^\star$ in the T-S scenario. Making use of the analogous formulation in (59), the average can simply be considered as a new replica $\mathbf{w}^0$ with index $a=0$ leading to a total of $r+1$ replicas. In the case of *random labels* (58), $\mathrm{P_y}$ is independent of $\mathbf{X}$ and therefore the computation is similar with only r replicas and an additional average over $\mathrm{P_y}$, see Appendix. B.2.

Average over the iid input data X We suppose that inputs are drawn from an i.i.d distribution, for example a Gaussian $\mathrm{P}_{\mathrm{x}}(\mathbf{x}) = \mathcal{N}_{\mathbf{x}}(\mathbf{0}, \mathbf{I}_d)$. More precisely, for $(i,j) \in [\![d]\!]^2$, $(\mu,\nu) \in [\![n]\!]^2$, $\mathbb{E}_{\mathbf{X}}[x_{\mu i} x_{\nu j}] = \delta_{\mu\nu}\delta_{ij}$. Hence $z_\mu^a = \frac{1}{\sqrt{d}} \sum_{i=1}^d x_{\mu i} w_i^a$ is the sum of i.i.d random variables. The CLT insures that in the thermodynamic limit $z_\mu^a \sim \mathcal{N}\left(\mathbb{E}_{\mathbf{X}}[z_\mu^a], \mathbb{E}_{\mathbf{X}}[z_\mu^a z_\mu^b]\right)$, with the two first moments given by:

$$\mathbb{E}_{\mathbf{X}}[z_\mu^a] = \frac{1}{\sqrt{d}} \sum_{i=1}^d \mathbb{E}_{\mathbf{X}}[x_{\mu i}] w_i^a = 0 ,$$
$$\mathbb{E}_{\mathbf{X}}[z_\mu^a z_\mu^b] = \frac{1}{d} \sum_{ij} \mathbb{E}_{\mathbf{X}}[x_{\mu i} x_{\mu j}] w_i^a w_j^b = \frac{1}{d} \sum_{ij} \delta_{ij} w_i^a w_j^b = \frac{1}{d} \mathbf{w}^a \cdot \mathbf{w}^b .$$

Note that averaging over the quenched disorder induces correlations between replicas, which were initially independent. In the following we introduce the symmetric *overlap* matrix that measures the correlations between the replicated vector $\mathbf{w}^a$: $\mathbf{Q}(\{\mathbf{w}^a\}) \equiv \left(\frac{1}{d} \mathbf{w}^a \cdot \mathbf{w}^b\right)_{a,b=0..r}$. Let us define $\tilde{\mathbf{z}}_\mu \equiv (z_\mu^a)_{a=0..r}$ and $\tilde{\mathbf{w}}_i \equiv (w_i^a)_{a=0..r}$ the replicated vectors. The vector $\tilde{\mathbf{z}}_\mu$ follows a multivariate Gaussian distribution $\tilde{\mathbf{z}}_\mu \sim \mathrm{P}_{\tilde{z}}(\tilde{\mathbf{z}}; \mathbf{Q}) = \mathcal{N}_{\tilde{\mathbf{z}}}(\mathbf{0}_{r+1}, \mathbf{Q})$ and as the i.i.d prior and channel distributions factorize $\mathrm{p}_{\mathrm{w}}(\mathbf{w}) = \prod_{i=1}^d \mathrm{p}_{\mathrm{w}}(w_i)$ and $\mathrm{p}_{\mathrm{out}}(\mathbf{y}|\mathbf{z}) = \prod_{\mu=1}^n \mathrm{p}_{\mathrm{out}}(y^{(\mu)} \mid z^{(\mu)})$, it follows

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathcal{Z}_d(\mathbf{y},\mathbf{X})^r\right] \\
&= \mathbb{E}_{\mathbf{X}} \int_{\mathbb{R}^n} \mathrm{d}\mathbf{y} \prod_{a=0}^r \int_{\mathbb{R}^n} \mathrm{d}\mathbf{z}^a \; \mathrm{p}_{\mathrm{out}^a}(\mathbf{y}|\mathbf{z}^a) \\
&\qquad\qquad \times \int_{\mathbb{R}^d} \mathrm{d}\mathbf{w}^a \; \mathrm{p}_{\mathrm{w}^a}(\mathbf{w}^a)\, \delta\left(\mathbf{z}^a - \frac{1}{\sqrt{d}} \mathbf{X}\mathbf{w}^a\right) \\
&= \left[\int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^{r+1}} \mathrm{d}\tilde{\mathbf{z}} \; \mathrm{p}_{\mathrm{out}}(y|\tilde{\mathbf{z}})\, \mathrm{p}_{\tilde{z}}(\tilde{\mathbf{z}}; Q(\tilde{\mathbf{w}}))\right]^n \left[\int_{\mathbb{R}^{r+1}} \mathrm{d}\tilde{\mathbf{w}} \; \mathrm{p}_{\tilde{\mathrm{w}}}(\tilde{\mathbf{w}})\right]^d ,
\end{aligned}$$

where we introduced $\mathrm{P}_{\tilde{\mathrm{w}}}(\tilde{\mathbf{w}}) = \prod_{a=0}^r \mathrm{P}_{\mathrm{w}}(w^a)$ the distribution of the replicated vector of weights. To finish decoupling the channels, we use the Fourier representation of a Dirac-delta function of a variable $x \in \mathbb{R}$ as a function of a purely imaginary parameter $\hat{x}$:

$$\delta(x) = \frac{1}{2i\pi} \int_{i\mathbb{R}} \mathrm{d}\hat{x}\, e^{-\hat{x}x} .$$

Applying the above identity to the following change of variable

$$\begin{aligned}
1 &= \int_{\mathbb{R}^{r+1\times r+1}} \mathrm{d}\mathbf{Q} \prod_{0\leq a\leq b\leq r} \delta\left(dQ_{ab} - \sum_{i=1}^d w_i^a w_i^b\right) \\
&\propto \int\int_{\mathbb{R}^{r+1\times r+1}} \mathrm{d}\mathbf{Q}\mathrm{d}\hat{\mathbf{Q}} \; \exp\left(-d\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right)\right) e^{\frac{1}{2}\sum_{i=1}^d \tilde{\mathbf{w}}_i^\intercal \hat{\mathbf{Q}} \tilde{\mathbf{w}}_i + \tilde{\mathbf{w}}_i^\intercal \mathrm{diag}\left(\hat{\mathbf{Q}}\right) \tilde{\mathbf{w}}_i} ,
\end{aligned}$$

that involves a new ad-hoc purely imaginary matrix parameter $\hat{\mathbf{Q}}$. Hence, multiplying the replicated partition function by 1, it becomes an integral over the symmetric matrices $\mathbf{Q} \in \mathbb{R}^{r+1 \times r+1}$ and $\hat{\mathbf{Q}} \in \mathbb{R}^{r+1 \times r+1}$, that can be evaluated using a Laplace method (Wong, 1989) in the $d \to \infty$ limit,

$$\begin{aligned}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r\right] &= \int_{\mathbb{R}^{r+1\times r+1}} \mathrm{d}\mathbf{Q} \int_{\mathbb{R}^{r+1\times r+1}} \mathrm{d}\hat{\mathbf{Q}} e^{d\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}})} \\ &\underset{d\to\infty}{\simeq} \exp\left(d \cdot \mathbf{extr}_{\mathbf{Q},\hat{\mathbf{Q}}}\left\{\Phi^{(r)}\left(\mathbf{Q},\hat{\mathbf{Q}}\right)\right\}\right),\end{aligned} \tag{65}$$

where we omitted the sub-leading factors and defined the free entropy potential

$$\begin{aligned}\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}}) &= -\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right) + \log \Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}}) + \alpha \log \Psi_{\mathrm{out}}^{(r)}(\mathbf{Q}), \\ \Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}}) &= \int_{\mathbb{R}^{r+1}} \mathrm{d}\tilde{\mathbf{w}}\ \mathrm{p}_{\tilde{\mathrm{w}}}(\tilde{\mathbf{w}}) e^{\frac{1}{2}\tilde{\mathbf{w}}^\intercal \hat{\mathbf{Q}}\tilde{\mathbf{w}} + \frac{1}{2}\tilde{\mathbf{w}}^\intercal \mathrm{diag}\left(\hat{\mathbf{Q}}\right)\tilde{\mathbf{w}}} \\ \Psi_{\mathrm{out}}^{(r)}(\mathbf{Q}) &= \int \mathrm{d}_{\mathbb{R}} y \int_{\mathbb{R}^{r+1}} \mathrm{d}\tilde{\mathbf{z}}\ \mathrm{p}_{\tilde{z}}(\tilde{\mathbf{z}};\mathbf{Q}) \mathrm{p}_{\mathrm{out}}(y|\tilde{\mathbf{z}}),\end{aligned} \tag{66}$$

and $\mathrm{P}_{\tilde{z}}(\tilde{\mathbf{z}};\mathbf{Q}) = e^{-\frac{1}{2}\tilde{\mathbf{z}}^\intercal \mathbf{Q}^{-1}\tilde{\mathbf{z}}} / \det\left(2\pi\mathbf{Q}\right)^{1/2}$. Recall that the average over the teacher vector has been merged as a new replica so that $\mathrm{p}_{\mathrm{out}^0} = \mathrm{p}_{\mathrm{out}^\star}$, $\mathrm{p}_{\mathrm{w}^0} = \mathrm{p}_{\mathrm{w}^\star}$. Finally switching the two limits $r \to 0$ and $d \to \infty$, the quenched free entropy Φ simplifies to a saddle point equation

$$\Phi(\alpha) = \mathbf{extr}_{\mathbf{Q},\hat{\mathbf{Q}}}\left\{\lim_{r\to 0} \frac{\partial \Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}})}{\partial r}\right\}, \tag{67}$$

over symmetric matrices $\mathbf{Q} \in \mathbb{R}^{r+1\times r+1}$ and $\hat{\mathbf{Q}} \in \mathbb{R}^{r+1\times r+1}$. In the following we will assume a simple Ansatz for these matrices in order to first obtain an analytic expression in r before taking the derivative with respect to r. Note that the partition function of this fully connected model can be expressed as a saddle point only because distributions $\mathrm{P}_{\mathrm{out}}$ and P_{w} factorize so that a prefactor scaling with the system size dominates the exponential distribution.

Replica Symmetric free entropy Let's compute the functional $\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}})$ in eq. (67) in the simplest Ansatz: the RS Ansatz. This latter assumes that all the replicas remain equivalent with a common overlap $q = \frac{1}{d}\mathbf{w}^a \cdot \mathbf{w}^b$ for $a \neq b$, a norm $Q = \frac{1}{d}\|\mathbf{w}^a\|_2^2$, and an overlap with the ground truth $m = \frac{1}{d}\mathbf{w}^a \cdot \mathbf{w}^\star$, leading to the following expressions of the replica symmetric matrices $\mathbf{Q}^{(\mathrm{rs})} \in \mathbb{R}^{r+1\times r+1}$ and $\hat{\mathbf{Q}}^{(\mathrm{rs})} \in \mathbb{R}^{r+1\times r+1}$:

$$\mathbf{Q}^{(\mathrm{rs})} = \begin{pmatrix} Q^\star & m & \cdots & m \\ m & Q & q & q \\ \vdots & q & \ddots & q \\ m & q & q & Q \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{Q}}^{(\mathrm{rs})} = \begin{pmatrix} -\frac{1}{2}\hat{Q}^\star & \hat{m} & \cdots & \hat{m} \\ \hat{m} & -\frac{1}{2}\hat{Q} & \hat{q} & \hat{q} \\ \vdots & \hat{q} & \ddots & \hat{q} \\ \hat{m} & \hat{q} & \hat{q} & -\frac{1}{2}\hat{Q} \end{pmatrix} \tag{68}$$

with $Q^\star = \frac{1}{d}\|\mathbf{w}^\star\|_2^2$. The factor $-\frac{1}{2}$ is not necessary but useful to recover commonly used formulations. The functional $\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}})$can be computed with this Ansatz: the first is a trace term, the second term $\Psi_{\mathrm{w}}^{(r)}$ depends on the prior distributions P_{w}, $\mathrm{P}_{\mathrm{w}^\star}$ and finally the third term $\Psi_{\mathrm{out}}^{(r)}$ depends on the channel distributions $\mathrm{P}_{\mathrm{out}^\star}$, $\mathrm{P}_{\mathrm{out}}$.

Replica trick $r \to 0$ limit The last step of the computation is to take properly the limit $r \to 0$. We obtain that

$$-\lim_{r\to 0} \partial_r \operatorname{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right)\big|_{\mathrm{rs}} = -m\hat{m} + \frac{1}{2}Q\hat{Q} + \frac{1}{2}q\hat{q}. \tag{69}$$

and

$$\begin{aligned} &\lim_{r\to 0} \partial_r \log \Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}})\Big|_{\mathrm{rs}} = \\ &\qquad \mathbb{E}_{\xi,w^\star} \mathcal{Z}_{\mathrm{w}^\star}\left(\hat{m}\hat{q}^{-1/2}\xi, \hat{m}^2\hat{q}^{-1}\right) \log \mathcal{Z}_{\mathrm{w}}\left(\hat{q}^{1/2}\xi, \hat{Q}+\hat{q}\right), \\ &\lim_{r\to 0} \partial_r \log \Psi_{\mathrm{out}}^{(r)}(\mathbf{Q})\Big|_{\mathrm{rs}} = \\ &\qquad \int \mathrm{d}y \mathbb{E}_{\xi} \mathcal{Z}_{\mathrm{out}^\star}\left(mq^{-1/2}\xi, Q^\star - m^2 q^{-1}\right) \log \mathcal{Z}_{\mathrm{out}}\left(q^{1/2}\xi, Q-q\right), \end{aligned} \tag{70}$$

with denoising functions $\mathcal{Z}_{\mathrm{out}^\star}, \mathcal{Z}_{\mathrm{out}}, \mathcal{Z}_{\mathrm{w}^\star}, \mathcal{Z}_{\mathrm{w}}$ defined in Appendix. A.4.

4.1.5.a SUMMARY

Gathering eq. (69, 70), we finally obtain the RS free entropy Φ_{rs}.

$$\begin{aligned} \Phi_{\mathrm{rs}}(\alpha) &\equiv \mathbf{extr}_{Q,\hat{Q},q,\hat{q},m,\hat{m}} \left\{ \lim_{r\to 0} \partial_r \Phi^{(r)}\left(\mathbf{Q}^{(\mathrm{rs})}, \hat{\mathbf{Q}}^{(\mathrm{rs})}\right) \right\} \\ &= \mathbf{extr}_{Q,\hat{Q},q,\hat{q},m,\hat{m}} \left\{ -m\hat{m} + \frac{1}{2}Q\hat{Q} + \frac{1}{2}q\hat{q} \right. \\ &\qquad \left. + \Psi_{\mathrm{w}}\left(\hat{Q},\hat{m},\hat{q}\right) + \alpha \Psi_{\mathrm{out}}\left(Q,m,q;\rho_{\mathrm{w}^\star}\right) \right\}, \end{aligned} \tag{71}$$

where $\rho_{\mathrm{w}^\star} = \lim_{d\to\infty} \mathbb{E}_{\mathbf{w}^\star} \frac{1}{d}\|\mathbf{w}^\star\|_2^2$ and the channel and prior integrals are defined by

$$\begin{aligned} \Psi_{\mathrm{w}}\left(\hat{Q},\hat{m},\hat{q}\right) &\equiv \mathbb{E}_{\xi}\left[\mathcal{Z}_{\mathrm{w}^\star}\left(\hat{m}\hat{q}^{-1/2}\xi, \hat{m}^2\hat{q}^{-1}\right) \log \mathcal{Z}_{\mathrm{w}}\left(\hat{q}^{1/2}\xi, \hat{Q}+\hat{q}\right)\right], \\ \Psi_{\mathrm{out}}\left(Q,m,q;\rho_{\mathrm{w}^\star}\right) &\equiv \mathbb{E}_{y,\xi}\left[\mathcal{Z}_{\mathrm{out}^\star}\left(y, mq^{-1/2}\xi, \rho_{\mathrm{w}^\star} - mq^{-1}m\right)\right. \\ &\qquad \left. \times \log \mathcal{Z}_{\mathrm{out}}\left(y, q^{1/2}\xi, Q-q\right)\right], \end{aligned} \tag{72}$$

for generic $\mathrm{P}_{\mathrm{out}^\star}, \mathrm{P}_{\mathrm{out}}$ and $\mathrm{P}_{\mathrm{w}^\star}, \mathrm{P}_{\mathrm{w}}$ distributions and corresponding update functions $\mathcal{Z}_{\mathrm{out}^\star}, \mathcal{Z}_{\mathrm{out}}, \mathcal{Z}_{\mathrm{w}^\star}, \mathcal{Z}_{\mathrm{w}}$ are defined in Appendix. A.4. As a conclusion, we notice remarkably that the behavior of the initial complex high-dimensional inference problem is charaterized by an optimization problem over only six scalar order parameters, and is therefore controlled by a set of six fixed point equations. Finally, let us mention that MMSE estimation can be performed in the Bayes-optimal setting for $\beta = 1$, while MAP estimation requires to take properly $\beta = \infty$ as detailed later on in Chap. 8.

Bayes-optimal free entropy In the Bayes-optimal setting, we perform inference using the knowledge of the *ground truth* distributions so that the student denoising functions are exactly the ones used to generate the dataset, namely $\mathrm{P}_{\mathrm{out}} = \mathrm{P}_{\mathrm{out}^\star}$ and $\mathrm{P}_{\mathrm{w}} = \mathrm{P}_{\mathrm{w}^\star}$ so that $\mathcal{Z}_{\mathrm{out}} = \mathcal{Z}_{\mathrm{out}^\star}$, $\mathcal{Z}_{\mathrm{w}} = \mathcal{Z}_{\mathrm{w}^\star}$. The Nishimori's conditions in the Bayes-optimal case, derived in Appendix. A.3, imply that $Q = Q^\star \equiv \rho_{\mathrm{w}^\star}$, $m = q \equiv q_{\mathrm{b}}$, $\hat{Q} = \hat{Q}^\star = 0$, $\hat{m} = \hat{q} \equiv \hat{q}_{\mathrm{b}}$. Therefore, in the Bayes-optimal setting, the free entropy of the high-dimensional inference problem eq. (71) simplifies as an optimization problem over scalar *overlaps* parameters $q_{\mathrm{b}}, \hat{q}_{\mathrm{b}}$:

$$\Phi_{\mathrm{rs}}^{\mathrm{b}}(\alpha) = \mathbf{extr}_{q_{\mathrm{b}},\hat{q}_{\mathrm{b}}} \left\{ -\frac{1}{2} q_{\mathrm{b}}\hat{q}_{\mathrm{b}} + \Psi_{\mathrm{w}}^{\mathrm{b}}(\hat{q}_{\mathrm{b}}) + \alpha \Psi_{\mathrm{out}}^{\mathrm{b}}(q_{\mathrm{b}};\rho_{\mathrm{w}^\star}) \right\}, \qquad (73)$$

with free entropy terms $\Psi_{\mathrm{w}}^{\mathrm{b}}$ and $\Psi_{\mathrm{out}}^{\mathrm{b}}$ given by

$$\begin{aligned} \Psi_{\mathrm{w}}^{\mathrm{b}}(\hat{q}_{\mathrm{b}}) &= \mathbb{E}_{\xi}\left[\mathcal{Z}_{\mathrm{w}^\star}\left(\hat{q}_{\mathrm{b}}^{1/2}\xi, \hat{q}_{\mathrm{b}}\right) \log \mathcal{Z}_{\mathrm{w}^\star}\left(\hat{q}_{\mathrm{b}}^{1/2}\xi, \hat{q}_{\mathrm{b}}\right)\right], \\ \Psi_{\mathrm{out}}^{\mathrm{b}}(q_{\mathrm{b}};\rho_{\mathrm{w}^\star}) &= \mathbb{E}_{y,\xi}\left[\mathcal{Z}_{\mathrm{out}^\star}\left(y, q_{\mathrm{b}}^{1/2}\xi, \rho_{\mathrm{w}^\star} - q_{\mathrm{b}}\right) \right. \\ &\qquad\qquad \left. \log \mathcal{Z}_{\mathrm{out}^\star}\left(y, q_{\mathrm{b}}^{1/2}\xi, \rho_{\mathrm{w}^\star} - q_{\mathrm{b}}\right)\right]. \end{aligned}$$

Notice that the above Bayes-optimal replica symmetric free entropy for the GLM class has been rigorously proven in (Barbier et al., 2019b). Taking the derivatives with respect to $q_{\mathrm{b}}, \hat{q}_{\mathrm{b}}$, we obtain the stationary conditions verified by the overlap parameters

$$\begin{aligned} q_{\mathrm{b}} &= \alpha \mathbb{E}_{y,\xi} \mathcal{Z}_{\mathrm{out}^\star}\left(y, q_{\mathrm{b}}^{1/2}\xi, \rho_{\mathrm{w}^\star} - q_{\mathrm{b}}\right) f_{\mathrm{out}^\star}\left(y, q_{\mathrm{b}}^{1/2}\xi, \rho_{\mathrm{w}^\star} - q_{\mathrm{b}}\right)^2 \\ \hat{q}_{\mathrm{b}} &= \mathbb{E}_{\xi} \mathcal{Z}_{\mathrm{w}^\star}\left(\hat{q}_{\mathrm{b}}^{1/2}\xi, \hat{q}_{\mathrm{b}}\right) f_{\mathrm{w}^\star}\left(\hat{q}_{\mathrm{b}}^{1/2}\xi, \hat{q}_{\mathrm{b}}\right)^2, \end{aligned} \qquad (74)$$

that will turn out to be strongly connected to the infinite-size behavior of the AMP algorithm, the so-called *state evolution* equations.

In this section, we have presented the heuristic replica method which provides a powerful technique to directly compute the free entropy, associated to a complex JPD, and to describe the statistical thresholds of the corresponding phase diagram. Next, we present other mean-field methods to perform approximate inference of this same JPD. Interestingly, even though these techniques do not directly yield the result like the replica method, however, they have the profound advantage of leading to interesting algorithmic perspectives and insights to complete the phase diagram.

4.2 ON VARIATIONAL MEAN-FIELD METHODS

Assume we consider a statistical model associated to a JPD $\mathrm{P}_d(\boldsymbol{\sigma};\beta)$ and an Hamiltonian energy function $\mathscr{H}_d(\boldsymbol{\sigma})$. The main challenge is to compute its marginal probabilities, moments or even more complex observable of the JPD.

Intractability of exact inference However, computing analytically the posterior, with or without the replica method, is very rarely possible. In general, even though the replica method provides a quick and strong tool to calculate it in some particular cases, computing the marginal probabilities of a high-dimensional JPD $\mathrm{P}_d(\boldsymbol{\sigma};\beta)$ according to $\mathrm{P}(\sigma_i) = \int_{\chi_{d-1}} \mathrm{d}\boldsymbol{\sigma}_{\setminus i}\mathrm{P}_d(\boldsymbol{\sigma};\beta)$, for some $i \in [\![d]\!]$, remains a hard task. Indeed computing the corresponding continuous or discrete sum requires very often a number of operations that scales exponentially with the size of the system and becomes critical in the high-dimensional regime that we consider $d \to \infty$. Of course in the case where the spins are restricted to one-body interactions and do not interact, the JPD distribution factorizes and the sum over $\mathbb{R}^d$ reduces to a sum over $\mathbb{R}$ and deeply simplify the computation. Yet, this kind of simplification remains very limited and, moreover,complex and interesting behaviors arise very often only when *interactions* come on stage.

On Tree factor graphs Let us first draw attention on very simple factor graphs and corresponding JPD. In the case where the factor graph under consideration is a *tree*, as an illustration see for instance Fig. 15 **(Right)** , the computation of the JPD can be performed in linear time complexity, in contrast with the exponential complexity mentioned above. Indeed using the Markov property and conditional expectation (Pearl, 1982; Pearl, 1986), it is possible to compute the whole JPD as a product of $\Theta(d)$ terms. Moreover, this procedure may be turned into a dynamical algorithm called the *sum-product algorithm* or BP equations that, as we just stressed, is *exact on tree* factor graphs. The corresponding algorithm reaches the fixed point of a well designed free energy approximation, the *Bethe free energy*, detailed in Sec. 4.3. More interestingly, it can approximate correctly the target JPD on loopy *factor graphs* as well, even though it is not guaranteed to converge.

Approximate variational methods On general factor graphs, we therefore have to resort to *approximate inference*, to circumvent this difficulty and compute *approximately* and *efficiently* the marginal probabilities $\mathrm{P}(\sigma_i)$. Sampling methods relying on MCMC algorithms, see Sec. 1.2.10.b, are widely used in practice. Yet, they are not very performant especially in the high-dimensional inference regime of interest. To address this issue, instead trying to sample a huge number of examples, other *variational mean-field method* have been designed, in particular in physics, to compute a good approximation of the posterior distribution (Opper et al., 2001a).

The design of such mean-field approximations requires, first, to recall and introduce some useful IT quantities in Sec. 4.2.1, that naturally lead to the *Gibbs free energy* and its variational formulation presented in Sec. 4.2.2. Finally, in Sec. 4.2.3, we recall the *naive* mean-field approach and its extension to more complex approximations, such as the TAP approach. For an extended introduction on variational mean-field methods, let us mention the comprehensive review (Blei et al., 2017).

4.2.1 INFORMATION THEORY QUANTITIES

In the perspective of comparing and constructing approximations of the complex JPD associated to interacting systems, we introduce the classical tools from IT to compare distribution families such as the Shanon entropy, the KL divergence and the mutual information. More details can be found in (MacKay et al., 2003; Koller et al., 2009).

4.2.1.A SHANON ENTROPY

Let X be a RV with probability distribution P and density $\mathrm{p}(s) \equiv \mathrm{dP}/\mathrm{d}x$ on a set $\mathbb{X}$, the Shannon entropy H(X) measures the quantity of information carried by the RV X and is defined by

$$\mathrm{H}(\mathrm{X}) = -\mathbb{E}_{\mathrm{X}\sim\mathrm{P}}\left[\log \mathrm{P}(x)\right] = -\int_{\mathbb{X}} \mathrm{d}x\, \mathrm{p}(x) \log \mathrm{p}(x)\,. \tag{75}$$

4.2.1.B THE KULLBACK-LEIBLER DIVERGENCE

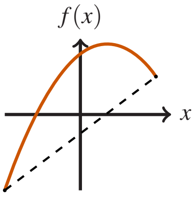
Consider two probability distributions Q and P, with densities q, p on a set $\mathbb{X}$. The KL divergence is used to compare two arbitrary distributions P and Q, defined as

$$\begin{aligned} \mathscr{D}_{\mathrm{KL}}\left(\mathrm{Q}\parallel\mathrm{P}\right) &= \mathbb{E}_{\mathrm{X}\sim\mathrm{Q}}\left[\log \mathrm{Q}(x) - \log \mathrm{P}(x)\right] \\ &= \int_{\mathbb{X}} \mathrm{d}x\, \mathrm{q}(x) \log\left(\frac{\mathrm{q}(x)}{\mathrm{p}(x)}\right), \end{aligned} \tag{76}$$

with densities p, q defined by $\mathrm{dP} \equiv \mathrm{p}(x)\mathrm{d}x$, $\mathrm{dQ} \equiv \mathrm{q}(x)\mathrm{d}x$. Because it is not symmetric under the exchange of Q and P, $\mathscr{D}_{\mathrm{KL}}\left(\mathrm{Q}\parallel\mathrm{P}\right) \neq \mathscr{D}_{\mathrm{KL}}\left(\mathrm{P}\parallel\mathrm{Q}\right)$ and does not verify the triangle inequality, the KL divergence is not formally a distance in the rigorous mathematical sense. However, it plays exactly the role of a distance in the space of probability densities as it is always positive as stated by the *Gibb's inequality*:

Proposition 4.2.1 (From (Cover et al., 2012)). *Consider two distributions* P, Q *with densities* p, q, *then* $\mathscr{D}_{\mathrm{KL}}\left(\mathrm{Q}\parallel\mathrm{P}\right) \geq 0$ *and* $\mathscr{D}_{\mathrm{KL}}\left(\mathrm{Q}\parallel\mathrm{P}\right) = 0 \Leftrightarrow \mathrm{Q} = \mathrm{P}$.

Recall that a function f is concave if for $x_1 \leq x \leq x_2$ the point $(x, f(x))$ is above the line joining the points $(x_1, f(x_1))$, $(x_2, f(x_2))$.



Proof. As the *logarithm* is concave, from the Jensen inequality we obtain

$$-\mathscr{D}_{\mathrm{KL}}\left(\mathrm{Q}\,\|\,\mathrm{P}\right) = \int_{\mathbb{X}} \mathrm{d}x\, \mathrm{q}(x) \log\left(\frac{\mathrm{p}(x)}{\mathrm{q}(x)}\right) \leq \int_{\mathbb{X}} \mathrm{d}x\, \mathrm{q}(x) - \mathrm{p}(x) = 0\,.$$

□

4.2.1.c THE MUTUAL INFORMATION

Consider two random variables X and Y jointly distributed according to $\mathrm{P}_{\mathrm{X,Y}}$, the mutual information specifically measures the KL divergence from the product $\mathrm{P}_{\mathrm{X}}\mathrm{P}_{\mathrm{Y}}$ to the joint distribution $\mathrm{P}_{\mathrm{X,Y}}$:

$$\begin{aligned}\mathscr{I}\left(\mathrm{X};\mathrm{Y}\right) = \mathscr{D}_{\mathrm{KL}}\left(\mathrm{P}_{\mathrm{X,Y}}\,\|\,\mathrm{P}_{\mathrm{X}}\mathrm{P}_{\mathrm{Y}}\right) &= \int_{\mathbb{X}^2} \mathrm{d}x\mathrm{d}y\, \mathrm{p}(x,y) \log\left(\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y))}\right) \\ &= \mathrm{H}\left(\mathrm{X}\right) - \mathrm{H}\left(\mathrm{X}|\mathrm{Y}\right) = \mathrm{H}\left(\mathrm{Y}\right) - \mathrm{H}\left(\mathrm{Y}|\mathrm{X}\right) \\ &= \mathrm{H}\left(\mathrm{X}\right) + \mathrm{H}\left(\mathrm{Y}\right) - \mathrm{H}\left(\mathrm{X},\mathrm{Y}\right)\,,\end{aligned} \tag{77}$$

where the marginal densities write $\mathrm{p}(x) = \int_{\mathbb{X}} \mathrm{d}y\, \mathrm{p}(x,y)$ and $\mathrm{p}(y) = \int_{\mathbb{X}} \mathrm{d}x\, \mathrm{p}(x,y)$.

4.2.2 GIBBS FREE ENERGY AND VARIATIONAL PRINCIPLE

Let us consider a JPD that we aim to approximate, for instance $\mathrm{P}_d(\boldsymbol{\sigma};\beta) \equiv e^{-\beta\mathscr{H}_d(\boldsymbol{\sigma})} / \mathscr{Z}_d(\beta)$, associated to the Hamiltonian $\mathscr{H}_d(\boldsymbol{\sigma})$ for some spins $\boldsymbol{\sigma} \in \chi_d$. For any arbitrary probability distribution Q, we define the Gibbs free energy as the trade-off between the variational energy $\mathrm{U}[\mathrm{Q}] \equiv \mathbb{E}_{\boldsymbol{\sigma}\sim\mathrm{Q}}\left[\mathscr{H}_d(\boldsymbol{\sigma})\right]$ and the entropy of the distribution $\mathrm{H}[\mathrm{Q}]$ according to

$$\varphi_d^{\mathrm{gibbs}}[\mathrm{Q}] \equiv \mathrm{U}[\mathrm{Q}] - \frac{1}{\beta}\mathrm{H}[\mathrm{Q}]\,, \tag{78}$$

where β is a free inverse temperature parameter. In order to find a good mean-field approximation, we introduce the Gibbs variational principle that states that the Gibbs free energy is minimal when the mean-field approximation equals the target JPD distribution P_d.

4.2.2.a GIBBS VARIATIONAL PRINCIPLE

The Gibbs variational principle follows from the fact that for any arbitrary distribution Q, the Gibbs free energy may be rewritten as

$$\begin{aligned}\varphi_d^{\mathrm{gibbs}}[\mathrm{Q}] &= \mathbb{E}_{\boldsymbol{\sigma}\sim\mathrm{Q}}\left[\mathscr{H}_d(\boldsymbol{\sigma})\right] \\ &\qquad + \frac{1}{\beta}\int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\, \mathrm{q}(\boldsymbol{\sigma}) \left(\log\frac{\mathrm{q}(\boldsymbol{\sigma})}{\mathrm{p}(\boldsymbol{\sigma})} + \mathrm{q}(\boldsymbol{\sigma})\log\mathrm{p}(\boldsymbol{\sigma})\right) \\ &= \frac{1}{\beta}\mathscr{D}_{\mathrm{KL}}\left(\mathrm{Q}||\mathrm{P}_d\right) + \varphi_d(\beta) \geq \varphi_d(\beta)\,,\end{aligned} \tag{79}$$

where we first used the definition of the target free energy $\varphi_d(\beta) = -\frac{1}{\beta} \log \mathscr{Z}_d(\beta)$ and the positivity of the KL divergence. The last inequality is known as the *Gibbs variational principle* also called the *Gibbs-Bogoliubov-Feynman inequality*. As a consequence the Gibbs free energy of any approximate distribution Q is larger than the true free energy $\varphi_d(\beta)$ associated to the P_d, namely $\varphi_d^{\text{gibbs}}[\mathrm{Q}] \geq \varphi_d(\beta)$. Moreover, the inequality is saturated if the approximation exactly equals the Gibbs distribution $\mathrm{Q} = \mathrm{P}_d$. This variational principle allows to measure the correctness of a given approximation. However, this variational principle cannot be solved in full generality, and we therefore need to restrict the set of possible probability distributions to a practical set. Instead of choosing arbitrarily a potential set, we present in the next section the *maximum entropy principle* which allows to restrict approximations to simple distributions families.

4.2.2.B MAXIMUM ENTROPY PRINCIPLE

To restrict the space of probability densities, first we consider only the ones that verify the *moments matching conditions* of a set of moments $\{\phi_k\}_{k\in\mathbb{K}}$, such that $\mathbb{E}_{\boldsymbol{\sigma}\sim\mathrm{Q}}[\phi_k(\boldsymbol{\sigma})] = \mu_k$. In other words, in expectation the k-th moments in the index set $\mathbb{K}$ of the approximated distribution Q should match the true moments $\{\mu_k\}_{k\in\mathbb{K}}$ of the target distribution P_d. Yet, imposing the moments matching constraints does not determine uniquely the distribution, since in general it exists an infinite number of solutions verifying them. The *maximum entropy principle*, introduced by (Jaynes, 1957; Jaynes, 2003; Wainwright et al., 2008) and very close to the *least action principle* in analytical mechanics, allows to prescribe *good choices* for the approximation Q of the JPD P_d. Assume we have access to n observations $\{\boldsymbol{\sigma}^{(1)}, \cdots, \boldsymbol{\sigma}^{(n)}\}$ drawn from the target Gibbs distribution, or any other distribution we try to approximate, the maximum entropy principle simply states that the probability distribution which best represents the *current state of knowledge* is the one with the *largest entropy*. In more details, imposing the normalization and the moments matching, the least action principle can be formulated mathematically as a Lagrangian problem over the distribution Q:

The maximum entropy principle is similar to the middle age philosophical principle known as the Ockham Razor. It has been slightly modified and popularized by the Shadoks *from Jacques Rouvel.*

$$\mathscr{L}[\mathrm{Q}] = \mathrm{H}[\mathrm{Q}] - \sum_{k\in\mathbb{K}} \lambda_k \left(\int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\, \mathrm{q}(\boldsymbol{\sigma}) \phi_k(\boldsymbol{\sigma}) - \mu_k \right) \\ - \lambda_0 \left(\int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\, \mathrm{q}(\boldsymbol{\sigma}) - 1 \right)$$

Minimizing the action, i. e. the Lagrangian $\mathscr{L}[\mathrm{Q}]$, yields

$$0 = \frac{\partial \mathscr{L}}{\partial \mathrm{Q}} = \log \mathrm{Q}(\boldsymbol{\sigma}) + 1 - \sum_{k\in\mathbb{K}} \lambda_k \phi_k(\boldsymbol{\sigma}) - \lambda_0 .$$

Imposing the normalization $\int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\, \mathrm{q}(\boldsymbol{\sigma}) = 1$ we finally recover the so-called *exponential family* (Jordan et al., 1999)

$$\mathrm{Q}(\boldsymbol{\sigma}) = \frac{1}{\mathscr{Z}} \exp\left(- \sum_{k \in \mathbb{K}} \lambda_k \phi_k(\boldsymbol{\sigma}) \right) , \tag{80}$$

where $\mathscr{Z} = e^{1-\lambda_0} = \int_{\chi_d} \mathrm{d}\boldsymbol{\sigma}\, \mathrm{e}^{-\sum_{k \in \mathbb{K}} \lambda_k \phi_k(\boldsymbol{\sigma})}$ with potential additional cut-offs to avoid the normalizing constant to diverge, especially for $k = 1$. To summarize, the least biased distribution to consider to approximate the Gibbs distribution are the ones belonging to the exponential family.

4.2.3 NAIVE AND TAP MEAN-FIELD APPROXIMATION

We present two simple approximations considered in the physics literature (Opper et al., 2001a), starting with the naive mean-field approximation and the TAP approach, which can be derived from the Gibbs variational principle.

4.2.3.a NAIVE MEAN-FIELD APPROXIMATION

The *naive* mean-field approximation consists in a simple factorized density approximation $\mathrm{Q}^{\mathrm{naive}}(\boldsymbol{\sigma}) = \prod_{i=1}^{d} \mathrm{Q}_i(\sigma_i)$ of d independent spins. It has been introduced in classical physics long time ago in the celebrated Curie-Weiss model (Curie, 1895; Weiss, 1907) to study magnetic properties of materials. Hundreds years later, the naive mean-field has been largely democratized and used in various communities (Jordan et al., 1999; Jaakkola et al., 2000; Wainwright et al., 2008). Computing its Gibbs free energy by injecting the naive mean-field approximation in (78) yields

$$\begin{aligned}
\varphi_d^{\mathrm{gibbs}}[\mathrm{Q}^{\mathrm{naive}}] &= \mathbb{E}_{\boldsymbol{\sigma} \sim \mathrm{Q}^{\mathrm{naive}}}\left[\mathscr{H}_d(\boldsymbol{\sigma})\right] + \frac{1}{\beta} \sum_{j=1}^{d} \int_{\chi} \mathrm{d}\sigma_j\, \mathrm{q}_j(\sigma_j) \log\left(\mathrm{q}_j(\sigma_j)\right) \\
&= \frac{1}{\beta} \int_{\chi} \mathrm{d}\sigma_i\, \mathrm{q}_i(\sigma_i) \log\left(\mathrm{q}_i(\sigma_i)\right) + \frac{1}{\beta} \sum_{j \neq i} \int_{\chi} \mathrm{d}\sigma_j\, \mathrm{q}_j(\sigma_j) \log\left(\mathrm{q}_j(\sigma_j)\right) \\
&+ \int_{\chi} \mathrm{d}\sigma_i\, \mathrm{q}_i(\sigma_i) \underbrace{\left[\int_{\chi} \prod_{j \neq i} \mathrm{d}\sigma_j\, \mathrm{q}_j(\sigma_j) \mathscr{H}_d(\boldsymbol{\sigma}) \right]}_{\equiv \mathbb{E}_{\boldsymbol{\sigma}_{\setminus i}}[\mathscr{H}_d(\boldsymbol{\sigma})]}
\end{aligned}$$

where we denote $\boldsymbol{\sigma}_{\setminus i}$ the vector formed by deleting the i-th component of the spin configuration $\boldsymbol{\sigma}$ and $\mathbb{E}_{\boldsymbol{\sigma}_{\setminus i}}\left[\mathscr{H}_d(\boldsymbol{\sigma})\right]$ the conditional expectation of $\mathscr{H}_d(\boldsymbol{\sigma})$ when we fix σ_i. Defining $\mathrm{Q}_{\setminus i}(\sigma_i) \equiv \frac{e^{-\beta \mathbb{E}_{\boldsymbol{\sigma}_{\setminus i}}[\mathscr{H}_d(\boldsymbol{\sigma})]}}{\mathscr{Z}_{\setminus i}(\beta)}$ with $\mathscr{Z}_{\setminus i}(\beta) \equiv$

$\int_\chi \mathrm{d}\sigma_i \, e^{-\beta \mathbb{E}_{\boldsymbol{\sigma}_{\setminus i}}[\mathcal{H}_d(\boldsymbol{\sigma})]}$, the Gibbs free energy $\varphi_d^{\mathrm{gibbs}}[\mathrm{Q}^{\mathrm{naive}}]$ can be ingeniously decomposed as

$$\begin{aligned}
\varphi_d^{\mathrm{gibbs}}[\mathrm{Q}^{\mathrm{naive}}] &= \frac{1}{\beta}\sum_{j\neq i}\int \mathrm{d}\sigma_j \, \mathrm{q}_j(\sigma_j)\log\left(\mathrm{q}_j(\sigma_j)\right) \\
&\quad + \frac{1}{\beta}\int \mathrm{d}\sigma_i \, \mathrm{q}_i(\sigma_i)\left[\log\left(\mathrm{q}_i(\sigma_i)\right) - \log\left(e^{-\beta \mathbb{E}_{\boldsymbol{\sigma}_{\setminus i}}[\mathcal{H}_d(\boldsymbol{\sigma})]}\right)\right] \\
&= \frac{1}{\beta}\sum_{j\neq i}\int \mathrm{d}\sigma_j \, \mathrm{q}_j(\sigma_j)\log\left(\mathrm{q}_j(\sigma_j)\right) \\
&\quad + \frac{1}{\beta}\int \mathrm{d}\sigma_i \, \mathrm{q}_i(\sigma_i)\left[\log\left(\frac{\mathrm{q}_i(\sigma_i)}{\mathrm{q}_{\setminus i}(\sigma_i)}\right) - \log\left(\mathcal{Z}_{\setminus i}(\beta)\right)\right] \\
&= \underbrace{\frac{1}{\beta}\sum_{j\neq i}\int \mathrm{d}\sigma_j \, \mathrm{q}_j(\sigma_j)\log\left(\mathrm{q}_j(\sigma_j)\right) - \frac{1}{\beta}\log\left(\mathcal{Z}_{\setminus i}(\beta)\right)}_{\varphi_{\setminus i}} \\
&\quad + \frac{1}{\beta}\mathcal{D}_{\mathrm{KL}}\left(\mathrm{Q}_i || \mathrm{Q}_{\setminus i}\right) ,
\end{aligned}$$

where the first term $\varphi_{\setminus i}$ is independent of the marginal density q_i. Therefore, minimizing the Gibbs free energy $\varphi_d^{\mathrm{gibbs}}[\mathrm{Q}^{\mathrm{naive}}]$, the Gibbs variational principle (79) prescribes the marginal densities to

$$\mathrm{Q}_i(\sigma_i) \equiv \frac{1}{\mathcal{Z}_{\setminus i}(\beta)} e^{-\beta \mathbb{E}_{\boldsymbol{\sigma}_{\setminus i}}[\mathcal{H}_d(\boldsymbol{\sigma})]} . \tag{81}$$

Applied to the Curie-Weiss model, which is only the mean-field Ising model, see Sec. 2.2.3.b, the naive mean-field approximation (81) allows in particular to recover the well-known set of implicit equations verified by the magnetizations

$$m_i \equiv \mathbb{E}_{\mathrm{Q}_i}[\sigma_i] = \tanh\left(\beta\left(h_i + \sum_{j=1}^{d} J_{ij} m_j\right)\right) . \tag{82}$$

As a conclusion, the naive mean-field approximation has the advantage to treat the surrounding interactions of each spin σ_i as an average interaction of all the other spins, but at the cost of discarding, eventually, important statistical correlations. In the case where interactions between spins are weak enough, this naive mean-field approach might be exact, as for instance in the case of the Curie-Weiss model they vanish in the thermodynamic limit $d \to \infty$. Consequently, the naive mean-field approximation can only poorly describe the behavior of finite-size systems or strongly interacting systems and it is more of pedagogical interest than of real practical utility.

4.2.3.b TAP, PLEFKA, GEORGES-YEDIDIA HIGH-TEMPERATURE EXPANSION

In fact, it turns out that the naive mean-field approximation can be recovered from the truncation of more complex approximations (Opper et al., 2001a).

Especially in the context of disordered systems with densely connected spin glass, namely the SK model (Sherrington et al., 1975) with Gaussian random $J_{ij} \sim \mathcal{N}(0, J_0/d)$ couplings, the TAP equations (Thouless et al., 1977) provide a more accurate approximation than the naive mean-field equations. Their derivation is closely related to the cavity method (Mézard et al., 1987) or equivalently the Bethe approximation presented in Sec. 4.3.1. Similarly to the cavity method, the idea is to approximate the marginal probability Q_i by considering a reduced set of $d-1$ spins $\boldsymbol{\sigma}_{\setminus i}$ where the spin σ_i has been removed. It is finally possible to write a consistent set of non-linear equations of the form

$$m_i^{t+1} = \tanh\left(\beta\left(h + \sum_{j=1}^{d} J_{ij} m_j^t\right) - \beta^2 m_i^{t-1} \sum_{j=1}^{d} J_{ij}^2 (1 - (m_j^t)^2)\right), \quad (83)$$

called, without the time indices, the TAP equations. We immediately observe that these equations are very similar, yet, more complex than the naive ones in (82). They simply include a correction term to the effective local field, known as the *Onsager reaction term*, to take into account the absence of the spin variable σ_i, which was not present in the oversimplified naive approximation. Indeed, later on, the corresponding TAP free energy has been derived with the so-called Plefka expansion (Plefka, 1982). It turned out that it was simply the second term of a high-temperature expansion proposed in a more general setting (Georges et al., 1991). In the context of the SK model, keeping only the first term leads therefore to the naive mean-field approximation, whereas truncating at the second order, by incorporating the Onsager term, turned out to be exact since other contributions are sub-leading and vanish in the thermodynamic limit. Moreover, the latter derivations insured that the fixed points of the TAP equations are the stationary points of the TAP free energy. However, finding a stationary solution is often achieved by turning them into an iterative procedure until convergence towards fixed points. Unfortunately neither the TAP equations, the Plefka expansion nor the Georges-Yedidia high-temperature expansion include the time indices to iteratively solve them correctly and gives free rein to interpretation. By simply and naturally assuming times $t+1$ on the left hand side of (83) and t on the other magnetizations on the right hand side led to convergence issues of the TAP equations first observed in (Kabashima, 2003). It turns out that this simple arbitrary prescription of the time indices was wrong and responsible for the convergence issues. The time indices were corrected in (Bolthausen, 2014) that finally leads to (83).

In the following, we present an alternative mean-field method based on the BP equations, that provides by construction the correct time indices of the iterative procedure and leads especially to performant algorithms.

4.3 BELIEF PROPAGATION AND THE BETHE FREE ENERGY

In this dissertation, we make deeply use of *message passing algorithms* such as AMP that can be simply derived from the more general set of BP iterative equations. The BP equations have a long history and have started in physics with the Bethe-Peierls approximation (Bethe, 1935; Peierls, 1936). Very interestingly it can be seen as an extended version of the the TAP (Thouless et al., 1977) and Plefka approach (Plefka, 1982), Georges-Yedidia expansions (Georges et al., 1991) presented in the previous section. Moreover as inference problems arose in many various fields, local message passing algorithms have been discovered simultaneously under different names. The BP approach was first introduced in information theory (Gallager, 1962) and in Bayesian inference (Pearl, 1982), whereas it was known under the name of *cavity method* in statistical physics of disordered systems (Mézard et al., 1987; Mézard et al., 2009). The different approaches are reviewed in (Aji et al., 2000; Yedidia et al., 2001a) that connects especially the BP to variational mean-field approach.

In this section, we review the main results of (Yedidia et al., 2001a; Yedidia et al., 2002; Yedidia et al., 2005; Wainwright et al., 2008; Mézard et al., 2009) starting by presenting in Sec. 4.3.1 the Bethe approximation and the Bethe free energy. Theses latter naturally give rise to the set of BP iterative equations presented in Sec. 4.3.3. Finally, in the perspective to derive the AMP algorithm for the GLM class, we present the BP equations for this model class in Sec. 4.3.4.

4.3.1 THE BETHE APPROXIMATION

The Bethe approximation plays a central role among approximations that take into account interactions between spins. In particular, it allows to incorporate correlations between the variables to describe more complex models. Consider the JPD described by a factor graph $\mathcal{G}(\mathrm{V},\mathrm{F},\mathrm{E})$ represented in Fig. 28, the Bethe approximation assumes that the JPD can be written as

$$\mathrm{Q}^{\text{bethe}}(\boldsymbol{\sigma}) = \frac{\prod_{\mu=1}^{n} \tilde{m}_\mu(\boldsymbol{\sigma}_{\partial_\mu})}{\prod_{i=1}^{d} m_i(\sigma_i)^{|\partial_i|-1}}, \tag{84}$$

where $m_i(\sigma_i)$ denotes the marginals of the variable σ_i, and $\tilde{m}_\mu$ the marginals of the cliques $\boldsymbol{\sigma}_{\partial_\mu}$ around the factors μ. The Bethe approximation can be easily derived on tree-like factor graphs, such as the one in Fig. 28, by simply taking the product of the marginals of all the cliques $\tilde{m}_\mu(\boldsymbol{\sigma}_{\partial_\mu})$ and dividing by the variables marginals $m_i(\sigma_i)$ to remove the marginals already taken into account. Therefore, the number of neighbouring factors of the variable σ_i, $|\partial_i|$, is present in the denominator to avoid counting repetitions. Consequently, the formulation (84) has the main advantage to be rigorously exact on tree-like connected factor graphs with no loops, and to be strongly connected to the BP algorithmic procedure (Kabashima et al., 1998) and the

TAP equations. Moreover, this latter formulation can also be applied to more general factor graphs, providing a powerful approximation but loosing in return its exactness.

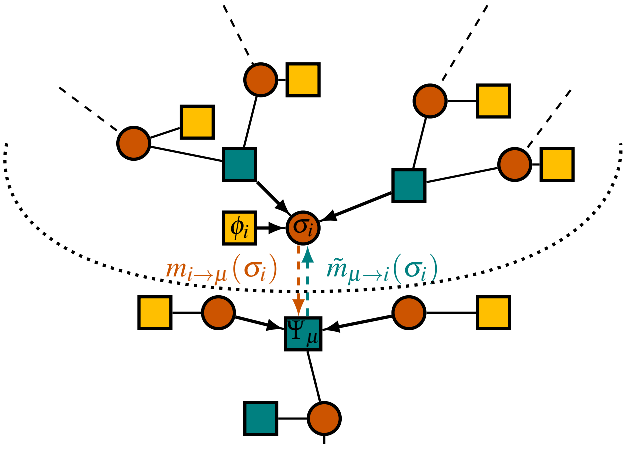


Figure 28: Tree-like factor graph on which the Belief Propagation iterative equations can be decomposed.

4.3.2 THE BETHE FREE ENERGY

Plugging the Bethe approximation (84) in the Gibbs free energy at $\beta = 1$ (78) leads to the corresponding Bethe free energy (Yedidia et al., 2001a; Mézard et al., 2009) which can be written as a functional over the marginals $\{\tilde{m}_\mu\}_{\mu=1}^n \cup \{m_i\}_{i=1}^d$

$$\varphi_d^{\mathrm{bethe}}\left[\{\tilde{m}_\mu\}_\mu, \{m_i\}_i\right] = \mathrm{U}^{\mathrm{bethe}}\left[\{\tilde{m}_\mu\}_\mu, \{m_i\}_i\right] - \mathrm{H}^{\mathrm{bethe}}\left[\{\tilde{m}_\mu\}_\mu, \{m_i\}_i\right], \tag{85}$$

where $\mathrm{U}^{\mathrm{bethe}}$, $\mathrm{H}^{\mathrm{bethe}}$ denote the variational energy and entropy

$$\begin{aligned}\mathrm{U}^{\mathrm{bethe}}\left[\{\tilde{m}_\mu\}_\mu, \{m_i\}_i\right] &\equiv \sum_\mu \int \mathrm{d}\boldsymbol{\sigma}_{\partial_\mu} \tilde{m}_\mu(\boldsymbol{\sigma}_{\partial_\mu}) \log \Psi_\mu(\boldsymbol{\sigma}_{\partial_\mu}) \\ &\quad + \sum_i \int \mathrm{d}\sigma_i m_i(\boldsymbol{\sigma}_i) \log \phi_i(\sigma_i), \\ \mathrm{H}^{\mathrm{bethe}}\left[\{\tilde{m}_\mu\}_\mu, \{m_i\}_i\right] &\equiv \sum_\mu \mathrm{H}\left[\tilde{m}_\mu\right] + \sum_i (|\partial_i| - 1) \mathrm{H}\left[m_i\right],\end{aligned}$$

and $\mathrm{H}[\mathrm{p}]$ the entropy of the probability density p defined in (75). Enforcing the self-consistency marginalization and normalization constraints $m_i(\sigma_i) = \int \mathrm{d}\boldsymbol{\sigma}_{\partial_\mu \setminus i} \tilde{m}_\mu(\boldsymbol{\sigma}_{\partial_\mu})$, $\int \mathrm{d}\sigma_i m_i(\sigma_i) = 1 = \int \mathrm{d}\boldsymbol{\sigma}_{\partial_\mu} \tilde{m}_\mu(\boldsymbol{\sigma}_{\partial_\mu})$, with some Lagrange multipliers (Yedidia et al., 2001a; Yedidia et al., 2005; Wainwright et al., 2008), the extremization of the Lagrangian leads to the following expressions of the marginals estimate

$$\tilde{m}_\mu(\boldsymbol{\sigma}_{\partial_\mu}) \propto \Psi_\mu(\boldsymbol{\sigma}_{\partial_\mu}) \prod_{i\in\partial\mu} m_{i\to\mu}(\sigma_i)\,, \qquad m_i(\sigma_i) \propto \prod_{\mu\in\partial_i} \tilde{m}_{\mu\to i}(\sigma_i)\,,$$

that involve approximate beliefs $\{m_{i\to\mu}, \tilde{m}_{\mu\to i}\}$ over the variable σ_i if we respectively cut the edge $(i\mu) \in \mathrm{E}$ of the factor graph between the variable σ_i and the factor Ψ_μ, as illustrated in Fig. 28. In the context of pairwise MRF, the above conditions are crucial to understand the link between BP, introduced in Sec. 4.3, and the Bethe approximation as stressed in (Kabashima et al., 1998; Yedidia et al., 2001a). Indeed, since the BP marginal densities are obtained by extremizing the Bethe free energy, the fixed point of the BP algorithm are by construction the stationary points of the Bethe free energy. Finally, under the Bethe approximation (84), the Bethe free energy can be written as a function of the one and two-body interactions $\{\phi_i, \Psi_\mu\}$ and the beliefs $\{m_{i\to\mu}, \tilde{m}_{\mu\to i}\}$:

$$\varphi_d^{\text{bethe}} = -\sum_{i\in\mathrm{V}} \log \mathcal{Z}_i - \sum_{\mu\in\mathrm{F}} \log \mathcal{Z}_\mu + \sum_{(i\mu)\in\mathrm{E}} \log \mathcal{Z}_{i\mu}\,, \tag{86}$$

with

$$\mathcal{Z}_i = \int \mathrm{d}\sigma_i \phi_i(\sigma_i) \prod_{i\in\partial_\mu} \tilde{m}_{\mu\to i}(\sigma_i)\,, \quad \mathcal{Z}_{i\mu} = \int \mathrm{d}\sigma_i \tilde{m}_{\mu\to i}(\sigma_i) m_{i\to\mu}(\sigma_i)\,,$$
$$\mathcal{Z}_\mu = \Psi_\mu(\boldsymbol{\sigma}_{\partial_\mu}) \int \prod_{i\in\partial_\mu} \mathrm{d}\sigma_i \prod_{i\in\partial_\mu} m_{i\to\mu}(\sigma_i)\,.$$

4.3.3 BELIEF PROPAGATION EQUATIONS

The BP algorithm is an inference algorithm that computes an approximation of the marginal densities of a complex JPD. In particular, it makes use of the fact that many JPD are locally factorizable to reduce the estimation of the full complex problem into tractable sub-problems on each factor of the factor graph. The set of BP iterative equations can be obtained directly from the variational principle and the Bethe free energy as presented in the previous section. Yet, for a more intuitive and practical perspective they can be directly obtained from the factor graph as we detail in the following. Depending on the problem under consideration, the BP approach can be expressed in two variants: the *sum-product* or the *max-sum* equations. The sum-product approach estimates the marginal densities and directly leads to MMSE estimation. In contrast, the max-sum approach is more suitable to MAP estimation and the corresponding equations can be found in (Mézard et al., 2009; Advani et al., 2016b).

4.3.3.A SUM-PRODUCT EQUATIONS

Let us present the sum-product version of the BP procedure for a general MRF illustrated in Fig. 28. Importantly, we first attach two kinds of auxiliary functions $\{m_{i\to\mu}, \tilde{m}_{\mu\to i}\}$ to the edges of the factor graph, called *messages*. These messages are interpreted (Mézard et al., 2009; Yedidia et al., 2005) as the estimates of the marginal $\mathrm{Q}(\sigma_i)$ if we remove the edges $(i\mu) = \{i \to \mu, \mu \to i\}$. In other words, $m_{i\to\mu}(\sigma_i)$ denotes the message from the variable σ_i to the factor node Ψ_μ delivering the estimation of the marginal density $\mathrm{Q}(\sigma_i)$ in the partial visited graph represented by the top part of the graph in Fig. 28 and delimited by the dotted line. Similarly, $\tilde{m}_{\mu\to i}(\sigma_i)$ is the message from the factor node Ψ_μ to the variable σ_i that transmits an estimation of the marginal density $\mathrm{Q}(\sigma_i)$ in the bottom part of the graph, called the *cavity graph*. Essentially, the BP algorithm consists in letting the variables and factors communicate their *beliefs* to their neighbours based on the informations captured from the nodes and factors already visited along the tree. Iterating the procedure, we expect qualitatively that the beliefs converge to an average value of their neighbouring beliefs. To formalize this procedure, the sum-product equations (Gallager, 1962; Pearl, 1982; Wainwright et al., 2008; Mézard et al., 2009) make use of the crucial tree-like assumption, originating from the Bethe approximation, that guarantees that the incoming messages to the variable σ_i are independent. Thereby, the messages are given by the self-consistency rule for the messages $m^{t+1}_{i\to\mu}$ and $\tilde{m}^t_{\mu\to i}$, $\forall i \in [\![d]\!]$, $\mu \in [\![n]\!]$ according to

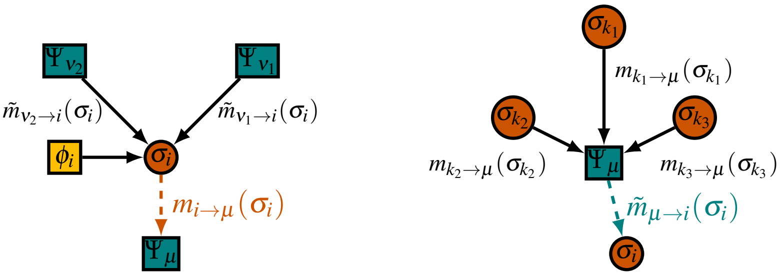

Ψ_{ν_2}
Ψ_{ν_1}
$\tilde{m}_{\nu_2\to i}(\sigma_i)$
$\tilde{m}_{\nu_1\to i}(\sigma_i)$
ϕ_i
σ_i
$m_{i\to\mu}(\sigma_i)$
Ψ_μ
σ_{k_1}
$m_{k_1\to\mu}(\sigma_{k_1})$
σ_{k_2}
σ_{k_3}
Ψ_μ
$m_{k_2\to\mu}(\sigma_{k_2})$
$m_{k_3\to\mu}(\sigma_{k_3})$
$\tilde{m}_{\mu\to i}(\sigma_i)$
σ_i

Figure 29: Local representation of the factor graph around the variable σ_i and the factor Ψ_μ.

$$\begin{aligned} m^{t+1}_{i\to\mu}(\sigma_i) &= \frac{1}{\mathcal{Z}_{i\to\mu}} \phi_i(\sigma_i) \prod_{\nu\in\partial_i\setminus\mu} \tilde{m}^t_{\nu\to i}(\sigma_i) \\ \tilde{m}^t_{\mu\to i}(\sigma_i) &= \frac{1}{\mathcal{Z}_{\mu\to i}} \sum_{\boldsymbol{\sigma}_{\partial_\mu\setminus i}} \Psi_\mu\left(\boldsymbol{\sigma}_{\partial_\mu\setminus i}\right) \prod_{k\in\partial_\mu\setminus i} m^t_{k\to\mu}(\sigma_k)\,. \end{aligned} \tag{87}$$

These update rules can be easily understood by looking at the local decomposition of the factor graph Fig. 29. The message $m_{i\to\mu}$ in dashed orange is built from the incoming messages of the neighbouring factors of the spin σ_i if we remove the edge $i \to \mu \in \mathrm{E}$. Similarly the message $\tilde{m}_{\mu\to j}$ in dashed green is obtained by summing over all the possible values of the messages coming from the neighbouring variables if we remove the edge $\mu \to i \in \mathrm{E}$.

4.3.3.B BP ALGORITHM AND PROPERTIES

The BP algorithm is the procedure that consists in iterating the set of dynamical equations (87) over time. Eventually if it converges, it provides at convergence an estimation of the Bethe free energy (86) and especially of the marginal probabilities given by $\forall i$, $\mathrm{Q}(\sigma_i) \propto \phi_i(\sigma_i) \prod_{\mu \in \partial_i} \tilde{m}_{\mu \to i}(\sigma_i)$ where $\partial_i = \{\mu : (\mu i) \in \mathrm{E}\}$ represents all the neighbouring factors of the variable i. However the *messages independence* is crucial for writing the BP equations (87), such that the obtained marginal estimation and the Bethe free energy are exact only in the case of DAG factor graphs for which there is no correlation between the incoming messages. In other words, by construction the convergence of BP to the fixed points of the Bethe free energy is guaranteed only for tree-like factor graphs.

Loopy BP Nevertheless, the powerful BP algorithm can be used as an approximation in more complex MRF that do not factorize as DAG and thus contain some loops. The influence of *loops* in the graph can induce strong correlations and harm the convergence of BP. Violating the messages independence hypothesis breaks the convergence guarantees, but provides anyway an *approximate* algorithmic procedure that, hoping for the best, may still converge. Notice that there exists some cases for which the presence of loops may not be that harmful. In particular, in the case of *long enough loops* and if the *correlations decrease fast enough* with the Hamming distance, since the factor graph remains locally tree-like, we expect the message independence to still hold. In this context, under the name of *loppy*-BP, the algorithm may sometimes succeed converging and provide good approximations of the marginals, loosing in return convergence guarantees. As an alternative, since the BP algorithm is not guaranteed to converge, we could instead directly find the minimum of the Bethe free energy (Yuille, 2001), even though it is much slower and the Bethe free energy does not provide anymore a variational Gibbs free energy upper-bound.

State Evolution In addition of being a general procedure adaptable to any MRF, the main interest of BP lies in the possibility to predict its asymptotic performances. Indeed, in the thermodynamic limit $d \to \infty$, it is possible to fully characterize the dynamics of the BP fixed point equations, known as the State Evolution (SE) equations . They have been introduced in (Bayati et al., 2011b) and their interest considerably increased with the regain of activity in the high-dimensional regime. In the next section, we will show in the context of the GLM class that the SE equations of AMP, which is nothing more than a Gaussian simplification of the BP algorithm, can be equivalently obtained from the replica computation, which therefore provides an efficient way for obtaining the asymptotic behaviour of such message passing algorithms.

4.3.4 APPLICATION - BP EQUATIONS FOR THE GLM

As a central illustration, we present the instructive and systematic derivation of the relaxed Belief Propagation (rBP) equations starting with the BP equations, before deriving the corresponding AMP algorithm in the next section, for the GLM class. The corresponding JPD can be written as

$$\mathrm{P}_d\left(\mathbf{w}|\mathbf{y},\mathbf{X}\right) = \frac{\prod_{\mu=1}^{n} \mathrm{P}_{\mathrm{out}}(y_\mu|z_\mu) \prod_{i=1}^{d} \mathrm{P}_{\mathrm{w}}(w_i)}{\mathcal{Z}_d(\mathbf{y},\mathbf{X})}, \tag{88}$$

already considered in Sec. 4.1.5, and where we assumed that the channel and prior distributions factorize over factors $\Psi_\mu = \mathrm{P}_{\mathrm{out},\mu}$ and spin variables $\phi_i = \mathrm{P}_{\mathrm{w},i}$. The posterior distribution can be naturally represented by the factor graph Fig. 30. We define the auxiliary variable $z_\mu = \frac{1}{\sqrt{d}}\mathbf{w}^\intercal\mathbf{x}_\mu = \Theta(1)$ which is of order one thanks to the crucial rescaling pre-factor $1/\sqrt{d}$. Indeed, even though the factor graph is fully-connected and contains short loops, this weak coupling insures that the messages remain slightly correlated and the BP equations hold true. The details of the computations for the more general committee machine hypothesis class can be found in Appendix. C.

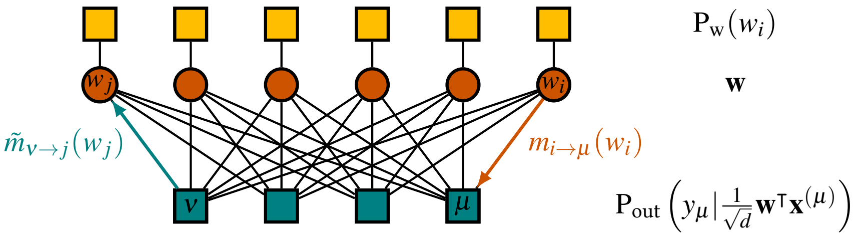

$\mathrm{P}_{\mathrm{w}}(w_i)$
w_j
w_i
$\mathbf{w}$
$\tilde{m}_{\nu\to j}(w_j)$
$m_{i\to\mu}(w_i)$
ν
μ
$\mathrm{P}_{\mathrm{out}}\left(y_\mu|\frac{1}{\sqrt{d}}\mathbf{w}^\intercal\mathbf{x}^{(\mu)}\right)$

Figure 30: Factor graph corresponding to the posterior distribution (88) associated to the GLM hypothesis class. The variable w_i send a message $m_{i\to\mu}(w_i)$ to the factor μ, and reciprocally it sends back a message $\tilde{m}_{\mu\to i}(w_i)$ to the variable based on the corresponding cavity graph.

BP equations for the GLM Let us consider a set of messages $\{m_{i\to\mu}, \tilde{m}_{\mu\to i}\}$ on the edges of the bipartite factor graph Fig. 30. These messages correspond to the marginal probabilities of w_i if we remove the edges $(i \to \mu)$ and $(\mu \to i)$. The sum-product BP equations (87) are simply follow as

$$\begin{aligned} m_{i\to\mu}^{t+1}(w_i) &= \frac{1}{\mathcal{Z}_{i\to\mu}} \mathrm{p}_{\mathrm{w}}(w_i) \prod_{\nu\neq\mu}^{n} \tilde{m}_{\nu\to i}^{t}(w_i), \\ \tilde{m}_{\mu\to i}^{t}(w_i) &= \frac{1}{\mathcal{Z}_{\mu\to i}} \int \prod_{j\neq i}^{d} \mathrm{d}w_j \; \mathrm{p}_{\mathrm{out}}\left(y_\mu | \frac{1}{\sqrt{d}} \sum_{j=1}^{d} x_{\mu j} w_j\right) m_{j\to\mu}^{t}(w_j). \end{aligned} \tag{89}$$

Towards relaxed-Belief Propagation equations The idea is to expand, in the limit $d \to \infty$, the set of $\Theta(d^2)$ messages $\{\tilde{m}_{\mu\to i}\}_{i,\mu}$ before plugging them back in the messages $\{m_{i\to\mu}\}_{i,\mu}$. Truncating the expansion and keeping only terms of order $\Theta(1/d)$, messages become *Gaussian* and therefore messages can be parametrized only by the mean $\hat{w}^t_{i\to\mu}$ and the variance $\hat{c}^t_{i\to\mu}$ of the marginal distribution estimate $m^t_{i\to\mu}$ at time t:

$$\begin{aligned} \hat{w}^t_{i\to\mu} &\equiv \int_{\mathbb{R}} \mathrm{d}w_i\, m^t_{i\to\mu}(w_i) w_i\,, \\ \hat{c}^t_{j\to\mu} &\equiv \int_{\mathbb{R}} \mathrm{d}w_i\, m^t_{i\to\mu}(w_i) w_i^2 - (\hat{w}^t_{i\to\mu})^2\,. \end{aligned} \tag{90}$$

Using a Fourier representation of $\mathrm{P}_{\mathrm{out}}$ in $\tilde{m}^t_{\mu\to i}$ in (89) to decouple its fully-connected argument, and expanding it in the large size limit $d \to \infty$, we obtain that the set of BP equations finally closes over Gaussian beliefs $\{m_{i\to\mu}\}_{i,\mu}$

$$m^{t+1}_{i\to\mu}(w_i) = \frac{1}{\mathcal{Z}_{i\to\mu}} \mathrm{p}_{\mathrm{w}}(w_i) \prod_{\nu\neq\mu}^{n} \sqrt{\frac{A^t_{\nu\to i}}{(2\pi)}} e^{-\frac{A^t_{\nu\to i}}{2}\left(w_i - (A^t_{\nu\to i})^{-1} b^t_{\nu\to i}\right)^2}\,, \tag{91}$$

with natural parameters $b^t_{\mu\to i}$ and the precision $A^t_{\mu\to i}$ defined as

$$b^t_{\mu\to i} \equiv \frac{x_{\mu i}}{\sqrt{d}} f_{\mathrm{out}}(y_\mu, \omega^t_{i\mu}, V^t_{i\mu})\,, \qquad A^t_{\mu\to i} \equiv -\frac{x^2_{\mu i}}{d} \partial_\omega f_{\mathrm{out}}(y_\mu, \omega^t_{i\mu}, V^t_{i\mu}) \tag{92}$$

with the *channel denoising functions* $f_{\mathrm{out}}, \partial_\omega f_{\mathrm{out}}$, defined in Appendix. A.4, which depend on the mean and variance of the *channel belief*

$$\omega^t_{\mu\to i} \equiv \frac{1}{\sqrt{d}} \sum_{j\neq i}^{d} x_{\mu j} \hat{w}^t_{j\to\mu}\,, \qquad V^t_{\mu\to i} \equiv \frac{1}{d} \sum_{j\neq i}^{d} x^2_{\mu j} \hat{c}^t_{j\to\mu}\,. \tag{93}$$

Finally the mean and variance (90) of the message $m_{i\to\mu}$ are updated by

$$\hat{w}^{t+1}_{i\to\mu} = f_{\mathrm{w}}(\gamma^t_{\mu\to i}, \Lambda^t_{\mu\to i})\,, \qquad \hat{c}^{t+1}_{i\to\mu} = \partial_\gamma f_{\mathrm{w}}(\gamma^t_{\mu\to i}, \Lambda^t_{\mu\to i})\,, \tag{94}$$

with the *prior denoising functions* $f_{\mathrm{w}}, \partial_\gamma f_{\mathrm{w}}$, defined in Appendix. A.4, where the mean $\gamma^t_{\mu\to i}$ and variance $\Lambda^t_{\mu\to i}$ of the *prior belief* are defined by

$$\gamma^t_{\mu\to i} = \sum_{\nu\neq\mu}^{n} b^t_{\nu\to i}\,, \qquad \Lambda^t_{\mu\to i} = \sum_{\nu\neq\mu}^{n} A^t_{\nu\to i}\,. \tag{95}$$

The set of equations (92, 93, 94, 95) form the set of $\Theta(d^2)$ rBP equations, which are simply the projection of the BP equations over any parametrized family, namely the Gaussian family in the presented case.

4.4 APPROXIMATE MESSAGE PASSING

AMP algorithms start to emerge (Boutros et al., 2002; Montanari et al., 2006) and being popular when applied to dense models such as CS (Donoho et al., 2009; Bayati et al., 2011b) and later to GLM with the Generalized Approximate Message Passing (GAMP) algorithm (Rangan, 2011). These algorithms are closely related to the so-called TAP equations in the context of the spin glass theory and the SK model (Thouless et al., 1977; Sherrington et al., 1975) as the latter mean-field equations and TAP free energy can be recovered from the Bethe free energy. However, AMP algorithms largely overtook these previous mean-field methods presented in Sec. 4.2.3 as they naturally provide the correct time indices to iterate the self-consistent set of fixed point equations. In contrast, the TAP approach struggled to solve them as no explicit iteration scheme is prescribed by the method. This connection with statistical mechanics was recently renewed with notably (Tanaka, 2002; Guo et al., 2005b; Rangan et al., 2009; Krzakala et al., 2012b) and this manuscript falls within this same approach by applying AMP to the theoretical understanding of ANN. Moreover, being popularized to various applications, AMP algorithms underwent various extensions such that BiGAMP for bilinear estimation (Parker et al., 2014) or ML-AMP for multi-layer estimation (Manoel et al., 2017). Informally the AMP algorithms can be seen as a Taylor expansion of the loopy-BP approach (Mézard et al., 1987; Mézard et al., 2009; Wainwright et al., 2008). The general procedure starts with the set of loopy-BP equations (87) associated to the corresponding JPD and factor graph. After performing the asymptotic expansion and parametrize the beliefs with Gaussians that allows to track only two parameters, the mean and variance, per message, we finally end up with a set of $\Theta(d^2)$ messages, the so-called set of rBP equations. Their latter computational cost can be reduced with additional expansions around the *full messages* to remove the target-node dependency at the cost of making appear *Onsager terms* at time previous steps. Finally, in the large size limit $d \to \infty$, keeping only the leading terms, the set of equations can be reduced to a set of only $\Theta(d)$ messages. Notice that the discrepancy between the BP algorithm and the obtained AMP algorithm are not quantified rigorously as anyway, for general factor graphs with loops, the loopy-BP provides only an approximate estimation, so does AMP. However, the resulting AMP has the strong advantage to be rigorously provable in a roundabout way, from the so-called SE equations that can be obtained from the replica computation, proven with a Guerra-like interpolation (Guerra, 2003).

For the sake of clarity, we present a pedagogical and instructive derivation of the GAMP algorithm, following closely the one of (Zdeborová et al., 2016a).

4.4.1 APPLICATION - AMP FOR THE GLM

The rBP set of equations for the GLM (92, 93, 94, 95) contain $\Theta(d^2)$ messages of the form $x_{i\to\mu}$. However it is worth observing that the messages *depend weakly* on the *target node* μ, as the missing message in the sum vanishes in the limit $d \to \infty$. This crucial observation allows to expand the previous rBP equations around the *full* messages by completing the sum that do not show anymore the target-node dependence:

$$\begin{aligned} \omega_\mu^t &\equiv \sum_{j=1}^{d} \frac{x_{\mu j}}{\sqrt{d}} \hat{w}_{j\to\mu}^t\,, \quad & V_\mu^t &\equiv \sum_{j=1}^{d} \frac{x_{\mu j}^2}{d} \hat{c}_{j\to\mu}^t\,, \\ \gamma_i^t &\equiv \sum_{\nu=1}^{n} b_{\nu\to i}^t\,, & \Lambda_i^t &\equiv \sum_{\nu=1}^{n} A_{\nu\to i}^t\,. \end{aligned} \tag{96}$$

Performing the expansion of the rBP (92,93, 94, 95), the set can be reduced to $\Theta(d)$ iterative equations, at the cost of introducing *memory terms* at previous time steps, the Onsager terms. The lengthy, yet straightforward, computation is shown for the committee machines hypothesis class in Appendix. C.3. We finally end up with the GAMP algorithm (Rangan, 2011) as a set of $\Theta(d)$ messages presented in Algo. 1. The GAMP algorithm can be interpreted in a series of iterative steps starting by the estimation of the mean $\boldsymbol{\omega}$ and variance $\mathbf{V}$ of the variable $\mathbf{z} = \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}$. The estimate of $\mathbf{z}$ provides a potential output that is compared to the true output vector $\mathbf{y}$ through the denoising functions f_{out}, $\partial_\omega f_{\text{out}}$. This comparison gives a feedback to update the mean $\boldsymbol{\gamma}$ and variance $\boldsymbol{\Lambda}$ of the variable $\mathbf{w}$ which is updated to provide a new estimation with the denoising functions f_{w} and $\partial_\gamma f_{\text{w}}$. Moreover, AMP provides a general inference algorithm valid on single instance of finite size for generic prior and channel distributions. As a consequence, it is valid in the Bayes-optimal case for MMSE

```
Input: vector $\mathbf{y} \in \mathbb{R}^n$ and matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$:
Initialize: $\hat{w}_i$, $f_{\text{out},\mu} \in \mathbb{R}$ and $\hat{c}_i$, $\partial_\omega f_{\text{out},\mu} \in \mathbb{R}$ for $1 \le i \le d$ and $1 \le \mu \le n$ at $t = 0$.
repeat
    Channel: Update the mean $\omega_\mu \in \mathbb{R}$ and variance $V_\mu \in \mathbb{R}^+$:
    $\omega_\mu^t = \sum_{i=1}^{d} \frac{x_{\mu i}}{\sqrt{d}} \hat{w}_i^t - V_\mu^t f_{\text{out},\mu}^{t-1}\,,\; V_\mu^t = \sum_{i=1}^{d} \frac{x_{\mu i}^2}{d} \hat{c}_i^t$
    Update $f_{\text{out},\mu} \in \mathbb{R}$ and $\partial_\omega f_{\text{out},\mu} \in \mathbb{R}^+$:
    $f_{\text{out},\mu}^t = f_{\text{out}}(y_\mu, \omega_\mu^t, V_\mu^t)\,,\; \partial_\omega f_{\text{out},\mu}^t = \partial_\omega f_{\text{out}}(y_\mu, \omega_\mu^t, V_\mu^t)$
    Prior: Update the mean $\gamma_i \in \mathbb{R}$ and variance $\Lambda_i \in \mathbb{R}^+$:
    $\gamma_i^t = \sum_{\mu=1}^{n} \frac{x_{\mu i}}{\sqrt{d}} f_{\text{out},\mu}^t + \Lambda_i^t \hat{w}_i^t\,,\; \Lambda_i^t = -\sum_{\mu=1}^{n} \frac{x_{\mu i}^2}{d} \partial_\omega f_{\text{out},\mu}^t$
    Update the estimated marginals $\hat{w}_i \in \mathbb{R}$ and $\hat{c}_i \in \mathbb{R}^+$:
    $\hat{w}_i^{t+1} = f_{\text{w}}(\gamma_i^t, \Lambda_i^t)\,,\; \hat{c}_i^{t+1} = \partial_\gamma f_{\text{w}}(\gamma_i^t, \Lambda_i^t)$
    $t = t + 1$
until Convergence on $\hat{\mathbf{w}}$, $\hat{\mathbf{c}}$.
Output: $\hat{\mathbf{w}}$ and $\hat{\mathbf{c}}$.
```

Algorithme 1 : Approximate Message Passing algorithm for Generalized Linear Models.

estimation when $\mathrm{P}_{\mathrm{out}} = \mathrm{P}_{\mathrm{out}^\star}$ and $\mathrm{P}_{\mathrm{w}} = \mathrm{P}_{\mathrm{w}^\star}$ or the mismatched setting for arbitrary distribution such as for MAP estimation and ERM. Finally looking beyond the cumbersome appearances of Algo. 1, in contrast with most of state-of-the-art gradient-based algorithms that suffer theoretical understanding, AMP algorithms have the main advantage that their asymptotic behavior and convergence performances can be rigorously tracked for large i.i.d input matrices through their SE equations. At the heart of this manuscript, we should stress that these SE equations connect surprisingly to the results obtained by the replica computation.

4.4.2 STATE EVOLUTION EQUATIONS - CONNECTION WITH REPLICAS

One of the main interests of the AMP algorithm is certainly that we can analyze its average behavior in the thermodynamic limit. Indeed, taking the average over the quenched disorder and introducing proper order parameters, the so-called *overlaps*, we can obtain an asymptotic closed set of equations, called the SE equations. They characterize the performances of the AMP algorithm (Donoho et al., 2009; Bayati et al., 2011b; Javanmard et al., 2013) in the large size limit $d \to \infty$. The derivation usually starts with the set of rBP equations. By assuming the fundamental message independence, using the CLT, and defining a correctly chosen set of order parameters, the statistical analysis ends up to the set of SE equations. Very importantly, under the strong i.i.d assumption and in the Bayes-optimal case, these SE equations converge to the stationary points of the RS replica free entropy. We illustrate the derivation again on the GLM hypothesis class and draw the intimate connection with the replica computation. The full computation for the committee hypothesis class can be found in Appendix. C.4.

4.4.3 APPLICATION - SE FOR THE GLM

The derivation of the SE equations starts by defining a series of order parameters, called *overlaps*

$$m^t \equiv \lim_{d\to\infty} \mathbb{E}_{\mathbf{w}^\star,\mathbf{X}} \left[\frac{1}{d}\hat{\mathbf{w}} \cdot \mathbf{w}^\star\right], \qquad q^t \equiv \lim_{d\to\infty} \mathbb{E}_{\mathbf{w}^\star,\mathbf{X}} \left[\frac{1}{d}\hat{\mathbf{w}}^t \cdot \hat{\mathbf{w}}^t\right],$$
$$\rho_{\mathrm{w}^\star} \equiv \lim_{d\to\infty} \mathbb{E}_{\mathbf{w}^\star} \left[\frac{1}{d}\mathbf{w}^\star \cdot \mathbf{w}^\star\right], \qquad \Sigma^t \equiv \lim_{d\to\infty} \mathbb{E}_{\mathbf{w}^\star,\mathbf{X}} \left[\frac{1}{d}\hat{\mathbf{c}}^t \cdot \mathbf{1}\right], \tag{97}$$

that measure the correlations between the ground truth vector $\mathbf{w}^\star$ and the estimator $\hat{\mathbf{w}}^t$ at time t of the AMP algorithm in Algo. 1. In a T-S scenario, they allow to quantify properly the reconstruction performance of the AMP algorithm. For the purpose of the derivation, we need to define other ad-hoc overlaps which have less direct physical meaning

$$\begin{aligned}
\hat{q}^t &\equiv \alpha \mathbb{E}_{\omega,z} \left[f_{\text{out}}(\varphi_{\text{out}^\star}(z), \omega^t, \Sigma^t)^2 \right] , \\
\hat{m}^t &\equiv \alpha \mathbb{E}_{\omega,z} \left[\partial_z f_{\text{out}}(\varphi_{\text{out}^\star}(z), \omega^t, \Sigma^t) \right] , \\
\hat{\chi}^t &\equiv \alpha \mathbb{E}_{\omega,z} \left[-\partial_\omega f_{\text{out}}(\varphi_{\text{out}^\star}(z), \omega^t, \Sigma^t) \right] .
\end{aligned} \tag{98}$$

Under the BP independent messages assumption and using the CLT, we obtain in Appendix. C.4 the message statistics of the rBP messages. Computing the average of the overlaps defined above and making use of the latter statistics, we finally obtain a set of six SE equations, that can be reduced using the Nishimori conditions in the Bayes-optimal case (Opper et al., 1991a; Iba, 1999)

$$m^t = q^t , \qquad \hat{q}^t = \hat{m}^t = \hat{\chi}^t , \qquad \Sigma^t = \rho_{\text{w}^\star} - q^t ,$$

to only two SE equations

$$\begin{aligned}
q^{t+1} &= \mathbb{E}_\xi \left[\mathcal{Z}_{\text{w}^\star}\left((\hat{q}^t)^{1/2} \xi, \hat{q}^t \right) f_{\text{w}^\star}\left((\hat{q}^t)^{1/2} \xi, \hat{q}^t \right)^2 \right] , \\
\hat{q}^t &= \alpha \int_{\mathbb{R}} \mathrm{d}y \, \mathbb{E}_\xi \, \mathcal{Z}_{\text{out}^\star}\left(y, (q^t)^{1/2} \xi, \rho_{\text{w}^\star} - q^t \right)^2 \\
&\qquad \times f_{\text{out}^\star}\left(y, (q^t)^{1/2} \xi, \rho_{\text{w}^\star} - q^t \right)^2 .
\end{aligned} \tag{99}$$

As a crucial conclusion, we finally observe that the set of SE equations (99), which characterize the asymptotic behavior of the AMP algorithm in the Bayes-optimal setting, are connected to the ones obtained by the i.i.d replica computation in (74). Indeed, similarly to the TAP approach, while the replica result does not provide the time indices to solve the fixed point equation, the SE (99) fully determine the dynamics of the AMP algorithm at any time t. As as consequence, it turns out that, under a T-S scenario the SE of the AMP algorithm follows exactly the gradient of the RS free entropy (321) in the Bayes-optimal setting, that intrinsically grasp the importance of the overlaps in the considered JPD. Importantly, the connection between AMP, the replica formalism and the possibility to prove them rigorously with Guerra-like interpolation breaks down in the *mismatched setting*. In this case, the prediction of the replica method fails delivering the correct behavior of the AMP algorithm under the simple RS assumption and reveal a more complex RSB structure of the phase space. Yet, it is not the case for convex optimization as shown in Chap. 8, where the RS turned out to hold rigorously correct even in the mismatched MAP estimation setting.

4.4.4 BEYOND I.I.D MATRICES AND AMP

Even though the derivation of AMP in the case of the GLM does not assume any hypothesis on the input matrix $\mathbf{X}$, it has been observed that AMP may experience divergences for non-i.i.d inputs, even for non-pathological matrices (Rangan et al., 2019b). In practice, to circumvent this issue, we can try to improve the stability by using mean removal, *damping* (Vila et al., 2015; Rangan et al., 2019b), sequential updating or other tricks. These stability techniques are partially successful but convergence may still fail and often needs specific tuning. Moreover even the convergence of AMP is proven (Bayati et al., 2011b; Bolthausen, 2014) only under particular restrictive conditions, such as the i.i.d hypothesis. As a consequence it is not surprising that correlated statistics can therefore breaks down the message independence assumption in AMP and leads to divergences for more complex input matrices. To overcome this fundamental limitation, many efforts have been made to generalize the mean-field approaches to more complex matrix statistics such as the high-temperature expansion for orthogonal matrices (Parisi et al., 1995) or the ADA-TAP approach for dense graphical models with generic weight statistics (Opper et al., 2001b; Opper et al., 2001c). The corresponding approaches were understood later as a particular case of the Expectation Propagation (EP) approximate inference algorithm (Minka, 2001b; Minka, 2001a; Heskes et al., 2005; Heskes et al., 2012). Similarly to BP with the Bethe free energy, the EP procedure is associated to an approximate free energy called the Expectation Consistency (EC) (Opper et al., 2005) and solution of the Gibbs variational principle by enforcing the moments matching constraints. Analogously to AMP with BP, the EP procedure was applied by projecting the messages to a Gaussian parametrization leading to the Vector Approximate Message Passing (VAMP) algorithm (Rangan et al., 2019a). For the sake of clarity, the main difference with AMP lies on the fact that the factor graph is *vector valued*, meaning that a vector variable with a separable prior is represented by a single variable and factor nodes, which insight comes from (Cakmak et al., 2014). It reduces the global inference problems to sub-vectorial-problems by imposing connecting Dirac-delta constraints. AMP and VAMP are conjectured to be equivalent for i.i.d matrices, and asymptotically are proved to be rigorously identical as the corresponding RS free energies are equivalent. However, VAMP experienced less convergence issues than the classical AMP. VAMP converges for orthogonally invariant matrices and is more stable and robust to ill-conditioned matrices at the cost of computing matrix inversion or SVD. Very importantly, similarly to AMP algorithms, the VAMP algorithm asymptotic performances are remarkably characterized by a set of SE equations for orthogonally invariant matrices. The SE are again related to the stationary point of the replica free energy computed for this matrix statistics as observed in (Tulino et al., 2013) and shown rigorously in (Barbier et al., 2018b) in linear estimation and (Gabrié et al., 2018; Reeves, 2017) in DNN.

Part II

MAIN CONTRIBUTIONS

Part II A.

BAYES-OPTIMAL, EMPIRICAL RISK MINIMIZATION AND WORST-CASE ANALYSIS IN SIMPLE FEED-FORWARD NEURAL NETWORKS

OUTLINE AND MOTIVATIONS

The mean-field methods originating from statistical physics presented in Chap. 4 have been extensively used in the past to analyze the equilibrium behavior of common estimators for simple model classes such as single-layer neural networks. This statistical physics approach, seeking to understand the *typical* behavior of such systems, focused essentially on a simple data generative process: the *teacher-student* scenario glimpsed in Chap. 3.

In this part, we essentially revisit this setting in light of the modern challenge of algorithmic complexity, and second, we develop and extend rigorous proofs of earlier heuristic analysis. With these contributions, we propose an overview of the complementary approaches used in different communities to understand simple classes of feed-forward neural networks. Especially, in the context of the classical *perceptron* and its multi-layer generalization, the *committee machine*, we try to reconcile the *Bayes-optimal* setting, the *wort-case* analysis and *empirical risk minimization methods* in a unified statistical physics framework.

Bayes-optimal analysis and computational complexity In Chap. 5, we revisit the T-S scenario to a more sophisticated class of two-layers neural networks, the *committee machines*. It naturally extends and encompasses the classical GLM class (Barbier et al., 2016), which was restricted to linearly separable data, to higher complexity models. We analyze the Bayes-optimal setting, in the case where the dataset has been generated by a ground truth *teacher* committee machine with weights $\mathbf{W}^\star \in \mathbb{R}^{d\times K}$ and prior distributions $\mathrm{P}_{\mathrm{out}^\star}$ and $\mathrm{P}_{w^\star}$. Under the Bayes-optimal hypothesis, the *student* seeks to fit the dataset based on the exact model architecture and by having access to the ground truth prior distributions $\mathrm{P}_{\mathrm{out}} = \mathrm{P}_{\mathrm{out}^\star}$ and $\mathrm{P}_{\mathrm{w}} = \mathrm{P}_{\mathrm{w}^\star}$. This idealistic setting provides nonetheless a crucial information theoretical lower-bound of optimal statistical estimation. It naturally provides answers to the questions of knowing *under what conditions*, and without any algorithmic consideration, if *it is possible to recover the structure in the data*? And *how many examples* are needed in that case? In the asymptotic regime where the size of the input vector d and the number of training examples n diverge, using Guerra interpolation we first provide a rigorous justification that the RS free entropy, initially derived with the heuristic replica method (Schwarze et al., 1992; Schwarze et al., 1993), is *exact* for i.i.d input data $\mathbf{X} \in \mathbb{R}^{n\times d}$. Moreover we provide expressions for the corresponding optimal generalization error learning curves. Secondly, we develop an extension of the polynomial time GAMP algorithm (Rangan, 2011) for the committee machine hypothesis class. This algorithmic perspective allows us to answer the burning questions of computational complexity that focuses on knowing *if the algorithm is efficient with respect to the information-theoretical baseline* and *how many examples it requires to achieve it*. By locating the phase transitions, we highlight hard regions where the best algorithm fails delivering the optimal predictions.

To summarize, our approach capitalizes on the trade-off between information theoretical statistical inference and the computational efficiency of the conjectured best polynomial AMP algorithm for i.i.d data. By making an intense use of the description of metastability, first and second order phase transitions phenomenology, borrowed to statistical physics and depicted in Chap. 2, we can study in details the phase statistical and algorithmic phase transitions. Especially, we unveil the existence of large computational gaps, even in the Bayes-optimal case, for small and extensive hidden-layer sizes K.

Worst case analysis: from the storage capacity and ground state energies to the VC dimension and the Rademacher complexity In practice, the previous Bayes-optimal approach is harshly criticized for its lack of fairness: the analysis requires a strong prior knowledge with the access to the prior distributions involved in the ground truth generative process. An alternative approach from the statistical learning theory literature consists instead in evaluating the *worst-case* performances of statistical inference, by quantifying generalization error upper-bounds. This is classically done with the VC dimension and the Rademacher complexity (Vapnik et al., 1994; Bartlett et al., 2002). Alternatively, the physics approach deeply focused on the *Gardner capacity* (Gardner et al., 1988; Krauth et al., 1989; Engel et al., 1993) that provides essentially a lower-bound of its twin from statistical learning theory, the VC dimension. The Gardner storage capacity is essentially the maximum number of examples that a model, namely the perceptron, is able to *memorize*. Indeed, under a randomly quenched disorder, input vectors $\mathbf{x}$ and output labels y are uncorrelated. Therefore, in a rCSP language the storage capacity is equivalently the maximum number of random input-output constraints the model parameters $\mathbf{w}$ can satisfy simultaneously. As a consequence, above this critical SAT-UNSAT threshold, the perceptron can no longer satisfy all the random constraints without making a prediction error.

In Chap. 6, we present the Gardner-like computation of the storage capacity for the binary perceptron with various activation functions. We show that, unlike for the usual step-function-binary-perceptron (Gardner et al., 1988), the critical capacity in simple symmetric variants is rigorously given by the annealed computation. Moreover by studying the structure of the configuration space, we unveil a frozen 1-step Replica Symmetry Breaking (f1RSB) structure using simple first and second moment methods.

By definition, above the SAT-UNSAT threshold the best configuration of the model parameters cannot satisfy all the random constraints and inevitably makes classification errors. Counting this minimal number of mistakes for a given constraint density is equivalent of computing the ground state energy of the system: below the storage capacity the energy vanishes whereas it becomes strictly positive above it. This rCSP approach can be naturally extended above the SAT-UNSAT transition to compute the corresponding ground state energies within the same framework.

In Chap. 7, we reveal the deep connection between the Rademacher worst-case generalization bound, which measures if a function can fit random noise, and the ground state energies from statistical physics. Finally, while statistical learning theory computes the generalization bounds up to asymptotic scalings, we are able to explicitly compute the Rademacher complexity for the spherical and binary perceptrons.

Empirical risk minimization in Generalized Linear Models for synthetic i.i.d data The two previous approaches provide the optimal and worst-case predictions that define the operating range of any statistical estimator. As a consequence, it turns out that *practical* machine learning estimators and algorithms are not described correctly neither by the pessimistic worst case analysis nor the idealistic Bayes-optimal.

In Chap. 8, we present how to analyze rigorously the behavior of practical ERM for regularized linear models, such as ridge, logistic or hinge regression. We focus on a common supervised classification task of a synthetic dataset, whose labels are generated by feeding a one-layer neural network with random i.i.d inputs. In this convex optimization task, the replica computation, under the RS ansatz, turns out to be correct and matches exactly the results of the Gordon convex Gaussian min-max theorem. After observing that, unlike ridge regression, logistic and hinge regressions surprisingly approach closely the Bayes-optimal generalization error, we design an optimal loss and regularizer that provably lead to Bayes-optimal generalization error performances.

As a conclusion, we summarize and reconcile the different approaches in a global picture. We conclude that, unlike the generalization error bounds, the Bayes-optimal analysis, even though its requires strong prior knowledges, captures the good scaling behaviors of the practical algorithms.

5 THE COMMITTEE MACHINE: COMPUTATIONAL TO STATISTICAL GAPS IN LEARNING A TWO-LAYERS NEURAL NETWORK

While the traditional approach to learning and generalization follows the VC (Vapnik, 1998) and Rademacher (Bartlett et al., 2002) worst-case type bounds, there has been a considerable body of theoretical work on calculating the generalization ability of neural networks for data arising from a probabilistic model within the framework of statistical mechanics (Seung et al., 1992; Watkin et al., 1993; Monasson et al., 1995a; Monasson et al., 1995b; Engel et al., 2001). In the wake of the need to understand the effectiveness of neural networks and also the limitations of the classical approaches (Zhang et al., 2016), it is of interest to revisit the results that have emerged thanks to the physics perspective. This direction is currently experiencing a strong revival, see e.g. (Chaudhari et al., 2017; Martin et al., 2017; Barbier et al., 2019b; Baity-Jest et al., 2018).

Of particular interest is the so-called T-S approach, where labels are generated by feeding i.i.d random samples to a neural network architecture (the *teacher*) and are then presented to another neural network (the *student*) that is trained using these data. Early studies computed the information theoretic limitations of the supervised learning abilities of the teacher weights by a student who is given n independent d-dimensional examples with $\alpha \equiv n/d = \Theta(1)$ and $d \to \infty$ (Seung et al., 1992; Watkin et al., 1993; Engel et al., 2001). These works relied on non-rigorous heuristic approaches, such as the replica and cavity methods (Mézard et al., 1987; Mézard et al., 2009). Additionally no provably efficient algorithm was provided to achieve the predicted learning abilities, and it was thus difficult to test those predictions, or to assess the computational difficulty.

Recent developments in statistical estimation and information theory —in particular of AMP (Donoho et al., 2009; Rangan, 2011; Bayati et al., 2011b; Javanmard et al., 2013), and a rigorous proof of the replica formula for the optimal generalization error (Barbier et al., 2019b)— allowed to settle these two missing points for single-layer neural networks (i.e. without any hidden variables). In the present chapter, we leverage on these works, and provide rigorous asymptotic predictions and corresponding message passing algorithm for a class of two-layers networks.

5.1 MAIN CONTRIBUTIONS AND RELATED WORKS

While our results hold for a rather large class of non-linear activation functions, we illustrate our findings on a case considered most commonly in the early literature: the *committee machine*. This is possibly the simplest version of a two-layers neural network where all the weights in the second layer are fixed to unity, and we illustrate it in Fig. 31. Denoting $\forall \mu \in [\![n]\!]$, y_μ the label associated with a d-dimensional sample $\mathbf{x}_\mu$, and $w^\star_{il}$ the weight connecting the i-th coordinate of the input to the k-th node of the hidden layer, it is defined by:

$$y_\mu = \text{sign}\left[\sum_{k=1}^{K} \text{sign}\left(\sum_{i=1}^{d} x_{\mu i} w^\star_{ik}\right)\right] = \text{sign}\left[\sum_{k=1}^{K} \text{sign}\left(\mathbf{x}_\mu^\intercal \mathbf{W}^\star\right)\right] , \quad (100)$$

where $\mathbf{W}^\star \in \mathbb{R}^{d\times K}$. We concentrate here on the T-S scenario: The teacher generates i.i.d data samples with i.i.d standard Gaussian coordinates $x_{\mu i} \sim \mathcal{N}(0,1)$, then she/he generates the associated labels y_μ using a committee machine as in (100), with i.i.d weights $w^\star_{il}$ unknown to the student. In the proof though, we will consider the more general case of a distribution for the weights of the form $\prod_{i=1}^n \mathrm{P}_\mathrm{w}(\{w^\star_{il}\}_{k=1}^K)$, but in practice we consider the fully separable case. The student is then given the n input-output pairs $(\mathbf{x}_\mu, y_\mu)_{\mu=1}^n$ and knows the distribution P_w used to generate $w^\star_{il}$. The goal of the student is to learn the weights $w^\star_{il}$ from the available examples $(\mathbf{x}_\mu, y_\mu)_{\mu=1}^n$ in order to reach the smallest possible generalization error, i. e. to be able to predict the label the teacher would generate for a new sample not present in the training set.

There have been several studies of this model within the non-rigorous statistical physics approach in the limit where $\alpha \equiv n/d = \Theta(1)$, $K = \Theta(1)$ and $d \to \infty$ (Schwarze, 1993; Schwarze et al., 1992; Schwarze et al., 1993; Mato et al., 1992; Monasson et al., 1995b; Engel et al., 2001). A particularly interesting result in the T-S setting is the *specialization of hidden neurons* (see sec. 12.6 of (Engel et al., 2001), or (Saad et al., 1995b) in the context of online learning): For $\alpha < \alpha_\text{spec}$, where α_spec is a certain critical value of the sample complexity, the permutational symmetry between hidden neurons remains conserved even after an optimal learning, and the learned weights of each of the hidden neurons are identical. For $\alpha > \alpha_\text{spec}$, however, this symmetry gets broken as each of the hidden units correlates strongly with one of the hidden units of the teacher. Another remarkable result is the calculation of the optimal generalization error as a function of α.

Our first contribution consists in a proof of the replica formula conjectured in the statistical physics literature, using the adaptive interpolation method of (Barbier et al., 2018a; Barbier et al., 2019b), that allows to put several of these results on a rigorous basis. However, this proof uses a technical unproven assumption. Our second contribution is the design of an AMP-type of algorithm that is able to achieve the optimal generalization error in the above limit of large dimensions for a wide range of parameters. The study of AMP —that is widely believed to be optimal between all polynomial algorithms in the above setting (Donoho et al., 2013b; Zdeborová et al., 2016a; Deshpande et al., 2015; Bandeira et al., 2018)— unveils, in the case of the committee machine with a large number of hidden neurons $K \to \infty$ with $K = o(d)$, the existence a large *hard phase* in which learning is information-theoretically possible, leading to a good generalization error decaying asymptotically as $1.25K/\alpha$ (in the $\alpha = \Theta(K)$ regime), but where AMP fails and provides only a poor generalization that does not go to zero when increasing α. This strongly suggests that no efficient algorithm exists in this hard region and therefore there is a computational gap in learning such neural networks. In other problems where a hard phase was identified its study boosted the development of algorithms that are able to match the predicted thresholds and we anticipate this will translate to the present model.

We also want to comment on a related line of work that studies the loss-function landscape of neural networks. While a range of works show under various assumptions that spurious local minima are absent in neural networks, others show under different conditions that they do exist, see e.g. (Safran et al., 2018). The regime of parameters that is hard for AMP must have spurious local minima, but the converse is not true in general. It might be that there are spurious local minima, yet the AMP approach succeeds. Moreover, in all previously studied models in the Bayes-optimal setting the generalization error obtained with the AMP is the best known and other approaches, e. g. noisy gradient-based, spectral algorithms or semidefinite programming, are not better in generalizing even in cases where the *student* models are over-parametrized. Of course in order to be in the Bayes-optimal setting one needs to know the model used by the teacher which is not the case in practice.

5.2 TECHNICAL RESULTS

5.2.1 A GENERAL MODEL

While in the illustration of our results we shall focus on the model (100), all our formulas are valid for a broader class of models: Given n input samples $\mathbf{X} = \{x_{\mu i}\}_{\mu,i=1}^{n,d}$, we denote $\mathbf{W}^\star = \{w_{ik}^\star\}_{i=1..d}^{k=1..K}$ the teacher-weight connecting for all $(i,k) \in [\![d]\!] \times [\![K]\!]$, the i-th input, i. e. the visible unit, to the k-th node of the hidden layer, represented in Fig. 31. For a generic function $\varphi_{\text{out}} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ one can formally write the output as

$$y_\mu = \varphi_{\text{out}}\Big(\Big\{\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_{\mu i} w_{ik}^\star\Big\}_{k=1}^{K}, a_\mu\Big) \ \text{ or } \ y_\mu \sim \mathrm{P}_{\text{out}}\Big(\cdot\Big|\Big\{\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_{\mu i} w_{ik}^\star\Big\}_{k=1}^{K}\Big),$$

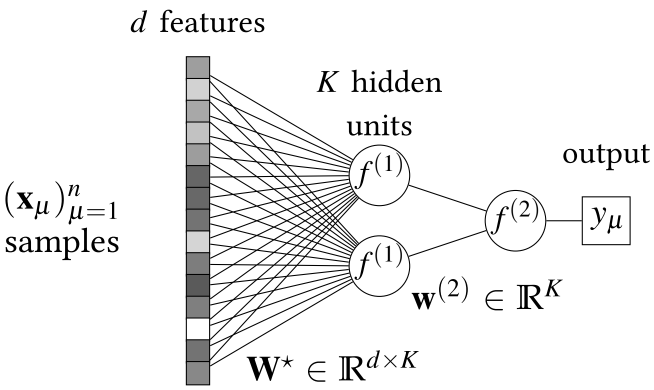

d features
K hidden units
output
$(\mathbf{x}_\mu)_{\mu=1}^n$ samples
$f^{(1)}$
$f^{(1)}$
$f^{(2)}$
y_μ
$\mathbf{w}^{(2)} \in \mathbb{R}^K$
$\mathbf{W}^\star \in \mathbb{R}^{d\times K}$

Figure 31: Illustration of the *committee machine*: it is one of the simplest models belonging to the considered model class (101), and on which we focus to illustrate our results. It is a two-layers neural network with sign activation functions $f^{(1)}, f^{(2)} = \text{sign}$ and weights $\mathbf{w}^{(2)}$ fixed to unity. It is represented for $K = 2$.

(101)

where $(a_\mu)_{\mu=1}^n$ are i.i.d real valued random variables with known distribution P_a, that form the probabilistic part of the model, generally accounting for noise. For deterministic models the second argument is simply absent and the distribution P_{out} is a Dirac mass. We can view alternatively (101) as a channel where the transition kernel P_{out} is directly related to φ_{out}. As discussed above, we focus on the T-S scenario where the teacher generates Gaussian i.i.d data $x_{\mu i} \sim \mathcal{N}(0,1)$, and i.i.d weights $w_{ik}^\star \sim \mathrm{P}_{\mathrm{w}}$. The student then learns $\mathbf{W}^\star \in \mathbb{R}^{d\times K}$ from the data $(\mathbf{x}_\mu, y_\mu)_{\mu=1}^n$ by computing marginal means of the posterior probability distribution (104).

Different scenarii fit into this general framework. Among those, the committee machine is obtained when choosing $\varphi_{\text{out}}(h) = \text{sign}(\sum_{k=1}^K \text{sign}(h_k))$ while another model considered previously is given by the parity machine, when $\varphi_{\text{out}}(h) = \prod_{k=1}^K \text{sign}(h_k)$, see e.g. (Engel et al., 2001) and Sec. 5.3.2 for the numerical results in the case $K = 2$. A number of layers beyond two has also been considered, see (Mato et al., 1992). Other activation functions can be used, and many more problems can be described, e.g. compressed pooling (El Alaoui et al., 2016; El Alaoui et al., 2017) or multi-vector compressed sensing (Zhu et al., 2017b).

5.2.2 TWO AUXILIARY INFERENCE PROBLEMS

Denote $\mathcal{S}_K$ the finite dimensional vector space of $K \times K$ matrices, $\mathcal{S}_K^+$ the convex set of semi-definite positive $K \times K$ matrices, $\mathcal{S}_K^{++}$ for positive definite $K \times K$ matrices, and $\forall \mathbf{N} \in \mathcal{S}_K^+$ we set $\mathcal{S}_K^+(\mathbf{N}) \equiv \{\mathbf{M} \in \mathcal{S}_K^+ \text{ s.t. } \mathbf{N} - \mathbf{M} \in \mathcal{S}_K^+\}$. Note that $\mathcal{S}_K^+(\mathbf{N})$ is convex and compact. Exceptionally in this section, parameters denoted with lowercase letters such as $\mathbf{q}, \hat{\mathbf{q}}, \boldsymbol{\rho}^\star$ represent matrices of size $K \times K$. Stating our results requires introducing two simpler auxiliary

K-dimensional estimation problems:

- The first one consists in retrieving a K-dimensional input vector $\mathbf{w} \sim \mathrm{P}_{\mathrm{w}}$ from the output of a Gaussian vector channel with K-dimensional observations

$$\mathbf{y}_0 = \hat{\mathbf{q}}^{1/2}\mathbf{w} + \mathbf{z}_0 \,,$$

$\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ and the *channel gain* matrix $\hat{\mathbf{q}} \in \mathcal{S}_K^+ \subseteq \mathbb{R}^{K\times K}$. The posterior distribution on $\mathbf{w} = (w_k)_{k=1}^K \in \mathbb{R}^K$ is

$$\mathrm{P}(\mathbf{w}|\mathbf{y}_0) = \frac{1}{\mathcal{Z}_{\mathrm{w}}}\, \mathrm{P}_{\mathrm{w}}(\mathbf{w})\, \exp\left(\mathbf{y}_0^\intercal \hat{\mathbf{q}}^{1/2}\mathbf{w} - \frac{1}{2}\mathbf{w}^\intercal \hat{\mathbf{q}} \mathbf{w}\right), \tag{102}$$

and the associated *free entropy* (or minus *free energy*) is given by the expectation over $\mathbf{y}_0$ of the log-partition function $\Psi_{\mathrm{w}}(\hat{\mathbf{q}}) \equiv \mathbb{E} \log \mathcal{Z}_{\mathrm{w}}$ and involves K dimensional integrals.

- The second problem considers K-dimensional i.i.d vectors $\mathbf{v}, \mathbf{u}^\star \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ where $\mathbf{v}$ is considered to be known and one has to retrieve $\mathbf{u}^\star$ from a scalar observation obtained as

$$\tilde{y}_0 \sim \mathrm{P}_{\mathrm{out}}(\,\cdot\,|\mathbf{q}^{1/2}\mathbf{v} + (\boldsymbol{\rho}^\star - \mathbf{q})^{1/2}\mathbf{u}^\star)$$

where the second moment matrix $\boldsymbol{\rho}^\star \equiv \mathbb{E}[\mathbf{w}\mathbf{w}^\intercal] \in \mathcal{S}_K^+$, where $\mathbf{w} \in \mathbb{R}^K \sim \mathrm{P}_{\mathrm{w}}$, and the so-called *overlap matrix* $\mathbf{q}$ is in $S_K^+(\boldsymbol{\rho}^\star)$. The associated posterior is

$$\mathrm{P}(\mathbf{u}|\tilde{y}_0, \mathbf{v}) = \frac{1}{\mathcal{Z}_{\mathrm{out}}} \frac{e^{-\frac{1}{2}\mathbf{u}^\intercal\mathbf{u}}}{(2\pi)^{K/2}}\, \mathrm{P}_{\mathrm{out}}\left(\tilde{y}_0 | \mathbf{q}^{1/2}\mathbf{v} + (\boldsymbol{\rho}^\star - \mathbf{q})^{1/2}\mathbf{u}\right), \tag{103}$$

and the free entropy reads this time $\Psi_{\mathrm{out}}(\mathbf{q}; \boldsymbol{\rho}^\star) \equiv \mathbb{E} \ln \mathcal{Z}_{\mathrm{out}}$, with the expectation over $\tilde{y}_0$ and $\mathbf{v}$, and also involves K dimensional integrals.

5.2.3 THE FREE ENTROPY

The central object of study leading to the optimal learning and generalization errors in the present setting is the posterior distribution of the weights:

$$\begin{aligned}\mathrm{P}(\{\mathbf{w}_k\}_{k=1}^K \mid \{\mathbf{x}, y_\mu\}_\mu^n) = \frac{1}{\mathcal{Z}_d} \prod_{i=1}^d \mathrm{P}_{\mathrm{w}}(\{w_{ik}\}_{k=1}^K) \\ \times \prod_{\mu=1}^n \mathrm{P}_{\mathrm{out}}\left(y_\mu \middle| \left\{\frac{1}{\sqrt{d}} \sum_{i=1}^d x_{\mu i} w_{ik}\right\}_{k=1}^K\right),\end{aligned} \tag{104}$$

where the normalization factor is nothing else than a *partition function*, i. e. the integral of the numerator over $\{w_{ik}\}_{i,l=1}^{d,K}$. The symbol $\mathbb{E}$ will generally denote an expectation over all random variables in the ensuing expression (here

$\{\mathbf{X},\mathbf{y}\}$). Subscripts will be used only when we take partial expectations or if there is an ambiguity. The expected free entropy is by definition

$$\Phi_d \equiv \frac{1}{d}\mathbb{E}\ln \mathcal{Z}_d\,. \tag{105}$$

The replica formula gives an explicit (conjectural) expression of Φ_d in the high-dimensional limit $d,n \to \infty$ with $\alpha = n/d$ fixed. We show in Appendix. B.1.1 how the heuristic replica method (Mézard et al., 1987; Mézard et al., 2009) yields the formula. This computation was first performed, to the best of our knowledge, by (Schwarze, 1993) in the case of the committee machine. Our first contribution is a rigorous proof of the corresponding free entropy formula using an interpolation method (Guerra, 2003; Talagrand, 2003; Barbier et al., 2018a), under a technical assumption, see Sec. 5.3 of (Aubin et al., 2018b).

In order to formulate our results, we add an arbitrarily small Gaussian regularization noise $z_\mu\sqrt{\Delta}$ to the first expression of the model (101), where $\Delta > 0$, $z_\mu \sim \mathcal{N}(0,1)$, which thus becomes

$$y_\mu = \varphi_{\text{out}}\left(\left\{\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_{\mu i}w^\star_{ik}\right\}_{k=1}^{K}, a_\mu\right) + z_\mu\sqrt{\Delta}\,, \tag{106}$$

so that the channel kernel is for $\mathbf{u} \in \mathbb{R}^K$,

$$\mathrm{P}_{\text{out}}(y|\mathbf{u}) = \frac{1}{\sqrt{2\pi\Delta}}\int_{\mathbb{R}} \mathrm{dP}_a(a) e^{-\frac{1}{2\Delta}(y-\varphi_{\text{out}}(\mathbf{u},a))^2}\,. \tag{107}$$

Let us define the RS *potential* as

$$\Phi^{(\text{rs})}(\mathbf{q},\hat{\mathbf{q}};\boldsymbol{\rho}^\star) \equiv -\frac{1}{2}\mathrm{Tr}\left(\hat{\mathbf{q}}\mathbf{q}\right) + \Psi_{\text{w}}(\hat{\mathbf{q}}) + \alpha\Psi_{\text{out}}(\mathbf{q};\boldsymbol{\rho}^\star)\,, \tag{108}$$

where $\alpha \equiv n/d$, and $\Psi_{\text{out}}(\mathbf{q};\boldsymbol{\rho}^\star)$ and $\Psi_{\text{w}}(\hat{\mathbf{q}})$ are the free entropies of the two simpler K-dimensional estimation problems (102) and (103). Notice that the expression is obtained from the replica computation in (321).

All along this chapter, we assume the following hypotheses for our rigorous statements:

1. The prior P_{w} has bounded support in $\mathbb{R}^K$.
2. The activation $\varphi_{\text{out}} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ is a bounded $\mathcal{C}^2$ function with bounded first and second derivatives with respect to its first argument, in $\mathbb{R}^K$-space.
3. For all $\mu \in [\![n]\!]$ and $i \in [\![d]\!]$ we have i.i.d $x_{\mu i} \sim \mathcal{N}(0,1)$.

We finally rely on a technical hypothesis, stated as Assumption 1 in Sec. 5.3 of (Aubin et al., 2018b).

Theorem 5.2.1 (Replica formula)**.** *Suppose 1, 2 and 3, and Assumption 1. Then for the model* (106) *with kernel* (107) *the limit of the free entropy is:*

$$\Phi_{\rm rs} \equiv \lim_{d\to\infty} \Phi_d \equiv \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\log \mathcal{Z}_d = \sup_{\hat{\mathbf{q}}\in\mathcal{S}_K^+} \inf_{\mathbf{q}\in\mathcal{S}_K^+(\rho)} \Phi^{(\rm rs)}(\mathbf{q},\hat{\mathbf{q}};\boldsymbol{\rho}^\star)\,. \tag{109}$$

This theorem extends the recent progress for generalized linear models of (Barbier et al., 2019b), which includes the case $K = 1$ of the present contribution, to the phenomenologically richer case of two-layers problems such as the committee machine. The proof sketch based on an *adaptive interpolation method* recently developed in (Barbier et al., 2018a) is outlined in Sec. 5 of (Aubin et al., 2018b) and the details can be found in the corresponding Sec. A.

Remark 5.2.2 (Relaxing the hypotheses)**.** *Note that, following similar approximation arguments as in (Barbier et al., 2019b), the hypothesis 1 can be relaxed to the existence of the second moment of the prior; thus covering the Gaussian case, 2 can be dropped and thus include model* (100) *and its* sign$(\cdot)$ *activation and 3 extended to data matrices* $\mathbf{X}$ *with i.i.d entries of zero mean, unit variance and finite third moment. Moreover, the case* $\Delta = 0$ *can be considered when the outputs are discrete, as in the committee machine* (100)*, see (Barbier et al., 2019b). The channel kernel becomes in this case* $P_{\rm out}(y|\mathbf{u}) = \int dP_a(a)\mathbb{1}\left[y - \varphi_{\rm out}(\mathbf{u},a)\right]$ *and the replica formula is the limit* $\Delta \to 0$ *of the one provided in Theorem 5.2.1. In general this regularizing noise is needed for the free entropy limit to exist.*

5.2.4 LEARNING THE TEACHER WEIGHTS AND OPTIMAL GENERALIZATION ERROR

A classical result in Bayesian estimation is that the estimator $\hat{\mathbf{W}}$ that minimizes the mean-square error with the ground-truth $\mathbf{W}^\star$ is given by the expected mean of the posterior distribution. Denoting $\mathbf{q}^\star$ the extremizer in the replica formula (109), we expect from the replica method that in the limit $d \to \infty$, $n/d = \alpha$, and with high probability, $\hat{\mathbf{W}}^\intercal\mathbf{W}^\star/d \to \mathbf{q}^\star$. We refer to proposition 5.3 and to the proof in Sec. A of (Aubin et al., 2018b) for the precise statement, that remains rigorously valid *only* in the presence of an additional (possibly infinitesimal) side-information. This condition is similar to the *small magnetic field* used to select a given Gibbs state in the Ising model in statistical physics. From the overlap matrix $\mathbf{q}^\star$, one can compute the Bayes-optimal generalization error when the student tries to classify a new, yet unseen, sample $\mathbf{x}_{\rm new} \in \mathbb{R}^{1\times d}$. The estimator of the new label $\hat{y}_{\rm new}$ that minimizes the mean-square error with the true label is given by computing the posterior mean of $\varphi_{\rm out}(\mathbf{x}_{\rm new}\mathbf{W})$ ($\mathbf{x}_{\rm new}$ is a row vector). Given the new sample, the optimal generalization error is then

$$\frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\left(\mathbb{E}_{\mathbf{w}|\mathbf{X},\mathbf{y}}\left[\varphi_{\rm out}(\mathbf{x}_{\rm new}\mathbf{W})\right] - \varphi_{\rm out}(\mathbf{x}_{\rm new}\mathbf{W}^\star)\right)^2\right] \xrightarrow[d\to\infty]{} \varepsilon_g(\mathbf{q}^\star), \tag{110}$$

where $\mathbf{W}$ is distributed according to the posterior measure (104). Note that this Bayes-optimal computation differs from the so-called Gibbs estimator by a factor 2. Indeed, one can naturally define the *Gibbs generalization error* as:

$$\varepsilon_g^{\text{gibbs}} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{W}^\star,\mathbf{X}}\left\langle \left[\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right) - \varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\right]^2 \right\rangle, \tag{111}$$

and define the *Bayes-optimal generalization error* as:

$$\varepsilon_g^{\text{bayes}} \equiv \frac{1}{2}\mathbb{E}_{\mathbf{W}^\star,\mathbf{X}}\left[\left(\left\langle\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)|\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)\right\rangle - \varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\right)^2\right]. \tag{112}$$

Using the Nishimori identity A.3.1, one obtains:

$$\begin{aligned}\varepsilon_g^{\text{bayes}} &= \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)^2\right] + \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\left\langle\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)|\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)\right\rangle^2\right] \\ &\quad - \mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left\langle\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)|\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)\right\rangle, \\ &= \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)^2\right] \\ &\quad - \frac{1}{2}\mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left\langle\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)|\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)\right\rangle .\end{aligned}$$

Using again the Nishimori identity one can write:

$$\begin{aligned}\varepsilon_g^{\text{gibbs}} &= \mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left[\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)^2\right] \\ &\quad - \mathbb{E}_{\mathbf{X},\mathbf{W}^\star}\left\langle\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)|\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)\right\rangle,\end{aligned}$$

which shows that $\varepsilon_g^{\text{gibbs}} = 2\varepsilon_g^{\text{bayes}}$. Note finally that since the distribution of $\mathbf{X}$ is rotationally invariant, the quantity $\mathbb{E}_{\mathbf{X}}\left[\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}^\star\right)\varphi_{\text{out}}\left(\mathbf{x}\mathbf{W}\right)\right]$ only depends on the *overlap* $\mathbf{q} \equiv \mathbf{W}^\intercal\mathbf{W}^\star$. As the overlap is shown to concentrate under the Gibbs measure, and as we expect that the value it concentrates on is the optimum $\mathbf{q}^\star$ of the replica formula (such fact is proven, e. g. , for random linear estimation problems in (Barbier et al., 2017)), the generalization error can itself be evaluated as a function of $\mathbf{q}^\star$. Examples where it is done include (Opper et al., 1996a; Seung et al., 1992; Schwarze, 1993; Barbier et al., 2019b).

In particular, when the data $\mathbf{X}$ is drawn from the standard Gaussian distribution on $\mathbb{R}^{n\times d}$, and is thus rotationally invariant, it follows that this error only depends on $\mathbf{W}^\intercal\mathbf{W}^\star/d$, which converges to $\mathbf{q}^\star$. Then a direct algebraic computation gives a lengthy but explicit formula for $\varepsilon_g(\mathbf{q}^\star)$ presented below.

5.2.4.a THE GENERALIZATION ERROR AT $K = 2$

From the definition of the generalization error, one can directly give an explicit expression of this error in the $K = 2$ case. Recall our committee-symmetric assumption on the overlap matrix, which here reads

$$\mathbf{q} = \begin{pmatrix} q_d + \frac{q_a}{2} & \frac{q_a}{2} \\ \frac{q_a}{2} & q_d + \frac{q_a}{2} \end{pmatrix}.$$

For concision, we denote here $\text{sign}(x) = \sigma(x)$. One obtains from (112):

$$\begin{aligned}
&\frac{1}{2} - 2\varepsilon_g^{\text{bayes},K=2} \\
&\equiv \mathbb{E} \int \mathrm{D}\mathbf{x}\, \sigma\left(\sigma\left(\frac{1}{\sqrt{d}}\mathbf{x}\cdot\mathbf{w}_1^\star\right) + \sigma\left(\frac{1}{\sqrt{d}}\mathbf{x}\cdot\mathbf{w}_2^\star\right)\right) \\
&\qquad \times \sigma\left(\sigma\left(\frac{1}{\sqrt{d}}\mathbf{x}\cdot\mathbf{w}_1\right) + \sigma\left(\frac{1}{\sqrt{d}}\mathbf{x}\cdot\mathbf{w}_2\right)\right) \\
&= \mathbb{E}\frac{1}{(2\pi)^4}\int_{\mathbb{R}^4} \mathrm{d}\mathbf{x}\sigma\left(\sigma(x_1)+\sigma(x_2)\right)\sigma\left(\sigma(x_3)+\sigma(x_4)\right) \\
&\qquad \times \int_{\mathbb{R}^4} \mathrm{d}\hat{\mathbf{x}}\, e^{i\hat{\mathbf{x}}^\intercal\mathbf{x}} \int \mathrm{D}\mathbf{x}\, e^{-\frac{i}{\sqrt{d}}\hat{\mathbf{x}}^\intercal\tilde{\mathbf{W}}^\intercal\mathbf{x}} \qquad\qquad (113) \\
&= \mathbb{E}\frac{1}{(2\pi)^4}\int_{\mathbb{R}^4} \mathrm{d}\mathbf{x}\,\sigma\left(\sigma(x_1)+\sigma(x_2)\right)\sigma\left(\sigma(x_3)+\sigma(x_4)\right) \\
&\qquad \times \int_{\mathbb{R}^4} \mathrm{d}\hat{\mathbf{x}}\, e^{i\hat{\mathbf{x}}^\intercal\mathbf{x}} e^{-\frac{1}{2}\hat{\mathbf{x}}^\intercal\boldsymbol{\Sigma}\hat{\mathbf{x}}} \\
&= \int_{\mathbb{R}^4} \mathrm{d}\mathbf{x}\,\sigma\left(\sigma(x_1)+\sigma(x_2)\right)\sigma\left(\sigma(x_3)+\sigma(x_4)\right)\mathcal{N}_{\mathbf{x}}(\mathbf{0},\boldsymbol{\Sigma})
\end{aligned}$$

where $\tilde{\mathbf{W}} = (\mathbf{w}_1^\star, \mathbf{w}_2^\star, \mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^{d\times K}$ with $\boldsymbol{\Sigma} = \frac{1}{d}\mathbb{E}\tilde{\mathbf{W}}^\intercal\tilde{\mathbf{W}} \underset{d\to\infty}{\longrightarrow} \begin{bmatrix} \mathbf{I}_2 & \mathbf{q} \\ \mathbf{q} & \mathbf{I}_2 \end{bmatrix}$. This expression can be reformulated also as

$$\begin{aligned}
\frac{1}{2} - 2\varepsilon_g^{\text{bayes},K=2} &= \int_{\mathbb{R}^4} \mathrm{D}\mathbf{x}\,\sigma\left[\sigma(x_1)+\sigma(x_2)\right] \\
\times\sigma\Bigg\{ &\sigma\left[(\frac{q_a}{2}+q_d)x_1 + \frac{q_a}{2}x_2 + x_3\sqrt{1-\frac{q_a^2}{2}-q_aq_d-q_d^2}\right] \\
&+\sigma\left[\frac{q_a}{2}x_1 + (\frac{q_a}{2}+q_d)x_2 - x_3\frac{q_a(q_d+\frac{q_a}{2})}{\sqrt{1-\frac{q_a^2}{2}-q_aq_d-q_d^2}}\right. \\
&\left.\quad + x_4\sqrt{\frac{(1-q_d^2)(1-(q_a+q_d)^2)}{1-\frac{q_a^2}{2}-q_aq_d-q_d^2}}\right]\Bigg\}. \qquad (114)
\end{aligned}$$

Note that one could possibly simplify this expression by using an appropriate orthogonal transformation on $\mathbf{x}$. These integrals were then computed using MC methods to obtain the generalization error in the left and middle plots of Fig. 32.

5.2.4.B THE GENERALIZATION ERROR AT LARGE K

Recall the definition of the generalization error in (112), one can compute it using (111) after a tedious, yet straightforward, calculation:

$$\varepsilon_g^{\text{bayes}} = \frac{1}{2}\varepsilon_g^{\text{gibbs}} = \frac{1}{\pi}\arccos\left[\frac{2}{\pi}\left(q_a + \arcsin q_d\right)\right] + \Theta(K^{-1}). \qquad (115)$$

This expression is the one used in the computation of the generalization error in the left panel of Fig. 33.

5.2.5 APPROXIMATE MESSAGE PASSING AND ITS STATE EVOLUTION

Our next result is based on an adaptation of a popular algorithm to solve random instances of generalized linear models, the AMP algorithm (Donoho et al., 2009; Rangan, 2011), for the case of the committee machine and models described by (101). The AMP algorithm can be obtained as a Taylor expansion of loopy belief-propagation, see Appendix. C for the derivation, and also originates in earlier statistical physics works (Thouless et al., 1977; Mézard, 1989; Opper et al., 1996b; Kabashima, 2008; Baldassi et al., 2007; Zdeborová et al., 2016a). It is conjectured to perform the best among all polynomial algorithms in the framework of these models. It thus gives us a tool to evaluate both the intrinsic algorithmic hardness of the learning and the performance of existing algorithms with respect to the optimal one in this model.

```
Input: vector $\mathbf{y} \in \mathbb{R}^n$ and matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$:
Initialize: $\hat{\mathbf{w}}_i$, $\mathbf{f}_{\text{out},\mu} \in \mathbb{R}^K$ and $\hat{\mathbf{C}}_i, \hat{\mathbf{V}}_i$, $\partial_{\boldsymbol{\omega}}\mathbf{f}_{\text{out},\mu} \in \mathbb{R}^{K\times K}$ for $1 \leq i \leq d$ and $1 \leq \mu \leq n$ at $t = 0$.
repeat
  Channel: Update the mean $\omega_\mu \in \mathbb{R}^K$ and variance $V_\mu \in \mathbb{R}^{K\times K}$:
  $\mathbf{V}^t_\mu = \sum_{i=1}^{d} \frac{x^2_{\mu j}}{d} \hat{\mathbf{C}}^t_i$
  $\boldsymbol{\omega}^t_\mu = \sum_{i=1}^{d} \frac{x_{\mu i}}{\sqrt{d}} \hat{\mathbf{w}}^t_i - \mathbf{V}^t_\mu \mathbf{f}^{t-1}_{\text{out},\mu}$ ,
  Update $\mathbf{f}_{\text{out},\mu}$ and $\partial_{\boldsymbol{\omega}}\mathbf{f}_{\text{out},\mu}$:
  $\mathbf{f}^t_{\text{out},\mu} = \mathbf{f}_{\text{out}}\left(y_\mu, \boldsymbol{\omega}^t_\mu, \mathbf{V}^t_\mu\right)$ , $\partial_{\boldsymbol{\omega}}\mathbf{f}^t_{\text{out}\cdot\mu} = \partial_{\boldsymbol{\omega}}\mathbf{f}_{\text{out}}\left(y_\mu, \boldsymbol{\omega}^t_\mu, \mathbf{V}^t_\mu\right)$
  Prior: Update the mean $\boldsymbol{\gamma}_i \in \mathbb{R}^K$ and variance $\boldsymbol{\Lambda}_i \in \mathbb{R}^{K\times K}$:
  $\boldsymbol{\Lambda}^t_i = -\sum_{\mu=1}^{n} \frac{x^2_{\mu i}}{d} \partial_{\boldsymbol{\omega}}\mathbf{f}^t_{\text{out},\mu}$
  $\boldsymbol{\gamma}^t_i = \sum_{\mu=1}^{n} \frac{x_{\mu i}}{\sqrt{d}} \mathbf{f}^t_{\text{out},\mu} + \boldsymbol{\Lambda}^t_i \hat{\mathbf{w}}^t_i$ ,
  Update the estimated marginals $\hat{w}_i \in \mathbb{R}$ and $\hat{c}_i \in \mathbb{R}^+$:
  $\hat{\mathbf{w}}^{t+1}_i = \mathbf{f}_{\text{w}}\left(\boldsymbol{\gamma}^t_i, \boldsymbol{\Lambda}^t_i\right)$ , $\hat{\mathbf{C}}^{t+1}_i = \partial_{\boldsymbol{\gamma}}\mathbf{f}_{\text{w}}\left(\boldsymbol{\gamma}^t_i, \boldsymbol{\Lambda}^t_i\right)$
  $t \leftarrow t + 1$
until Convergence on $\hat{\mathbf{w}}_i$, $\hat{\mathbf{C}}_i$.
Output: $\{\hat{\mathbf{w}}\}_{i=1}^d$ and $\{\hat{\mathbf{C}}_i\}_{i=1}^d$.
```

Algorithme 2 : Approximate Message Passing for the committee machine.

The AMP algorithm is summarized by its pseudo-code in Algo. 2, where the update functions $\mathbf{f}_{\text{out}}$, $\partial_{\boldsymbol{\omega}}\mathbf{f}_{\text{out}}$, $\mathbf{f}_{\text{w}}$ and $\partial_{\boldsymbol{\gamma}}\mathbf{f}_{\text{w}}$ are related, again, to the two auxiliary problems (102) and (103) and defined in Appendix. A.4. The functions $\mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma}, \boldsymbol{\Lambda})$ and $\partial_{\boldsymbol{\gamma}} \mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma}, \boldsymbol{\Lambda})$ are respectively the mean and variance under the posterior distribution (102) when $\hat{\mathbf{q}} \to \boldsymbol{\Lambda}$ and $\mathbf{y} \to \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\gamma}$, while $\mathbf{f}_{\mathrm{out}}(y_\mu, \boldsymbol{\omega}_\mu, \mathbf{V}_\mu)$ is given by the product of $\mathbf{V}_\mu^{-1/2}$ and the mean of $\mathbf{u}$ under the posterior (103) using $\tilde{y}_0 \to y_\mu$, $\boldsymbol{\rho}^\star - \mathbf{q} \to \mathbf{V}_\mu$ and $\mathbf{q}^{1/2}\mathbf{v} \to \boldsymbol{\omega}_\mu$.

After convergence, $\hat{\mathbf{W}}$ estimates the weights of the teacher-neural network. The label of a sample $\mathbf{x}_{\mathrm{new}}$ not seen in the training set is estimated by the AMP algorithm as

$$y^t_{\mathrm{new}} = \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}\, y\, \mathrm{P}_{\mathrm{out}}(y|\mathbf{z}) \mathcal{N}_{\mathbf{z}}(\boldsymbol{\omega}^t_{\mathrm{new}}, \mathbf{V}^t_{\mathrm{new}}) \,, \tag{116}$$

where $\boldsymbol{\omega}^t_{\mathrm{new}} = \sum_{i=1}^d x_{\mathrm{new},i} \hat{\mathbf{w}}^t_i$ is the mean of the normally distributed variable $\mathbf{z} \in \mathbb{R}^K$, and $\mathbf{V}^t_{\mathrm{new}} = \boldsymbol{\rho}^\star - \mathbf{q}^t_{\mathrm{amp}}$ is the $K \times K$ covariance matrix (see below for the definition of $\mathbf{q}^t_{\mathrm{amp}}$). We provide a demonstration code of the algorithm on GitHub (Aubin et al., 2018a).

AMP is particularly interesting because its performance can be tracked rigorously, again in the asymptotic limit when $d \to \infty$, via a procedure known as SE, which is a rigorous version of the cavity method in physics (Mézard et al., 2009), see (Javanmard et al., 2013). SE tracks the value of the overlap between the hidden ground truth $\mathbf{W}^*$ and the AMP estimate $\hat{\mathbf{W}}^t$, defined as $\mathbf{q}^t_{\mathrm{amp}} \equiv \lim_{d\to\infty} (\hat{\mathbf{W}}^t)^\intercal \mathbf{W}^\star / d$, via the iteration of the following equations:

$$\mathbf{q}^{t+1}_{\mathrm{amp}} = 2\nabla\Psi_{\mathrm{w}}(\hat{\mathbf{q}}^t_{\mathrm{amp}}) \,, \qquad \hat{\mathbf{q}}^{t+1}_{\mathrm{amp}} = 2\alpha\nabla\Psi_{\mathrm{out}}(\mathbf{q}^t_{\mathrm{amp}}; \boldsymbol{\rho}^\star) \,. \tag{117}$$

See sec. G of (Aubin et al., 2018b) for more details and note that the fixed points of these equations correspond surprisingly to the critical points of the replica free entropy (109). Let us comment further on the convergence of the algorithm. In the large d limit, and if the integrals are performed without errors, then the algorithm is guaranteed to converge. This is a consequence of the SE combined with the Bayes-optimal setting. In practice, of course, d is finite and integrals are approximated. In that case convergence is not guaranteed, but is robustly achieved in all the cases presented in this paper. We also expect, by experience with the single layer case, that if the input-data matrix is not random, which is beyond our assumptions, then we will encounter convergence issues, which could be fixed by moving to some variant of the algorithm such as VAMP (Rangan et al., 2019a).

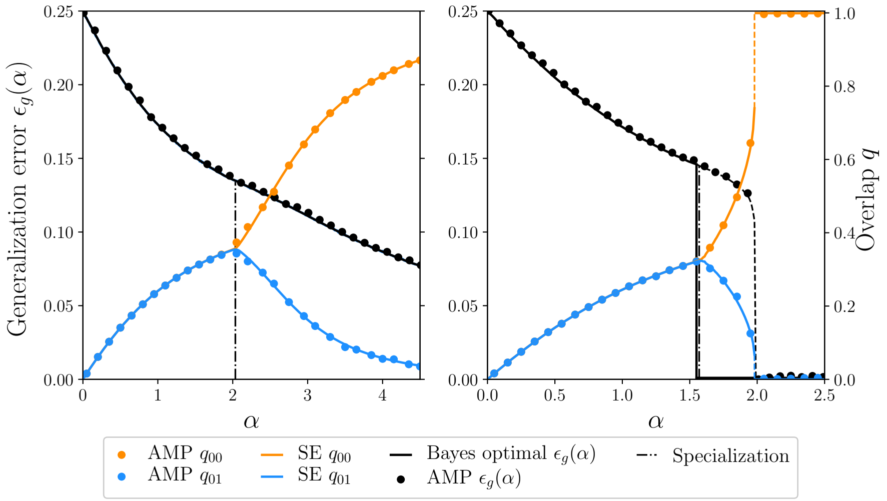


Figure 32: Generalization error and order parameter for a committee machine with two hidden neurons ($K = 2$) with (**Left**) Gaussian weights, (**Right**) binary/Rademacher weights. These are shown as a function of the ratio $\alpha = n/d$ between the number of samples n and the dimensionality d. Lines are obtained from the state evolution (SE) equations (dominating solution is shown in full line), data-points from the AMP algorithm averaged over 10 instances of the problem of size $d = 10^4$. q_{00} and q_{01} denote diagonal and off-diagonal overlaps of the matrices $\mathbf{q}^\star$ and $\mathbf{q}_{\text{amp}}$, and their values are given by the labels on the far-right of the figure.

5.3 FROM TWO TO MORE HIDDEN NEURONS AND THE SPECIALIZATION PHASE TRANSITION

5.3.1 TWO NEURONS COMMITTEE MACHINE $K = 2$

Let us now discuss how the above results can be used to study the optimal learning in the simplest non-trivial case of a two-layers neural network with two hidden neurons, that is when model (100) is simply

$$y_\mu = \text{sign}\Big[\text{sign}\Big(\sum_{i=1}^{d} x_{\mu i} w_{i1}^\star\Big) + \text{sign}\Big(\sum_{i=1}^{d} x_{\mu i} w_{i2}^\star\Big)\Big],$$

and is represented in Fig. 31, with the convention that $\text{sign}(0) = 0$. We remind that the input-data matrix $\mathbf{X}$ has i.i.d $\mathcal{N}(0,1)$ entries, and the teacher-weights $\mathbf{W}^\star$ used to generate the labels $\mathbf{y}$ are taken i.i.d from P_w. In Fig. 32 we plot the optimal generalization error as a function of the sample complexity $\alpha = n/d$. In the left panel the weights are Gaussian (for both the teacher and the student), while in the right panel they are binary/Rademacher. The full line is obtained from the fixed point of the SE of the AMP algorithm (117), corresponding to the extremizer of the replica free entropy (109). The points are results of the AMP algorithm run till convergence averaged over

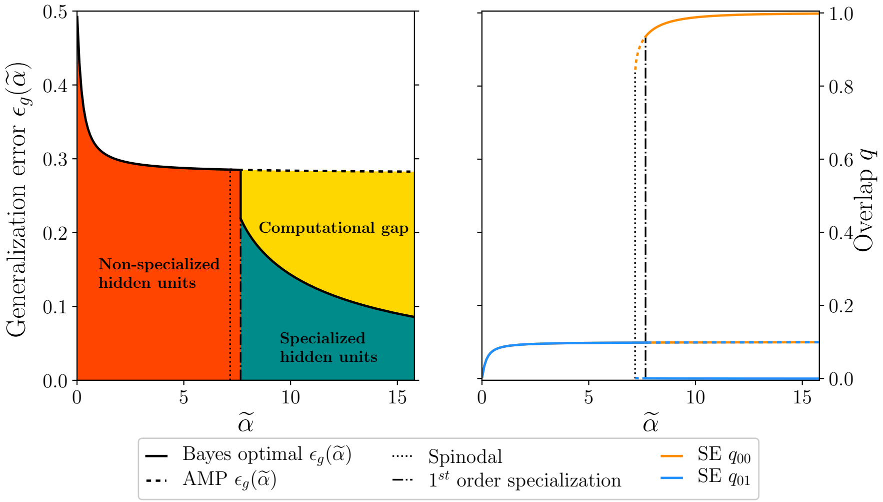


Figure 33: (**Left**) Bayes optimal and AMP generalization errors and (**Right**) diagonal and off-diagonal overlaps q_{00}, q_{01} for a committee machine with a large number of hidden neurons K and Gaussian weights, as a function of the rescaled parameter $\tilde{\alpha} = \alpha/K$. Solutions corresponding to global and local minima of the replica free entropy are respectively represented with full and dashed lines. The dotted line marks the spinodal at $\widetilde{\alpha}^G_{\text{spinodal}} \simeq 7.17$, i. e. the apparition of a local minimum in the replica free entropy, associated to a solution with specialized hidden units. The dotted-dashed line shows the first order specialization transition at $\widetilde{\alpha}^G_{\text{spec}} \simeq 7.65$, at which the specialized fixed point becomes the global minimum. For $\widetilde{\alpha} < \widetilde{\alpha}^G_{\text{spec}}$, AMP reaches the Bayes optimal generalization error and overlaps, corresponding to a non-specialized solution (red area). However, for $\widetilde{\alpha} > \widetilde{\alpha}^G_{\text{spec}}$, the AMP algorithm does not follow the optimal specialized solution (green area) and is stuck in the non-specialized solution plateau, represented with dashed lines. Hence it unveils a large computational gap (yellow area).

10 instances of size $d = 10^4$. In this case and with random initial conditions the AMP algorithm did converge in all our trials. As expected we observe excellent agreement between the SE and AMP.

In both left and right panels of Fig. 32 we observe the so-called *specialization* phase transition. Indeed (117) has two types of fixed points: a *n*on-specialized fixed point where every matrix element of the $K \times K$ order parameter $\mathbf{q}$ is the same (so that both hidden neurons learn the same function) and a *specialized* fixed point where the diagonal elements of the order parameter are different from the non-diagonal ones. We checked for other types of fixed points for $K = 2$ (one where the two diagonal elements are not the same), but have not found any. In terms of weight-learning, this means for the non-specialized fixed point that the estimators for both $\mathbf{w}_1$ and $\mathbf{w}_2$, with $\hat{\mathbf{W}} = (\mathbf{w}_1, \mathbf{w}_2)$ are the same, whereas in the specialized fixed point the estimators of the weights corresponding to the two hidden neurons are different, and that the network "figured out" that the data are better described by a model that is not linearly separable. The specialized fixed point is associated with lower error than the non-specialized one (as one can see in Fig. 32). The existence of this phase transition was discussed in statistical physics literature on the committee machine, see e.g. (Schwarze et al., 1992; Saad et al., 1995b).

For Gaussian weights (Fig. 32 left), the specialization phase transition arises continuously at $\alpha^G_{\rm spec}(K=2) \simeq 2.04$. This means that for $\alpha < \alpha^G_{\rm spec}(K=2)$ the number of samples is too small, and the student-neural network is not able to learn that two different teacher-vectors $\mathbf{w}_1^\star$ and $\mathbf{w}_2^\star$ were used to generate the observed labels. For $\alpha > \alpha^G_{\rm spec}(K=2)$, however, it is able to distinguish the two different weight-vectors and the generalization error decreases fast to low values (see Fig. 32). For completeness we remind that in the case of $K=1$ corresponding to single-layer neural network no such specialization transition exists. We show in sec. E of (Aubin et al., 2018b) that it is absent also in multi-layer neural networks as long as the activations remain linear. The non-linearity of the activation function is therefore an essential ingredient in order to observe a specialization phase transition.

The right part of Fig. 32 depicts the fixed point reached by the state evolution of AMP for the case of binary weights. We observe two phase transitions in the performance of AMP in this case: (a) the specialization phase transition at $\alpha^B_{\rm spec}(K=2) \simeq 1.58$, and for slightly larger sample complexity a transition towards *perfect generalization* (beyond which the generalization error is asymptotically zero) at $\alpha^B_{\rm perf}(K=2) \simeq 1.99$. The binary case with $K=2$ differs from the Gaussian one in the fact that perfect generalization is achievable at finite α. While the specialization transition is continuous here, the error has a discontinuity at the transition of perfect generalization. This discontinuity is associated with the 1st order phase transition, in the physics nomenclature, leading to a gap between algorithmic (AMP in our case) performance and information-theoretically optimal performance reachable by exponential algorithms. To quantify the optimal performance we need to evaluate the global extremum of the replica free entropy (not the local one reached by the state evolution). In doing so that we get that information theoretically there is a single discontinuous phase transition towards perfect generalization at $\alpha^B_{\rm IT}(K=2) \simeq 1.54$.

While the information-theoretic and specialization phase transitions were identified in the physics literature on the committee machine (Schwarze et al., 1992; Schwarze et al., 1993; Seung et al., 1992; Watkin et al., 1993), the gap between the information-theoretic performance and the performance of AMP —that is conjectured to be optimal among polynomial algorithms— was not yet discussed in the context of this model. Indeed, even its understanding in simpler models than those discussed here, such as the single layer case, is more recent (Donoho et al., 2009; Zdeborová et al., 2016a; Donoho et al., 2013b).

5.3.2 TWO NEURONS PARITY MACHINE $K=2$

Although we mainly focused on the committee machine, another classical two-layers neural network is the parity machine (Engel et al., 2001) and our proof applies to this case as well. While learning is known to be computationally

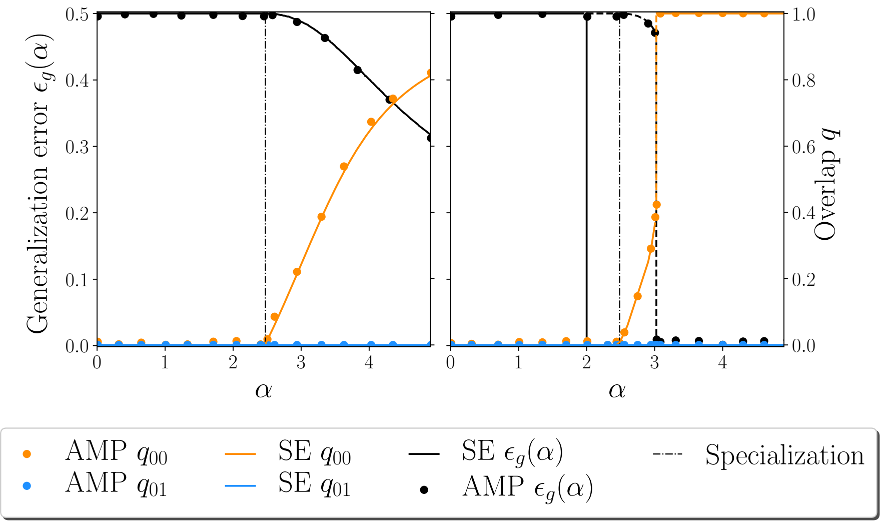


Figure 34: Similar plot as in Fig. 32 but for the parity machine with two hidden neurons. Value of the order parameter and the optimal generalization error for a parity machine with two hidden neurons with (**Left**) Gaussian weights and (**Right**) binary/Rademacher weights. SE and AMP overlaps are respectively represented in full line and points.

hard for general K, the case $K = 2$ is special, and in fact can be reformulated as a committee machine, where the sign activation function has been replaced by $f_2(z) = \mathbb{1}(z \neq 0) - \mathbb{1}(z = 0)$:

$$y_\mu = \text{sign}\left[\prod_{k=1}^{K} \text{sign}\left(\sum_{i=1}^{d} x_{\mu i} w^\star_{ik}\right)\right] = f_2\left[\sum_{k=1}^{K} \text{sign}\left(\sum_{i=1}^{d} x_{\mu i} w^\star_{ik}\right)\right]. \quad (118)$$

We have repeated our analysis for the $K = 2$ parity machine and the phase diagram is summarized in Fig. 34 where we show the generalization error and the elements of the overlap matrix for Gaussian (**Left**) and binary weights (**Right**), with the results of the AMP algorithm (points). Below the specialization phase transition $\alpha < \alpha_{\text{spec}}$, the symmetry of the output imposes the non-specialized fixed point $q_{00} = q_{01} = 0$ to be the only solution, with $\alpha^G_{\text{spec}}(K = 2) \simeq 2.48$ and $\alpha^B_{\text{spec}}(K = 2) \simeq 2.49$. Above the specialization transition α_{spec}, the overlap becomes specialized with a non-trivial diagonal term. Additionally, in the binary case, an information theoretical transition towards a perfect learning occurs at $\alpha^B_{\text{IT}}(K = 2) \simeq 2.00$, meaning that the perfect generalization fixed point ($q_{00} = 1, q_{01} = 0$) becomes the global optimizer of the free entropy. It leads to a first order phase transition of the AMP algorithm which retrieves the perfect generalization phase only at $\alpha^B_{\text{perf}}(K = 2) \simeq 3.03$. This is similar to what happens in single layer neural networks for the symmetric door activation function, see (Barbier et al., 2019b). Again, these results for the parity machine emphasize a gap between information-theoretical and computational performance.

5.3.3 MORE IS DIFFERENT $K \to \infty$

It becomes more difficult to study the replica formula for larger values of K as it involves (at least) K-dimensional integrals. Quite interestingly, it is possible to work out the solution of the replica formula in the large K limit (thus taken *a*fter the large d limit, so that K/d vanishes). It is indeed natural to look for solutions of the replica formula, as suggested in (Schwarze, 1993), of the form $\mathbf{q} = q_d \mathbf{I}_K + (q_a/K)\mathbf{1}_K \mathbf{1}_K^\intercal$, with the unit vector $\mathbf{1}_K = (1)_{l=1}^K$. Since both $\mathbf{q}$ and $\boldsymbol{\rho}^\star$ are assumed to be positive, this scaling implies that $0 \leq q_d \leq 1$ and $0 \leq q_a + q_d \leq 1$, as it should, see sec. D of (Aubin et al., 2018b). We also detail in this same section the corresponding large K expansion of the free entropy for the teacher-student scenario with Gaussian weights. Only the information-theoretically reachable generalization error was computed (Schwarze, 1993), thus we concentrated on the analysis of performance of AMP by tracking the SE equations. In doing so, we unveil a large computational gap.

In the right panel of Fig. 33 we show the fixed point values of the two overlaps $q_{00} = q_d + q_a/K$ and $q_{01} = q_a/K$ and the resulting generalization error, plotted in the left panel. As discussed in (Schwarze, 1993) it can be written in a closed form as $\epsilon_g = \arccos\left[2\left(q_a + \arcsin q_d\right)/\pi\right]/\pi$, represented in the left panel of Fig. 33. The specialization transition arises for $\alpha = \Theta(K)$ so we define $\widetilde{\alpha} \equiv \alpha/K$. The specialization is now a first order phase transition, meaning that the specialization fixed point first appears at $\widetilde{\alpha}^G_{\text{spinodal}} \simeq 7.17$ but the free entropy global extremizer remains the one of the non-specialized fixed point until $\widetilde{\alpha}^G_{\text{spec}} \simeq 7.65$. This has interesting implications for the optimal generalization error that gets towards a plateau of value $\epsilon_{\text{plateau}} \simeq 0.28$ for $\widetilde{\alpha} < \widetilde{\alpha}^G_{\text{spec}}$ and then jumps discontinuously down to reach a decay aymptotically as $1.25/\widetilde{\alpha}$. See left panel of Fig. 33.

AMP is conjectured to be optimal among all polynomial algorithms (in the considered limit) and thus analyzing its SE sheds light on possible computational-to-statistical gaps that come hand in hand with first order phase transitions. In the regime of $\alpha = \Theta(K)$ for large K the non-specialized fixed point is always stable implying that AMP will not be able to give a lower generalization error than $\epsilon_{\text{plateau}}$. Analyzing the replica formula for large K in more details, see sec. D of (Aubin et al., 2018b), we concluded that AMP will not reach the optimal generalization for any $\alpha < \Theta(K^2)$. This implies a rather sizable gap between the performance that can be reached information-theoretically and the one reachable tractably (see yellow area in Fig. 33). Such large computational gaps have been previously identified in a range of inference problems —most famously in the planted clique problem (Deshpande et al., 2015)— but the committee machine is the first model of a multi-layer neural network with realistic non-linearities (the parity machine is another example but use a very peculiar non-linearity) that presents such large gap.

CONCLUSION

In this chapter, we revisited a model for two-layer neural network known as the committee machine in the T-S scenario that allows for explicit evaluation of Bayes-optimal learning errors. This model has been solved in early statistical physics literature using the non-rigorous replica method. We built on recent progress in proving the replica formulas rigorous in the Bayes-optimal setting and extend these proof to the case of the committee machine.

One of our contributions is the design of an AMP-type algorithm that is able to achieve the Bayes-optimal learning error in the limit of large dimensions for a range of parameters out of the so-called hard phase. The hard phase is associated with first order phase transitions appearing in the solution of the model. In the case of the committee machine with a large number of hidden neurons we identify a large hard phase in which learning is possible information-theoretically but not efficiently. In other problems where such a hard phase was identified, its study boosted the development of algorithms that are able to match the predicted threshold. We anticipate this will also be the same for the present model. We should, however, note that for larger $K > 2$ the present AMP algorithm includes higher-dimensional integrals that hamper the speed of the algorithm. Our current strategy to tackle this is to combine the large-K expansion and use it in the algorithm. Detailed account of the corresponding results are left for future work.

We studied the Bayes-optimal setting where the student-network is the same as the teacher-network, for which the replica method can be readily applied. The method still applies when the number of hidden units in the student and teacher are different, while our proof does not generalize easily to this case. It is an interesting subject for future work to see how the hard phase evolves under over-parametrization and what is the interplay between the simplicity of the loss-landscape and the achievable generalization error. We conjecture that in the present model over-parametrization will not improve the generalization error achieved by AMP in the Bayes-optimal case.

Even though we focused on a two-layers neural network, the analysis and algorithm can be readily extended to a multi-layer setting, see (Mato et al., 1992), as long as the number of layers as well as the number of hidden neurons in each layer is held constant, and as long as one learns only weights of the first layer, for which the proof already applies. The numerical evaluation of the phase diagram would be more challenging than the cases presented in this paper as multiple integrals would appear in the corresponding formulas. In future works, we also plan to analyze the case where the weights of the second and subsequent layers (including the biases of the activation functions) are also learned. This could be done for instance with a combination of Expectation Maximization and AMP along the lines of (Krzakala et al., 2012b; Kamilov et al., 2012) where this is done for the simpler single layer case.

Concerning extensions of the present work, an important open case is the one where the number of samples per dimension $\alpha = \Theta(1)$ and also the size of the hidden layer per dimension $K/d = \Theta(1)$ as $d \to \infty$, while in this paper we treated the case $K = \Theta(1)$ and $d \to \infty$. This other scaling where $K/d = \Theta(1)$ is challenging even for the non-rigorous replica method.

6 STORAGE CAPACITY IN SYMMETRIC BINARY PERCEPTRONS

In this chapter, we revisit the problem of computing the capacity of the binary perceptron (Gardner et al., 1988; Krauth et al., 1989) for storing random patterns. This problem lies at the core of early statistical physics studies of neural networks and their learning and generalization properties, for reviews see e. g. (Watkin et al., 1993; Seung et al., 1992; Engel et al., 2001; Nishimori, 2001). While the perceptron problem is motivated by studies of simple artificial neural networks as discussed in detail in the above literature, in this paper we view it as a rCSP where the vector of binary weights $\mathbf{w} \in \{\pm 1\}^d$ (a *solution*) must satisfy n *step* constraints of the type

$$\sum_{i=1}^{d} x_{\mu i} w_i \geq K\,, \tag{119}$$

where $\mu \in [\![n]\!]$, $K \in \mathbb{R}$ is the *threshold*, the random variables $x_{\mu i}$ are i.i.d Gaussian variables with zero mean and variance $1/d$, and the rows of the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ are called patterns. We define an indicator function associated to the perceptron with a step constraint as $\varphi^s(z) = \mathbb{1}\left[z \geq K\right]$.

We say that a given vector $\mathbf{w}$ is a solution of the perceptron instance if all n constraints given by eq. (119) are satisfied. The *storage capacity* is then defined similarly to the satisfiability threshold in random constraint satisfaction problems: we denote the constraint density as $\alpha \equiv n/d$ and define the storage capacity $\alpha_c(K)$ as the infimum of densities α such that in the limit $d \to \infty$, with high probability over the choice of the matrix $\mathbf{X}$ there are no solutions. It is natural to conjecture that the converse also holds, i. e. the storage capacity $\alpha_c(K)$ equals the supremum of α such that in the limit $d \to \infty$ solutions exist with high probability. In this case we would say the storage capacity is a *sharp threshold* according to the definition:

$$\exists \varepsilon > 0 \,/\, \forall \alpha > \alpha_c + \varepsilon,\ \lim_{n,d\to\infty} \mathbb{P}\left[\nexists \mathbf{w} \,/\, \forall \mu \in [\![n]\!],\ \varphi\left(\mathbf{x}_\mu \cdot \mathbf{w}\right)\right] = 1\,. \tag{120}$$

Gardner and Derrida in their paper (Gardner et al., 1988) assume the storage capacity $\alpha_c(K)$ is a sharp threshold and they apply the replica calculation to compute it, but reach a result inconsistent with a simple upper bound obtained by the first moment method. Mézard and Krauth (Krauth et al., 1989) found a way to obtain a consistent prediction from the replica calculation and concluded that the storage capacity $\alpha_c^s(K)$ for the step binary perceptron

(SBP), i.e. associated to the constraint φ^s, is given by the largest α for which the following quantity, the *entropy* in physics, is positive:

$$\Phi_s^{(\mathrm{rs})}(\alpha,K) = \mathbf{extr}_{q_0,\hat{q}_0}\left\{\frac{1}{2}(q_0-1)\hat{q}_0 + \int \mathrm{D}\xi\ \log\left[2\cosh\left(\sqrt{\hat{q}_0}\xi\right)\right]\right. \\ \left. +\alpha\int \mathrm{D}\xi\ \log\left[\int_{\frac{K-\xi\sqrt{q_0}}{\sqrt{1-q_0}}}^{\infty}\mathrm{D}z\right]\right\}, \qquad (121)$$

where $\mathrm{D}\xi = \frac{e^{-\xi^2/2}}{\sqrt{2\pi}}\mathrm{d}\xi$ is a normal Gaussian measure, and **extr** means that the expression is evaluated where the derivatives on the curl-bracket, with respect to $q_0 \geq 0$ and $\hat{q}_0 \geq 0$, are zero.

Several decades of subsequent research in the statistical physics of disordered systems are consistent with the conjectured Mézard-Krauth formula for the storage capacity of the binary perceptron. Despite the simplicity of the above conjecture and decades of impressive progress in the mathematics of spin glasses and related problems, (see e.g. (Talagrand, 2006b; Talagrand, 2003; Mézard et al., 2009; Achlioptas et al., 2011; Panchenko, 2014; Ding et al., 2015) and many others), the storage capacity of the binary perceptron remains an open mathematical problem. In fact, even the very existence of a sharp threshold, i. e. the fact that in the limit $d \to \infty$ the probability that patterns can be stored drops sharply from one to zero at the capacity, is an open problem. Up to very recently only widely non-matching upper bounds and lower bounds for the storage capacity of the binary perceptron were available (Kim et al., 1998; Stojnic, 2013b). As the present work was being finalized Ding and Sun (Ding et al., 2019) proved in a remarkable paper a lower bound on the capacity that matches the Krauth and Mezard conjecture (note that much like Theorem 6.1.4 below, the main theorem in (Ding et al., 2019) depends on a numerical hypothesis). A matching upper bound remains an open challenge in mathematical physics and probability theory.

In this chapter, we introduce two simple *symmetric* variants of the binary perceptron problem. Let $z_\mu(\mathbf{w}) = \sum_{i=1}^{d} x_{\mu i} w_i = \mathbf{x}_\mu \cdot \mathbf{w}$. For a threshold $K \in \mathbb{R}^+$, we consider two different types of symmetric constraints:

- The rectangle binary perceptron (RBP) requires $|z_\mu| \leq K, \forall \mu \in [\![n]\!]$. Its associated indicator function is $\varphi^r(z) = \mathbb{1}\left[|z| \leq K\right]$.
- The *u*-function binary perceptron (UBP) requires $|z_\mu| \geq K, \forall \mu \in [\![n]\!]$. Its associated indicator function is $\varphi^u(z) = \mathbb{1}\left[|z| \geq K\right]$.

These constraints are symmetric in the sense that if $\mathbf{w}$ is a solution then $-\mathbf{w}$ is a solution as well. Our motivation behind these symmetric variants of the perceptron is that this symmetry simplifies greatly the mathematical treatment of the problem, while keeping the relevant physical properties intact. Thus, results that remain open questions for the canonical perceptron can be established rigorously for these symmetric versions. Symmetric perceptron models are also directly related to the problem of determining the discrepancy of a random matrix or set system (Bansal et al., 2019), a problem of interest in combinatorics.

Our main result, presented in Sec. 6.1, is a proof, subject to a numerical hypothesis, of a formula for the storage capacity, defined in the same way as for the step-function binary perceptron above. In particular, we show that in these symmetric variants the first moment upper bound, corresponding to the annealed capacity in physics, on the storage capacity is tight (except for $K > K^* \simeq 0.817$ for the UBP case). We prove this statement using the second moment method. We note that the existing physics literature on perceptron-like problem contains other cases of models where the first moment upper bound on the storage capacity was observed to be tight, in particular the parity machine (Opper, 1995), and the reversed-wedge binary perceptron (Bex et al., 1995; Hosaka et al., 2002). Those works, however, rely on the comparison of the first moment bound on the capacity with the result of the replica method, rather than providing a rigorous justification. To formally state our main result, let $z \sim \mathcal{N}(0,1)$, and for $K \in \mathbb{R}^+$ let $p_{r,K} = \mathbb{P}[|z| \leq K]$ and $p_{u,K} = \mathbb{P}[|z| \geq K]$.

- The storage capacity for the rectangle binary perceptron is:

$$\alpha_c^r(K) = \frac{-\log(2)}{\log(p_{r,K})} \qquad \forall K \in \mathbb{R}^+ . \tag{122}$$

- The storage capacity for the $u-$function binary perceptron is:

$$\alpha_c^u(K) = \frac{-\log(2)}{\log(p_{u,K})} \quad \text{for} \quad 0 < K < K^* \simeq 0.817 . \tag{123}$$

The constant $K^* \simeq 0.817$ stems from the properties of the second moment entropy eq. (129). In the physics terms it is defined as the point of intersection between the annealed capacity $\alpha_a^u(K)$ and the local stability of the RS solution $\alpha_{\mathrm{AT}}^u(K)$ eq. (137). That is, K^* is the solution of the following implicit equation:

$$\pi \; p_{u,K}^2 \; \exp\left(K^2\right) \; \log(p_{u,K}) = -2 \; \log(2) \; K^2 . \tag{124}$$

The two symmetric variants of the perceptron problem considered here share many of the intriguing geometric properties of the original step-function binary perceptron problem. Most significant is the conjectured f1RSB (Krauth et al., 1989) nature of the space of solutions that splits into well separated clusters of vanishing entropy at any $\alpha > 0$. Remarkably, this f1RSB property can be deduced from the form of the second moment entropy as we explain in section 6.2. Our justification of the f1RSB property does not rely on the replica method and is hence of independent interest.

For the UBP and $K > K^*$, the second-moment proof technique fails, and this failure marks tightly the onset of the replica symmetry breaking region. In that region, we evaluate the 1RSB approximation for the storage capacity, but conclude that FRSB would be needed to obtain the exact result. While the

Binary perceptron	Constraint	Constraint function	Range of K	Storage capacity
Step-function	$z \geq K$	$\varphi^s(z) = \mathbb{1}\left[z \geq K\right]$	$\forall K \in \mathbb{R}$	RS eq. (121)
Rectangle	$\lvert z\rvert \leq K$	$\varphi^r(z) = \mathbb{1}\left[\lvert z\rvert \leq K\right]$	$\forall K \in \mathbb{R}^+$	Annealed eq. (122)
U-function	$\lvert z\rvert \geq K$	$\varphi^u(z) = \mathbb{1}\left[\lvert z\rvert \geq K\right]$	$0 < K < K^* = 0.817$	Annealed eq. (123)
U-function	$\lvert z\rvert \geq K$	$\varphi^u(z) = \mathbb{1}\left[\lvert z\rvert \geq K\right]$	$\forall K > K^* = 0.817$	FRSB ?

Table 1: This table summarizes results for storage capacity in binary perceptrons with different types of constraints. The result for canonical step-function is from (Krauth et al., 1989). The results for the rectangle and u-function are obtained in this paper.

FRSB equations can be written along the lines of (Franz et al., 2017), they are more involved than the ones for the Sherrington-Kirkpatrick model (Parisi, 1979; Parisi, 1980c; Parisi, 1980d), and solving them numerically or getting additional insight from them is a challenging task left for future work. We present the replica analysis in section 6.3. Table 1 contains the summary of our main results along with the predictions for the step-function perceptron.

Finally let us comment on the simpler and more commonly considered case of spherical perceptron where the binary constraint on the vector $\mathbf{w}$ is replaced by the spherical constraint $\|\mathbf{w}\|_2^2 = \sum_{i=1}^d w_i^2 = d$. For $K = 0$ the spherical perceptron reduces to the famous problem of intersection of half-spaces with capacity $\alpha_c = 2$ as solved by Wendell (Wendel, 1962) and Cover (Cover, 1965). For $K > 0$ the Gardner-Derrida solution (Gardner et al., 1988) is correct as proven in (Shcherbina et al., 2003; Stojnic, 2013a). For $K < 0$ the situation is more challenging and FRSB is needed to compute the storage capacity; for recent progress in physics see (Franz et al., 2016; Franz et al., 2017), while mathematical considerations about this case were presented in (Stojnic, 2013c).

6.1 PROOF OF CORRECTNESS OF THE ANNEALED CAPACITY

To precisely state the main results, we introduce some definitions. Let $\mathbf{X}(d,n)$ be the random $n \times d$ pattern matrix. Define the partition functions

$$\begin{aligned} \mathcal{Z}_r(\mathbf{X}) &= \sum_{\mathbf{w}\in\{\pm1\}^d} \prod_{\mu=1}^{n} \varphi^r(z_\mu(\mathbf{w})) \\ \mathcal{Z}_u(\mathbf{X}) &= \sum_{\mathbf{w}\in\{\pm1\}^d} \prod_{\mu=1}^{n} \varphi^u(z_\mu(\mathbf{w}))\,, \end{aligned} \tag{125}$$

which count respectively the number of solutions for the rectangle and $u-$function constraints. Let $\mathcal{E}^r(d,n)$ and $\mathcal{E}^u(d,n)$ be the events that $\mathcal{Z}_r(\mathbf{X}) \geq 1$ and $\mathcal{Z}_u(\mathbf{X}) \geq 1$, we formally define the storage capacity as follows.

Definition 6.1.1. *The storage capacity $\alpha_c^r(K)$ is*

$$\alpha_c^r(K) = \inf\left\{\alpha : \lim_{d\to\infty} \mathbb{P}[\mathcal{E}^r(d, \lfloor \alpha d \rfloor)] = 0\right\},$$

and likewise for $\alpha_c^u(K)$.

It is believed that there is a sharp threshold for the existence of solutions.

Conjecture 6.1.2. *The storage capacity is a sharp threshold:*

$$\alpha_c^r(K) = \sup\left\{\alpha : \lim_{d\to\infty} \mathbb{P}[\mathscr{E}^r(d, \lfloor \alpha d \rfloor)] = 1\right\},$$

and likewise for $\alpha_c^u(K)$.

The corresponding conjecture for the random k-SAT model is the celebrated *satisfiability threshold conjecture* proved for k large by Ding, Sly, and Sun (Ding et al., 2015). Next, couple two standard Gaussians z_1, z_β by letting z and z' be independent standard Gaussians and setting $z_1 = \sqrt{\beta}z + \sqrt{1-\beta}z'$ and $z_\beta = \sqrt{\beta}z - \sqrt{1-\beta}z'$. Let

$$\begin{aligned} q_{r,K}(\beta) &\equiv \mathbb{P}[|z_1| \le K \wedge |z_\beta| \le K] = q_K(\beta), \\ q_{u,K}(\beta) &\equiv \mathbb{P}[|z_1| \ge K \wedge |z_\beta| \ge K] = 1 - 2p_{r,K} + q_K(\beta), \end{aligned} \tag{126}$$

with $q_K(\beta)$ the probability that two standard Gaussians with correlation $2\beta - 1$ are both at most K in absolute value, explicitly given by

$$q_K(\beta) = \frac{1}{2\pi}\int_{-K}^{K} \mathrm{d}y \int_{\frac{-K+(1-2\beta)y}{2\sqrt{\beta(1-\beta)}}}^{\frac{K+(1-2\beta)y}{2\sqrt{\beta(1-\beta)}}} \mathrm{d}x \, \exp\left(-\frac{x^2+y^2}{2}\right).$$

Note that $q_{t,K}(1) = p_{t,K}$ and $q_{t,K}(1/2) = p_{t,K}^2$ for $t \in \{r, u\}$. We now introduce the functions that dictate the effectiveness of the second moment bound. Let

$$\begin{aligned} F_{r,K,\alpha}(\beta) &= \mathrm{H}(\beta) + \alpha \log q_{r,K}(\beta), \\ F_{u,K,\alpha}(\beta) &= \mathrm{H}(\beta) + \alpha \log q_{u,K}(\beta) \end{aligned} \tag{127}$$

where $\mathrm{H}(\beta) = -\beta\log\beta - (1-\beta)\log(1-\beta)$ is the Shannon entropy function. We state a numerical hypothesis in terms of the derivatives of these two functions.

Hypothesis 6.1.3. *For all choices of* $K > 0$ *and* $\alpha > 0$ *so that* $F''_{r,K,\alpha}(1/2) < 0$, *there is exactly one* $\beta \in (1/2, 1)$ *so that* $F'_{r,K,\alpha}(\beta) = 0$. *The same holds for* $F_{u,K,\alpha}$.

Our main theorem is a proof, under Hypothesis 6.1.3, that the storage capacity is given by the annealed computation.

Theorem 6.1.4. *Under the assumption of Hypothesis 6.1.3, the following hold.*

1. *For all* $K > 0$, *we have* $\alpha_c^r(K) = -\log(2)/\log(p_{r,K})$.
2. *For all* $K \in (0, K^*)$, *we have* $\alpha_c^u(K) = -\log(2)/\log(p_{u,K})$.

Under our definition of $\alpha_c^r(K)$ and $\alpha_c^u(K)$, we must prove two statements to show that $\alpha_c^r(K) = -\log(2)/\log(p_{r,K})$ (and similarly for $\alpha_c^u(K)$). We use the first moment method to show that for $\alpha > -\log(2)/\log(p_{r,K})$, $\lim_{d\to\infty} \mathbb{P}[\mathscr{E}^r(d,n)] = 0$; then we use the second moment method to show that for $\alpha < -\log(2)/\log(p_{r,K})$, $\liminf_{d\to\infty} \mathbb{P}[\mathscr{E}^r(d,n)] > 0$ (a result analogous to what Ding and Sun prove for the more challenging step binary perceptron (Ding et al., 2019)). Conjecture 6.1.2 asserts the stronger statement that for $\alpha < -\log(2)/\log(p_{r,K})$, $\lim_{d\to\infty} \mathbb{P}[\mathscr{E}^r(d,n)] = 1$.

6.1.1 FIRST MOMENT UPPER BOUND

Proposition 6.1.5 (First moment upper bound)**.**

1. *If $\alpha > \alpha_a^r(K) = \frac{-\log(2)}{\log(p_{r,K})}$, then with high probability there is no satisfying assignment to the binary perceptron with the rectangle activation function.*
2. *If $\alpha > \alpha_a^u(K) = \frac{-\log(2)}{\log(p_{u,K})}$, then with high probability there is no satisfying assignment to the binary perceptron with the u-function activation function.*

Proof. We give the proof for the rectangle function as the proof for the u-function is identical. Let $\varepsilon = \alpha - \alpha_a^r(K) > 0$. Let $\mathbf{1}_d$ denote the vector of dimension d with all 1 entries.

$$\begin{aligned}\mathbb{P}[\mathscr{E}^r(d,\alpha d)] \le \mathbb{E}\left[\mathscr{Z}_r\left(\mathbf{X}(d,\alpha d)\right)\right] &= 2^d\ \mathbb{E}\left[\prod_{\mu=1}^{\alpha d} \mathbb{1}\left[\left|z_\mu(\mathbf{1})\right| \le K\right]\right] \\ &= 2^d\ p_{r,K}^{\alpha d} = \exp(d(\log(2) + \alpha\log(p_{r,K}))) \\ &= \exp(d\varepsilon\log(p_{r,K})) \to 0 \text{ as } d \to \infty.\end{aligned}$$

□

6.1.2 SECOND MOMENT LOWER BOUND

Proposition 6.1.6 (Second moment lower bound)**.**

1. *If $\alpha < \frac{-\log(2)}{\log(p_{r,K})}$, then $\liminf_{d\to\infty} \mathbb{P}[\mathscr{E}^r(d,\alpha d)] > 0$.*
2. *If $K < K^*$ and $\alpha < \frac{-\log(2)}{\log(p_{u,K})}$, then $\liminf_{d\to\infty} \mathbb{P}[\mathscr{E}^u(d,\alpha d)] > 0$.*

To prove Proposition 6.1.6 we will apply the second-moment method in a similar fashion to Achlioptas and Moore (Achlioptas et al., 2002) who determined the satisfiability threshold of random k-SAT to within a factor 2 by considering not-all-equal satisfying assignments (not-all-equal satisfiability (NAE-SAT) constraints are symmetric in the same way the rectangle and u-function constraints are symmetric). Recall the Paley-Zygmund inequality.

Lemma 6.1.7. *Let* X *be a non-negative random variable. Then*

$$\mathbb{P}[\mathrm{X} > 0] \ge \frac{\mathbb{E}[\mathrm{X}]^2}{\mathbb{E}[\mathrm{X}^2]}.$$

We will also use the following application of Laplace's method from Achlioptas and Moore (Achlioptas et al., 2002).

Lemma 6.1.8. *Let $g(\beta)$ be a real analytic function on $[0,1]$ and let*

$$G(\beta) = \frac{g(\beta)}{\beta^\beta (1-\beta)^{1-\beta}} .$$

If $G(1/2) > G(\beta)$ for all $\beta \neq 1/2$ and $G''(1/2) < 0$, then there exists constants c_1, c_2 so that for all sufficiently large d

$$c_1 G(1/2)^d \leq \sum_{l=0}^{d} \binom{d}{l} g(l/d)^d \leq c_2 G(1/2)^d .$$

6.1.2.A RECTANGLE BINARY PERCEPTRON

We calculate

$$\mathbb{E}[\mathscr{Z}_r(\mathbf{X})^2] = \sum_{\mathbf{w}_1,\mathbf{w}_2 \in \{\pm 1\}^d} \mathbb{P}[\mathbf{w}_1, \mathbf{w}_2 \text{ satisfying}]$$

$$= 2^d \sum_{\mathbf{w} \in \{\pm 1\}^d} \mathbb{P}[\mathbf{1}, \mathbf{w} \text{ satisfying}] = 2^d \sum_{l=0}^{d} \binom{d}{l} q_{r,K}(l/d)^{\alpha d}$$

$$= \exp\left(d(\log(2) + F_{r,K,\alpha}(\beta))\right) .$$

where we recall $q_{r,K}$ from eq. (126). Define

$$G_{r,K,\alpha}(\beta) \equiv \exp(F_{r,K,\alpha}(\beta)) = \frac{q_{r,K}(\beta)^\alpha}{\beta^\beta (1-\beta)^{1-\beta}} . \tag{128}$$

If we can show that $G_{r,K,\alpha}(1/2) > G_{r,K,\alpha}(\beta)$ for all $\beta \neq 1/2$ and $G''_{r,K,\alpha}(1/2) < 0$, then by Lemma 6.1.8, we have

$$\mathbb{E}[\mathscr{Z}_r(\mathbf{X})^2] \leq c_2\, 4^d\, q_{r,K}(1/2)^{\alpha d} = c_2\, 4^d\, p_{r,K}^{2\alpha d} .$$

Then since $\mathscr{Z}_r(\mathbf{X})$ is integer valued, we have

$$\mathbb{P}[\mathscr{Z}_r(\mathbf{X}) \geq 1] \geq \frac{\mathbb{E}[\mathscr{Z}_r(\mathbf{X})]^2}{\mathbb{E}[\mathscr{Z}_r(\mathbf{X})^2]} = \frac{(2^d p_{r,K}^{\alpha d})^2}{\mathbb{E}[\mathscr{Z}_r(\mathbf{X})^2]} \geq \frac{(2^d p_{r,K}^{\alpha d})^2}{c_2 4^d p_{r,K}^{2\alpha d}} = \frac{1}{c_2} > 0 .$$

It remains to show that when $\alpha < \frac{-\log(2)}{\log(p_{r,K})}$, then $G_{r,K,\alpha}(1/2) > G_{r,K,\alpha}(\beta)$ for all $\beta \neq 1/2$ and $G''_{r,K,\alpha}(1/2) < 0$. By eq. (128) and the fact that $G'_{r,K,\alpha}(1/2) = 0$, it is enough to show the same for $F_{r,K,\alpha}$. Certainly one necessary condition is that $F_{r,K,\alpha}(1/2) > F_{r,K,\alpha}(1)$. This reduces to the condition $2p_{r,K}^{2\alpha} > p_{r,K}^{\alpha}$ or $\alpha < \frac{-\log(2)}{\log(p_{r,K})}$ which is exactly the condition of Proposition 6.1.6. Next consider $F''_{r,K,\alpha}(1/2)$. A straightforward calculation shows that

$$F''_{r,K,\alpha}(1/2) = 4\left(-1 + \frac{2}{\pi} \frac{\alpha K^2 e^{-K^2}}{p_{r,K}^2}\right) .$$

In particular, $F''_{r,K,\alpha}(1/2) < 0$ if and only if $\alpha < \frac{\pi}{2}\frac{p_{r,K}^2}{K^2 \exp(-K^2)}$. But another calculation also shows that

$$-\frac{\log(2)}{\log(p_{r,K})} < \frac{\pi}{2}\frac{p_{r,K}^2}{K^2 e^{-K^2}}$$

for all $K > 0$ and so the condition of Proposition 6.1.6 implies that $F''_{r,K,\alpha}(1/2) < 0$. Moreover, since $F_{r,K,\alpha}(\beta)$ is symmetric around $\beta = 1/2$ and it has a local maximum at $\beta = 1/2$, Hypothesis 6.1.3 implies that the global maximum of $F_{r,K,\alpha}(\beta)$ occurs at either $1/2$ or 1, and since $F_{r,K,\alpha}(1/2) > F_{r,K,\alpha}(1)$, we have that $F_{r,K,\alpha}(1/2) > F_{r,K,\alpha}(\beta)$ for all $\beta \neq 1/2$, completing the proof of Proposition 6.1.6 for the rectangle binary perceptron.

6.1.2.B u-FUNCTION BINARY PERCEPTRON

The proof for the u-function is similar. We can calculate

$$\begin{aligned}\mathbb{E}[\mathscr{Z}_u(\mathbf{X})^2] &= 2^d \sum_{l=0}^{d} \binom{d}{l} q_{u,K}(l/d)^{\alpha d} \\ &= \exp\left(d(\log(2) + F_{u,K,\alpha}(\beta))\right),\end{aligned}$$

where we recall $q_{u,K}$ from eq. (126). Using Lemma 6.1.8 and Hypothesis 6.1.3 again, it suffices to show that for $0 < K < K^*$ and $\alpha < \frac{-\log(2)}{\log(p_{u,K})}$ we have $F_{u,K,\alpha}(1/2) > F_{u,K,\alpha}(1)$ and $F''_{u,K,\alpha}(1/2) < 0$. The first follows immediately from the fact that $\alpha < \frac{-\log(2)}{\log(p_{u,K})}$. For the second, we have

$$F''_{u,K,\alpha}(1/2) = 4\left(-1 + \frac{2}{\pi}\frac{\alpha K^2 e^{-K^2}}{p_{u,K}^2}\right)$$

and so $F''_{u,K,\alpha}(1/2) < 0$ if and only if $\alpha < \frac{\pi}{2}\frac{p_{u,K}^2}{K^2 e^{-K^2}}$. Unlike with the rectangle function it is not true that

$$-\frac{\log(2)}{\log(p_{u,K})} < \frac{\pi}{2}\frac{p_{u,K}^2}{K^2 e^{-K^2}} \tag{129}$$

for all K: the left and right sides of the inequality cross at $K = K^*$, which implicitly defines K^*. Thus for $K < K^*$ and $\alpha < -\frac{\log(2)}{\log(p_{u,K})}$ we have $F''_{u,K,\alpha}(1/2) < 0$, which completes the proof of Proposition 6.1.6 for the u-function binary perceptron.

6.1.2.C ILLUSTRATION

As an illustration, we plot the second moment entropy density $\lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}[\mathscr{Z}_t^2] = \log(2) + F_{t,K,\alpha}$ for $t \in \{r, u\}$ at $K = 1 > K^*$ in Fig. 35. For the rectangle function (**Left**), the second moment is tight: the maximum is reached for $\beta = 1/2$ for all α smaller than the first moment α_a^r (dashed pink). Exactly the same happens for the $u-$function with $K < K^*$. However for $K > K^*$, the

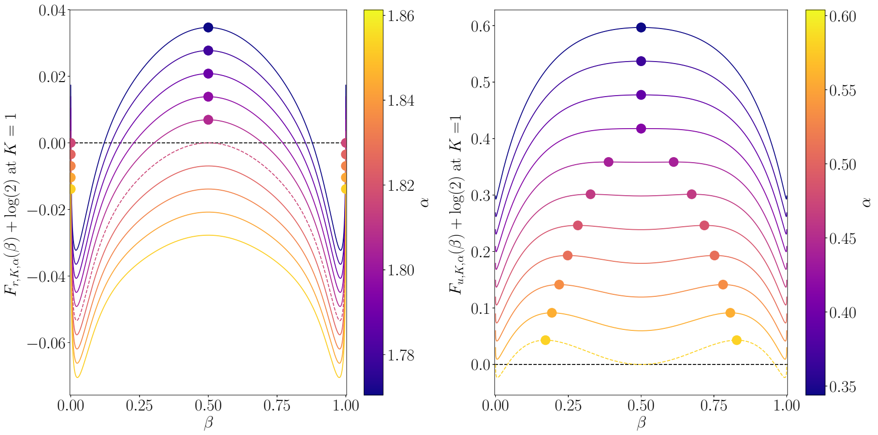


Figure 35: Second moment entropy densities. (**Left**) the rectangle binary perceptron for $\alpha \leq \alpha_a^r = 1.816$ (dashed pink), $\beta = \frac{1}{2}$ is the global maximizer. For $\alpha \geq \alpha_a^r$, $\beta = 0$ and $\beta = 1$ are the maximizers. (**Right**) the u-function binary perceptron for $\alpha \leq \alpha^* = 0.430$, $\beta = \frac{1}{2}$ is the maximizer while for $\alpha^* \leq \alpha \leq \alpha_a^u = 0.604$ (dashed yellow), the maximizer is non-trivial $\beta \neq 0$.

second moment method fails (**Right**): $\beta = 1/2$ becomes a minimum and the maximum is obtained for non trivial values $\beta \neq 1/2$ for constraint density smaller than the first moment α_a^u (dashed yellow).

6.2 FROZEN-1RSB STRUCTURE OF SOLUTIONS IN BINARY PERCEPTRONS

One of the most striking properties of the canonical step-function perceptron is the predicted f1RSB (Krauth et al., 1989) nature of the space of solutions. This means that the dominant, i. e. with measure tending to one, part of the space of solutions splits into well separated clusters each of which has vanishing entropy density at any $\alpha > 0$. This f1RSB scenario and quantitative properties of the solution space were studied in detail recently (H. Huang, 2013; Huang et al., 2014). Following up on conjectures that such a frozen structure of solutions implies computational hardness in diluted constraint satisfaction problems (Zdeborová et al., 2008a), it was argued that finding a satisfying assignment in the binary perceptron should also be algorithmically hard since its solution space is dominated by clusters of vanishing entropy density (Huang et al., 2014). Yet this conjecture contradicted empirical results of (Braunstein et al., 2006). This paradox was resolved in (Baldassi et al., 2015) where the authors identified that there are subdominant parts (i. e. parts of measure converging to zero as the system size diverges) of the solution space that form extended clusters with large local entropy and all the algorithms that work well always find a solution belonging to one of those large-local-entropy clusters. These sub-dominant clusters are not frozen and somewhat strangely are not captured in the canonical 1RSB calculation (Baldassi et al., 2015). It was argued that existence of these large-local-entropy clusters bears more general consequences on the dynamics of learning algorithms in neural networks, see e.g. (Baldassi et al., 2016).

While f1RSB structure has also been identified in CSP on sparse graphs (Zdeborová et al., 2008b; Zdeborová et al., 2011), we want to note that its nature in the binary perceptron is of a rather different nature. In sparse systems a simple argument using expansion properties of the underlying graph and properties of the constraints show that each cluster with high probability contains only one solution. In the perceptron model, which has a fully connected bipartite interaction graph, this argument from sparse models does not apply.

In the present work, we deduce from the second moment calculation of the previous section that the space of solutions in the symmetric binary perceptrons is also of the f1RSB type and this property moreover extends to any finite temperature (with energy being defined as the number of unsatisfied constraints). This is different from the locked CSP of (Zdeborová et al., 2008a; Zdeborová et al., 2011) living on diluted hyper-graphs, where the solution-clusters have extensive entropy at any non-zero temperature. Another difference is that whereas in the locked CSP the size of each cluster is one with high probability, in the binary perceptron there are still many solutions in the clusters, it is only their entropy density, i. e. the logarithm of their number per variable, that vanishes as $d \to \infty$. Investigation of the large local entropy clusters and their implications for learning in the symmetric perceptrons is also of great interest, but left for future work. Clearly since mathematically the symmetric perceptrons are simpler than the step-function one, they should also be the proper playground to deepen our understanding of the large local entropy clusters and their relation to learning and generalization.

We present the f1RSB scenario as a conjecture and then below indicate how the second moment calculation gives evidence for this conjecture. Given an instance $\mathbf{X}$ and a solution $\mathbf{w}$, let $\Gamma(\mathbf{w}, \tilde{d})$ denote the set of solutions $\mathbf{w}'$ with Hamming distance at most $\tilde{d}$ from $\mathbf{w}$.

Conjecture 6.2.1. *For every $K > 0$ and every $\alpha \in (0, \alpha_c^r(K))$ there exists a Hamming distance $\tilde{d}_{min} > 0$ so that with high probability over the choice of the random instance $\mathbf{X}$ from the RBP, the following property holds: for almost every solution $\mathbf{w}$,*

$$\frac{1}{d} \log |\Gamma(\mathbf{w}, \tilde{d}_{min})| \xrightarrow[d \to \infty]{} 0$$

The same holds for the UBP for all $K \leq K^*$.

6.2.1 THE LINK BETWEEN THE SECOND-MOMENT ENTROPY AND SIZE OF CLUSTERS

In this section we use $t \in \{r,u\}$ and note that the form of the second moment entropy density $\frac{1}{d}\log\mathbb{E}[\mathscr{Z}_t^2]$ has very direct implications on the structure of solutions in the corresponding models. As we defined it above, the second moment entropy is the normalized logarithm of the expected number of pairs of solutions of overlap β.

For problems such as the symmetric binary perceptrons where the quenched and annealed entropies are equal in leading order, there is a striking relation between the planted and the random ensemble of the model (Achlioptas et al., 2008; Krzakala et al., 2009). The *random ensemble* is the problem we have considered so far, while the *planted ensemble* is defined by starting with a configuration of the weights (a solution) and then including only constraints that are satisfied by this *planted* configuration. As long as the quenched and annealed entropies of the random ensemble are equal in leading order the planted and random ensembles should be contiguous, meaning that high-probability properties that hold in one ensemble also hold in the other. Moreover the planted configuration in the planted ensemble has all the properties of a configuration sampled uniformly at random in the random ensemble. These properties follow on the heuristic level from the cavity method reasoning (Krzakala et al., 2009). They were established fully rigorously in a range of models, see e.g. (Achlioptas et al., 2008; Mossel et al., 2015; Coja-Oghlan et al., 2018). In the present case of symmetric binary perceptrons we have not yet managed to prove contiguity between the random and the planted ensemble, and so we leave a rigorous mathematical result for future work. (In fact the missing ingredient is a version of Friedgut's sharp threshold result (Friedgut, 1999) suitable for perceptrons; such a result combined with Theorem 6.1.4 would also prove Conjecture 6.1.2). We hence rely on the above heuristic argument and assume it holds in what follows.

Given a planted solution $\mathbf{w}$ and a configuration $\mathbf{w}_\beta$ that agrees with $\mathbf{w}$ on βd coordinates, the probability that $\mathbf{w}_\beta$ is a solution in the planted model is $(q_{t,K}(\beta)/p_{t,K})^n$, and thus the expected number of solutions $\mathscr{Z}_\beta$ at Hamming distance βd from the planted solution in the planted ensemble is

$$\mathbb{E}[\mathscr{Z}_\beta] = \binom{d}{\beta d}(q_{t,K}(\beta)/p_{t,K})^n,$$

and its entropy density is

$$\omega_t(\beta) \equiv \lim_{d\to\infty}\frac{1}{d}\log\mathbb{E}[\mathscr{Z}_\beta] = F_{t,K,\alpha}(\beta) - \alpha\log p_{t,K} \text{ for } t\in\{r,u\}. \tag{130}$$

Recalling that *contiguity* implies that the planted solution has the properties of a uniformly chosen solution in the random ensemble then this entropy gives us direct access to properties of the solution space in the random ensemble at equilibrium. Most notably we notice (see derivation in section 6.2.2 below) that the derivative of $\omega_t(\beta)$ at $\beta = 1$ is $+\infty$ thus implying that $\forall \varepsilon > 0$ with high probability there are no solutions at overlap $\beta \in [\tilde{d}_{\min}(\alpha, K), (1-\varepsilon)]$. In turn, this means that the dominant (measure converging to one as $d \to \infty$) part of the solution space splits into clusters each of which has vanishing entropy density (i.e. logarithm of the number of solutions in the cluster divided by d goes to zero as $d \to \infty$). The missing ingredient in a full proof of Conjecture 6.2.1 is a proof of the contiguity statement.

6.2.2 FORM OF THE 2ND MOMENT ENTROPY IMPLYING FROZEN-1RSB

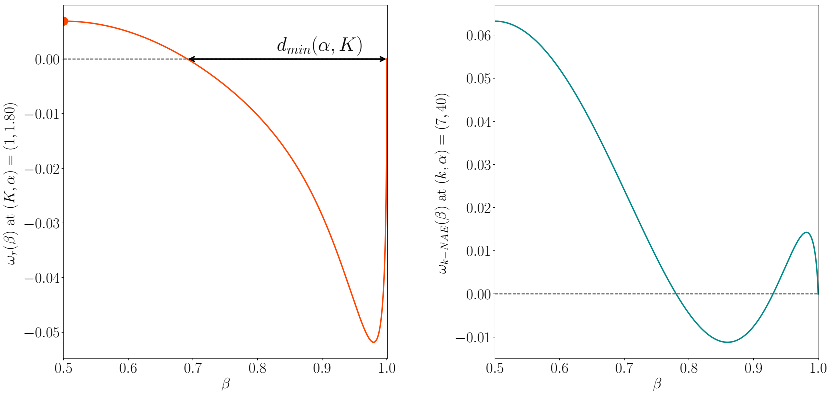


Figure 36: **(Left)** Density of the annealed entropy of solutions at overlap β from a random solution in the rectangle binary perceptron at $K = 1$, $\alpha = 1.80 \leq \alpha_c^r(K = 1)$. We see there are no solution in an interval of overlaps $(1 - \tilde{d}_{\min}, 1 - \varepsilon)$. This curve is obtained from the second moment entropy and contiguity between the random and planted ensembles. It implies the frozen-1RSB nature of the space of solutions. The same holds for the u−function. **(Right)** To compare we plot the density of the annealed entropy of solutions at overlap β from a random solution in the k-NAE SAT model (Achlioptas et al., 2002) at $k = 7$, $\alpha = 40$. We see the density is positive in a large region close to $\beta = 1$, showing the absence of frozen-1RSB structure in this problem.

In Fig. 36 (**Left**), we plot $\omega_r(\beta)$ for the rectangle binary perceptron, at $K = 1$, $\alpha = 1.80 \leq \alpha_c^r(K = 1)$. Thanks to the contiguity between the planted and random ensembles that holds as long as the second moment entropy density is twice the first moment entropy density, this curve represents also the annealed entropy of solutions at overlap β with a random reference solution. We see notably that there is an interval of distances in which no solutions are present. Analytically we can see from the properties of the functions $F_{t,K,\alpha}(\beta)$ and $\log p_{t,K}$ that $F_{t,K,\alpha}(1) = \alpha \log p_{t,K}$ and the derivative of $F_{t,K,\alpha}(\beta) \to \infty$. This is in contrast with, for instance, the satisfiability problems studied in (Achlioptas et al., 2002), where the function corresponding to $F_{t,K,\alpha}(\beta)$ would have a negative derivative in $\beta = 1$, see Fig. 36 (**Right**). There could still be an interval of *forbidden* distance, but the bump in entropy for $\beta \approx 1$ corresponds to the size of the clusters to which typical solutions belong and those would be extensive.

6.2.2.A FROZEN 1RSB IN RECTANGLE BINARY PERCEPTRON

In the rectangle binary perceptron, the random and planted ensembles are conjectured to be contiguous for all $K > 0$ and $\alpha \in (0, \alpha_c^r(K))$. Using eq. (127), the first derivative of $\omega_r(\beta)$, eq. (130), is given by

$$\frac{\partial \omega_r}{\partial \beta} = \frac{\partial F_{r,K,\alpha}}{\partial \beta} = \log\left(\frac{1-\beta}{\beta}\right) + \frac{\alpha}{q_{r,K,T}(\beta)} \frac{1}{\pi\sqrt{\beta(1-\beta)}} \left(e^{-\frac{K^2}{2(1-\beta)}}\left(e^{\frac{(2\beta-1)K^2}{2(1-\beta)\beta}} - 1\right)\right) \xrightarrow[\beta\to 1]{} +\infty ,$$

where the computation is detailed in Sec. E of (Aubin et al., 2019e). It diverges for all $K \in \mathbb{R}^+$, $\alpha > 0$ in the limit $\beta \to 1$, that implies vanishing entropy density of clusters to which typical solutions belong.

6.2.2.B FROZEN 1RSB IN THE u-FUNCTION BINARY PERCEPTRON

In the u-function binary perceptron, the random and planted ensembles are conjectured to be contiguous for all $0 < K \leq K^*$ and $\alpha \in (0, \alpha_c^u(K))$. Using eq. (127), the first derivative of $\omega_u(\beta)$ eq. (130), is given by

$$\frac{\partial \omega_u}{\partial \beta} = \frac{\partial F_{u,K,\alpha}}{\partial \beta} = \log\left(\frac{1-\beta}{\beta}\right) + \frac{\alpha}{q_{u,K,T}(\beta)} \frac{1}{\pi\sqrt{\beta(1-\beta)}} \left(e^{-\frac{K^2}{2(1-\beta)}}\left(e^{\frac{(2\beta-1)K^2}{2(1-\beta)\beta}} - 1\right)\right) \xrightarrow[\beta\to 1]{} +\infty ,$$

thus reaching the same conclusion on presence of f1RSB.

In Sec. E of (Aubin et al., 2019e), we extend the second moment calculation to finite temperature (for both the rectangle and $u-$function case). This means that we define the energy of a configuration $\mathscr{E}(\mathbf{w})$ as the number of constraints that are violated by this configurations. Then the corresponding partition function is defined $\mathscr{Z}(T) = \sum_{\mathbf{w}} e^{-\mathscr{E}(\mathbf{w})/T}$. There is a one-to-one mapping between the temperature T and energy density $e = \mathscr{E}/d$, consequently the corresponding finite-temperature second moment entropy density counts the number of pairs of solutions at overlap β and energy density e. In Sec. E of (Aubin et al., 2019e), we apply the same argument as here connecting the random and planted ensemble, and deduce that the finite-temperature solution space of the models is of also of the f1RSB type for any $T < \infty$.

6.2.3 FROZEN-1RSB AS DERIVED FROM THE REPLICA ANALYSIS

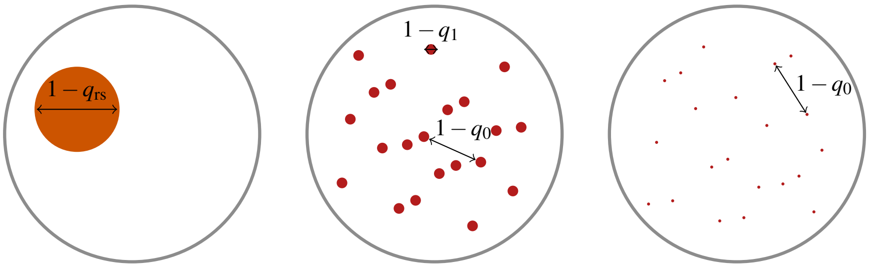


Figure 37: Illustration of the configuration space for the different phases (**Left**) RS: solutions are concentrated in a single cluster of typical size $1-q_{\rm rs}$. (**Center**) 1RSB: solutions form clusters of size $1-q_1$ at a distance $1-q_0$ from each other. (**Right**) f1RSB: clusters are point-like ($1-q_1 \simeq 0$) at a distance $1-q_0 = 1-q_{\rm rs}$ from each other.

We stress that we derived the f1RSB nature of the space of solutions without the use of replicas. For completeness we summarize here how this translates to the properties of the one-step-replica-symmetry breaking solution. This is the way this phenomena was originally discovered and described in (Martin et al., 2004; H. Huang, 2013). For readers not familiar with the replica method this section should be read after reading section 6.3. In general, three kinds of fixed points of the 1RSB equations are possible:

- The replica symmetric RS solution $q_0 = q_1 = q_{\rm rs} < 1$,
- The frozen-1RSB solution f1RSB $(q_0, q_1) = (q_{\rm rs}, 1)$,
- The 1RSB solution (q_0, q_1) with $q_1 \neq 1$.

The f1RSB is characterized by an inner-cluster overlap $q_1 = 1$ and an inter-cluster overlap $q_0 = q_{\rm rs}$, which means that clusters have vanishing entropy density and remain far from each other. Mathematically RS and f1RSB solutions are equivalent in the sense that these solutions have the same free entropy $\Phi^{(\rm 1rsb)}\{q_0 = q_{\rm rs}, q_1 = q_{\rm rs}\} = \Phi^{(\rm 1rsb)}\{q_0 = q_{\rm rs}, q_1 = 1\}$, and the complexity of the f1RSB solution equals the RS entropy $\Sigma(\Phi = 0) = \Phi^{(\rm rs)}$ eq. (145, 134). However, RS and f1RSB do not share the same configuration space. The RS phase is associated to a single cluster of solution with typical size $1-q_{\rm rs}$, while the f1RSB configuration space is composed of many point-like solutions of size $q_1 \simeq 1$ and at distance $1-q_0 = 1-q_{\rm rs}$ of each other, see Fig. 37. From this point of view f1RSB is the correct description of the phase space.

6.3 REPLICA CALCULATION OF THE STORAGE CAPACITY

In this section we provide the replica free entropies leading to the expression of the storage capacity in the step-function binary perceptron (121). We show that in the symmetric binary perceptrons the annealed calculation is reproduced by the replica symmetric result. For the $u-$function binary perceptron we show that K^* coincides with the onset of replica symmetry breaking and we evaluate the 1RSB capacity for $K > K^*$. The details of the computation is presented in Appendix. B.2 for the constraint function at zero temperature

$$\mathscr{C}(\mathbf{z}) \equiv \prod_{\mu=1}^{n} \varphi(z_\mu) \text{ with } z_\mu = \mathbf{x}_\mu \cdot \mathbf{w}, \tag{131}$$

and $\mathrm{P_y}(\mathbf{y}) = \delta(\mathbf{y}-\mathbf{1})$ if we use the Gauge transformation $\mathbf{x} \to y\mathbf{x}$, $y \to 1$ by symmetry of the labels and the data. The replica computation of the quenched average of the partition functions (125) $\mathbb{E}_{\mathbf{y},\mathbf{X}}[\log(\mathscr{Z}_d(\mathbf{X}))]$

$$\mathscr{Z}_d(\mathbf{y},\mathbf{X}) = \int_{\mathbb{R}} \mathrm{d}y\, \mathrm{P_y}(y) \int_{\mathbb{R}^d} \mathrm{d}\mathbf{w}\, \mathrm{P_w}(\mathbf{w}) \int \mathrm{d}\mathbf{z}\, \mathscr{C}(\mathbf{z})\delta(\mathbf{z}-\mathbf{X}\mathbf{w}),$$

and boils down to a free entropy formulation

$$\Phi(\alpha) = \mathbf{extr}_{\mathbf{Q},\hat{\mathbf{Q}}} \left\{ \lim_{r\to 0} \frac{\partial \Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha)}{\partial r} \right\}, \tag{132}$$

as a function of symmetric overlap matrices $\mathbf{Q} \in \mathbb{R}^{r\times r}$ and $\tilde{\mathbf{Q}} \in \mathbb{R}^{r\times r}$ in the limit $r \to 0$:

$$\begin{aligned} \Phi^{(r)}\left(\mathbf{Q},\hat{\mathbf{Q}},\alpha\right) &\equiv -\frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right) + \log \Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}}) + \alpha \log \Psi_{\mathrm{out}}^{(r)}(\mathbf{Q}), \\ \Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}}) &= \int_{\mathbb{R}^r} \mathrm{dP_w}(\tilde{\mathbf{w}}) e^{\frac{1}{2}\tilde{\mathbf{w}}^\intercal \hat{\mathbf{Q}} \tilde{\mathbf{w}}}, \\ \Psi_{\mathrm{out}}^{(r)}(\mathbf{Q}) &= \int_{\mathbb{R}^r} \mathrm{dP_z}(\tilde{\mathbf{z}},\mathbf{Q}) \mathscr{C}(\tilde{\mathbf{z}}). \end{aligned} \tag{133}$$

To obtain a tractable expression of the free entropy, in the following we perform the so-called RS and 1RSB ansatz.

6.3.1 RS CALCULATION AND STABILITY

6.3.1.A RS ENTROPY

The simplest ansatz is to assume that the overlap matrix $\mathbf{Q}$ is RS, which means that all replicas play the same role: the correlation between two arbitrary, but different, replicas is denoted q_0, and therefore the RS ansatz reads:

$$\forall (a,b) \in [\![r]\!]^2, \qquad \frac{1}{d}(\mathbf{w}^a \cdot \mathbf{w}^b) = \begin{cases} q_0 \text{ if } a \neq b, \\ Q = 1 \text{ if } a = b. \end{cases}$$

It enforces the matrix $\tilde{\mathbf{Q}}$ to present the same symmetry, respectively with parameters $\hat{q}_0$ and $\hat{Q} = 1$. Using this Ansatz and the $r \to 0$ limit, the RS entropy can be expressed as a set of saddle point equations over scalar parameters q_0 and $\hat{q}_0$, evaluated at the saddle point, see Appendix. B.2.4,

$$\Phi^{(\mathrm{rs})}(\alpha) = \underset{q_0,\hat{q}_0}{\mathbf{extr}} \left\{ \frac{1}{2}(q_0\hat{q}_0 - 1) + \Psi_{\mathrm{w}}^{(\mathrm{rs})}(\hat{q}_0) + \alpha \Psi_{\mathrm{out}}^{(\mathrm{rs})}(q_0) \right\}, \qquad (134)$$

with

$$\Psi_{\mathrm{w}}^{(\mathrm{rs})}(\hat{q}_0) \equiv \mathbb{E}_{\xi_0} \log\, g_0^w(\xi_0,\hat{q}_0), \qquad \Psi_{\mathrm{out}}^{(\mathrm{rs})}(q_0) \equiv \mathbb{E}_{\xi_0} \log\, f_0^z(\xi_0,q_0), \qquad (135)$$

$$\begin{aligned} g_i^w(\xi_0,\hat{q}_0) &\equiv \mathbb{E}_w \left[w^i \,\exp\left(\frac{(1-\hat{q}_0)}{2} w^2 + \xi_0 \sqrt{\hat{q}_0} w \right) \right], \\ f_i^z(\xi_0,q_0) &\equiv \mathbb{E}_z \left[z^i \,\varphi(\sqrt{q_0}\xi_0 + \sqrt{1-q_0} z) \right], \end{aligned} \qquad (136)$$

for $i \in \mathbb{N}$ and where $\xi_0, z \sim \mathcal{N}(0,1)$, $w \sim \mathrm{P_w}(.)$. In the binary perceptron case, the function $\mathrm{P_w}$ is defined as $\mathrm{P_w}(w) = [\delta(w-1) + \delta(w+1)]$ (note that this is not a probability distribution because of the normalization), and recall $\varphi(z)$ is the indicator function, checking that a constraint on the argument is satisfied (e.g in the step case, $\varphi^s(z) = 1$ if $z > K$).

While in the step binary perceptron (SBP) the fixed point solution $(q_0,\hat{q}_0)$ is non-trivial, the symmetry of the activation function in the RBP and UBP cases enforces the configuration space to be symmetric and the fixed point $(q_0,\hat{q}_0) = (0,0)$ to exist. If this symmetric fixed point is stable and has the lowest free energy, the RS free entropy matches the annealed entropy $\Phi_t^a(\alpha) = \log(2) + \alpha \log(p_{t,K}) = \lim_{d\to\infty} \frac{1}{d} \log \mathbb{E}_{\mathbf{X}}[\mathcal{Z}_t(\mathbf{X})]$ from section 6.1.1 and Appendix. B.2.2 with $t \in \{r,u\}$.

Rectangle Solving numerically the corresponding saddle point equations leads to the single symmetric fixed point $(q_0,\hat{q}_0) = (0,0)$. Hence the RS entropy saturates the first moment bound:

$$\Phi_r^{(\mathrm{rs})}(\alpha) = \log(2) + \alpha \log(p_{r,K}) = \Phi_r^a(\alpha),$$

and the RS capacity equals the annealed capacity eq. (6.1.1):

$$\alpha_{\mathrm{rs}}^r(K) = \alpha_a^r(K) = \frac{-\log(2)}{\log(p_{r,K})}.$$

U-**function**

- For $K \leq K^*$, only the symmetric fixed point $(q_0,\hat{q}_0) = (0,0)$ exists, which leads again to the annealed free entropy:

$$\Phi_u^{(\mathrm{rs})}(\alpha) = \log(2) + \alpha \log(p_{u,K}) = \Phi_u^a(\alpha),$$

and annealed capacity eq. (6.1.1):

$$\alpha_{\mathrm{rs}}^{u}(K) = \alpha_{a}^{u}(K) = \frac{-\log(2)}{\log\left(p_{u,K}\right)} .$$

- For $K > K^{*}$, the RS entropy does not match the annealed entropy because the fixed point $(q_0, \hat{q}_0) \neq (0,0)$ corresponds to a lower free energy than the symmetric fixed point $(0,0)$. The symmetric fixed point becomes unstable for $K > K^{*}$, where K^{*} is remarkably given by the same value as in the independent section 6.1.2.b. Hence it naturally verifies eq. (124) even though its definition derives from the stability of the RS solution, that we study in the next section.

6.3.1.B RS STABILITY

The local stability of the RS solution can be studied using dAT method (Almeida et al., 1978), based on the positivity of the Hessian of $-\Phi^{(r)}(\mathbf{Q}, \tilde{\mathbf{Q}})$. The replica symmetric dAT-line α_{at} is given by the solution of the following implicit equation, derived in Appendix. B.2.5:

$$\begin{aligned} \frac{1}{\alpha} = & \frac{1}{(1-q_0(\alpha))^2} \mathbb{E}_{\xi_0}\left[\frac{\left(f_0^z(f_0^z - f_2^z) + (f_1^z)^2\right)^2}{(f_0^z)^4}(\xi_0, q_0(\alpha)) \right] \\ & \times \mathbb{E}_{\xi_0}\left[\frac{\left(g_0^w g_2^w - (g_1^w)^2\right)^2}{(g_0^w)^4}(\xi_0, \hat{q}_0(\alpha)) \right] . \end{aligned}$$

As illustrated above, for the rectangle and $u-$function, the symmetry of the weights $\mathrm{P_w}$ and the constraint φ imposes the existence of the symmetric fixed point $(q_0, \hat{q}_0) = (0,0)$. This simplifies the previous condition and becomes equivalent to the linear stability condition of the symmetric fixed point $(q_0, \hat{q}_0) = (0,0)$, see Appendix. B.2.5,

$$\frac{1}{\alpha_{\mathrm{at}}} = \left(\frac{\tilde{f}_2^z - \tilde{f}_0^z}{\tilde{f}_0^z}\right)^2 \left(\frac{\tilde{g}_2^w}{\tilde{g}_0^w}\right)^2 ,$$

where for $i \in \mathbb{N}$

$$\tilde{g}_i^w = \mathbb{E}_{w \sim \mathrm{P_w}}\left[w^i \exp\left(w^2/2\right)\right] , \qquad \tilde{f}_i^z = \mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[z^i \ \varphi(z)\right] .$$

We plot the annealed capacity, the RS capacity and the dAT-line for the step, rectangle and u-function binary perceptrons as functions of K in Fig. 38, 39, 40.

Step binary perceptron We note that for the step binary perceptron the RS solution is always stable towards 1RSB, even for negative threshold $K < 0$. This is interesting in the view of recent work on the spherical perceptron with negative threshold where the replica symmetry breaks for all $K < 0$, and FRSB is needed to evaluate the storage capacity (Franz et al., 2017).

Rectangle As the RS capacity α^r_{rs} is always below the dAT-line α^r_{at}, the RS solution is always locally stable.

U**-function** There is a crossing between the values of the RS capacity α^u_{rs} and the dAT-line α^u_{at}, which defines implicitly the value $K^* \simeq 0.817$, and matches the equality in eq. (129):

$$\frac{-\log(2)}{\log(p_{u,K^*})} = \frac{\pi}{2} \frac{(p_{u,K^*})^2}{\exp(-(K^*)^2)\,(K^*)^2} \,. \tag{137}$$

For $K \leq K^*$, the RS solution is locally stable, while for $K > K^*$ the RS solution becomes unstable, and a symmetry breaking solution appears.

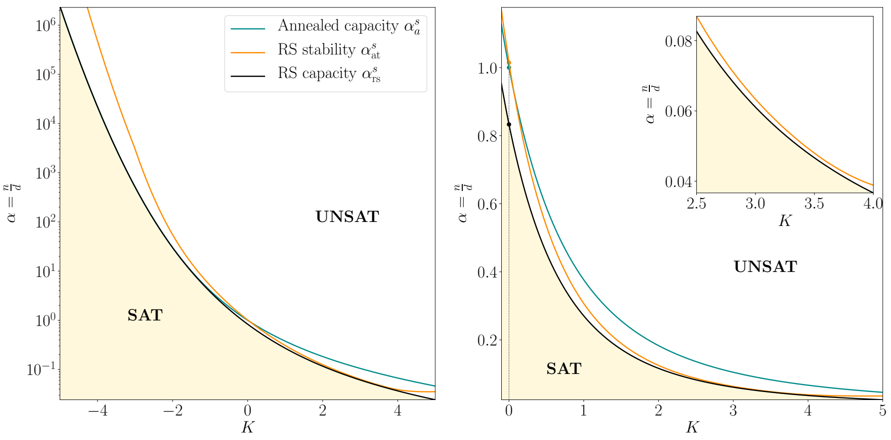


Figure 38: Step binary perceptron (SBP): the RS capacity α^s_{rs} (black) does not match the annealed capacity α^s_a (blue) and is always below the dAT-line α^s_{at} (orange). The dAT-line is closest to the annealed capacity for $K_{\text{min}} \simeq 3.62$ where the difference $\alpha^s_{\text{at}} - \alpha^s_a \simeq 0.0012$. For $K = 0$, we retrieve well known results (Krauth et al., 1989): $\alpha^r_{\text{rs}} \simeq 0.833$, $\alpha^r_{\text{at}} \simeq 1.015$ and $\alpha^r_a = 1$. The left and right hand sides, and the inset, represent the same data on different scales. The satisfiable (SAT) phase is represented by the beige shaded area and is located below the RS capacity, while the unsatisfiable (UNSAT) starts at the capacity (black line) and extends for a larger number of constraints.

6.3.2 1RSB CALCULATION AND STABILITY

6.3.2.A 1RSB ENTROPY

In the previous section we concluded that the replica symmetric solution is unstable in the u−function binary perceptron for $K > K^*$, we analyze therefore the first-step of replica symmetry breaking 1RSB Ansatz in this section. This ansatz and calculations is due to seminal works of G. Parisi and is classic in the field of disordered systems and well presented in the literature (Mézard et al., 1987; Parisi, 1979; Parisi, 1980c; Parisi, 1980d), we thus mainly give the key formulas and defer the details into Appendix. B.2.6.

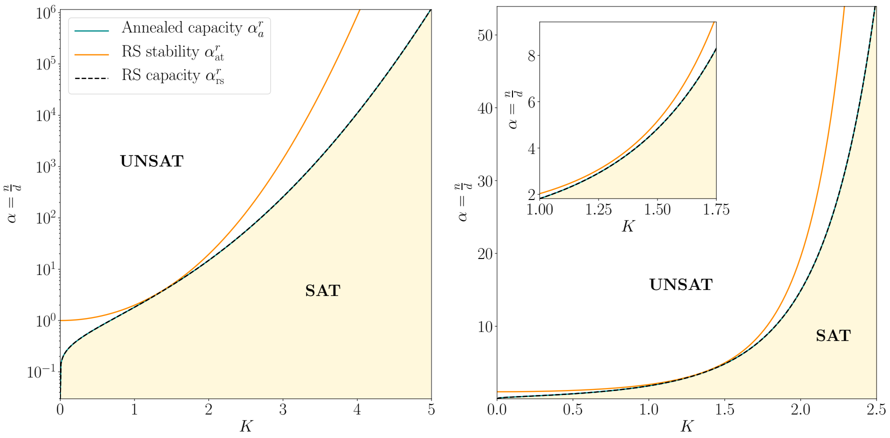


Figure 39: Rectangle binary perceptron (RBP): the RS capacity α_{rs}^r (black) matches the annealed bound α_a^r (blue), and the RS solution is locally stable for all K: $\alpha_{\text{rs}}^r < \alpha_{\text{at}}^r$. The dAT-line (orange) is closest to the annealed capacity for $K_{\text{min}} \simeq 1.24$ where the difference $\alpha_{\text{at}}^s - \alpha_a^s \simeq 0.15$. The left and right hand sides, and the inset, represent the same data on different scales. The satisfiable (SAT) phase is represented by the beige shaded area and is located below the RS capacity, while the unsatisfiable (UNSAT) starts at the capacity (black line) and extends for a larger number of constraints.

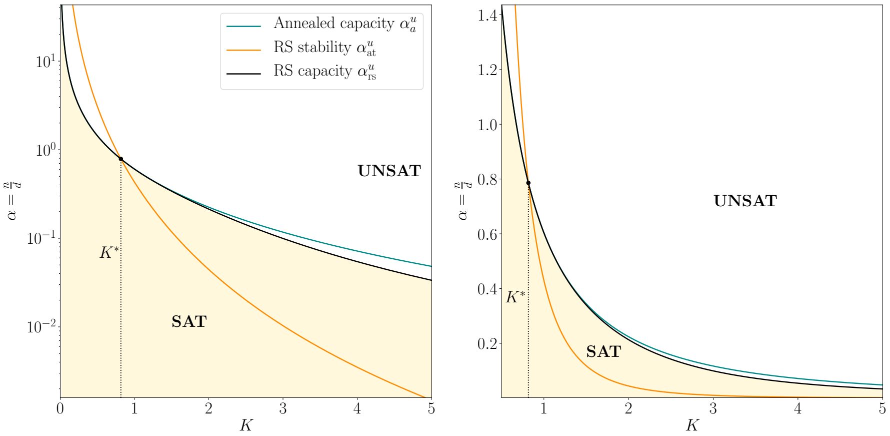


Figure 40: $U-$function binary perceptron (UBP): the RS capacity Â§black) matches the annealed bound (blue) for $K < K^*$. At $K = K^*$, the RS capacity crosses the dAT-line (orange). For $K > K^*$, the RS solution is unstable and the RS capacity deviates from the annealed capacity. The left and right hand sides, and the inset, represent the same data on different scales. The satisfiable (SAT) phase is represented by the beige shaded area and is located below the RS capacity, while the unsatisfiable (UNSAT) starts at the capacity (black line) and extends for a larger number of constraints.

The 1RSB Ansatz assumes that the space of configurations splits into states. Consequently replicas are not symmetric anymore and instead r replicas are organized in $\frac{r}{m}$ groups containing m replicas each:

$$\forall (a,b) \in [\![r]\!]^2, \qquad \frac{1}{d}(\mathbf{w}^a \cdot \mathbf{w}^b) = \begin{cases} q_1 \text{ if } a,b \text{ belong to the same state,} \\ q_0 \text{ if } a,b \text{ do not belong to the same state,} \\ Q = 1 \text{ if } a = b\,. \end{cases}$$

(138)

Following (Monasson et al., 1995a), the replicated partition function $\mathcal{Z}_m$ associated to m replicas falling in the same state is expressed as a sum over all possible states ψ weighted by their corresponding free entropy Φ:

$$\begin{aligned}\mathcal{Z}_m &= \sum_{\{\psi\}} \exp(dm\Phi(\psi)) = \sum_{\{\Phi\}} \mathcal{N}_\Phi \exp(dm\Phi) \\ &= \sum_{\{\Phi\}} \exp(d\Sigma(\Phi)) \exp(dm\Phi) \sim \int \mathrm{d}\Phi \exp(d(m\Phi + \Sigma(\Phi)))\end{aligned} \tag{139}$$

where we introduced the number of states at a given free entropy Φ: $\mathcal{N}_\Phi \equiv \exp(d\Sigma(\Phi))$ and the complexity $\Sigma(\Phi)$, also called the configurational entropy. Using the saddle point method in the $d \to \infty$ limit, the 1RSB replicated free entropy $\Phi_m^{(\text{1rsb})}$ is written as a function of the Parisi parameter m, the free entropy Φ and the complexity $\Sigma(\Phi)$:

$$\Phi_m^{(\text{1rsb})}(m, \alpha) \equiv \lim_{d \to \infty} \frac{1}{d} \mathbb{E}_{\mathbf{X}} \left[\log(\mathcal{Z}_m(\mathbf{X}))\right] = m\Phi + \Sigma(\Phi) \,. \tag{140}$$

Injecting the 1RSB ansatz eq. (138), the 1RSB replicated free entropy $\Phi_m^{(\text{1rsb})} = m\Phi^{(\text{1rsb})}$ is written as a saddle point equation over $\mathbf{q} = (q_0, q_1)$ and $\tilde{\mathbf{q}} = (\hat{q}_0, \hat{q}_1)$, see Appendix. B.2.6:

$$\begin{aligned}\Phi_m^{(\text{1rsb})}(m, \alpha) = m \cdot \underset{\mathbf{q}, \hat{\mathbf{q}}}{\mathbf{extr}} \Bigg\{ & \frac{1}{2}\left(q_1 \hat{q}_1 - Q\hat{Q}\right) + \frac{m}{2}\left(q_0 \hat{q}_0 - q_1 \hat{q}_1\right) \\ & + \tilde{\Psi}_{\mathrm{w}}^{(\text{1rsb})}(\hat{\mathbf{q}}, m) + \alpha \tilde{\Psi}_{\text{out}}^{(\text{1rsb})}(\mathbf{q}, m) \Bigg\} \,,\end{aligned} \tag{141}$$

with

$$\begin{aligned}\tilde{\Psi}_{\mathrm{w}}^{(\text{1rsb})}(\hat{\mathbf{q}}, m) &\equiv \frac{1}{m} \mathbb{E}_{\xi_0} \log\left(\mathbb{E}_{\xi_1} \left(g_0^w\right)^m\right) \,, \\ \tilde{\Psi}_{\text{out}}^{(\text{1rsb})}(\mathbf{q}, m) &\equiv \frac{1}{m} \mathbb{E}_{\xi_0} \log\left(\mathbb{E}_{\xi_1} \left(f_0^z\right)^m\right) \,,\end{aligned} \tag{142}$$

and

$$\begin{aligned}g_i^w(\boldsymbol{\xi}, \mathbf{q}) &= \mathbb{E}_w \left[w^i \exp\left(\frac{(1 - \hat{q}_1)}{2} w^2 + \left(\sqrt{\hat{q}_0}\xi_0 + \sqrt{\hat{q}_1 - \hat{q}_0}\xi_1 \right) w \right) \right] \,, \\ f_i^z(\boldsymbol{\xi}, \mathbf{q}) &= \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[z^i \varphi\left(\sqrt{1 - q_1} z + \sqrt{q_0}\xi_0 + \sqrt{q_1 - q_0}\xi_1 \right) \right] \,,\end{aligned} \tag{143}$$

for $\boldsymbol{\xi} = (\xi_0, \xi_1)$ and for $i \in \mathbb{N}$. Taking the derivative of $\Phi_m^{(1\mathrm{rsb})}$ with respect to m, the 1RSB free entropy $\Phi^{(1\mathrm{rsb})}$ and complexity Σ can be expressed as:

$$\begin{aligned}\Phi^{(1\mathrm{rsb})}(\alpha) &= \frac{\partial \Phi_m^{(1\mathrm{rsb})}(m,\alpha)}{\partial m} \\ &= \underset{\mathbf{q},\tilde{\mathbf{q}},m}{\mathbf{extr}} \left\{ \frac{1}{2}(q_1\hat{q}_1 - 1) + m(q_0\hat{q}_0 - q_1\hat{q}_1) \right. \\ &\qquad \left. + \Psi_{\mathrm{w}}^{(1\mathrm{rsb})}(\tilde{\mathbf{q}},m) + \alpha \Psi_{\mathrm{out}}^{(1\mathrm{rsb})}(\mathbf{q},m) \right\}, \end{aligned} \tag{144}$$

$$\begin{aligned}\Sigma(\Phi^{(1\mathrm{rsb})}) &= \Phi_m^{(1\mathrm{rsb})} - m\Phi^{(1\mathrm{rsb})} \\ &= \underset{\mathbf{q},\tilde{\mathbf{q}},m}{\mathbf{extr}} \left\{ \frac{m^2}{2}(q_1\hat{q}_1 - q_0\hat{q}_0) + m\left(\tilde{\Psi}_{\mathrm{w}}^{(1\mathrm{rsb})} - \Psi_{\mathrm{w}}^{(1\mathrm{rsb})}\right)(\tilde{\mathbf{q}},m) \right. \\ &\qquad \left. + m\alpha\left(\tilde{\Psi}_{\mathrm{out}}^{(1\mathrm{rsb})} - \Psi_{\mathrm{out}}^{(1\mathrm{rsb})}\right)(\mathbf{q},m) \right\}, \end{aligned} \tag{145}$$

with

$$\begin{aligned}\Psi_{\mathrm{w}}^{(1\mathrm{rsb})}(\tilde{\mathbf{q}},m) &= \partial_m\left(m\tilde{\Psi}_{\mathrm{w}}^{(1\mathrm{rsb})}\right) = \mathbb{E}_{\xi_0}\left[\frac{\mathbb{E}_{\xi_1}\left[\log\left(g_0^w(\boldsymbol{\xi},\mathbf{q})\right) g_0^w(\boldsymbol{\xi},\mathbf{q})^m\right]}{\mathbb{E}_{\xi_1}\left[g_0^w(\boldsymbol{\xi},\mathbf{q})^m\right]}\right], \\ \Psi_{\mathrm{out}}^{(1\mathrm{rsb})}(\mathbf{q},m) &= \partial_m\left(m\tilde{\Psi}_{\mathrm{out}}^{(1\mathrm{rsb})}\right) = \mathbb{E}_{\xi_0}\left[\frac{\mathbb{E}_{\xi_1}\left[\log\left(f_0^z(\boldsymbol{\xi},\mathbf{q})\right) f_0^z(\boldsymbol{\xi},\mathbf{q})^m\right]}{\mathbb{E}_{\xi_1}\left[f_0^z(\boldsymbol{\xi},\mathbf{q})^m\right]}\right]. \end{aligned}$$

6.3.2.B 1RSB RESULTS FOR UBP

From now on, we only consider the $u-$function binary perceptron, whose RS solution is unstable for $K > K^*$. To describe the equilibrium of the system in the SAT phase, we need to find the value of the Parisi parameter at equilibrium m_{eq}. The complexity $\Sigma(\Phi)$ is the entropy of clusters having internal entropy Φ. In order to capture clusters that carry almost all configurations, we need to maximize the total entropy $\Phi_{\mathrm{tot}} = \Sigma(\Phi) + \Phi$ under the constraint that the free entropy and complexity are both positive $\Phi \geq 0$ and $\Sigma(\Phi) \geq 0$. Hence from eq. (140), the equilibrium Parisi parameter m_{eq} verifies

$$\Phi_{\mathrm{eq}} = \underset{\Phi\geq 0, \Sigma \geq 0}{\mathrm{argmax}} \left\{\Phi + \Sigma(\Phi)\right\} \qquad \text{and} \qquad m_{\mathrm{eq}} = -\left.\frac{d\Sigma}{d\Phi}\right|_{\Phi_{\mathrm{eq}}}. \tag{146}$$

As a side remark, we note that in the Parisi's replica theory the more commonly known condition for m_{eq} is obtained by extremizing the (rescaled) replicated free entropy $\Phi_m^{(1\mathrm{rsb})}(m,\alpha)/m$ which leads using (140) to the condition $-\frac{\Sigma(\Phi_{\mathrm{eq}})}{m^2} = 0$. This extrema is in fact a minima as $\Sigma(\Phi)$ is concave and $m = -\frac{d\Sigma}{d\Phi}$. This is, however, only valid when $m_{\mathrm{eq}} < 1$, and is moreover highly counter-intuitive as physical systems maximize entropy whereas here one minimizes it. We hence prefer to use the formulation of eq. (146) which we find physically better justified. Using the expressions eq. (145) and varying the Parisi parameter $m \in [0;1]$, we obtain the curve of the complexity $\Sigma(\Phi)$ as shown in Fig. 41. At $m = 1$, the complexity is negative. Decreasing m, the complexity increases and becomes positive at the value m_{eq}. Besides for small

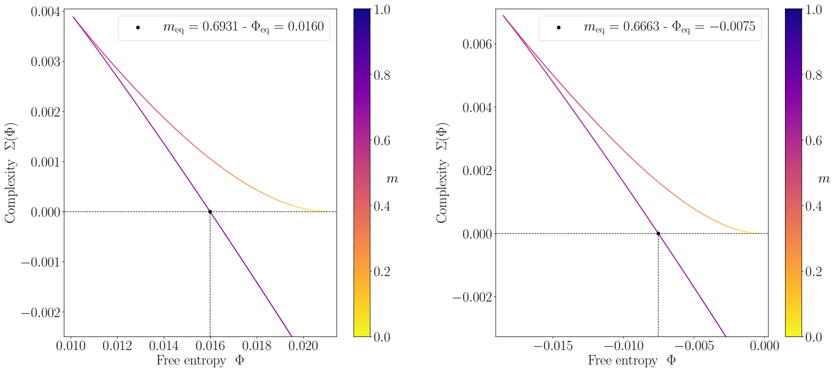


Figure 41: Complexity $\Sigma(\Phi)$ as a function of the free entropy Φ for the $u-$function binary perceptron at $K = 1.5 > K^*$. Complexity reaches $\Sigma = 0$ (black dot) at m_{eq}. For $K = 1.5$ and $\alpha = 0.33$ (**Left**) the free-entropy corresponding to m_{eq} is positive $\Phi_{\mathrm{eq}} > 0$, whereas for $\alpha = 0.34$, (**Right**) the free entropy at m_{eq} is negative $\Phi_{\mathrm{eq}} < 0$ and therefore there is no part of the curve where both complexity and free entropy are positive: thus this value of α is beyond the 1RSB storage capacity, and the capacity is in the interval $[0.33; 0.34]$.

values of m, an unphysical (convex) branch appears, as commonly observed in other systems solved by the replica method.

We note that as α increases both the equilibrium complexity and free entropy decrease. In CSP such as k-SAT or random graph coloring the mechanism in which the satisfiability threshold appears is that the maximum of the complexity becomes negative. In the present UBP problem it is actually both the free entropy and the complexity that vanish together, as illustrated in Fig. 41.

Computing the equilibrium value $m_{\mathrm{eq}}(\alpha)$, we have access to the corresponding equilibrium overlaps q_0^* and q_1^*, that we may compare with the RS solution q_{rs}. All these are depicted in Fig. 42. The function $m_{\mathrm{eq}}(\alpha)$ shows a non monotonic behaviour as it has been previously observed, e.g. in the SK model as a function of temperature (Mézard et al., 1987). We also compute the 1RSB entropy that verifies $\Phi_u^{(\mathrm{1rsb})} \leq \Phi_u^{(\mathrm{rs})}$ and which vanishes at the 1RSB capacity α_{1rsb}^u as depicted in Fig. 43 (**Left**). We note that the above inequality is as predicted by Parisi's replica theory (Mézard et al., 1987), taking into account that we are working at strictly zero energy, where the entropy becomes minus the free energy. The 1RSB solution provides a small correction to the RS result for storage capacity, as illustrated in Fig. 43 (**Right**), where we plotted the difference between the annealed upper bound and the capacity for the RS and 1RSB solutions: $\alpha_a^u - \alpha_{\mathrm{rs}}^u$ and $\alpha_a^u - \alpha_{\mathrm{1rsb}}^u$.

6.3.2.c 1RSB STABILITY

In the previous section we evaluated the 1RSB storage capacity of the $u-$function binary perceptron for $K > K^*$. In this section we will argue that this cannot be an exact solution to the problem. We could investigate the stability of 1RSB

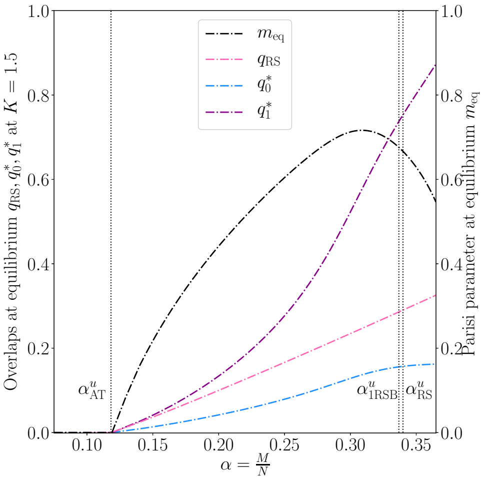


Figure 42: Equilibrium values of the overlap $q_0^* \neq q_{\mathrm{rs}}$, q_1^* and the Parisi parameter m_{eq} for the UBP at $K = 1.5$. For $K < K^*$, the RS solution is stable and the only fixed point is $q_0^* = q_1^* = q_{\mathrm{rs}} = 0$.

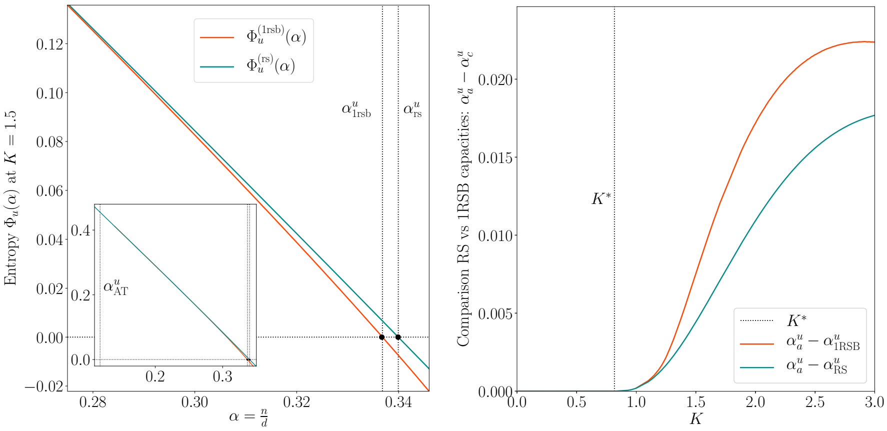


Figure 43: (**Left**) Comparison of the RS (blue) and 1RSB (orange) entropy for the UBP at $K = 1.5$. For $\alpha < \alpha_{\mathrm{at}} \simeq 0.118$, RS and 1RSB entropies are equalled. For $\alpha > \alpha_{\mathrm{at}}$, 1RSB entropy deviates slightly of the RS entropy before vanishing respectively at $\alpha_{\mathrm{1rsb}}^u \simeq 0.337$ and $\alpha_{\mathrm{rs}}^u \simeq 0.334$. The inset represents the same data on a different scale. (**Right**) Difference between the annealed upper bound and the 1RSB capacity $\alpha_a^u - \alpha_{\mathrm{1rsb}}^u$ (orange) and the RS capacity $\alpha_a^u - \alpha_{\mathrm{rs}}^u$ (blue). Below K^* the RS solution is stable: RS and 1RSB entropies match exactly. Above K^*, the RS solution is unstable: the 1RSB entropy deviates slightly from the RS solution.

towards further levels of replica symmetry breaking along the same lines we did for the RS solution. However, in the present case we do not need to do that to see that the obtained solution cannot be correct. The explanations lies in the breaking of the up-down symmetry in the problem. This symmetry must either be broken explicitly as in the ferromagnet, where the system would acquire an overall magnetization, but we have not observed any trace of this in the present problem. Or this up-down symmetry must be conserved in the final correct solution. The conservation of the up-down symmetry is manifested in the value $q_0 = 0$ in the replica symmetric phase. The fact that in the 1RSB solution evaluated above we do not observe $q_0 = 0$, but instead $q_0 > 0$ is a sign of the fact that we are evaluating a wrong solution. The only possible way to obtain an exact solution we foresee is to evaluate the full-step replica symmetry breaking with a continuity of overlaps $q(x)$, the smallest one of them should be 0 in order to restore the up-down symmetry. We let the evaluation of the FRSB for future work.

Finally let us note that the 1RSB solution obtained in the previous section can be interpreted as frozen-2RSB. In 2RSB we would have 3 kinds of overlaps, q_0, q_1 and q_2. In frozen 2RSB we would have $q_2 = 1$, $q_1 = q_1^{\text{1rsb}}$, $q_0 = q_0^{\text{1rsb}}$.

CONCLUSION

In this chapter we analyzed a class of symmetric binary perceptron problems that are simple variants of the canonical step-function binary perceptron. The step-function binary perceptron has thus far eluded a rigorous establishment of the conjectured storage capacity, eq. (121). This prediction is expected to be exact because of the f1RSB nature of the problem (Krauth et al., 1989). At the same time the work of (Baldassi et al., 2015) sheds light on the fact that the structure of the space of solutions is not fully described by the f1RSB picture, and that rare dense and unfrozen regions exist and in fact are amenable to dynamical procedures searching for solutions. It remains to be understood how is it possible that the 1RSB calculation does not capture these dense unfrozen regions of solutions (Baldassi et al., 2015). They do not dominate the equilibrium, but the RSB calculation is expected to describe rare events via their large deviations, which in this case it does not.

We focus on two cases of the binary perceptron with symmetric constraints, the rectangle binary perceptron and the $u-$function binary perceptron. We prove (up to a numerical assumption) using the second moment method that the storage capacity agrees in those cases with the annealed upper bound, except for the $u-$function binary perceptron for $K > K^*$ eq. (124). We analyze the 1RSB solution in that case and indeed obtain a lower prediction for the storage capacity. However, we do not expect the 1RSB to provide the exact solution because it does not respect the up-down symmetry of the problem. Though the precise nature of the satisfiable phase for the $u-$function binary perceptron for $K > K^*$ remains illusive, we can conjecture it is FRSB (Parisi, 1979; Parisi, 1980c; Parisi, 1980d). Establishing this rigorously would provide much deeper understanding and remains a challenging subject for future work.

7 RADEMACHER COMPLEXITY AND SPIN GLASSES: A LINK BETWEEN THE REPLICA AND STATISTICAL THEORIES OF LEARNING

ERM is the workhorse of most of modern supervised machine learning successes. Consider for instance a data-set $\{y_\mu, \mathbf{x}_\mu\}_{\mu=1}^{n}$ of n examples $\mathbf{x}_\mu \in \mathbb{R}^d$ assumed to be drawn from a distribution $\mathrm{P}_x(.)$, with labels $y_\mu \in \{-1,+1\}$ used for a binary classification task. We consider an estimator $f_{\mathbf{w}}(.)$ that belongs to a *hypothesis class* $\mathscr{F}$, for instance a neural network or a linear function, with respective weights or parameters $\mathbf{w}$. The latter are typically computed by minimizing the empirical risk

$$\mathscr{R}^n_{\text{empirical}}(f_{\mathbf{w}}) = \frac{1}{n}\sum_{\mu=1}^{n} \mathscr{L}\left(y_\mu, f_{\mathbf{w}}\left(\mathbf{x}_\mu\right)\right)$$

over $\mathbf{w}$, where $\mathscr{L}$ denotes a loss function, e.g. the mean-squared-loss $\mathscr{L}(a,b) = (a-b)^2$. The main theoretical issue of statistical learning theory concerns the performance of the estimator $f_{\mathbf{w}}(.)$ obtained by such a minimization on yet unseen data, namely the *generalization problem*. In fact, what we really hope to minimize is the population risk, defined as

$$\mathscr{R}_{\text{population}}(f_{\mathbf{w}}) = \mathbb{E}_{y,\mathbf{x}}\left[\mathscr{L}(y, f_{\mathbf{w}}(\mathbf{x}))\right].$$

Since we are optimizing the empirical risk instead, the difference between the two might be arbitrarily large. Bounding this difference between *empirical* and *population* risks is therefore a major problem of statistical learning theories.

In a large part of the literature, statistical learning analysis, see e.g. (Bartlett et al., 2002; Vapnik, 2013; Shalev-Shwartz et al., 2014) relies on the VC analysis and on the so-called *Rademacher complexity*. The latter is a measure of the complexity of $\mathscr{F}$, the hypothesis class spanned by $f_{\mathbf{w}}(.)$, to bound $\mathscr{R}_{\text{population}} - \mathscr{R}^n_{\text{empirical}}$, the *generalization gap*. A gem within the literature is the Uniform Convergence result which states the following: if the Rademacher complexity or the VC dimension is finite, then for a large enough number of samples the generalization gap will vanish uniformly over all possible values of parameters $\mathbf{w}$. Informally, uniform convergence tells us that with high probability, for any weights value $\mathbf{w}$, the generalization gap satisfies

$$\mathscr{R}_{\text{population}}(f_{\mathbf{w}}) - \mathscr{R}^{n}_{\text{empirical}}(f_{\mathbf{w}}) = \Theta\left(\sqrt{\frac{d_{\text{VC}}(\mathscr{F})}{n}}\right), \tag{147}$$

where $d_{\text{VC}}(\mathscr{F})$ denotes the VC dimension of the hypothesis class $\mathscr{F}$. Tighter bounds can be obtained using the Rademacher complexity. These bounds, although useful, do not seem to fully explain the success of current deep-learning architectures (Zhang et al., 2016).

Over the last four decades, a different vision of generalization — based on the analysis of *typical case* problems with synthetic data created from simple generative models — was developed to a large extent in the statistical physics literature, see e.g. (Seung et al., 1992; Watkin et al., 1993; Opper, 1995; Engel et al., 2001) for a review. The link with the VC dimension was discussed in many of these works, notably via its connection with its twin from statistical physics, the *Gardner capacity* (Gardner et al., 1988). In particular, one can show that the VC capacity is always larger than half of the Gardner one (Engel et al., 2001). We shall review this discussion later on in this paper. However, to the best of our knowledge the Rademacher complexity was absent from these considerations. This omission is unfortunate: not only does the Rademacher complexity give tighter bounds than the VC dimension, it also intrinsically connects with a quantity that physicists are familiar with and have been computing from the very beginning of their studies, namely the average *ground-state energy*.

The goal of the present chapter is to bridge this gap and unveil the deep link between ground-state energy and Rademacher complexity, and how this connection is valuable to both parties. The chapter is organized as follows: After giving proper definitions of common generalization bounds in sec. 7.1, we detail calculations of Rademacher complexities for simple function classes in sec. 7.2. These sections serve as an introduction to the readers not familiar with these notions. The subsequent sections 7.3 and 7.4 provide the original content of this work.

Here we summarize the main contributions of this work:

- We point out the one-to-one connections between the Rademacher complexity in statistical learning, and the ground-state energies and Gardner capacity from statistical physics.
- We show how the heuristic replica method from statistical physics can be used to compute the Rademacher complexity in the high-dimensional statistics limit and reinterpret classical results of the statistical physics literature as Rademacher bounds in the case of perceptron and committee machines models with i.i.d data.
- We contrast these results with the generalization in the teacher-student scenario, illustrating the worst-case nature of the Rademacher bound that fails to capture the typical-case behavior.

- We finally show *en passant*, that learning theory also bears consequences for the spin glass physics and the related replica symmetry breaking scheme by showing it implies strong constraint on the ground-state energy of some spin glass models.

7.1 A PRIMER ON RADEMACHER COMPLEXITY

The bound of the generalization gap involving the VC dimension is specific to binary classification, and does not depend on the data distribution. While this is a strong property, the Rademacher approach does depend on data distribution and allows for tighter bounds. Moreover, it generalizes to multi-class classification and regression problems. We recall the definition of the Rademacher complexity:

Definition 7.1.1. *Let $f_{\mathbf{w}}$ be any function in the hypothesis class $\mathscr{F}$, and let $\boldsymbol{\varepsilon} \in \{\pm 1\}^n$ be drawn uniformly at random. The* ***empirical Rademacher complexity*** *is defined as*

$$\hat{\mathfrak{R}}_n\left(\mathscr{F},\mathbf{X}\right) \equiv \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f_{\mathbf{w}}\in\mathscr{F}} \frac{1}{n}\sum_{\mu=1}^{n} \varepsilon_\mu f_{\mathbf{w}}\left(\mathbf{x}_\mu\right)\right], \tag{148}$$

and depends on the sample examples $\mathbf{X} = \{\mathbf{x}_1,\ldots \mathbf{x}_n\} \in \mathbb{R}^{d\times n}$. The ***Rademacher complexity*** *is defined as the population average*

$$\mathfrak{R}_n\left(\mathscr{F}\right) \equiv \mathbb{E}_{\mathbf{X}}\left[\hat{\mathfrak{R}}_n\left(\mathscr{F},\mathbf{X}\right)\right]. \tag{149}$$

In this chapter, we focus on binary classification and consider the corresponding loss function $\mathscr{L}(a,b) = \mathbb{1}\left[a \neq b\right]$ that counts the number of misclassified samples. We will be therefore interested in a hypothesis class $\mathscr{F} = \left\{f_{\mathbf{w}} : \mathbb{R}^d \longrightarrow \{\pm 1\}\right\}$. Defining the training $\varepsilon^n_{\text{train}}(.)$ and generalization errors $\varepsilon_{\text{gen}}(.)$ for any function $f_{\mathbf{w}} \in \mathscr{F}$ by

$$\begin{aligned} \varepsilon^n_{\text{train}}(f_{\mathbf{w}}) &\equiv \frac{1}{n}\sum_{\mu=1}^{n} \mathbb{1}\left[y_\mu \neq f_{\mathbf{w}}\left(\mathbf{x}_\mu\right)\right], \\ \varepsilon_{\text{gen}}(f_{\mathbf{w}}) &\equiv \mathbb{E}_{y,\mathbf{x}}\left[\mathbb{1}\left[y \neq f_{\mathbf{w}}\left(\mathbf{x}\right)\right]\right], \end{aligned} \tag{150}$$

the Rademacher complexity provides a generalization error bound as expressed by the following theorem, and many of its variants, see e.g. (Bartlett et al., 2002; Vapnik, 2013; Shalev-Shwartz et al., 2014; Mohri et al., 2012):

Theorem 7.1.2. *Uniform convergence bound - Binary classification*
Fix a distribution P_x and let $\delta > 0$. Let $\mathbf{X} = \{\mathbf{x}_1,\ldots \mathbf{x}_n\} \in \mathbb{R}^{d\times n}$ be drawn i.i.d from P_x. Then with probability at least $1-\delta$ (over the draw of $\mathbf{X}$),

$$\forall f_{\mathbf{w}} \in \mathscr{F}, \varepsilon_{gen}(f_{\mathbf{w}}) - \varepsilon^n_{\text{train}}(f_{\mathbf{w}}) \leq \mathfrak{R}_n(\mathscr{F}) + \sqrt{\frac{\log(1/\delta)}{n}}. \tag{151}$$

Thus, the Rademacher complexity is a uniform bound of the generalization gap. In the high-dimensional limit when both n and d goes to infinity that we will consider in the remaining of the paper, we shall see that we can discard the $\delta-$dependent term and that only the first term will remains finite. Note that this theorem can be used to recover the classical result (147). Indeed it can be shown (Massart, 2000; Ledoux et al., 2013; Dudley, 1967) that the Rademacher complexity can be bounded by the VC dimension so that for some constant value C,

$$\mathfrak{R}_n(\mathscr{F}) \leq C\sqrt{\frac{d_{\mathrm{vc}}(\mathscr{F})}{n}}. \tag{152}$$

We remind the reader that the VC dimension is the size of the set that can be fully shattered by the hypothesis class $\mathscr{F}$. Informally, if $n > d_{\mathrm{vc}}$ then for all set of n data points, there exists an assignment of labels that cannot be fully fitted by the function class (Vapnik, 2013).

Proof. Applying Massart's lemma (Massart, 2000) for $\mathscr{F}_{\mathbf{X}} = \{f(\mathbf{x}_1),..,f(\mathbf{x}_n)\} \subset \mathbb{R}^n$ with $f : \mathbb{R}^d \longrightarrow \{\pm 1\}$. Hence $\sup_{\mathbf{x}\in\mathscr{F}_{\mathbf{X}}} \|\mathbf{x}\|_2 = \sqrt{n}$ and it follows

$$\begin{aligned}
\mathfrak{R}_n(\mathscr{F}) &\equiv \mathbb{E}_{\mathbf{X}}\left[\hat{\mathfrak{R}}_n(\mathscr{F},\mathbf{X})\right] \leq \mathbb{E}_{\mathbf{X}}\left[\sup_{\mathbf{x}\in\mathscr{F}_{\mathbf{X}}} \|\mathbf{x}\|_2 \frac{\sqrt{2\log|\mathscr{F}_{\mathbf{X}}|}}{n}\right] \\
&\leq \mathbb{E}_{\mathbf{X}}\left[\sqrt{\frac{2\log\max_{\{\mathbf{x}_1,..\mathbf{x}_n\}}|\mathscr{F}_{\mathbf{X}}|}{n}}\right] \\
&= \sqrt{\frac{2\log\Pi_{\mathscr{F}}(n)}{n}} \leq \Theta\left(\sqrt{\frac{d_{\mathrm{vc}}(\mathscr{F})}{n}}\right),
\end{aligned}$$

where $\Pi_{\mathscr{F}}(n) \equiv \max_{\{\mathbf{x}_1,..\mathbf{x}_n\}}\{f(\mathbf{x}_1),..,f(\mathbf{x}_n)\} \leq 2^n$ is the growth function of the hypothesis class $\mathscr{F}$. The last inequality comes from the fact that the VC dimension of the hypothesis class $\mathscr{F}$ is defined as the maximum sample size dataset that can be shattered $d_{\mathrm{vc}}(\mathscr{F}) = \max_n\{n : \Pi_{\mathscr{F}}(n) = 2^n\}$. □

7.2 SYNTHETIC MODELS IN THE HIGH DIMENSIONAL STATISTICS LIMIT

In this section, we consider data generated by a simple generative model. We suppose that each vector of input data points $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathbb{R}^{d\times n}$ has been generated i.i.d from a factorized, e.g. Gaussian, distribution, that is $\forall\mu \in [\![n]\!], \mathrm{P}_x(\mathbf{x}_\mu) = \prod_{i=1}^d \mathrm{P}_x(x_{i\mu})$. In the following, we will focus on this simple data distribution, but sec. 7.4.5 presents a generalization to rotationally invariant data matrices $\mathbf{X}$ with arbitrary spectrum. The main interest of such settings is to use the analysis of *typical case* problems with synthetic data created from simple generative models as means of getting additional insight on real world applications where data are not worst case (Seung et al., 1992; Watkin et al., 1993; Opper, 1995; Engel et al., 2001; Zdeborová et al., 2016a). In particular, we shall be interested in the high-dimensional statistics limit when $n, d \longrightarrow \infty$, with $\alpha = \frac{n}{d} = \Theta(1)$. In the following, the aim is to compute exactly, rather than merely bounding, and asymptotically the Rademacher complexity for such problems.

7.2.1 LINEAR MODEL

As the simplest example, we first tackle the computation of the Rademacher complexity for a simple function class containing all linear models with weights $\mathbf{w} \in \mathbb{R}^d$,

$$\mathscr{F}_{\text{linear}} = \left\{ f_{\mathbf{w}} : \begin{cases} \mathbb{R}^d \longrightarrow \mathbb{R} \\ \mathbf{x} \longrightarrow \frac{1}{\sqrt{d}} \mathbf{w}^\intercal \mathbf{x} \end{cases}, \mathbf{w} \in \mathbb{R}^d \;\; / \;\; \|\mathbf{w}\|_2 = \Gamma \sqrt{d} \right\} . \tag{153}$$

From eq. (149), computing the empirical Rademacher complexity amounts to finding the vector $\mathbf{w}^\star$ that maximizes the scalar product between $\mathbf{y}$ (that replaces the variable $\boldsymbol{\varepsilon}$) and $\mathbf{X}^\intercal \mathbf{w}$. It is thus sufficient to take $\mathbf{w}^\star = \frac{\mathbf{X}\mathbf{y}}{\|\mathbf{X}\mathbf{y}\|_2} \|\mathbf{w}\|_2$ and the empirical Rademacher complexity (149) thus reads

$$\begin{aligned} \mathfrak{R}_n\left(\mathscr{F}_{\text{linear}}\right) &= \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\sup_{f \in \mathscr{F}_{\text{linear}}} \frac{1}{n} \sum_{\mu=1}^{n} y_\mu f\left(\mathbf{x}_\mu\right) \right] \qquad (154) \\ &= \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\frac{1}{n} \mathbf{y}^\intercal \left(\frac{1}{\sqrt{d}} \mathbf{X}^\intercal \mathbf{w}^\star \right) \right] = \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\frac{1}{n} \mathbf{y}^\intercal \left(\frac{1}{\sqrt{d}} \mathbf{X}^\intercal \frac{\mathbf{X}\mathbf{y}}{\|\mathbf{X}\mathbf{y}\|_2} \|\mathbf{w}\|_2 \right) \right] \\ &= \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\frac{1}{n} \frac{1}{\sqrt{d}} \|\mathbf{X}\mathbf{y}\|_2 \|\mathbf{w}\|_2 \right] . \end{aligned}$$

$\mathbf{X}$ having i.i.d entries, we can apply the CLT, which enforces $\forall i \in [\![d]\!]$, $(\mathbf{X}\mathbf{y})_i = \sum_{\mu=1}^{n} x_{i\mu} y_\mu \sim \mathcal{N}(0, n)$ hence $\mathbb{E}_{\mathbf{y},\mathbf{X}} \|\mathbf{X}\mathbf{y}\|_2 = \sqrt{dn}$. Assuming that weights are restricted to lie on the sphere of radius Γ in $\mathbb{R}^d$, we set $\|\mathbf{w}\|_2 = \Gamma\sqrt{d}$ and finally obtain

$$\mathfrak{R}_n\left(\mathscr{F}_{\text{linear}}\right) = \frac{\Gamma}{\sqrt{\alpha}} , \tag{155}$$

where recall $\alpha = \frac{n}{d}$. The above result for the simple linear function hypothesis class allows to grasp the meaning of the Rademacher complexity: At fixed input dimension d, it decreases with the number of samples as $\alpha^{-1/2}$, closing the generalization gap in the infinite α limit. Illustrating the bias-variance trade-off, we also see that increasing the radius of the weights expands the function complexity (and might help for fitting the data-set), but unfortunately leads to a looser generalization bound.

Note also that the fact that the Rademacher complexity is $\Theta(\alpha^{-1/2})$ shows that it remains finite in the high-dimensional statistics limit. In this case, we see indeed that we can disregard the term $\sqrt{\log(1/\delta)/n}$ that goes to zero as $n \to \infty$ in eq. (151).

7.2.2 PERCEPTRON MODEL

The scaling of Rademacher complexity inverse as $\sqrt{\alpha}$ in the high-dimensional statistics limit is actually *not* restricted to the linear model but appears to be a universal property, at least at large enough α. To see this we now focus on a different hypothesis class: the perceptron, denoted $\mathscr{F}_{\text{sign}}$. This class contains linear classifiers which output binary variables, and will fit much better labels in the binary classification task. The class writes

$$\mathscr{F}_{\text{sign}} = \left\{ f_{\mathbf{w}} : \begin{cases} \mathbb{R}^d \longrightarrow \{\pm 1\} \\ \mathbf{x} \longrightarrow \text{sign}\left(\frac{1}{\sqrt{d}}\mathbf{w}^\intercal \mathbf{x}\right) \end{cases}, \mathbf{w} \in \mathbb{R}^d \right\}. \tag{156}$$

Let us consider a sample i.i.d matrix $\mathbf{X} \in \mathbb{R}^{d\times m}$ with $\mathbf{x}_\mu \sim \mathscr{N}(\mathbf{0}, \mathbf{I}_d)$.

Theorem 7.2.1. *For the perceptron model class eq. (156) with random i.i.d. input data in the high-dimensional limit,* $\mathfrak{R}_n\left(\mathscr{F}_{sign}\right) = \Theta\left(\frac{1}{\sqrt{\alpha}}\right)$.

Proof. In a nutshell, the proof uses the fact that the Rademacher complexity is upper-bounded by the VC dimension divided by $\alpha^{1/2}$, and lower-bounded by one particular example of its function class, when the weights are chosen according to Hebb's rule, which also gives a behavior scaling as $\alpha^{-1/2}$.

Upper bound

For a linear classifier with binary outputs such as the perceptron, the VC dimension is easy to compute and $d_{\text{vc}} = d$. Hence we know from Massart theorem's (Massart, 2000) that

$$\mathfrak{R}_n(\mathscr{F}_{\text{sign}}) \leq \Theta\left(\sqrt{\frac{d_{\text{vc}}(\mathscr{F}_{\text{sign}})}{n}}\right) = \Theta\left(\sqrt{\frac{d}{n}}\right) = \Theta\left(\alpha^{-1/2}\right).$$

Lower bound

Let us consider the following estimator, known as the Hebb's rule (Hebb, 1962): $\mathbf{w}^\star = \frac{1}{\sqrt{d}} \sum_{\nu=1}^{n} y_\nu \mathbf{x}_\nu$. Hence for a given sample $\mathbf{x}_\mu$ the above estimator outputs

$$f_{\mathbf{w}^\star}(\mathbf{x}_\mu) = \text{sign}\left(\frac{1}{\sqrt{d}}\mathbf{w}^{\star\intercal}\mathbf{x}_\mu\right) = \text{sign}\left(\left(\frac{1}{d}\sum_{\nu=1}^{n} y_\nu \mathbf{x}_\nu\right)^\intercal \mathbf{x}_\mu\right).$$

Injecting its expression in the definition the Rademacher complexity eq. (149) one obtains:

$$\begin{aligned}
\mathfrak{R}_n(\mathscr{F}_{\text{sign}}) &\equiv \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\sup_{\mathbf{w}} \frac{1}{n}\sum_{\mu=1}^{n} y_\mu f_{\mathbf{w}}(\mathbf{x}_\mu)\right] \\
&\geq \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\frac{1}{n}\sum_{\mu=1}^{n} y_\mu f_{\mathbf{w}^\star}(\mathbf{x}_\mu)\right] \\
&= \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\frac{1}{n}\sum_{\mu=1}^{n} \text{sign}\left(y_\mu \frac{1}{d}\left(\sum_{\nu=1}^{n} y_\nu \mathbf{x}_\nu\right)^\intercal \mathbf{x}_\mu\right)\right] \\
&= \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\frac{1}{n}\sum_{\mu=1}^{n} \text{sign}\left(1 + \frac{1}{d}\sum_{\nu\neq\mu} y_\mu y_\nu \mathbf{x}_\nu^\intercal \mathbf{x}_\mu\right)\right].
\end{aligned}$$

As $\mathbf{x}_\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and the labels are drawn uniformly $y_\mu \sim \mathcal{U}(\pm 1)$, $\mathbf{z}_\mu \equiv y_\mu \mathbf{x}_\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Hence let us define the Gaussian random variable

$$\theta_\mu \equiv \frac{1}{d}\sum_{\nu\neq\mu} y_\mu y_\nu \mathbf{x}_\nu^\intercal \mathbf{x}_\mu = \frac{1}{d}\sum_{\nu\neq\mu} \mathbf{z}_\nu^\intercal \mathbf{z}_\mu,$$

and compute its two first moments

$$\begin{aligned}
\mathbb{E}[\theta_\mu] &= \mathbb{E}_{\mathbf{z}}\left[\frac{1}{d}\sum_{\nu\neq\mu}^{n} \mathbf{z}_\nu^\intercal \mathbf{z}_\mu\right] = \mathbb{E}_{\mathbf{z}}\left[\frac{1}{d}\sum_{\nu\neq\mu}^{n}\sum_{i=1}^{d} z_{i\nu} z_{i\mu}\right] = 0, \\
\mathbb{E}\left[\theta_\mu^2\right] &= \mathbb{E}\left[\frac{1}{d^2}\left(\sum_{\nu\neq\mu}^{n} \mathbf{z}_\nu^\intercal \mathbf{z}_\mu\right)^2\right] = \frac{(n-1)}{d} \underset{n\to\infty}{\longrightarrow} \alpha.
\end{aligned}$$

Hence because of the CLT, in the high-dimensional limit $\theta_\mu \sim \mathcal{N}(0, \alpha)$,

$$\begin{aligned}
\mathfrak{R}_n(\mathscr{F}_{\text{sign}}) \geq \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n}\sum_{\mu=1}^{n} \text{sign}(1+\theta_\mu)\right] &= \mathbb{E}_\theta[\text{sign}(1+\theta)] \\
&= \mathbb{P}[\theta \geq -1] - \mathbb{P}[\theta \leq -1] = 2\mathbb{P}[\theta \geq -1] - 1.
\end{aligned}$$

Finally, noting that

$$\mathbb{P}[\theta \geq -1] = \int_{-\frac{1}{\sqrt{\alpha}}}^{\infty} \mathrm{D}_\theta = \frac{1}{2}\text{erfc}\left(-\frac{1}{\sqrt{2\alpha}}\right) \underset{\alpha\to\infty}{\simeq} \frac{1}{2} - \frac{1}{\sqrt{2\pi\alpha}},$$

we obtain a lower bound for the Rademacher complexity

$$\mathfrak{R}_n(\mathscr{F}_{\text{sign}}) \geq \sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{\alpha}} = \Theta\left(\frac{1}{\sqrt{\alpha}}\right).$$

□

Heuristically, this result generalizes as well to a two-layer neural network with K hidden neurons. Indeed, the two-layer function class contains, as a particular case, the single layer one, so the lower bounds goes through. The upper bound is however harder to control rigorously. Since neural networks have a finite VC dimension, the Rademacher complexity is again lower-bounded by $\Theta(1/\sqrt{n})$; However, we do not know of any theorem that would ensure that the VC dimension is bounded by $\Theta(d)$ (Bartlett et al., 2003). Nevertheless, anticipating on the statistical physics approach, we indeed expect from the concentration (self-averaging) properties of the ground-state energy (Talagrand, 2003) in the high-dimensional limit that it will yield a Rademacher complexity that is a function of $\alpha = n/d$ *only* at fixed K. From this argument, we expect that the $\Theta\left(\frac{1}{\sqrt{\alpha}}\right)$ dependence of the Rademacher complexity to be very generic in the high-dimensional limit.

7.3 THE STATISTICAL PHYSICS APPROACH

7.3.1 AVERAGE CASE PROBLEMS: STATISTICAL PHYSICS OF LEARNING

As anticipated in the previous chapter, the approach inspired by statistical physics to understand neural networks considers a set of data points coming from known distributions. Again, for the purpose of this presentation we focus on a simple example, where $\mathbf{x} \sim \mathrm{P}_x(.)$ with $\mathrm{P}_x(\mathbf{x}) = \mathcal{N}_{\mathbf{x}}(\mathbf{0}, \mathbf{I}_d)$. Sec. 7.4.5 is devoted to a generalization to random input data corresponding to random matrices with arbitrary singular value density.

Consider a function class, for instance we can again use the *perceptron* one $\mathscr{F}_{\text{sign}}$: $\{f_{\mathbf{w}} : \mathbf{x} \to \text{sign}\left(\frac{1}{\sqrt{d}}\mathbf{w}^\intercal \mathbf{x}\right)\}$; a typical question in the literature was to compute how many misclassified examples can be obtained for a given rule used to generate the labels (Engel et al., 2001). Given n samples $\{y_\mu, \mathbf{x}_\mu\}_{\mu=1}^n$, in order to count the number of wrongly classified training samples, we define the Hamiltonian, or *energy* function (Mézard et al., 1986):

$$\begin{aligned}\mathscr{H}_d\left(\{\mathbf{y}, \mathbf{X}\}, \mathbf{w}\right) &\equiv \sum_{\mu=1}^{n} \mathbb{1}\left[y_\mu \neq f_{\mathbf{w}}\left(\mathbf{x}_\mu\right)\right] \\ &= \frac{1}{2}\left(n - \sum_{\mu=1}^{n} y_\mu f_{\mathbf{w}}\left(\mathbf{x}_\mu\right)\right).\end{aligned} \tag{157}$$

A classical problem in statistical physics is to compute the random capacity also called *Gardner capacity* α_c (Gardner et al., 1989) studied in Chap. 6: given n examples $\{\mathbf{x}_\mu\}_{\mu=1}^n$ and labels $\{y_\mu\}_{\mu=1}^n$ randomly chosen between ±1, it consists in finding how many samples n_c can be correctly classified.

It turns out there exists a deep connection between the Gardner capacity and the VC dimension, as their common aim is to measure the maximum number of points n_c such that there exists a function in the hypothesis class being able to fit the data set. In particular, using Sauer's lemma (Sauer, 1972) in the large size limit $n, d \longrightarrow \infty$, keeping $\alpha_c = \frac{n_c}{d} = \Theta(1)$ and $\alpha_{\text{vc}} = \frac{d_{\text{vc}}}{d} = \Theta(1)$, it is possible to show that the Gardner capacity α_c provides a lower-bound of the VC dimension (Engel et al., 2001):

$$\alpha_c \leq 2\alpha_{\text{vc}} \,. \tag{158}$$

To illustrate this inequality, let us consider again the perceptron classifier hypothesis class $\mathscr{F}_{\text{sign}}$ for which the above inequality is saturated. In fact, the VC dimension is in this case (linear classification with binary outputs) simply $d_{\text{vc}} = d$. Hence on one hand $\alpha_{\text{vc}} = 1$, on the other hand the Gardner capacity amounts to $\alpha_c = 2$ (Cover, 1965; Gardner et al., 1989).

It is fair to say that a large part of the statistical physics literature focused mainly on the Gardner capacity, in particular in a series of works in the 90's (Gardner et al., 1989; Krauth et al., 1989) that led to more recent rigorous works (Talagrand, 2003; Talagrand, 2006b; Ding et al., 2019; Aubin et al., 2019e).

7.3.2 THE RADEMACHER COMPLEXITY AND THE GROUND STATE ENERGY

As we shall see now, computing the Rademacher complexity for random input data can be directly reduced to a more natural object in the physics literature: the *ground state energy*. Defining the Gibbs measure at inverse temperature β, that weighs configurations with their respective cost, as

$$\langle \ldots \rangle_\beta \equiv \frac{\int_{\mathbb{R}^d} \mathbf{dw} \, \ldots \, e^{-\beta \mathscr{H}_d(\{\mathbf{y},\mathbf{X}\},\mathbf{w})}}{\int_{\mathbb{R}^d} \mathbf{dw} \, e^{-\beta \mathscr{H}_d(\{\mathbf{y},\mathbf{X}\},\mathbf{w})}} \,, \tag{159}$$

we observe that averaging the Hamiltonian in eq. (157) over $\{\mathbf{y}, \mathbf{X}\}$ and the Gibbs measure for any function $f_{\mathbf{w}} \in \mathscr{F}$ provides

$$\mathbb{E}_{\mathbf{y},\mathbf{X}} \left\langle \frac{\mathscr{H}_d(\{\mathbf{y},\mathbf{X}\},\mathbf{w})}{d} \right\rangle_\beta = \frac{\alpha}{2} \left[1 - \mathbb{E}_{\mathbf{y},\mathbf{X}} \left\langle \frac{1}{n} \sum_{\mu=1}^{n} y_\mu f_{\mathbf{w}}(\mathbf{x}_\mu) \right\rangle_\beta \right] , \tag{160}$$

where $\alpha = \frac{n}{d} = \Theta(1)$. Taking the zero temperature limit, i. e. $\beta \to \infty$, in the above equation, we finally obtain the ground state energy e_{gs}, a quantity commonly used in physics. Interestingly, we recognize the definition of the Rademacher complexity $\mathfrak{R}_n(\mathscr{F})$

$$\begin{aligned} e_{\text{gs}} &\equiv \lim_{\beta\to\infty} \lim_{d\to\infty} \mathbb{E}_{\mathbf{y},\mathbf{X}} \left\langle \frac{\mathscr{H}_d(\{\mathbf{y},\mathbf{X}\},\mathbf{w})}{d} \right\rangle_\beta \\ &= \frac{\alpha}{2} \left[1 - \mathbb{E}_{\mathbf{y},\mathbf{X}} \sup_{f_{\mathbf{w}} \in \mathscr{F}} \frac{1}{n} \sum_{\mu=1}^{n} y_\mu f_{\mathbf{w}}(\mathbf{x}_\mu) \right] \\ &= \frac{\alpha}{2} \left[1 - \mathfrak{R}_n(\mathscr{F}) \right] , \end{aligned} \tag{161}$$

where random labels $\mathbf{y}$ play the role of the Rademacher variable $\boldsymbol{\varepsilon}$ in (149). The above equation shows a simple correspondence between the ground state energy on the perceptron model with randomly quenched disorder and the Rademacher complexity of the corresponding hypothesis class, and shall bring insights from both the machine learning and statistical physics communities. Consequently, as we shall see, this connection means that the Rademacher complexity can be computed, rather than bounded, for many models using the replica method from statistical physics. As far as we are aware, this basic connection between the ground state energy and Rademacher complexity was not previously stated in literature.

7.3.3 AN INTUITIVE UNDERSTANDING ON THE RADEMACHER BOUNDS ON GENERALIZATION

At this point, the Rademacher complexity becomes a more familiar object to the physics-minded reader. However, could we understand more intuitively why the Rademacher complexity, or equivalently the ground state energy, is involved in the generalization gap bound? Let us present an intuitive hand-waving explanation. Consider the fraction of mistakes performed by a classifier $f_{\mathbf{w}}$ on unknown samples, namely the generalization error $\varepsilon_{\text{gen}}(f_{\mathbf{w}})$, and on the training set the training error $\varepsilon^{n}_{\text{train}}(f_{\mathbf{w}})$. The worst case scenario that could occur is trying to fit while there exists no underlying rule, meaning that labels are purely random uncorrelated from input. The estimator will purely overfit and its generalization error will remain constant to $1/2$ in any case. This leads to the following heuristic generalization bound:

$$
\begin{aligned}
&\varepsilon_{\text{gen}}(f_{\mathbf{w}}) - \varepsilon^{n}_{\text{train}}(f_{\mathbf{w}}) \leq \varepsilon^{\text{random labels}}_{\text{gen}}(f_{\mathbf{w}}) - \varepsilon^{\text{random labels},n}_{\text{train}}(f_{\mathbf{w}}) \\
&= \frac{1}{2} - \varepsilon^{\text{random labels},n}_{\text{train}}(f_{\mathbf{w}}) = \frac{1}{2}\left(1 - 2\varepsilon^{\text{random labels},n}_{\text{train}}(f_{\mathbf{w}})\right) \\
&= \frac{1}{2}\hat{\mathfrak{R}}_n(\mathscr{F}) \,.
\end{aligned}
\tag{162}
$$

Note that this heuristic reasoning does not give the *exact* Rademacher generalization bound. In fact, the actual stronger and uniform over all possible $\mathbf{w} \in \mathbb{R}^d$ bound does not have a factor $1/2$, and surely cannot be fully captured by the simple above argument. Nevertheless, this argument reflects the crux of the Rademacher bound: it provides a very pessimistic bound by assuming the worst possible scenario: i. e. fitting data and trying to make predictions while the labels are random. Of course, in real data problems the rule is not random; it is then no surprise that the Rademacher bound is not tight (Zhang et al., 2016). Indeed, real problems labels are *not* randomly correlated with the inputs.

7.4 CONSEQUENCES AND BOUNDS FOR SIMPLE MODELS

In this section, we illustrate our previous arguments and the connection between the spin glass approach and the Rademacher complexity still for the case of Gaussian i.i.d input data matrix $\mathbf{X}$ in the high-dimensional limit when $n, d \to \infty$.

7.4.1 GROUND STATE ENERGIES OF THE PERCEPTRON

For a number of samples smaller than the Gardner capacity α_c, also called the SAT-UNSAT threshold, it is by definition possible to fit all random labels $\mathbf{y}$. Accordingly, the number of misclassified examples is zero and the ground state energy $e_{\rm gs} = 0$. This means that the Rademacher complexity is asymptotically equal to 1 for $\alpha < \alpha_c$. However above the Gardner capacity $\alpha > \alpha_c$, the estimator $f_{\mathbf{w}}$ cannot perfectly fit the random labels and will misclassify some of them, equivalently $e_{\rm gs} > 0$. From the arguments given in sec. 7.2, we thus expect

$$\begin{aligned} &\mathfrak{R}_n\left(\mathscr{F}\right) = 1 \text{ for } \alpha < \alpha_c\,, \\ &\mathfrak{R}_n\left(\mathscr{F}\right) \approx \Theta\left(\sqrt{\frac{\alpha_c}{\alpha}}\right) \text{ for } \alpha \gg \alpha_c\,. \end{aligned} \tag{163}$$

This relation is already non-trivial, as it yields a link between the Gardner capacity and the Rademacher complexity. Using the replica method from spin glass analysis, and the mapping with ground state energies (161), we shall now see how one can go beyond these simple arguments, and compute the actual precise asymptotic value of the Rademacher complexity.

7.4.2 COMPUTING THE GROUND-STATE ENERGY WITH THE REPLICA METHOD

Knowing that statistical physics literature focused mainly on the Gardner capacity, the connection between the ground-state energy and the Rademacher complexity suggests that it would be worth looking at these old results in a new light. In fact, the replica method allows for an *exact* computation of the Rademacher complexity for random input data in the large size limit. In the following, we handle computations by focusing on a simple generalization of the linear functions hypothesis class. Fix any activation function $\varphi : \mathbb{R} \longrightarrow \{\pm 1\}$, we define the following hypothesis class

$$\mathscr{F}_\varphi \equiv \left\{ f_{\mathbf{w}} : \left\{ \begin{array}{l} \mathbb{R}^d \longrightarrow \{\pm 1\} \\ \mathbf{x} \longrightarrow \varphi\left(\frac{1}{\sqrt{d}} \mathbf{w}^\intercal \mathbf{x}\right) \end{array} \right. , \mathbf{w} \in \mathbb{R}^d \right\} . \tag{164}$$

Starting with the posterior distribution

$$\mathbb{P}(\mathbf{w}|\mathbf{y},\mathbf{X}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w},\mathbf{X})\mathbb{P}(\mathbf{w})}{\mathbb{P}(\mathbf{y},\mathbf{X})} = \frac{e^{-\beta \mathscr{H}_d(\{\mathbf{y},\mathbf{X}\},\mathbf{w})} \mathrm{P}_{\mathrm{w}}(\mathbf{w})}{\mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)} , \tag{165}$$

we introduced the partition function associated to the Hamiltonian eq. (157) at inverse temperature β

$$\mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta) = \int_{\mathbb{R}^d} \mathrm{dP}_{\mathrm{w}}(\mathbf{w})\ e^{-\beta \mathscr{H}_d(\{\mathbf{y},\mathbf{X}\},\mathbf{w})} . \tag{166}$$

In the large size limit $d \to \infty$, the posterior distribution becomes highly peaked in particular regions of parameters. In physics we are interested in these dominant regions and focus on the free energy at inverse temperature β defined as

$$\varphi_{\mathbf{y},\mathbf{X}}(\{\mathbf{y},\mathbf{X}\},\alpha,\beta) \equiv -\lim_{d\to\infty} \frac{1}{d\beta} \log \mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta) . \tag{167}$$

The free energy is closely related to the free entropy that can be equivalently considered according to $\varphi_{\mathbf{y},\mathbf{X}} = -\Phi_{\mathbf{y},\mathbf{X}}$. However, as we are interested in computing quantities in the *typical case*, we want to average over all potential training sets $\{\mathbf{y},\mathbf{X}\}$ and compute instead the averaged free energy

$$\varphi(\alpha,\beta) \equiv \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\varphi_{\mathbf{y},\mathbf{X}}(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)\right] . \tag{168}$$

Computing directly this average rigorously is difficult, hence we will carry out the computation using the so-called *replica method*, starting by writing the *replica trick*

$$-\frac{1}{d\beta}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\log \mathscr{Z}_d\right] = -\frac{1}{d\beta} \lim_{r\to 0} \frac{\partial \log \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^r\right]}{\partial r} , \tag{169}$$

which replaces the expectation of $\log \mathscr{Z}_d$ by the moments of $\mathscr{Z}_d$, which are easier to compute. Taking the limit $d \to \infty$, and assuming that we can revert it with the limit $r \to 0$, we finally obtain

$$\varphi(\alpha,\beta) = \lim_{r\to 0}\left[\lim_{d\to\infty} -\frac{1}{d\beta} \frac{\partial \log \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^r\right]}{\partial r}\right] . \tag{170}$$

We give some details on the replica computation in the context of the GLM with randomly quenched disorder in Appendix. B.2, and we also refer the reader to the relevant literature in physics (Mézard et al., 1986; Hertz et al., 1993; Engel et al., 2001; Mézard et al., 2009; Zdeborová et al., 2016a) and in mathematics (Talagrand, 2003; Talagrand, 2006b; Bolthausen et al., 2007; Panchenko et al., 2004; Panchenko et al., 2018). Notice that in this randomly quenched setting where the labels are uncorrelated from the input vector, the replica computation is exactly the same than the one used for the storage capacity problem in Chap. 6. The computation of the free energy by the replica method is done by deriving a hierarchy of approximate ansatz, named RS, 1RSB, 2RSB ...While in some problems the RS or the 1RSB ansatz is sufficient, in others only the infinite step solution FRSB gives the exact ansatz (Mézard, 1989; Talagrand, 2003; Talagrand, 2006b), although the 1RSB approach is usually an accurate approximation.

Computing the ground state energy consists in taking the zero temperature limit $\beta \to \infty$ above the capacity $\alpha > \alpha_c$ in the replica free energy $\varphi(\alpha,\beta) = e(\alpha,\beta) - \beta^{-1}s(\alpha,\beta)$; where e, s denote respectively the densities of the energy and entropy contributions. The simplest form of the replica computation is known as RS and the next simplest is 1RSB which plugged in eq. (170) leads to expressions (Majer et al., 1993; Erichsen et al., 1993; Whyte et al., 1996)

$$\begin{aligned}
\varphi_{\text{iid}}^{(\text{rs})}(\alpha,\beta) &= -\frac{1}{\beta}\,\underset{q_0,\hat{q}_0}{\mathbf{extr}}\left\{\frac{1}{2}(q_0\hat{q}_0 - 1) + \Psi_{\text{w}}^{(\text{rs})}(\hat{q}_0) + \alpha\Psi_{\text{out}}^{(\text{rs})}(q_0,\beta)\right\}, \\
\varphi_{\text{iid}}^{(\text{1rsb})}(\alpha,\beta) &= -\frac{1}{\beta}\,\underset{q_0,q_1,\hat{q}_0,\hat{q}_1,x}{\mathbf{extr}}\left\{\frac{1}{2}(q_1\hat{q}_1 - 1) + \frac{x}{2}(q_0\hat{q}_0 - q_1\hat{q}_1)\right. \\
&\qquad \left. + \Psi_{\text{w}}^{(\text{1rsb})}(\hat{q}_0,\hat{q}_1,x) + \alpha\Psi_{\text{out}}^{(\text{1rsb})}(q_0,q_1,\beta,x)\right\},
\end{aligned} \tag{171}$$

with auxiliary functions

$$\begin{aligned}
\Psi_{\text{w}}^{(\text{rs})}(\hat{q}_0) &\equiv \mathbb{E}_{\xi_0}\log\mathbb{E}_w\left[\exp\left(\frac{(1-\hat{q}_0)}{2}w^2 + \xi_0\sqrt{\hat{q}_0}w\right)\right], \\
\Psi_{\text{out}}^{(\text{rs})}(q_0,\beta) &\equiv \mathbb{E}_y\mathbb{E}_{\xi_0}\log\mathbb{E}_z\left[\mathscr{C}\left(y\middle|\sqrt{1-q_0}z + \sqrt{q_0}\xi_0,\beta\right)\right], \\
\Psi_{\text{w}}^{(\text{1rsb})}(\hat{q}_0,\hat{q}_1,x) &\equiv \frac{1}{x}\mathbb{E}_{\xi_0}\log\Bigg(\\
&\quad \mathbb{E}_{\xi_1}\mathbb{E}_w\left[\exp\left(\frac{(1-\hat{q}_1)}{2}w^2 + \left(\sqrt{\hat{q}_0}\xi_0 + \sqrt{\hat{q}_1 - \hat{q}_0}\xi_1\right)w\right)\right]^x\Bigg), \\
\Psi_{\text{out}}^{(\text{1rsb})}(q_0,q_1,\beta,x) &\equiv \frac{1}{x}\mathbb{E}_y\mathbb{E}_{\xi_0}\log\Bigg(\\
&\quad \mathbb{E}_{\xi_1}\mathbb{E}_z\left[\mathscr{C}(y|\sqrt{q_0}\xi_0 + \sqrt{q_1 - q_0}\xi_1 + \sqrt{1-q_1}z,\beta)\right]^x\Bigg).
\end{aligned} \tag{172}$$

We introduced a temperature-dependent constraint function $\mathscr{C}(y|z) = \exp(-\beta V(y|z))$ where the generic cost function V reads in our case $V(y|z) = \mathbb{1}\left[y \neq \varphi(z)\right]$ and $y \sim \mathrm{P}_y(.)$ the distribution of the random labels. Above expressions are valid for any generic weight distribution $\mathrm{P}_{\text{w}}(.)$ and non-linearity φ. The detailed computation is left in Appendix. B.2, in particular eq. (345)

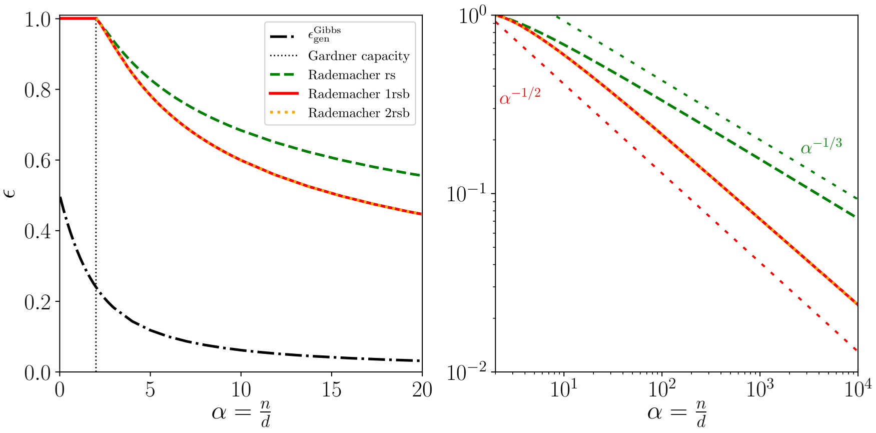


Figure 44: Explicit Rademacher complexity for the spherical perceptron ($\alpha_c = 2$). (**Left**) For $\alpha < \alpha_c$ the problem is satisfiable so the number of error is zero and the Rademacher complexity is constant to unity. For $\alpha > \alpha_c$, the problem becomes unsatisfiable and $e_{\mathrm{gs}} > 0$. (**Right**) In the case of the spherical perceptron, RS (dashed green) and 1RSB (red) ansatz provide really different results that scale respectively with $\alpha^{-1/3}$ and $\alpha^{-1/2}$ (scaling are represented with colored dashed lines). Performing 2RSB ansatz (dashed orange) does not change the scaling and difference with respect to 1RSB is visually imperceptible. The black dotted-dashed curve is the generalization error in the teacher-student scenario (Barbier et al., 2019b). Note the large gap between the worst case Rademacher bound and the actual teacher-student generalization error.

and eq. (362). Then the general method to find the ground state energy it to take the zero temperature limit $\beta \to \infty$

$$e_{\mathrm{gs,iid}}\left(\alpha\right) \equiv \lim_{\beta \to \infty} \varphi_{\mathrm{iid}}\left(\alpha, \beta\right), \tag{173}$$

while handling carefully the scaling of the optimized order parameters in this limit.

7.4.2.a SPHERICAL PERCEPTRON

The most commonly studied model (Gardner et al., 1988; Gardner et al., 1989; Gardner et al., 1988) with continuous weights is the spherical model with $\mathbf{w} \in \mathbb{R}^d$ such that $\|\mathbf{w}\|_2^2 = d$. The spherical constraint allows to have a well-defined model which excludes diverging or vanishing weights. In this case, the Gardner capacity is rigorously known to be equal to $\alpha_c = 2$ (Cover, 1965).

We computed both the RS and 1RSB free energies (Majer et al., 1993; Erichsen et al., 1993; Whyte et al., 1996), see also Appendix. B.2.7. Taking the zero temperature limits $\beta \to \infty, q_0 \to 1$ and $q_1 \to 1, x \to 0$ in the 1RSB case, while keeping $\chi \equiv \beta(1-q_0)$ and $\Omega_0 \equiv \frac{\beta x}{\chi}$ finite leads to the following expressions of the ground states energies:

$$e^{(\mathrm{rs})}_{\mathrm{rs,iid}} = \mathbf{extr}_{\chi}\left\{-\frac{1}{2\chi} + \alpha\mathbb{E}_{y,\xi_0}\min_{z}\left[V(y|z) + \frac{(z-\xi_0)^2}{2\chi}\right]\right\} \tag{174}$$

$$\begin{aligned} e^{(\mathrm{1rsb})}_{\mathrm{rs,iid}} = \mathbf{extr}_{\chi,\Omega_0,q_0}&\left\{\frac{1}{2\Omega_0\chi}\log\left(1+\Omega_0(1-q_0)\right)\right. \\ &+\frac{q_0}{2\chi\left(1+\Omega_0(1-q_0)\right)} \\ &\left.+\frac{\alpha}{\chi\Omega_0}\mathbb{E}_{\xi_0}\log\mathbb{E}_{\xi_1}e^{-\Omega_0\chi\min_z\left[V(y|z)+\frac{1}{2\chi}\left(z-\sqrt{q_0}\xi_0-\sqrt{1-q_0}\xi_1\right)^2\right]}\right\}, \end{aligned} \tag{175}$$

where the cost function $V(y|z) = \mathbb{1}\left[y \neq \varphi(z)\right]$. The details of the derivation via the replica methods are given in Appendix B.2.7. The results for Rademacher variable y with $\varphi(z) = \mathrm{sign}(z)$ are depicted in Fig. 44.

Interestingly, the bounds on the Rademacher complexity also imply consequences for the ground state energy. Indeed the Rademacher complexity scales as $\alpha^{-1/2}$ for large values of α — namely there exists a constant $\mathscr{C}$ such that $\mathfrak{R}_n(\mathscr{F}) \underset{\alpha\to\infty}{\approx} \frac{C}{\sqrt{\alpha}}$ — therefore the ground state energy behaves for large α as

$$e_{\mathrm{gs}}(\alpha) = \frac{\alpha}{2}\left(1-\mathfrak{R}_n(\mathscr{F})\right) \underset{\alpha\to\infty}{\longrightarrow} \frac{\alpha}{2}\left(1-\frac{C}{\sqrt{\alpha}}\right). \tag{176}$$

We first notice that the replica symmetric RS solution complexity fails to deliver the correct scaling as sketched in Fig. 44, so the scaling in eq. (176) must not be entirely trivial. On the other hand, the 1RSB solution we used, which is expected to be numerically very close to the harder to evaluate FRSB one, seems to yield the correct scaling, see Fig. 44. It is rather striking that the statistical learning connection allows to predict, through eq. (176), the scaling of the energy in the large α regime, that is only satisfied with replica symmetry breaking ansatz. This yields an open question for replica theory: in practice, can one compute exactly the value of the constant C? Given the FRSB solution is notoriously hard to evaluate, this might be an issue worth investigating in mathematical physics.

7.4.2.B BINARY PERCEPTRON

Another common choice for the weights distribution is the binary prior $\mathrm{P}_{\mathrm{w}}(w) = \delta(w-1) + \delta(w+1)$ studied e.g. in (Krauth et al., 1989). In this case, the Gardner capacity is predicted to be $\alpha_c \approx 0.83\ldots$, a prediction which, remarkably, is still not entirely rigorously proven, but see (Ding et al., 2019; Aubin et al., 2019e). To see this, we use eq. (171). In the binary perceptron, the landscape of the model is said to be f1RSB, i. e. clustered in point-like dominant solutions as discussed in Chap. 6, and the RS and 1RSB free energies are the same, even though their entropies are different $\varphi(\alpha,\beta) = e(\alpha,\beta) - \beta^{-1}s(\alpha,\beta)$. In this case computing the ground state can be tackled via finding the effective temperature $\beta^\star$ such that the $s(\alpha,\beta^\star) = 0$, that can be plugged

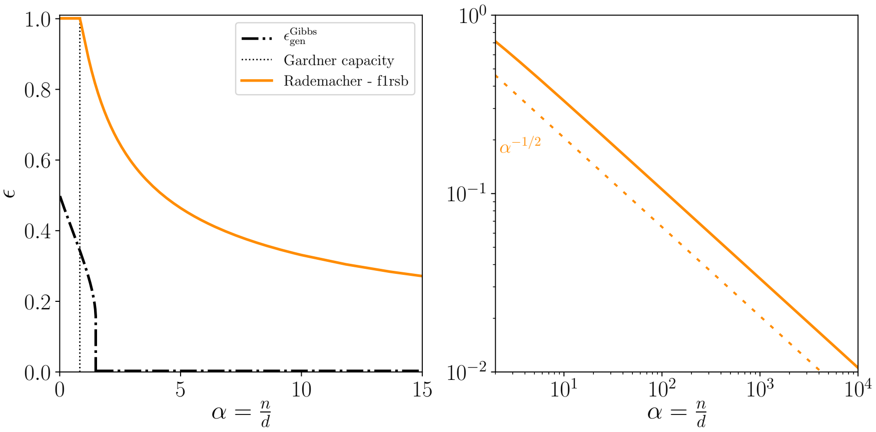


Figure 45: Explicit Rademacher complexity for (**Left**) the binary perceptron ($\alpha_c = 0.83\ldots$). The replica solution (orange) leads again (**Right**) to a $\alpha^{-1/2}$ scaling (dashed orange) of the Rademacher complexity at large α. The dotted-dashed black curve is the generalization error in the teacher-student scenario. Note the gap between the worst case bound (Rademacher) and the teacher-student generalization error.

back to find the ground state energy $e_{\text{gs}}(\alpha) = \varphi(\alpha, \beta^\star)$. Again, we note that even though the 1RSB ansatz is unstable and should be replaced by a more complex (and ultimately FRSB) solution, it already gives the good scaling $\mathfrak{R}_n(\mathscr{F}) \sim \alpha^{-1/2}$, and satisfies the scaling eq. (176) for large α, as in the case of the spherical model, see Fig. 45.

7.4.3 TEACHER-STUDENT SCENARIO VERSUS WORST CASE RADEMACHER

The Rademacher bounds are really interesting as they depend only on the data distribution, and are valid for *any rule* used to generate the labels, no matter how complicated. In this sense, it is a worst-case scenario on the rule that prescribes labels to data. A different approach, again pioneered in statistical physics (Gardner et al., 1989), is to focus on the behavior for a given rule, called the *teacher* rule. Given the Rademacher bounds tackle the worst case with respect to that rule, it is interesting to consider the generalization error one actually gets for the *best case*, i.e. fitting the labels according to the same teacher rule.. This is the so-called T-S approach. In the wake of the need to understand the effectiveness of neural networks, and the limitations of the classical approaches, it is of interest to revisit the results that have emerged thanks to the physics perspective.

We shall thus assume that the *actual labels* are given by the rule

$$y = \text{sign}\left(\frac{1}{\sqrt{d}}\mathbf{w}^{\star\top}\mathbf{x}\right), \tag{177}$$

with $\mathbf{w}^\star$, the *teacher weights* that can be taken as Rademacher ± 1 variables, or Gaussian ones. Now that labels are generated by feeding i.i.d random samples to a neural network architecture (the teacher) and are then presented to another neural network (the student) that is trained using this data, it is interesting to compare the worst case Rademacher bound with the actual generalization error of this student on such synthetic data.

We now consider the error of a *typical* solution $\mathbf{w}$ from the posterior distribution (this is often called the Gibbs rule) for the student. Given the rule is outputting ± 1 variables, this yields

$$\varepsilon_{\text{gen}}^{\text{gibbs}} = 1 - \mathbb{E}_{\mathbf{x},\mathbf{w}^\star}\left[\langle f_{\mathbf{w}^\star}(\mathbf{x}) \times f_{\mathbf{w}}(\mathbf{x})\rangle\right] = 1 - q^\star \tag{178}$$

where $q^\star = \mathbb{E}_{\mathbf{x},\mathbf{w}^\star}\left[\langle f_{\mathbf{w}^\star}(\mathbf{x}) \times f_{\mathbf{w}}(\mathbf{x})\rangle\right]$. Computing $q^\star$ can be done within the statistical mechanics approach (Seung et al., 1992; Watkin et al., 1993; Opper, 1995; Engel et al., 2001) and can be rigorously done as well (Barbier et al., 2019b). Notice that this error is equal to the Bayes-optimal error for the quadratic loss, see as well (Barbier et al., 2019b).

The two *optimistic* (teacher-student) and *pessimistic* (Rademacher) errors can be seen in Fig. 44 for spherical and in Fig. 45 for binary weights. In this case, since a perfect fit is always possible, the training error is zero and the Rademacher complexity is itself the bound on the generalization error. These two figures show how different the worst and teacher-student case can be in practice, and demonstrate that one should perhaps not be surprised by the fact that the empirical Rademacher complexity does not always give the correct answer (Zhang et al., 2016), as after all it deals only with worst case scenarios.

7.4.4 COMMITTEE MACHINE WITH GAUSSIAN WEIGHTS

Given the large gap between the Rademacher bound and the teacher-student setting, we can ask whether we can find a case where the Rademacher bound is void in the sense that the Rademacher complexity is 1 yet generalization is good for the teacher-student setting? This can be done by moving to two-layer networks. Consider a simple version of this function class, namely the committee machine (Engel et al., 2001). It is a two-layer network where the second layer has been fixed, such that only weights of the first layer $\mathbf{W} = \{\mathbf{w}_1, \cdots, \mathbf{w}_K\} \in \mathbb{R}^{d\times K}$ are learnt. The function class for a committee machine with K hidden units, already considered in Chap. 5, is defined by

$$\mathscr{F}_{\text{com}} \equiv \left\{ f_{\mathbf{W}} : \begin{cases} \mathbb{R}^d \longrightarrow \{-1,1\} \\ \mathbf{x} \longrightarrow \operatorname{sign}\left(\sum_{k=1}^K \operatorname{sign}\left(\frac{1}{\sqrt{d}}\mathbf{w}_k^\intercal \mathbf{x}\right)\right) \end{cases} , \mathbf{W} \in \mathbb{R}^{d\times K} \right\}. \tag{179}$$

Instead of computing the Rademacher complexity with the replica method, it is sufficient for the purpose of this section to understand its rough behavior. As discussed in sec. 7.4.1, this requires knowing the Gardner capacity. A generic bound by (Mitchison et al., 1989) states that it is upper bounded by $\Theta(K\log(K))$. Additionally, the Gardner capacity has been computed by the replica method in (Monasson et al., 1995b; Urbanczik, 1997; Xiong et al., 1998) who obtained that $\alpha_c = \Theta(K\sqrt{\log(K)})$. We thus expect that

$$\begin{aligned} &\mathfrak{R}_n\left(\mathscr{F}_{\mathrm{com}}\right) = 1 \text{ for } \alpha < \Theta\left(K\sqrt{\log(K)}\right), \\ &\mathfrak{R}_n\left(\mathscr{F}_{\mathrm{com}}\right) \approx \Theta\left(\sqrt{\frac{K\sqrt{\log K}}{\alpha}}\right) \text{ for } \alpha \gg \Theta\left(K\sqrt{\log K}\right). \end{aligned} \tag{180}$$

To compare with the teacher-student case, when the labels are produced by a teacher committee machine as

$$y = \mathrm{sign}\left(\sum_{k=1}^{K} \mathrm{sign}\left(\frac{1}{\sqrt{d}}\mathbf{w}_k^{\star\intercal}\mathbf{x}\right)\right), \tag{181}$$

the error of the Gibbs algorithm reads

$$\varepsilon_{\mathrm{gen}}^{\mathrm{gibbs}} = 1 - \mathbb{E}_{\mathbf{x},\mathbf{w}^\star}\left[\langle f_{\mathbf{w}^\star}(\mathbf{x}) \times f_{\mathbf{w}}(\mathbf{x})\rangle\right] = 1 - q^\star \tag{182}$$

where, again $q^\star = \mathbb{E}_{\mathbf{x},\mathbf{w}^\star}\left[\langle f_{\mathbf{w}^\star}(\mathbf{x}) \times f_{\mathbf{w}}(\mathbf{x})\rangle\right]$, has been computed in a series of papers in statistical physics (Hertz et al., 1993; Schwarze, 1993), and using the Guerra interpolation method in Chap. 5 and (Aubin et al., 2018b). Interestingly, in this case, one can get an error that decays as $1/\alpha$ as soon as $\alpha \gg \Theta(K)$. One thus observes a huge gap between the Rademacher bound that scales as $\mathfrak{R}_n\left(\mathscr{F}_{\mathrm{com}}\right) = \Theta\left(\sqrt{K\sqrt{\log(K)}/\alpha}\right)$ and the actual generalization error $\varepsilon_{\mathrm{gen}} = \Theta(K/\alpha)$ for large sample size. This large gap further illustrates the considerable difference in behavior one can get between the worst case and teacher-student case analysis, see Fig. 46.

7.4.5 EXTENSION TO ROTATIONALLY INVARIANT MATRICES

The previous free entropy and ground state energy computation for i.i.d data matrix $\mathbf{X}$ can be generalized to Rotationally Invariant (RI) random matrices $\mathbf{X} = \mathbf{USV}$ with rotation matrices $\mathbf{U} \in \mathbf{O}(d)$, $\mathbf{V} \in \mathbf{O}(n)$ independently sampled from the Haar measure, and $\mathbf{S} \in \mathbb{R}^{d\times n}$ a diagonal matrix of singular values. Computation for this kind of matrices can be handled again using the

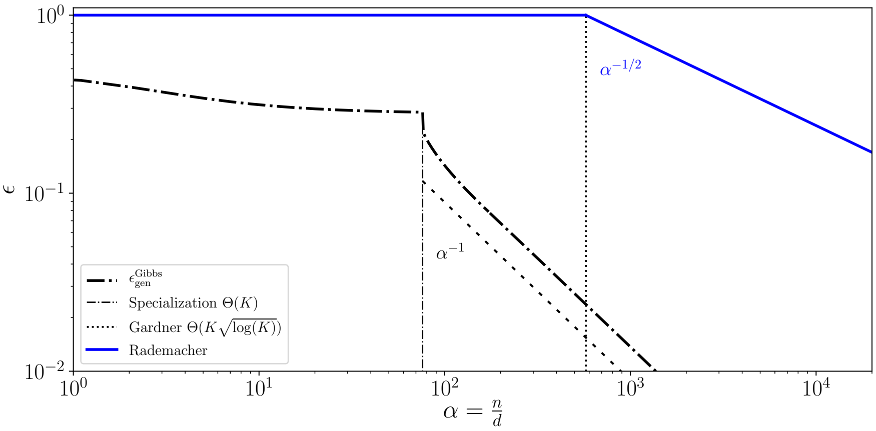


Figure 46: Illustration of the scaling of the Rademacher complexity (blue) for the fully connected committee machine, drawn together with the exact generalization error in the teacher-student scenario (dotted-dashed black), scaling as α^{-1} at large α. Notice the large gap between the worst case bound (Rademacher) and the teacher-student result.

replica method (Kabashima, 2008; Barbier et al., 2018b; Gabrié et al., 2018) and leads to RS and 1RSB free energies

$$\begin{aligned}
\varphi_{\rm RI}^{(\rm rs)}(\alpha,\beta) &= -\frac{1}{\beta}\textbf{extr}_{\chi_w,\chi_u,q_w,q_u}\Big\{\mathscr{A}_0^{(\rm rs)}(\chi_w,\chi_u,q_w,q_u) \\
&\quad +\mathscr{A}_w^{(\rm rs)}(\chi_w,q_w)+\alpha\mathscr{A}_u^{(\rm rs)}(\chi_u,q_u,\beta)\Big\}\,, \\
\varphi_{\rm RI}^{(\rm 1rsb)}(\alpha,\beta) &= -\frac{1}{\beta}\textbf{extr}_{\chi_w,\chi_u,v_w,v_u,q_w,q_u,x}\{ \\
&\quad +\mathscr{A}_0^{(\rm 1rsb)}(\chi_w,\chi_u,v_w,v_u,q_w,q_u,x) \\
&\quad +\mathscr{A}_w^{(\rm 1rsb)}(\chi_w,v_w,q_w,x)+\alpha\mathscr{A}_u^{(\rm 1rsb)}(\chi_u,v_u,q_u,x,\beta)\Big\}\,,
\end{aligned} \tag{183}$$

where each term is properly defined in Appendix B.2 of (Abbara et al., 2020). Note that taking $\mathbf{X}$ a random Gaussian i.i.d matrix, the eigenvalue density $\rho(\lambda)$ follows the Marchenko-Pastur distribution and (183) matches free energies eq. (171), and ground states energies eq. (174) in the spherical case. The ground state energy (and therefore the Rademacher complexity) can be again computed as in the i.i.d case, taking the zero temperature limit $\beta \to \infty$

$$e_{\rm gs,RI}(\alpha) = \lim_{\beta\to\infty}\varphi_{\rm RI}(\alpha,\beta)\,, \tag{184}$$

keeping in particular $\beta\chi_w$ and $\omega = x\beta$ finite in the limits $\beta\to\infty, x\to 0, \chi_w \to 0$.

CONCLUSION

In this chapter, we discussed the deep connection between the Rademacher complexity and some of the classical quantities studied in the statistical physics literature on neural networks, namely the Gardner capacity, the ground state energy of the random perceptron model Chap. 6, and the generalization error in the T-S model discussed in Chap. 5. We believe it is rather interesting to draw the link with approaches inspired by statistical physics, and compare its findings with the worst-case results. In the wake of the need to understand the effectiveness of neural networks and also the limitations of the classical approaches, it is of interest to revisit the results that have emerged thanks to the physics perspective. This direction is currently experiencing a strong revival, see e.g. (Chaudhari et al., 2017; Martin et al., 2017; Advani et al., 2017; Baity-Jest et al., 2018). The connection discussed in the paper opens the way to a unified presentation of these often contrasted approaches, and we hope this paper will help bridging the gap between researchers in traditional statistics and in statistical physics. There are many possible follow-ups, the more natural one being the computation of Rademacher complexities from statistical physics methods for more complicated and realistic models of data, starting for instance with correlated matrices discussed in Sec. 7.4.5.

8 GENERALIZATION ERROR IN HIGH-DIMENSIONAL PERCEPTRONS: APPROACHING BAYES ERROR WITH CONVEX OPTIMIZATION

High-dimensional statistics, where the ratio $\alpha = n/d$ is kept finite while the dimensionality d and the number of samples n grow, often display interesting non-intuitive features. Asymptotic generalization performances for such problems in the so-called T-S setting, with synthetic data, have been the subject of intense investigations spanning many decades (Seung et al., 1992; Watkin et al., 1993; Engel et al., 2001; Bayati et al., 2011a; El Karoui et al., 2013; Donoho et al., 2016). To understand the effectiveness of modern machine learning techniques, and also the limitations of the classical statistical learning approaches (Zhang et al., 2016; Belkin et al., 2019a), it is of interest to revisit this line of research. Indeed, this direction is currently the subject to a renewal of interests, as testified by some very recent, yet already rather influential papers (Candès et al., 2020; Barbier et al., 2019b; Hastie et al., 2019; Belkin et al., 2019b; Mei et al., 2019). The present work subscribes to this line of work and studies high-dimensional classification within one of the simplest models considered in statistics and machine learning: convex linear estimation with data generated by a teacher *perceptron* (Gardner et al., 1989). We will focus on the generalization abilities in this problem, and compare the performances of Bayes-optimal estimation to the more standard ERM. We then compare the results with the prediction of standard generalization bounds that illustrate in particular their limitation even in this simple, yet non-trivial, setting.

Synthetic data model — We consider a supervised machine learning task, whose dataset is generated by a single layer neural network, often named a *teacher* (Seung et al., 1992; Watkin et al., 1993; Engel et al., 2001), that belongs to the GLM class. Therefore we assume the n samples are drawn according to

$$\mathbf{y} = \varphi^\star_{\text{out}}\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}^\star\right) \Leftrightarrow \mathbf{y} \sim \mathrm{P}^\star_{\text{out}}\left(.\right) , \tag{185}$$

where $\mathbf{w}^\star \in \mathbb{R}^d$ denotes the ground truth vector drawn from a probability distribution $\mathrm{P}_{\mathbf{w}^\star}$ with second moment $\rho_{\mathbf{w}^\star} \equiv \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\left[\|\mathbf{w}^\star\|_2^2\right]$ and $\varphi^\star_{\text{out}}$ represents a deterministic or stochastic activation function equivalently associated to a distribution $\mathrm{P}_{\mathrm{out}}^{\star}$. The input data matrix $\mathbf{X} = (\mathbf{x}_\mu)_{\mu=1}^{n} \in \mathbb{R}^{n\times d}$ contains i.i.d Gaussian vectors, i. e. $\forall \mu \in [\![n]\!]$, $\mathbf{x}_\mu \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Even though the framework we use and the theorems and results we derived are valid for a rather generic channel in eq. (185) —including regression problems— we will mainly focus the presentation on the commonly considered perceptron case: a binary classification task with data given by a sign activation function $\varphi_{\mathrm{out}}^{\star}(\mathbf{z}) = \mathrm{sign}(\mathbf{z})$, with a Gaussian weight distribution $\mathrm{P}_{\mathrm{w}^\star}(\mathbf{w}^\star) = \mathcal{N}_{\mathrm{w}^\star}(\mathbf{0}, \rho_{\mathrm{w}^\star}\mathbf{I}_d)$. The ±1 labels are thus generated as

$$\mathbf{y} = \mathrm{sign}\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}^\star\right), \quad \text{with} \quad \mathbf{w}^\star \sim \mathcal{N}_{\mathrm{w}^\star}(\mathbf{0}, \rho_{\mathrm{w}^\star}\mathbf{I}_d). \tag{186}$$

Empirical Risk Minimization — The workhorse of machine learning is ERM, where one minimizes a *loss function* in the corresponding high-dimensional parameter space $\mathbb{R}^d$. To avoid overfitting of the training set one often adds a *regularization term* $r(\mathbf{w})$. ERM then corresponds to estimating $\hat{\mathbf{w}}_{\mathrm{erm}} = \mathrm{argmin}_{\mathbf{w}}\left[\mathcal{L}(\mathbf{w};\mathbf{y},\mathbf{X})\right]$ where the regularized training loss $\mathcal{L}$ is defined by, using the notation $z_\mu(\mathbf{w},\mathbf{x}_\mu) \equiv \frac{1}{\sqrt{d}}\mathbf{x}_\mu^\intercal \mathbf{w}$,

$$\mathcal{L}(\mathbf{w};\mathbf{y},\mathbf{X}) = \sum_{\mu=1}^{n} l(y_\mu, z_\mu(\mathbf{w},\mathbf{x}_\mu)) + r(\mathbf{w}). \tag{187}$$

The goal of the present chapter is to discuss the generalization performance of these estimators for the classification task (186) in the high-dimensional limit. We focus our analysis on commonly used loss functions l, namely the square $l^{\mathrm{square}}(y,z) = \frac{1}{2}(y-z)^2$, logistic $l^{\mathrm{logistic}}(y,z) = \log(1+\exp(-yz))$ and hinge losses $l^{\mathrm{hinge}}(y,z) = \max(0, 1-yz)$. We will mainly illustrate our results for the ℓ_2 regularization $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2/2$, where we introduced a regularization strength hyper-parameter λ. The same analysis can be performed for any other convex-separable regularization.

Related works — The above learning problem has been extensively studied in the statistical physics community using the heuristic replica method (Gardner et al., 1989; Seung et al., 1992; Watkin et al., 1993; Opper et al., 1996a; Engel et al., 2001). Due to the interest in high-dimensional statistics, they have experienced a resurgence in popularity in recent years. In particular, rigorous works on related problems are much more recent. The authors of (Barbier et al., 2019b) established rigorously the replica-theory predictions for the Bayes-optimal generalization error. Here we focus on standard ERM estimation and compare it to the results obtained in (Barbier et al., 2019b). Authors of (Thrampoulidis et al., 2018) analyzed rigorously M-estimators for the regression case where data are generated by a linear-activation teacher. Here we analyze classification with a more general and non-linear teacher, focusing in particular on the sign-teacher. The case of max-margin loss was studied in (Montanari et al., 2019) with a technically closely related proof, but with a focus on the over-parametrized regime, thus not addressing the questions that we focus on. A range of unregularized losses was also analyzed for a sigmoid teacher (that is very similar to a sign-teacher) again in the context of the double-descent behavior in (Deng et al., 2019; Kini et al., 2020). Here we focus instead on the regularized case as it drastically improves generalization performances of the ERM and that allows us to compare with the Bayes-optimal estimation as well as to standard generalization bounds. Our proof, as in the above mentioned works and (Mignacco et al., 2020a), is based on Gordon's minimax formalism, including in particular the effect of the regularization.

Main contributions — Our first main contribution is to provide rigorously, in Sec. 8.1, the classification generalization performances of empirical risk minimization with the loss given by (187) in the high-dimensional limit, for any convex loss and an ℓ_2 regularization. Note that the proof is easily extended to any convex separable regularization. Additionally, we provide a proof of the equivalence between the results of our paper and the ones initially obtained by the replica method, which is of additional interest given the wide range of application of these heuristics statistical-physics technics in machine learning and computer science (Mézard et al., 2009; Zdeborová, 2020). In particular, the replica predictions in (Opper et al., 1990; Opper et al., 1991b; Opper et al., 1996a) follow from our results. Another approach that originated in physics are the so-called TAP equations (Mézard, 1989; Kabashima, 2003; Kabashima et al., 2004) that lead to the AMP algorithm for solving linear and generalized linear problems with Gaussian matrices (Donoho et al., 2009; Rangan, 2011). This algorithm can be analyzed with the so-called SE method (Bayati et al., 2011b), and it is widely believed, and in fact proven for linear problems (Bayati et al., 2011a; Gerbelot et al., 2020) that the fixed-point of the SE gives the optimal error in high-dimensional convex optimization problems. The SE equations are in fact equivalent to the one given by the replica theory and therefore our results vindicate this approach as well. We also demonstrate numerically that these asymptotic results are very accurate even for moderate system sizes, and they have been performed with the scikit-learn library (Pedregosa et al., 2011).

Secondly, and more importantly, we provide in Sec. 8.2 a detailed analysis of the generalization error for standard losses such as square, hinge (or equivalently support vector machine) and logistic, as a function of the regularization strength λ and the number of samples per dimension α. We observe, in particular, that while the ridge regression never closely approaches the Bayes-optimal performance, the logistic regression with optimized ℓ_2 regularization gets extremely close to optimal. And so does, to a lesser extent, the hinge regression and the max-margin estimator to which the unregularized logistic and hinge converge (Rosset et al., 2004). It is quite remarkable that these canonical losses are able to approach the error of the Bayes-optimal estimator for which, in principle, the marginals of a high-dimensional probability distribution need to be evaluated. Notably, all the later losses give —for a *good choice* of the regularization strength λ— generalization errors scaling as $\Theta\left(\alpha^{-1}\right)$ for large α, just as the Bayes-optimal generalization error (Barbier et al., 2019b). This is found to be at variance with the prediction of Rademacher and max-margin-based bounds that predict instead a $\Theta\left(\alpha^{-1/2}\right)$ rate (Vapnik, 2006; Shalev-Shwartz et al., 2014), which therefore appear to be vacuous in the high-dimensional regime.

Third, in Sec. 8.3, we design a custom, non-convex, loss and regularizer that provably gives a plug-in estimator that efficiently achieves Bayes-optimal performances, including the optimal $\Theta\left(\alpha^{-1}\right)$ rate for the generalization error. Our construction is related to the one discussed in (Gribonval, 2011; Gribonval et al., 2013; Advani et al., 2016a), but is not restricted to convex losses.

8.1 MAIN TECHNICAL RESULTS

In the formulas that arise for this statistical estimation problem, the correlations between the estimator $\hat{\mathbf{w}}$ and the ground truth vector $\mathbf{w}^\star$ play a fundamental role and we thus define two scalar overlap parameters to measure the statistical reconstruction:

$$m \equiv \frac{1}{d}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\hat{\mathbf{w}}^\intercal \mathbf{w}^\star\right], \qquad q \equiv \frac{1}{d}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\|\hat{\mathbf{w}}\|_2\right]^2. \tag{188}$$

In particular, the generalization error of the estimator $\hat{\mathbf{w}}(\alpha) \in \mathbb{R}^d$, obtained by performing ERM on the training loss $\mathscr{L}$ in eq. (187) with $n = \alpha d$ samples,

$$e_{\mathrm{g}}^{\mathrm{erm}}(\alpha) \equiv \mathbb{E}_{y,\mathbf{x}}\mathbb{1}\left[y \neq \hat{y}\left(\hat{\mathbf{w}}(\alpha);\mathbf{x}\right)\right], \tag{189}$$

where $\hat{y}\left(\hat{\mathbf{w}}(\alpha);\mathbf{x}\right)$ denotes the predicted label, has both at finite size d and in the asymptotic limit an explicit expression depending only on the above overlaps m and q:

Proposition 8.1.1 (Generalization error of classification)**.** *In our synthetic binary classification task, the generalization error of ERM (or equivalently the test error) is given by*

$$e_g^{erm}(\alpha) = \frac{1}{\pi} acos\left(\sqrt{\eta}\right), \tag{190}$$

with

$$\eta \equiv \frac{m^2}{\rho_d\, q}, \qquad \rho_d \equiv \frac{1}{d}\mathbb{E}\left[\|\mathbf{w}^\star\|_2^2\right].$$

Proof. The proof is a simple computation based on Gaussian integration. The generalization error e_{g} is the prediction error of the estimator $\hat{\mathbf{w}}$ on new samples $\{\mathbf{y},\mathbf{X}\}$, where $\mathbf{X}$ is a Gaussian matrix with i.i.d entries and $\mathbf{y}$ are ± 1 labels generated according to eq. (185) $\mathbf{y} = \varphi_{\mathrm{out}^\star}(\mathbf{z})$ with $\mathbf{z} = \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}^\star$. As the model fitted by ERM may not lead to binary outputs, we add a non-linearity $\varphi : \mathbb{R} \mapsto \{\pm 1\}$ (for example a sign or a soft-sign) on top of it to ensure to obtain binary outputs $\hat{\mathbf{y}} \pm 1$ according to $\hat{\mathbf{y}} = \varphi_{\mathrm{out}}(\hat{\mathbf{z}})$ with $\hat{\mathbf{z}} = \frac{1}{\sqrt{d}}\mathbf{X}\hat{\mathbf{w}}$.

The classification generalization error is given by the probability that the predicted labels $\hat{\mathbf{y}}$ and the true labels $\mathbf{y}$ do not match. To compute it, first note that the vectors $(\mathbf{z},\hat{\mathbf{z}})$ averaged over all possible ground truth vectors $\mathbf{w}^\star$ (or equivalently labels y) and input matrix $\mathbf{X}$ follow in the large size limit a joint Gaussian distribution with zero mean and covariance matrix

$$\boldsymbol{\sigma} = \frac{1}{d}\mathbb{E}_{\mathbf{w}^\star,\mathbf{X}}\begin{bmatrix} \|\mathbf{w}^\star\|_2^2 & \mathbf{w}^{\star\intercal}\hat{\mathbf{w}} \\ \mathbf{w}^{\star\intercal}\hat{\mathbf{w}} & \|\hat{\mathbf{w}}\|_2^2 \end{bmatrix} \equiv \begin{bmatrix} \rho_d & \sigma_{\mathrm{w}^\star\hat{\mathrm{w}}} \\ \sigma_{\mathrm{w}^\star\hat{\mathrm{w}}} & \sigma_{\hat{\mathrm{w}}} \end{bmatrix}. \tag{191}$$

The asymptotic generalization error depends only on the covariance matrix $\boldsymbol{\sigma}$ and as the samples are i.i.d it reads

$$\begin{aligned} e_{\mathrm{g}}(\alpha) &= \mathbb{1}\left[y \neq \hat{y}\left(\hat{\mathbf{w}}(\alpha);\mathbf{x}\right)\right] = 1 - \mathbb{P}[y = \hat{y}\left(\hat{\mathbf{w}}(\alpha);\mathbf{x}\right)] \\ &= 1 - 2\int_{(\mathbb{R}^+)^2} \mathrm{d}\mathbf{x}\, \mathcal{N}_{\mathbf{x}}\left(\mathbf{0},\boldsymbol{\sigma}\right) \\ &= 1 - \left(\frac{1}{2} + \frac{1}{\pi}\operatorname{atan}\left(\sqrt{\frac{\sigma^2_{\mathrm{w}^\star\hat{\mathrm{w}}}}{\rho_d\sigma_{\hat{\mathrm{w}}} - \sigma^2_{\mathrm{w}^\star\hat{\mathrm{w}}}}}\right)\right) = \frac{1}{\pi}\operatorname{acos}\left(\eta\right), \end{aligned} \tag{192}$$

where we used the fact that $\operatorname{atan}(x) = \frac{1}{2}\left(\pi - \operatorname{acos}\left(\frac{x^2-1}{1+x^2}\right)\right)$, $\frac{1}{2}\operatorname{acos}\left(2x^2-1\right) = \operatorname{acos}(x)$ and defined $\eta = \frac{\sigma_{\mathrm{w}^\star\hat{\mathrm{w}}}}{\sqrt{\rho_d\sigma_{\hat{\mathrm{w}}}}}$. For the ERM estimator, the parameters $\sigma_{\hat{\mathrm{w}}} = \frac{1}{d}\mathbb{E}_{\mathbf{w}^\star,\mathbf{X}}\|\hat{\mathbf{w}}\|_2^2 = q$ and $\sigma_{\mathrm{w}^\star\hat{\mathrm{w}}} = \frac{1}{d}\mathbb{E}_{\mathbf{w}^\star,\mathbf{X}}\left(\hat{\mathbf{w}}\right)^\intercal\mathbf{w}^\star = m$, such that the generalization error for classification is given by (192) with $\eta \equiv \frac{m^2}{\rho_d q}$. □

To obtain the generalization performances of ERM, it remains to obtain the asymptotic values of m, q (and thus of η), in the limit $d \to \infty$. For any $\tau > 0$, let us first recall the definitions of the Moreau-Yosida regularization $\mathcal{M}_\tau$ and the proximal operator $\mathcal{P}_\tau$ of a convex loss function $(y,z) \mapsto \ell(y\cdot z)$:

$$\begin{aligned} \mathcal{M}_\tau(z) &= \min_x\left\{\ell(x) + \frac{(x-z)^2}{2\tau}\right\}, \\ \mathcal{P}_\tau(z) &= \operatorname{argmin}_x\left\{\ell(x) + \frac{(x-z)^2}{2\tau}\right\}. \end{aligned} \tag{193}$$

With the ℓ_2 regularization, the asymptotic overlaps are characterized by a set of fixed point equations and follow from the Gordon's Convex Gaussian Min-max Theorem (CGMT) states in the next theorems.

Theorem 8.1.2 (Gordon's min-max fixed point - Regression/Classification with ℓ_2 regularization)**.** *As $n,d \to \infty$ with $n/d = \alpha = \Theta(1)$, the overlap parameters m,q concentrate to*

$$m \underset{d\to\infty}{\longrightarrow} \sqrt{\rho_{\mathrm{w}^\star}}\,\nu^*, \qquad q \underset{d\to\infty}{\longrightarrow} (\nu^*)^2 + (\delta^*)^2 \qquad \rho_d \underset{d\to\infty}{\longrightarrow} \rho_{\mathrm{w}^\star}. \tag{194}$$

*For **regression** the parameters ν^*,δ^* are the solutions of*

$$\begin{aligned} (\nu^*,\delta^*) = \underset{\nu,\delta\geq 0}{\arg\min}\ \sup_{\tau>0}\Bigg\{ & \frac{\lambda(\nu^2+\delta^2)}{2} - \frac{\delta^2}{2\tau} \\ & + \alpha\mathbb{E}_{g,s}\mathcal{M}_\tau[l(\varphi_{out^\star}(\sqrt{\rho_{\mathrm{w}^\star}}s),.)](\nu s + \delta g)\Bigg\}, \end{aligned} \tag{195}$$

while for ***classification****,* v^*, δ^* *are the solutions of*

$$(v^*,\delta^*) = \underset{v,\delta\geq 0}{\arg\min}\ \sup_{\tau>0} \Bigg\{ \frac{\lambda(v^2+\delta^2)}{2} - \frac{\delta^2}{2\tau} \\ + \alpha\mathbb{E}_{g,s}\mathscr{M}_\tau[\delta g + v s\varphi_{out^\star}(\sqrt{\rho_{w^\star}}s)]\Bigg\} . \tag{196}$$

Here, g, s *are two* i.i.d *standard Gaussian normal random variables. The solutions* (v^*,δ^*) *of* (196) *for classification can be reformulated as a set of fixed point equations*

$$\begin{aligned} v^* &= \frac{\alpha}{\lambda\tau^* + \alpha}\mathbb{E}_{g,s}[s \times \varphi_{out^\star}(\sqrt{\rho_{w^\star}}s) \\ &\qquad \times \mathscr{P}_{\tau^*}(\delta^* g + v^* s\varphi_{out^\star}(\sqrt{\rho_{w^\star}}s))] , \\ \delta^* &= \frac{\alpha}{\lambda\tau^* + \alpha - 1}\mathbb{E}_{g,s}[g \times \mathscr{P}_{\tau^*}(\delta^* g + v^* s\varphi_{out^\star}(\sqrt{\rho_{w^\star}}s))] , \\ (\delta^*)^2 &= \alpha\mathbb{E}_{g,s}[((\delta^* g + v^* s\varphi_{out^\star}(\sqrt{\rho_{w^\star}}s)) \\ &\qquad - \mathscr{P}_{\tau^*}(\delta^* g + v^* s\varphi_{out^\star}(\sqrt{\rho_{w^\star}}s)))^2] . \end{aligned} \tag{197}$$

Proof. Since the teacher weight vector $\mathbf{w}^\star$ is independent of the input data matrix $\mathbf{X}$, we can assume without loss of generality that $\mathbf{w}^\star = \sqrt{d\,\rho_d}\mathbf{e}_1$ where $\mathbf{e}_1$ is the first natural basis vector of $\mathbb{R}^d$, and $\rho_d = \|\mathbf{w}^\star\|_2^2/d$. As $d \to \infty$, $\rho_d \to \rho_{w^\star}$. Accordingly, it will be convenient to split the data matrix into two parts $\mathbf{X} = [\mathbf{s}, \mathbf{B}]$, where $\mathbf{s} \in \mathbb{R}^{n\times 1}$ and $\mathbf{B} \in \mathbb{R}^{n\times(d-1)}$ are two submatrices of i.i.d standard normal entries. The weight vector $\mathbf{w}$ can also be written as $\mathbf{w} = [\sqrt{d}v, \mathbf{v}^\intercal]^\intercal$, where $v \in \mathbb{R}$ denotes the projection of $\mathbf{w}$ onto the direction spanned by the teacher weight vector $\mathbf{w}^\star$, and $\mathbf{v} \in \mathbb{R}^{d-1}$ is the projection of $\mathbf{w}$ onto the complement subspace. These representations serve to simplify the notations in our subsequent derivations. For example, we can now write the output as $y_\mu = \varphi_{\mathrm{out}^\star}(\sqrt{\rho_d}s_\mu)$ where s_μ is the μ-th entry of the Gaussian vector $\mathbf{s}$. Let Φ_d denote the cost of the ERM according to the loss (187), normalized by d. Using our new representations introduced above, we have

$$\Phi_d = \min_{v,\mathbf{v}} \frac{1}{d}\sum_{\mu=1}^{n} l\left(y_\mu, v s_\mu + \tfrac{1}{\sqrt{d}}\mathbf{b}_\mu^\intercal \mathbf{v}\right) + \frac{\lambda(dv^2 + \|\mathbf{v}\|^2)}{2d}, \tag{198}$$

where $\mathbf{b}_\mu^\intercal$ denotes the i-th row of $\mathbf{B}$. Since the loss function $l(y_\mu, z)$ is convex with respect to z, we can rewrite it as $l(y_\mu, z) = \sup_q\{qz - l^*(y_\mu, q)\}$, where $l^*(y_\mu, q) = \sup_z\{qz - l(y_\mu, z)\}$ is its convex conjugate. Substituting l into (198), we obtain

$$\Phi_d = \min_{v,\mathbf{v}} \sup_{\mathbf{q}} \Bigg\{ \frac{v\mathbf{q}^\intercal\mathbf{s}}{d} + \frac{1}{d^{3/2}}\mathbf{q}^\intercal\mathbf{B}\mathbf{v} \\ - \frac{1}{d}\sum_{\mu=1}^{n} l^*(y_\mu, q_\mu) + \frac{\lambda\left(dv^2 + \|\mathbf{v}\|^2\right)}{2d} \Bigg\} .$$

Now consider a new optimization problem

$$\tilde{\Phi}_d = \min_{v,\mathbf{v}} \sup_{\mathbf{q}} \left\{ \frac{v\mathbf{q}^\top \mathbf{s}}{d} + \frac{\|\mathbf{q}\|}{\sqrt{d}} \frac{\mathbf{h}^\top \mathbf{v}}{d} + \frac{\|\mathbf{v}\|}{\sqrt{d}} \frac{\mathbf{g}^\top \mathbf{q}}{d} \right.$$
$$\left. - \frac{1}{d} \sum_{\mu=1}^{n} l^*(y_\mu, q_\mu) + \frac{\lambda \left(dv^2 + \|\mathbf{v}\|^2 \right)}{2d} \right\},$$

where $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d-1})$ and $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ are two independent standard normal vectors. It follows from Gordon's minimax comparison inequality, see e. g. (Gordon, 1985; Thrampoulidis et al., 2015), that

$$\mathbb{P}(|\Phi_d - c| \geq \varepsilon) \leq 2\mathbb{P}\left(\left| \tilde{\Phi}_d - c \right| \geq \varepsilon \right), \tag{199}$$

for any constants c and $\varepsilon > 0$. This implies that $\tilde{\Phi}_d$ serves as a surrogate of Φ_d. Specifically, if $\tilde{\Phi}_d$ concentrates around some deterministic limit c as $d \to \infty$, so does Φ_d. In what follows, we proceed to solve the surrogate problem for $\tilde{\Phi}_d$. First, let $\delta = \|\mathbf{v}\| / \sqrt{d}$. It is easy to see that $\tilde{\Phi}_d$ can be simplified as

$$\tilde{\Phi}_d = \min_{v,\delta \geq 0} \sup_{\mathbf{q}} \left\{ \frac{\mathbf{q}^\top (v\mathbf{s} + \delta \mathbf{g})}{d} - \delta \frac{\|\mathbf{q}\|}{\sqrt{d}} \frac{\|\mathbf{h}\|}{\sqrt{d}} \right.$$
$$\left. - \frac{1}{d} \sum_{\mu=1}^{n} l^*(y_\mu, q_\mu) + \frac{\lambda (v^2 + \delta^2)}{2} \right\}$$
$$\overset{(a)}{=} \min_{v,\delta \geq 0} \sup_{\tau > 0} \sup_{\mathbf{q}} \left\{ -\frac{\tau \|\mathbf{q}\|^2}{2d} - \frac{\delta^2 \|\mathbf{h}\|^2}{2\tau d} + \frac{\mathbf{q}^\top (v\mathbf{s} + \delta \mathbf{g})}{d} \right.$$
$$\left. - \frac{1}{d} \sum_{\mu=1}^{n} l^*(y_\mu, q_\mu) + \frac{\lambda (v^2 + \delta^2)}{2} \right\}$$
$$= \min_{v,\delta \geq 0} \sup_{\tau > 0} \left\{ \frac{\lambda (v^2 + \delta^2)}{2} - \frac{\delta^2 \|\mathbf{h}\|^2}{2\tau d} \right.$$
$$\left. - \frac{\alpha}{n} \inf_{\mathbf{q}} \left[\frac{\tau \|\mathbf{q}\|^2}{2} - \mathbf{q}^\top (v\mathbf{s} + \delta \mathbf{g}) + \sum_{\mu=1}^{n} l^*(y_\mu, q_\mu) \right] \right\}$$
$$\overset{(b)}{=} \min_{v,\delta \geq 0} \sup_{\tau > 0} \left\{ \frac{\lambda (v^2 + \delta^2)}{2} - \frac{\delta^2 \|\mathbf{h}\|^2}{2\tau d} \right.$$
$$\left. - \frac{\alpha}{n} \sum_{\mu=1}^{n} \mathscr{M}_\tau[l(y_\mu, .)](v s_\mu + \delta g_\mu) \right\}.$$

In (a), we have introduced an auxiliary variable τ to rewrite $-\delta \frac{\|\mathbf{q}\|}{\sqrt{d}} \frac{\|\mathbf{h}\|}{\sqrt{d}}$ as

$$-\delta \frac{\|\mathbf{q}\|}{\sqrt{d}} \frac{\|\mathbf{h}\|}{\sqrt{d}} = \sup_{\tau > 0} \left\{ -\frac{\tau \|\mathbf{q}\|^2}{2d} - \frac{\delta^2 \|\mathbf{h}\|^2}{2\tau d} \right\},$$

and to get (b), we use the identity

$$\inf_q \left\{ \frac{\tau}{2} q^2 - qz + \ell^*(q) \right\} = -\inf_x \left\{ \frac{(z-x)^2}{2\tau} + \ell(x) \right\},$$

that holds for any z and for any convex function $\ell(x)$ and its conjugate $\ell^*(q)$. As $d \to \infty$, standard concentration arguments give us $\frac{\|\mathbf{h}\|^2}{d} \to 1$ and $\frac{1}{n}\sum_{\mu=1}^{n} \mathscr{M}_\tau[l(y_\mu, .)](\nu s_\mu + \delta g_\mu) \to \mathbb{E}_{g,s} \mathscr{M}_\tau[l(y, .)](\nu s + \delta g)$ uniformly over τ, ν and δ. Using (199), we can conclude that the normalized cost of the ERM Φ_d converges to the optimal value of the deterministic optimization problem in (195). Finally, since $\lambda > 0$, one can show that the cost function of (195) has a unique global minima at ν^* and δ^*. It follows that the empirical values of (ν, δ) also converge to their corresponding deterministic limits (ν^*, δ^*).

To obtain the result for *classification*, we note that

$$\begin{aligned} \mathscr{M}_\tau[l(y, .)](z) &= \min_x \left\{ l(y; x) + \frac{(x-z)^2}{2\tau} \right\} = \min_x \left\{ \ell(yx) + \frac{(x-z)^2}{2\tau} \right\} \\ &= \min_x \left\{ \ell(x) + \frac{(x-yz)^2}{2\tau} \right\} = \mathscr{M}_\tau(yz), \end{aligned}$$

where to reach the last equality we have used the fact that $y \in \{\pm 1\}$. Substituting this special form into (195) and recalling $y_\mu = \varphi_{\text{out}^\star}(\sqrt{\rho_d} s_\mu)$, we obtain the result. Finally, to obtain the fixed point equations, we simply take the partial derivatives of the cost function with respect to ν, δ, τ, and use the following well-known calculus rules for the Moreau-Yosida regularization (Hiriart-Urruty et al., 1993):

$$\frac{\partial \mathscr{M}_\tau(z)}{\partial z} = \frac{z - \mathscr{P}_\tau(z)}{\tau}, \qquad \frac{\partial \mathscr{M}_\tau(z)}{\partial \tau} = -\frac{(z - \mathscr{P}_\tau(z))^2}{2\tau^2}.$$

□

Interestingly, this set of fixed point equations (197) can be finally mapped to the ones obtained by the heuristic *replica* method from statistical physics, whose heuristic derivation is shown in SM. III.3 of (Aubin et al., 2020a), as well as the SE of the AMP (Kabashima, 2003; Rangan, 2011; Zdeborová et al., 2016a). Thus their validity for this convex estimation problem is rigorously established by the following theorem:

Corollary 8.1.3 (Equivalence Gordon-replicas). *As $n,d \to \infty$ with $n/d = \alpha = \Theta(1)$, the overlap parameters m, q concentrate to the fixed point of the following set of equations:*

$$
\begin{aligned}
m &= \alpha\Sigma\rho_{\mathrm{w}^\star} \times \mathbb{E}_{y,\xi}\left[\mathcal{Z}_{out^\star}\left(y, \sqrt{\rho_{\mathrm{w}^\star}\eta}\xi, \rho_{\mathrm{w}^\star}(1-\eta)\right)\right.\\
&\qquad\qquad \left.\times f_{out^\star}\left(y, \sqrt{\rho_{\mathrm{w}^\star}\eta}\xi, \rho_{\mathrm{w}^\star}(1-\eta)\right) \times f_{out}\left(y, q^{1/2}\xi, \Sigma\right)\right]\\
q &= m^2/\rho_{\mathrm{w}^\star} + \alpha\Sigma^2 \times \mathbb{E}_{y,\xi}\left[\mathcal{Z}_{out^\star}\left(y, \sqrt{\rho_{\mathrm{w}^\star}\eta}\xi, \rho_{\mathrm{w}^\star}(1-\eta)\right)\right.\\
&\qquad\qquad\qquad \left.\times f_{out}\left(y, q^{1/2}\xi, \Sigma\right)^2\right]\\
\Sigma &= \left(\lambda - \alpha \times \mathbb{E}_{y,\xi}\left[\mathcal{Z}_{out^\star}\left(y, \sqrt{\rho_{\mathrm{w}^\star}\eta}\xi, \rho_{\mathrm{w}^\star}(1-\eta)\right)\right.\right.\\
&\qquad\qquad\qquad \left.\left.\times \partial_\omega f_{out}\left(y, q^{1/2}\xi, \Sigma\right)\right]\right)^{-1}
\end{aligned}
\tag{200}
$$

with $\eta \equiv \frac{m^2}{\rho_{\mathrm{w}^\star} q}$ and where ξ, z denote two i.i.d standard normal random variables, and $\mathbb{E}_y$ the continuous or discrete sum over all possible values y according to $\mathrm{P}_{out^\star}$. The corresponding functions $\mathcal{Z}_{out^\star}$, $f_{out^\star}$ and f_{out}, $\partial_\omega f_{out}$ are defined in Appendix. A.4.1.a-A.4.1.b.

For clarity, the proof is left in SM. III.3 of (Aubin et al., 2020a). Moreover, an equivalent set of six equations for the whole GLM class (classification and regression) with any separable and convex regularizer different than ℓ_2 are shown in Appendix. B.1.2.a and in SM. III.2 of (Aubin et al., 2020a).

Bayes optimal baseline — Finally, we shall compare the ERM performances to the Bayes-optimal generalization error. Being the information-theoretically best possible estimator, we will use it as a reference baseline for comparison. The expression of the Bayes-optimal generalization was derived in (Opper et al., 1991b) and proven in (Barbier et al., 2019b) and we recall here the result:

Theorem 8.1.4 (Bayes asymptotic performance, from (Barbier et al., 2019b)). *For the model (185) with $\mathrm{P}_{\mathrm{w}^\star}(\mathbf{w}^\star) = \mathcal{N}_{\mathrm{w}^\star}(\mathbf{0}, \rho_d \mathbf{I}_d)$, such that $\rho_d \underset{d\to\infty}{\longrightarrow} \rho_{\mathrm{w}^\star}$, the Bayes-optimal generalization error is quantified by two scalar parameters q_{b} and $\hat{q}_{\mathrm{b}}$ that verify asymptotically the set of fixed point equations*

$$
\begin{aligned}
q_{\mathrm{b}} &= \frac{\hat{q}_{\mathrm{b}}}{1+\hat{q}_{\mathrm{b}}},\\
\hat{q}_{\mathrm{b}} &= \alpha\mathbb{E}_{y,\xi}\left[\mathcal{Z}_{out^\star}\left(y, q_{\mathrm{b}}^{1/2}\xi, \rho_{\mathrm{w}^\star} - q_{\mathrm{b}}\right) \cdot f_{out^\star}\left(y, q_{\mathrm{b}}^{1/2}\xi, \rho_{\mathrm{w}^\star} - q_{\mathrm{b}}\right)^2\right],
\end{aligned}
\tag{201}
$$

and is expressed by

$$
e_g^{bayes}(\alpha) = \frac{1}{\pi} acos\left(\sqrt{\eta_{\mathrm{b}}}\right) \quad with \quad \eta_{\mathrm{b}} = \frac{q_{\mathrm{b}}}{\rho_{\mathrm{w}^\star}}. \tag{202}
$$

Proof. The Bayes estimator $\hat{\mathbf{w}}$ is the average over the posterior distribution, denoted $\langle . \rangle$, knowing the teacher prior $\mathrm{P}_{\mathrm{w}^\star}$ and channel $\mathrm{P}_{\mathrm{out}^\star}$ distributions so that $\hat{\mathbf{w}} = \langle \mathbf{w} \rangle$. Hence we obtain $m = q = q_{\mathrm{b}}$ and the parameters $\lim_{d\to\infty} \sigma_{\hat{\mathrm{w}}} =$

$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{\mathbf{w}^\star,\mathbf{X}} \|\langle \mathbf{w}\rangle\|_2^2 \equiv q_{\mathrm{b}}$ and $\lim_{d\to\infty} \sigma_{\mathrm{w}^\star\hat{\mathrm{w}}} = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{\mathbf{w}^\star,\mathbf{X}} \langle \mathbf{w}\rangle^\intercal \mathbf{w}^\star \equiv m_{\mathrm{b}}$. Using the Nishimori identity, see Appendix. A.3, the generalization error (190) simplifies in the Bayes-optimal setting to (192) with $\eta_{\mathrm{b}} = \frac{q_{\mathrm{b}}}{\rho_{\mathrm{w}^\star}}$. □

8.2 GENERALIZATION ERRORS

We now move to the core of the contribution and analyze the set of fixed point equations (197), or equivalently (200), leading to the generalization performances given by (190), for common classifiers on our synthetic binary classification task. As already stressed, even though the results are valid for a wide range of regularizers, we focus on estimators based on ERM with ℓ_2 regularization $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2/2$, and with square loss (ridge regression) $l^{\text{square}}(y,z) = \frac{1}{2}(y-z)^2$, logistic loss (logistic regression) $l^{\text{logistic}}(y,z) = \log(1+\exp(-yz))$ or hinge loss (SVM) $l^{\text{hinge}}(y,z) = \max(0, 1-yz)$. In particular, we study the influence of the hyper-parameter λ on the generalization performances and the different large α behavior generalization rates in the high-dimensional regime, and compare with the Bayes results. We show the solutions of the set of fixed point equations eqs. (200) in Figs. 47a, 47b, 47c respectively for ridge, hinge and logistic ℓ_2 regressions. Ridge regression is a special case, for which its quadratic loss allows to derive and fully solve the equations, see SM. V.3 of (Aubin et al., 2020a). However in general the set of equations has no analytical closed form and needs therefore to be solved numerically. It is in particular the case for logistic and hinge, whose Moreau-Yosida regularization is, however, analytical.

First, to highlight the accuracy of the theoretical predictions, we compare in Figs. 47a-47b-47c the ERM asymptotic ($d \to \infty$) generalization error with the performances of numerical simulations ($d = 10^3$, averaged over $n_s = 20$ samples) of ERM of the training loss eq. (187). Presented for a wide range of number of samples α and of regularization strength λ, we observe a perfect match between theoretical predictions and numerical simulations so that the error bars are barely visible and have been therefore removed. This shows that the asymptotic predictions are valid even with very moderate sizes. As an information theoretical baseline, we also show the Bayes-optimal performances (black) given by the solution of eq. (201).

8.2.1 RIDGE ESTIMATION

As we might expect the square loss gives the worst performances. For low values of the generalization, it leads to an interpolation-peak at $\alpha = 1$. The limit of vanishing regularization $\lambda \to 0$ leads to the *least-norm* or *pseudo-inverse* estimator $\hat{\mathbf{w}}_{\text{pseudo}} = \left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{y}$. The corresponding generalization error presents the largest interpolation-peak and achieves a maximal generalization error $e_{\mathrm{g}} = 0.5$. These are well known observations, discussed as early as in (Opper et al., 1996a; Opper et al., 1990), that are object of a renewal of interest under the name *double descent*, following a recent series of papers (Hastie et al., 2019; Geiger et al., 2019; Geiger et al., 2020; Belkin et al., 2019a; Mitra, 2019; Mei et al., 2019; Gerace et al., 2020; d'Ascoli et al., 2020). This double descent behavior for the pseudo-inverse is shown in Fig. 47a with a yellow line. On the contrary, larger regularization strengths do not suffer this peak at $\alpha = 1$, but their generalization error performance is significantly worse than the Bayes-optimal baseline for larger values of α. Indeed, as we might expect, for a large number of samples, a large regularization biases wrongly the training. However, even with optimized regularizations, performances of the ridge estimator remains far away from the Bayes-optimal performance.

8.2.2 HINGE AND LOGISTIC ESTIMATION

Both these losses, which are the classical ones used in classification problems, improve drastically the generalization error. First of all, let us notice that they do not display a double-descent behavior. This is due to the fact that our results are illustrated in the noiseless case and that our synthetic dataset is always linearly separable. Optimizing the regularization, our results in Fig. 47a-47b-47c show both hinge and logistic ERM-based classification approach very closely the Bayes error. To offset these results, note that performances of logistic regression on non-linearly separable data are however very poor, as illustrated by our analysis of a *rectangle door* teacher, see SM. V.6 of (Aubin et al., 2020a).

8.2.3 MAX-MARGIN ESTIMATION

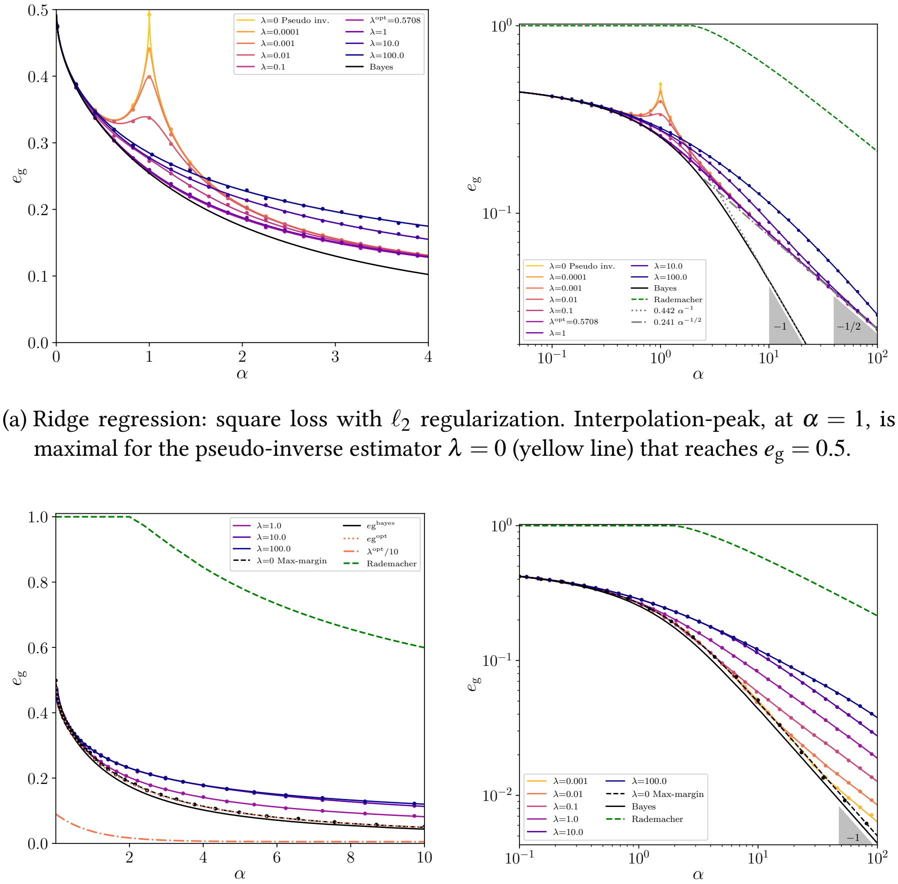
As discussed in (Rosset et al., 2004), both the logistic and hinge estimator converge, for vanishing regularization $\lambda \to 0$, to the *max-margin* solution. Taking the $\lambda \to 0$ limit in our equations, we thus obtain the *max-margin* estimator performances. While this is not what gives the best generalization error (as can be seen in Fig.47c the logistic with an optimized λ has a lower error), the max-margin estimator gives very good results, and gets very close to the Bayes-error.

8.2.4 OPTIMAL REGULARIZATION

Defining the regularization value that optimizes the generalization as

$$\lambda^{\text{opt}}(\alpha) = \text{argmin}_{\lambda}\, e_{\text{g}}^{\text{erm}}(\alpha, \lambda)\,, \tag{203}$$

we show in Figs. 47a-47b-47c that both optimal values $\lambda^{\text{opt}}(\alpha)$ (dashed-dotted orange) for logistic and hinge regression decrease to 0 as α grows and more data are given. Somehow surprisingly, we observe in particular that the generalization performances of logistic regression with optimal regularization are *extremely close* to the Bayes performances. The difference with the optimized logistic generalization error is barely visible by eye, so



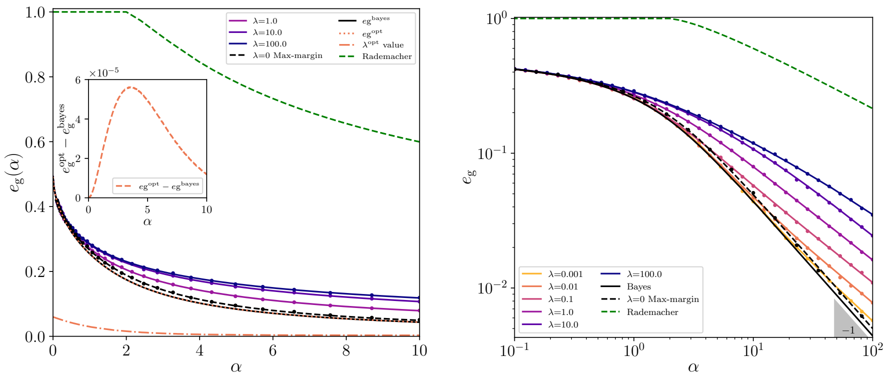
(a) Ridge regression: square loss with ℓ_2 regularization. Interpolation-peak, at $\alpha = 1$, is maximal for the pseudo-inverse estimator $\lambda = 0$ (yellow line) that reaches $e_g = 0.5$.

(b) Hinge regression: hinge loss with ℓ_2 regularization. For clarity the rescaled value of $\lambda^{\text{opt}}/10$ (dotted-dashed orange) is shown as well as its generalization error e_g^{opt} (dotted orange) that is slightly below and almost indistinguishable of the max-margin performances (dashed black).

that we explicitly plotted the difference, which is roughly of order 10^{-5}. Ridge regression Fig. 47a shows a singular behavior: there exists an optimal value (purple) which is moreover independent of α achieved for $\lambda^{\text{opt}} \simeq 0.5708$. This value was first found numerically and confirmed afterwards semi-analytically in SM. V.3 in (Aubin et al., 2020a).

8.2.5 GENERALIZATION RATES AT LARGE α

Finally, we turn to the very instructive behavior at large values of α when a large amount of data is available. First, we notice that the Bayes-optimal generalization error, whose large α analysis is performed in SM. V.1 of (Aubin et al., 2020a), decreases as $e_g^{\text{bayes}} \underset{\alpha\to\infty}{\sim} 0.4417\alpha^{-1}$. Compared to this optimal value, ridge regression gives poor performances in this regime. For any value of the regularization λ — and in particular for both the pseudo-inverse case at $\lambda = 0$ and the optimal estimator λ^{opt} — its generalization performances decrease much slower than the Bayes rate, and goes only as $e_g^{\text{ridge}} \underset{\alpha\to\infty}{\sim} 0.2405\alpha^{-1/2}$, see SM. V.3 of (Aubin et al., 2020a) for the derivation.



(c) Logistic regression: logistic loss with ℓ_2 regularization - The value of λ^{opt} (dotted-dashed orange) is shown as well as its generalization error $e_{\text{g}}^{\text{opt}}$ (dotted orange). Visually indistinguishable from the Bayes-optimal line, their difference $e_{\text{g}}^{\text{opt}} - e_{\text{g}}^{\text{bayes}}$ is shown as an inset (dashed orange).

Figure 47: Asymptotic generalization error for ℓ_2 regularization ($d \to \infty$) as a function of α for different regularizations strengths λ, compared to numerical simulation (points) of ridge regression for $d = 10^3$ and averaged over $n_s = 20$ samples. Numerics has been performed with the default methods *Ridge*, *LinearSVC*, *LogisticRegression* of scikit-learn package (Pedregosa et al., 2011). Bayes optimal performances are shown with a black line and goes as $\Theta\left(\alpha^{-1}\right)$, while the Rademacher complexity (dashed green) decrease as $\Theta\left(\alpha^{-1/2}\right)$. Both hinge and logistic converge to max-margin estimator (limit $\lambda = 0$) which is shown in dashed black and deceases as $\Theta(\alpha^{-1})$, while Ridge decreases as $\Theta(\alpha^{-1/2})$.

Hinge and logistic regressions present a radically different, and more favorable, behavior. Fig. 47b-47c show that keeping λ finite when α goes to ∞, does not yield the Bayes-optimal rates. However the max-margin solution (that corresponds to the $\lambda \to 0$ limit of these estimators) gives extremely good performances $e_{\text{g}}^{\text{logistic,hinge}} \underset{\lambda \to 0}{\sim} e_{\text{g}}^{\text{max-margin}} \underset{\alpha \to \infty}{\sim} 0.500\alpha^{-1}$ see derivation in SM. V.4 of (Aubin et al., 2020a). This is the same rate as the Bayes one, only that the constant is slightly higher.

8.2.6 COMPARISON WITH VC AND RADEMACHER STATISTICAL BOUNDS

Given the fact that both the max-margin estimator and the optimized logistic achieve optimal generalization rates going as $\Theta\left(\alpha^{-1}\right)$, it is of interest to compare those rates to the prediction of statistical learning theory bounds. Statistical learning analysis (see e.g. (Vapnik, 2006; Bartlett et al., 1998; Shalev-Shwartz et al., 2014)) relies to a large extent on the VC dimension analysis and on the so-called *Rademacher complexity*. The uniform convergence result states that if the Rademacher complexity or the VC dimension d_{vc} is finite, then for a large enough number of samples the generalization gap will vanish uniformly over all possible values of parameters. Informally, uniform convergence tells us that with high probability, for any value of the weights $\mathbf{w}$, the generalization gap satisfies $\mathscr{R}_{\text{population}}(\mathbf{w}) - \mathscr{R}^n_{\text{empirical}}(\mathbf{w}) = \Theta\left(\sqrt{d_{\text{vc}}/n}\right)$ where $d_{\text{vc}} = d - 1$ for our GLM hypothesis class. Therefore, given that the empirical risk can go to *zero* (since our data are separable), this provides a generalization error upper-bound $e_{\text{g}} \leq \Theta(\alpha^{-1/2})$. This is much worse that what we observe in practice, where we reach the Bayes rate $e_{\text{g}} = \Theta(\alpha^{-1})$. Tighter bounds can be obtained using the Rademacher complexity, and this was studied recently, using the aforementioned *replica method* (Abbara et al., 2020) for the very same problem as presented in Chap. 7. We reproduced their results and plotted the Rademacher complexity generalization bound in Fig.47 (dashed-green) that decreases as $\Theta\left(\alpha^{-1/2}\right)$ for the binary classification task eq. (186).

One may wonder if this could be somehow improved. Another statistical-physics heuristic computation, however, suggests that, unfortunately, uniform bound are plagued to a slow rate $\Theta\left(\alpha^{-1/2}\right)$. Indeed, the authors of (Engel et al., 1993) showed with a replica method-style computation that *there exists* some set of weights, in the binary classification task. (186), that lead to $\Theta\left(\alpha^{-1/2}\right)$ rates: the uniform bound is thus tight. The gap observed between the uniform bound and the almost Bayes-optimal results observed in practice in this case is therefore not a paradox, but an illustration that the price to pay for uniform convergence is the inability to describe the optimal rates one can sometimes get in practice. Therefore, we believe, that the fact this phenomena can be observed in a such simple problem sheds an interesting light on the current debate in understanding generalization in deep learning (Zhang et al., 2016).

Remarking our synthetic dataset is linearly separable, we may try to take this fact into consideration to improve the generalization rate. In particular, it can be done using the max-margin based generalization error for separable data:

Theorem 8.2.1 (Hard-margin generalization bound (Vapnik, 2006; Bartlett et al., 1998; Shalev-Shwartz et al., 2014)). *Given a set $S = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ such that $\forall \mu \in [\![n]\!], \|\mathbf{x}_\mu\| \leq r$. Let $\hat{\mathbf{w}}$ the hard-margin SVM estimator on S drawn with distribution D. With probability $1-\delta$, the generalization error is bounded by*

$$e_{\text{g}}(\alpha) \underset{\alpha\to\infty}{\leq} \left(4r\|\hat{\mathbf{w}}\| + \sqrt{\log(4/\delta)\log_2\|\hat{\mathbf{w}}\|}\right)/\sqrt{n}. \tag{204}$$

In our case one has $r^2 \simeq \frac{1}{d}\mathbb{E}_{\mathbf{x}}\|\mathbf{x}\|_2^2 = \frac{1}{d}\sum_{i=1}^d \mathbb{E}\, x_i^2 \underset{d\to\infty}{\longrightarrow} 1$. On the other hand, in the large size limit, the norm of the estimator $\|\hat{\mathbf{w}}\|_2/\sqrt{d} \underset{d\to\infty}{\longrightarrow} \sqrt{q}$, that yields $e_{\text{g}}(\alpha) \leq 4\sqrt{\frac{q}{\alpha}}$. We now need to plug the values of the norm q obtained by our max-margin solution to finally obtain the results. Unfortunately, this bound turns out to be even worse than the previous one. Indeed the norm of the hard margin estimator q is found to grow with α in the solution of the fixed point equation, and therefore the margin decay rather fast, rendering the bound vacuous. For small values of α, one finds that $q \sim \alpha$ that provides a vacuous constant generalization bound $e_{\text{g}} \leq \Theta(1)$, while for large α, $q \sim \alpha^2$

that yields an even worse bound $e_g \leq \Theta\left(\sqrt{\alpha}\right)$. Clearly, max-margin based bounds do not perform well in this high-dimensional example.

8.3 REACHING BAYES OPTIMALITY

Given the fact that logistic and hinge losses reach values extremely close to Bayes optimal generalization performances, one may wonder if by somehow slightly altering these losses one could actually reach the Bayesian values with a plug-in estimator obtained by ERM. This is what we achieve in this section, by constructing a non-convex optimization problem with a specially tuned loss and regularization, whose solution yields Bayes-optimal generalization. Recent insights have shown that indeed one can sometime re-interpret Bayesian estimation as an optimization program in inverse problems (Gribonval, 2011; Gribonval et al., 2013; Gribonval et al., 2018; Gribonval et al., 2019). In particular, (Advani et al., 2016a) showed explicitly, on the basis of the non-rigorous replica method of statistical mechanics, that some Bayes-optimal reconstruction problems could be turned into convex M-estimation.

Matching ERM and Bayes-optimal generalization errors eqs. (190)-(202) with overlaps respectively solutions of eq. (200)-(201) and assuming that $\mathcal{Z}_{w^\star}(\gamma,\Lambda)$ and $\mathcal{Z}_{\text{out}^\star}(y,\omega,V)$, defined in Appendix. A.4.1.a, are log-concave in γ and ω, we define the optimal loss and regularizer l^{opt}, r^{opt}:

$$\begin{aligned} l^{\text{opt}}(y,z) &= -\min_{\omega}\left(\frac{(z-\omega)^2}{2(\rho_{w^\star}-q_b)} + \log \mathcal{Z}_{\text{out}^\star}(y,\omega,\rho_{w^\star}-q_b)\right), \\ r^{\text{opt}}(w) &= -\min_{\gamma}\left(\frac{1}{2}\hat{q}_b w^2 - \gamma w + \log \mathcal{Z}_{w^\star}(\gamma,\hat{q}_b)\right), \end{aligned} \tag{205}$$

with $(q_b, \hat{q}_b)$ solution of eq. (201). Following these considerations, we provide the following theorem:

Theorem 8.3.1. *The result of empirical risk minimization eq. (187) with l^{opt} and r^{opt} in eq. (205), leads to Bayes optimal generalization error in the high-dimensional regime.*

Proof. The derivation is largely inspired by (Opper et al., 1991b; Kinouchi et al., 1996; Gribonval, 2011; Bean et al., 2013; Advani et al., 2016a; Advani et al., 2016b; Donoho et al., 2016; Gribonval et al., 2013; Gribonval et al., 2018; Gribonval et al., 2019). First we note that the so called Bayes-optimal AMP algorithm (Rangan, 2011), presented in Algo. 2 in Chap. 5, for $K=1$ in the context of the GLM, is provably convergent. With Bayes-optimal update functions $f_{\text{out}}^{\text{bayes}}(y,\omega,V) = \partial_\omega \log\left(\mathcal{Z}_{\text{out}^\star}\right)$ and $f_{w}^{\text{bayes}}(\gamma,\Lambda) = \partial_\gamma \log\left(\mathcal{Z}_{w^\star}\right)$, it indeed reaches Bayes-optimal performances, see (Barbier et al., 2019b). Instead performing Bayes-optimal (MMSE) estimation, we can simply use the

AMP algorithm and change the denoising functions to perform ERM (MAP) estimation with

$$\begin{aligned} f_{\text{out}}^{\text{erm},l}(y,\omega,V) &= -\partial_\omega \mathscr{M}_V[l(y,.)](\omega)\,, \\ f_{\text{w}}^{\text{erm},r}(\gamma,\Lambda) &= \Lambda^{-1}\gamma - \Lambda^{-1}\partial_{\Lambda^{-1}\gamma}\mathscr{M}_{\Lambda^{-1}}\left[r(.)\right](\Lambda^{-1}\gamma)\,, \end{aligned}$$

detailed in Appendix. A.4.2. The corresponding GAMP algorithms achieve potentially different fixed points and performances. As GAMP provably converges to the optimal generalization error with Bayes-optimal updates, it is sufficient to enforce that at each time step t the Bayes-optimal and ERM denoising functions are equal $f^{\text{bayes}} = f^{\text{erm}}$, to insure that GAMP algorithm for ERM estimation matches the same performances. Enforcing the constraint $f_{\text{out}}^{\text{bayes}}(y,\omega,V) = f_{\text{out}}^{\text{erm},l}(y,\omega,V)$ yields

$$\partial_\omega \log\left(\mathscr{Z}_{\text{out}^\star}\right)(y,\omega,V) = -\partial_\omega \mathscr{M}_V\left[l^{\text{opt}}(y,.)\right](\omega)\,.$$

Integrating, leaving aside the constant that will not influence the final result, and taking the Moreau-Yosida regularization on both sides, we obtain:

$$\mathscr{M}_V\left[\log \mathscr{Z}_{\text{out}^\star}(y,.,V)\right](\omega) = \mathscr{M}_V\left[-\mathscr{M}_V\left[l^{\text{opt}}(y,.)\right](\omega)\right] = -l^{\text{opt}}(y,\omega)\,,$$

where we invert the Moreau-Yosida regularization in the last equality that is valid as long as $\mathscr{Z}_{\text{out}^\star}(y,\omega,V)$ is assumed to be log-concave in ω, see (Advani et al., 2016a) for a derivation. We finally obtain the *optimal loss* l^{opt}

$$\begin{aligned} l^{\text{opt}}(y,z) &= -\mathscr{M}_V\left[\log\left(\mathscr{Z}_{\text{out}^\star}\right)(y,.,V)\right](z) \\ &= -\min_\omega\left(\frac{(z-\omega)^2}{2V} + \log \mathscr{Z}_{\text{out}^\star}(y,\omega,V)\right)\,. \end{aligned} \tag{206}$$

Introducing a rescaled prior denoising distribution

$$\begin{aligned} \tilde{Q}_{\text{w}^\star}(w;\gamma,\Lambda) &\equiv \frac{1}{\tilde{\mathscr{Z}}_{\text{w}^\star}(\gamma,\Lambda)} \text{P}_{\text{w}^\star}(w) e^{-\frac{1}{2}\Lambda\left(w-\Lambda^{-1}\gamma\right)^2}\,, \\ \log\left(\tilde{\mathscr{Z}}_{\text{w}^\star}(\gamma,\Lambda)\right) &= \log\left(\mathscr{Z}_{\text{w}^\star}(\gamma,\Lambda)\right) - \frac{1}{2}\Lambda^{-1}\gamma^2\,, \end{aligned}$$

so that the the prior updates read

$$\begin{aligned} f_{\text{w}}^{\text{bayes}}(\gamma,\Lambda) &= \Lambda^{-1}\gamma + \Lambda^{-1}\partial_{\Lambda^{-1}\gamma}\log\left(\tilde{\mathscr{Z}}_{\text{w}^\star}\right)\,, \\ f_{\text{w}}^{\text{erm},r}(\gamma,\Lambda) &= \Lambda^{-1}\gamma - \Lambda^{-1}\partial_{\Lambda^{-1}\gamma}\mathscr{M}_{\Lambda^{-1}}[r](\Lambda^{-1}\gamma)\,. \end{aligned}$$

Imposing the equivalence $f_{\text{w}}^{\text{bayes}}(\gamma,\Lambda) = f_{\text{w}}^{\text{erm},r}(\gamma,\Lambda)$ yields

$$\partial_{\Lambda^{-1}\gamma}\log\left(\tilde{\mathscr{Z}}_{\text{w}^\star}\right) = -\partial_{\Lambda^{-1}\gamma}\mathscr{M}_{\Lambda^{-1}}\left[r^{\text{opt}}\right](\Lambda^{-1}\gamma)\,,$$

and assuming that $\mathscr{Z}_{\mathrm{w}}(\gamma, \Lambda)$ is log-concave in γ, we may invert the Moreau-Yosida regularization, that leads to the expression for the optimal regularizer r^{opt}

$$\begin{aligned} r^{\mathrm{opt}}\left(\Lambda^{-1}\gamma\right) &= -\mathscr{M}_{\Lambda^{-1}}\left[\log\left(\tilde{\mathscr{Z}}_{\mathrm{w}^\star}\right)\left(.,\Lambda^{-1}\right)\right](w) \\ &= -\min_{\gamma}\left(\frac{1}{2}\Lambda w^2 - \gamma w + \log \mathscr{Z}_{\mathrm{w}^\star}(\gamma, \Lambda)\right). \end{aligned} \tag{207}$$

Finally, enforcing the equivalence between the AMP algorithm for the minimization of the ERM and the Bayes-optimal AMP lead to the expressions for the optimal loss l^{opt} and regularizer r^{opt} in (205). The last step is to characterize the undetermined variances V and Λ involved in (206) and (207). To achieve the Bayes-optimal performances, we therefore use the variances V and Λ solutions of the Bayes-optimal GAMP algorithm. In the large size limit $d \to \infty$, taking the expectation over the ground truth $\mathbf{w}^\star$ and the input data $\mathbf{X}$ the parameters V and Λ concentrate and are given by the SE of the GAMP algorithm (Barbier et al., 2019b)

$$\lim_{d\to\infty} \mathbb{E}_{\mathbf{w}^\star,\mathbf{X}}[V] = \rho_{\mathrm{w}^\star} - q_{\mathrm{b}}, \qquad \lim_{d\to\infty} \mathbb{E}_{\mathbf{w}^\star,\mathbf{X}}[\Lambda] = \hat{q}_{\mathrm{b}}, \tag{208}$$

where q_{b} and $\hat{q}_{\mathrm{b}}$ are solutions of the Bayes-optimal set of fixed point equations eq. (201). This shows that AMP applied to the ERM problem corresponding to (205) both converge to its fixed point and reach Bayes-optimal performances. The theorem finally follows by noting, see (Montanari, 2012; Gerbelot et al., 2020), that the AMP fixed point corresponds to the extremization conditions of the loss. □

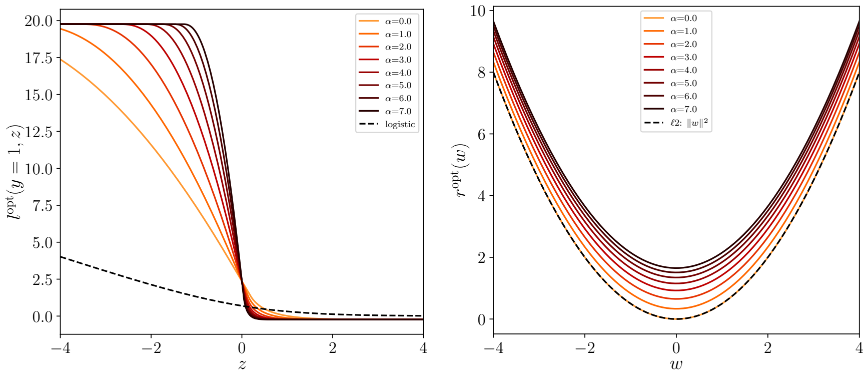


Figure 48: Optimal loss $l^{\mathrm{opt}}(y=1,z)$ and regularizer $r^{\mathrm{opt}}(w)$ for model eq. (186) as a function of α. The logistic loss and the ℓ_2 regularizer are plotted in dashed black for comparison.

The optimal loss and regularizer λ^{opt} and r^{opt} for the model (186) are illustrated in Fig. 48. Notice in particular that even though the loss l^{opt} is not convex (but seems quasi-convex), numerical simulations of ERM (black dots) presented in Fig. 49 show that ERM achieves indeed the Bayes-optimal performances (black line) even at finite dimension.

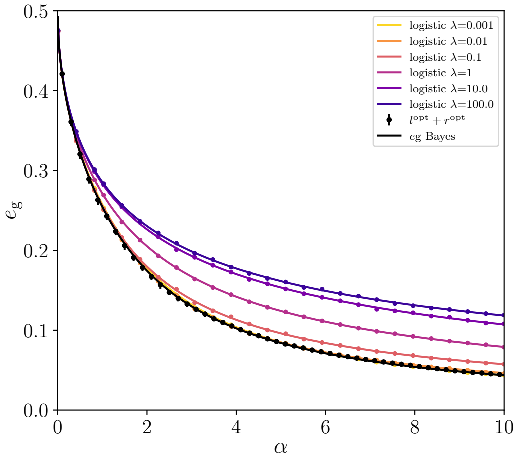


Figure 49: Generalization error obtained by optimization of the optimal loss l^{opt} and r^{opt} for the model (186), compared to ℓ_2 logistic regression and Bayes-optimal performances. Numerics has been performed with scipy.optimize.minimize with the L-BFGS-B solver for $d = 10^3$ and averaged over $n_s = 10$ instances. The error bars are barely visible.

Part II B.

THEORY FOR THE STATISTICAL ESTIMATION WITH RANDOM MULTI-LAYER NEURAL NETWORKS GENERATIVE PRIORS

OUTLINE AND MOTIVATIONS

Another recent ongoing direction of research aims to extend the mean-field methods to the combination of known and already analyzed elementary models such as the GLM (Barbier et al., 2019b) or the low-rank matrix factorization (Lesieur et al., 2017a). Combining the corresponding graphical models leads naturally to the description of more complex JPD. However, understanding how and when this *plug-in* of different models is justified is a promising research direction. In particular, this approach was successfully applied to the inference in multi-layer GLM estimation (Manoel et al., 2017), for i.i.d weight matrices. It was later generalized to orthogonally invariant weight matrices with the corresponding VAMP algorithm (Fletcher et al., 2018).

Within this general plug-in approach, we consider estimation problems of the form $\mathbf{y} = \Gamma(\mathbf{x}^\star)$ where the operator Γ represents different noisy channels such as linear inverse problems, spiked matrix estimation or phase retrieval. The ground truth signal $\mathbf{x}^\star$ must be estimated from the noisy observations $\mathbf{y}$ and the knowledge of the operator Γ. To perform this statistical reconstruction, in signal processing we often use a low-dimensional parametrization of the signal $\mathbf{x}^\star$ with for instance a *sparse* dimensionality reduction technique. Naturally, exploiting the structure of the signal drastically helps to achieve better accuracy for larger signal-to-noise ratio. Recently, this sparsity structure has been challenged and successfully replaced by generative priors based on neural networks, such as GAN or VAE, that demonstrated to be particularly performant in various estimation applications.

In Chap. 9 and Chap. 10 we respectively investigate the low-rank matrix factorization and the phase retrieval and compressed sensing estimation problems with a multi-layer feed-forward DNN generative prior with i.i.d random weights. In this series of works, we investigate and provide a theory of estimation with random generative priors. Especially, we derive sharp asymptotics for the information-theoretically optimal performances and also for the algorithmic performances of a structured polynomial AMP algorithm naturally built from the AMP algorithms on the sub-models. In the analyzed cases, we observed that generative priors have smaller statistical-to-algorithmic gaps than sparse priors, giving theoretical support to previous experimental observations that generative priors might be advantageous in terms of algorithmic performance compared to classical sparse separable priors.

Additionally, in the context of the low-rank matrix factorization, we also take advantage of the structured model to design a new enhanced spectral algorithm Linearized Approximate Message Passing (LAMP) based on the *linearization* of the AMP algorithm and that beats PCA on synthetic and real data.

Finally, in this general *plug-in* approach, instead deriving and implementing from scratch the corresponding structured AMP algorithms, we developed the tramp python package, standing for *TRee Approximate Message Passing*. The package provides an implementation of EP for modular compositional inference in high-dimensional tree-structured models. We do not reprint the corresponding paper (Baker et al., 2020) but the source code is publicly available at *https://github.com/sphinxteam/tramp*. Nevertheless, similarly to previous works (Tramel et al., 2016b; Bora et al., 2017; Fletcher et al., 2018), in Sec. 10.3, we empirically explore the reconstruction of tramp on common estimation tasks on real datasets by making use of VAE generative priors learned on the MNIST dataset (LeCun et al., 2010).

9 THE SPIKED MATRIX MATRIX MODEL WITH GENERATIVE PRIORS

Exploiting structure for efficient signal reconstruction is a central endeavor in modern signal processing. Notable technological advances - such as JPEG and MP3 compression for example - stem from the fact that images and sound admit a sparse representation in wavelet and Fourier bases. In a seminal work, Donoho, Candès and Tao have shown that underparametrized linear systems can be inverted if the signal is assumed to be sparse. This result opened the door for novel sub-Nyquist sampling strategies leveraged by sparsity which are at the heart of CS (Donoho, 2006). But interest in sparse representations reaches far beyond CS, and similar results have been derived for other signal processing tasks, such as sparse coding and sparse PCA. Despite the remarkable success of these results, they broadly assume the latent sparse representation is given, thus relying on expert knowledge for signal pre-processing.

Recent progress in deep learning has witnessed a surge of interest in neural network-based generative models. Opposed to sparsity, generative networks are trained to learn a latent representation of the structured signal. The expressiveness of neural networks allied with the capacity to capture hierarchical representations led to impressive results in signal modelling, the most notable perhaps being GAN or VAE, which can be trained to generate realistic images of human faces (Goodfellow et al., 2014). An important and natural question to ask is whether signals from generative models enjoy the same aforementioned interesting properties as sparse signals in reconstruction tasks. Early results in regression-related problems suggest that the latent structure in generative models can be leveraged to improve signal reconstruction (Tramel et al., 2016b; Bora et al., 2017; Manoel et al., 2017; Hand et al., 2018b; Fletcher et al., 2018; Hand et al., 2018a; Mixon et al., 2018), indeed suggesting that (Villar, 2018):

> *Generative models are the new sparsity.*

In this chapter we give a further step in this direction by analyzing a class of random-neural generative priors in an unsupervised task: rank-one (a.k.a. spiked) matrix factorization. Given a "data" matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$, the problem consists in finding two vectors, also called the *spikes*, $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^p$ such that $\mathbf{Y}$ can be factorized as $\mathbf{Y} = \mathbf{u}\mathbf{v}^\intercal + \sqrt{\Delta}\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is an i.i.d noise matrix of unit variance. This model is widely studied as a prototype for PCA, since for small noise ($\Delta < 1$) and Gaussian spikes $\mathbf{u}, \mathbf{v}$, the optimal estimator is given by the leading principal component of $\mathbf{Y}$ (Baik et al., 2005). Optimality relies on the assumption of unstructured spikes, and no longer hold if one of the spikes is sparse. In a similar spirit to CS, the investigation of sparse spikes in this model resulted into bespoke algorithms widely studied under the umbrella of sparse-PCA, e.g. (Jenatton et al., 2010).

An important conclusion of the aforementioned works is the existence of an algorithmic gap for sparse signal reconstruction. In other words, even if signal reconstruction is *a priori* possible, no polynomial-time algorithm is known. For spiked-matrix factorization, this means that even though the best known sparse-PCA algorithm perform better than vanilla PCA, it doesn't reach the optimal threshold set by the theoretical, and practically intractable, Bayesian estimator. As we will show, this is in sharp contract to the class of neural generative models we study, for which we provide a polynomial time algorithm reaching the optimal theoretical performance, suggesting instead that:

> *Generative models are better than sparsity.*

Before moving to the bulk of the technical analysis, we give a detailed introduction of the model and regime will study, followed by an account of our main contributions.

9.1 MODEL AND STUDIED REGIME

We will focus on the following two widely studied models in the sparse-PCA literature (Rangan et al., 2012; Deshpande et al., 2014a; Lesieur et al., 2015; Barbier et al., 2016; Perry et al., 2016; Lelarge et al., 2019; Miolane, 2017):

Spiked Wigner model ($\mathbf{vv}^\intercal$) Consider an unknown vector, the spike, $\mathbf{v}^\star \in \mathbb{R}^p$ drawn from a distribution $\mathrm{P_v}$; we observe a matrix $\mathbf{Y} \in \mathbb{R}^{p\times p}$ with a symmetric noise term $\boldsymbol{\xi} \in \mathbb{R}^{p\times p}$ and $\Delta > 0$:

$$\mathbf{Y} = \frac{1}{\sqrt{p}}\mathbf{v}^\star\mathbf{v}^{\star\intercal} + \sqrt{\Delta}\boldsymbol{\xi}\,, \tag{209}$$

where $\xi_{ij} \sim \mathcal{N}(0,1)$ i.i.d. The aim is to recover the hidden spike $\mathbf{v}^\star$ from the knowledge of $\mathbf{Y}$, up to a global sign.

Spiked Wishart (or spiked covariance) model ($\mathbf{uv}^\intercal$) Consider two unknown vectors $\mathbf{u}^\star \in \mathbb{R}^n$ and $\mathbf{v}^\star \in \mathbb{R}^p$ drawn from distributions $\mathrm{P_u}$ and $\mathrm{P_v}$ and let $\boldsymbol{\xi} \in \mathbb{R}^{n\times p}$ with $\xi_{\mu i} \sim \mathcal{N}(0,1)$ i.i.d and $\Delta > 0$, we observe

$$\mathbf{Y} = \frac{1}{\sqrt{p}}\mathbf{u}^\star\mathbf{v}^{\star\intercal} + \sqrt{\Delta}\boldsymbol{\xi}\,, \tag{210}$$

the goal is to find back the hidden spikes $\mathbf{u}^\star$ and $\mathbf{v}^\star$ from $\mathbf{Y} \in \mathbb{R}^{n\times p}$.

The noisy high-dimensional limit that we consider in this work, also called the *thermodynamic limit*, is $p, n \to \infty$ while $\beta \equiv n/p = \Theta(1)$, and the noise $\boldsymbol{\xi}$ has a variance $\Delta = \Theta(1)$. The prior $\mathrm{P_v}$ is representing the spike $\mathbf{v}$ via a k-dimensional parametrization with $\alpha \equiv p/k = \Theta(1)$. In the sparse case, k is the number of non-zeros components of $\mathbf{v}^\star$, while in generative models k is the number of latent variables.

9.1.1 CONSIDERED GENERATIVE MODELS

The simplest non-separable prior $\mathrm{P_v}$ that we consider is the Gaussian model with a covariance matrix $\boldsymbol{\Sigma}$, that is $\mathrm{P_v}(\mathbf{v}) = \mathcal{N}_{\mathbf{v}}(\mathbf{o}, \boldsymbol{\Sigma})$. This prior is not compressive, yet it captures some structure and can be simply estimated from data via the empirical covariance. We use this prior later to produce Fig. 55.

To exploit the practically observed power of generative models, it would be desirable to consider models (e.g. GAN, VAE, restricted Boltzmann machines, or others) trained on datasets of examples of possible spikes. Such training, however, leads to correlations between the weights of the underlying neural networks for which the theoretical part of the present work does not apply readily. To keep tractability in a closed form, and subsequent theoretical insights, we focus on multi-layer generative models where all the weight matrices $\mathbf{W}^{(l)}$, $l = 1, \dots, L$, are fixed, layer-wise independent, i.i.d Gaussian with zero mean and unit variance. Let $\mathbf{v} \in \mathbb{R}^p$ be the output of such a generative model

$$\mathbf{v} = \varphi^{(L)}\left(\frac{1}{\sqrt{k_L}}\mathbf{W}^{(L)} \dots \varphi^{(1)}\left(\frac{1}{\sqrt{k}}\mathbf{W}^{(1)}\mathbf{z}\right) \dots\right), \qquad (211)$$

with $\mathbf{z} \in \mathbb{R}^k$ a latent variable drawn from separable distribution $\mathrm{P_z}$, with $\rho_z = \mathbb{E}_{\mathrm{P_z}}\left[z^2\right]$. $\forall l \in [\![L]\!]$, $\varphi^{(l)}$ are the element-wise activation functions that can be either deterministic or stochastic. It will be useful to define the hidden variables $\mathbf{h}^{(l)} \in \mathbb{R}^{k_l}$ obtained from the output of layer $l-1$. The hidden variable $\mathbf{h}^{(l+1)} \in \mathbb{R}^{k_{l+1}}$ is then given by

$$\mathbf{h}^{(l+1)} = \varphi^{(l)}\left(\frac{1}{\sqrt{k_l}}\mathbf{W}^{(l)}\mathbf{h}^{(l)}\right) \quad \Leftrightarrow \quad \mathbf{h}^{(l+1)} \sim \mathrm{P}^{(l)}_{\text{out}}\left(\cdot \,\middle|\, \frac{1}{\sqrt{k_l}}\mathbf{W}^{(l)}\mathbf{h}^{(l)}\right)$$

with $\mathbf{h}^{(0)} = \mathbf{z}$ and $\mathbf{h}^{(L+1)} = \mathbf{v}$. The densities $\mathrm{P}^{(l)}_{\text{out}}$ over $\mathbb{R}^{k_{l+1}}$ parametrize the input/output relationship at each layer of the generative network. Note that since $\varphi^{(l)}$ act component-wise $\mathrm{P}^{(l)}_{\text{out}}$ is a separable distribution, and factorize in a product of identical k_{l+1} scalar distributions over $\mathbb{R}$ which, abusing notation, we we will denote by $\mathrm{P}^{(l)}_{\text{out}}$. For instance, a deterministic layer l with non-linearity $\varphi^{(l)}$ is fully characterized by the scalar density $\mathrm{P}^{(l)}_{\text{out}}(x|z) = \delta(x - \varphi(z))$.

In the setting considered in this work the ground-truth spike $\mathbf{v}^\star$ is generated using a ground-truth value of the latent variable $\mathbf{z}^\star$. The spike is then estimated from the knowledge of the data matrix $\mathbf{Y}$, and the known form of the spiked-matrix and of the generative model. In particular the matrices

$\mathbf{W}^{(l)} \in \mathbb{R}^{k_{l+1} \times k_l}$ are known, as are the parameters β, Δ, $\mathrm{P_z}$, $\mathrm{P_u}$, $\mathrm{P_v}$ and $\varphi^{(l)}$. Only the spikes $\mathbf{v}^\star$, $\mathbf{u}^\star$ and the latent vector $\mathbf{z}^\star$ are unknown, and are to be inferred.

For concreteness and simplicity, the generative model that will be analyzed in most examples given in the present work is the single-layer case of (211) with $L = 1$. We define the total compression ratio $\alpha \equiv p/k$. In what follows we will illustrate our results for φ being linear, sign and ReLU functions.

9.1.2 SUMMARY OF MAIN CONTRIBUTIONS

First, we provide an information-theoretical analysis for the performance of the optimal estimator for the spiked-matrix models (209) and (210). This analysis is based on a rigorous expression for the mutual information between the matrix $\mathbf{Y}$ and a general spike $\mathbf{v}^\star$ from a non-separable distribution $\mathrm{P_v}$ in $\mathbb{R}^p$, and holds in the afore defined thermodynamic limit. Evaluating this expression on the generative priors discussed in Sec. 9.1.1, we obtain the optimal statistical threshold Δ_c below which the spike $\mathbf{v}^\star$ can be reconstructed. On a second moment, we derive an AMP algorithm for the models (209) and (210), and show that, for the all the generative architectures analysed, they attain the same performance previously derived for the Bayesian optimal estimator. Next, we propose a simple spectral method derived from our AMP algorithm reaching the same statistical threshold Δ_c. Finally, we show that this same spectral method can be, in certain cases, rigorously derived from a Random Matrix Theory.

Our main findings are in stark contrast to the known results for sparse-PCA, and therefore it is useful to present them in this context. We draw two main conclusions from the present work:

(i) No algorithmic gap with generative-model priors: Sharp and detailed results are known in the thermodynamic limit (as defined above) when the spike $\mathbf{v}^\star$ is sampled from a separable distribution $\mathrm{P_v}$. A detailed account of several examples can be found in (Lesieur et al., 2017a). The main finding for sparse priors $\mathrm{P_v}$ is that when the sparsity $\rho = k/p = 1/\alpha$ is large enough then there exist optimal algorithms (Deshpande et al., 2014a), while for ρ small enough there is a striking gap between statistically optimal performance and the one of best known algorithms (Lesieur et al., 2015). The small-ρ expansion studied in (Lesieur et al., 2017a) is consistent with the well-known results for exact recovery of the support of $\mathbf{v}^\star$ (Amini et al., 2009; Berthet et al., 2013), which is one of the best-known cases in which gaps between statistical and best-known algorithmic performance were described.

Our analysis of the spiked-matrix models with generative priors reveals that in this case known algorithms are able to obtain (asymptotically) optimal performance even when the dimension is greatly reduced, i.e. $\alpha \gg 1$. Analogous conclusion about the lack of algorithmic gaps was reached for the problem of phase retrieval under a generative prior in (Hand et al., 2018a). This result suggests that plausibly generative priors are better than sparsity as they lead to algorithmically easier problems.

(ii) Spectral algorithms reaching statistical threshold: Arguably the most basic algorithm used to solve the spiked-matrix model is based on the leading singular vectors of the matrix $\mathbf{Y}$. We will refer to this as PCA. Previous work on spiked-matrix models (Perry et al., 2016; Lesieur et al., 2017a) established that in the thermodynamic limit and for separable priors of zero mean PCA reaches the best performance of all known efficient algorithms in terms of the value of noise Δ below which it is able to provide positive correlation between its estimator and the ground-truth spike. While for sparse priors positive correlation is statistically reachable even for larger values of Δ (Perry et al., 2016; Lesieur et al., 2017a), no efficient algorithm beating the PCA threshold is known. Notice that this result holds only for sparsity $\rho = \Theta(1)$. A line of works shows that when sparsity k scales slower than linearly with p, algorithms more performant than PCA exist (Amini et al., 2009; Deshpande et al., 2014b).

In the case of generative priors we find in this contribution that other spectral methods improve on the canonical PCA. We design a spectral method, called LAMP, that under certain assumptions, e.g. zero mean of the spikes, reach the statistically optimal threshold, meaning that for larger values of noise variance no other (even exponential) algorithm is able to reach positive correlation with the spike. Again this is a striking difference with the sparse separable prior, making the generative priors algorithmically more attractive. We demonstrate the performance of LAMP on the spiked-matrix model when the spike is taken to be one of the fashion-MNIST images (Xiao et al., 2017) showing considerable improvement over canonical PCA. Each of the following sections is dedicated to one of the results above.

9.2 ANALYSIS OF INFORMATION THEORETICALLY OPTIMAL ESTIMATION

In this section, we derive a set of fixed point equations, known as SE equations, that fully characterize the performance of the optimal estimator for the spike $\mathbf{v}^{\star}$. For the sake of concreteness, the results in this section are given for the Wigner model, and can be fully generalized to the Wishart case presented in Appendix. B.2 of (Aubin et al., 2019c).

9.2.1 RIGOROUS MUTUAL INFORMATION

From an optimization perspective, the problem we want to solve is to find the estimator $\mathbf{v}^{\star}$ that minimizes the MSE

$$\mathrm{mse}(\Delta) = \mathbb{E}||\hat{\mathbf{v}} - \mathbf{v}^{\star}||_2^2. \tag{212}$$

Since the information about the generative model P_{v} of the spike is given, we know that the estimator minimizing eq. (212) is given by the mean of the posterior distribution of the spike, i.e. $\hat{\mathbf{v}}^{\text{opt}} = \mathbb{E}_{\mathrm{P}(\mathbf{v}^\star|\mathbf{Y})}\mathbf{v}$, where $\mathrm{P}(\mathbf{v}^\star|\mathbf{Y})$ is written from Bayes rule as

$$\mathrm{P}(\mathbf{v}^\star|\mathbf{Y}) = \frac{\mathrm{P_v}(\mathbf{v}^\star)}{\mathrm{P}(\mathbf{Y})} \prod_{1\leq i<j\leq p} \frac{1}{\sqrt{2\pi\Delta}} \exp\left(-\frac{1}{2\Delta}\left(y_{ij} - \frac{v_i^\star v_j^\star}{\sqrt{p}} \right)^2 \right) . \quad (213)$$

The expression above is written in full generality, and for the time being we have not assumed anything about $\mathrm{P_v}$. The naive approach of estimating $\hat{\mathbf{v}}^{\text{opt}}$ from exact sampling of the posterior is intractable numerically, specially in the large-dimensional limit $p \to \infty$ of interest. However, it is still possible to track the performance of the optimal estimator without direct sampling through the I-MMSE theorem connecting the MMSE to a derivative of the mutual information between the signal and the data (Guo et al., 2005a). Following this rationale, our first main result is a rigorous expression for the mutual information between the ground-truth spike $\mathbf{v}^\star$ and the observation $\mathbf{Y}$, defined as $\mathscr{I}(\mathbf{Y};\mathbf{v}^\star) = \mathscr{D}_{\mathrm{KL}}(\mathrm{P}_{(\mathrm{v}^\star,\mathrm{Y})}\|\mathrm{P}_{\mathrm{v}^\star}\mathrm{P}_{\mathrm{Y}})$, valid in the thermodynamic limit defined in Sec. 9.1.

Theorem 9.2.1 (Mutual information for the spiked Wigner model with structured spike)**.** *Informally, assume the spike $\boldsymbol{v}^\star$ come from a sequence (of growing dimension p) of a generic structured prior $\mathrm{P_v}$ on $\mathbb{R}^p$, then*

$$\lim_{p\to\infty} i_p \equiv \lim_{p\to\infty} \frac{\mathscr{I}(\mathbf{Y};\boldsymbol{v}^\star)}{p} = \inf_{\rho_v \geq q_v \geq 0} i_{\mathrm{rs}}(\Delta, q_v), \quad (214)$$

$$\textit{with} \quad i_{\mathrm{rs}}(\Delta,q_v) \equiv \frac{(\rho_v - q_v)^2}{4\Delta} + \lim_{p\to\infty} \frac{\mathscr{I}\left(\boldsymbol{v};\boldsymbol{v}+\sqrt{\frac{\Delta}{q_v}}\boldsymbol{\xi}\right)}{p} \quad (215)$$

and $\boldsymbol{\xi}$ being a Gaussian vector with zero mean, unit diagonal variance and $\rho_v = \lim_{p\to\infty} \mathbb{E}_{\mathrm{P_v}}[\boldsymbol{v}^\intercal \boldsymbol{v}]/p$.

The proof for this theorem is left in Appendix. C of (Aubin et al., 2019c), and instead we draw its consequences. Our theorem connects the asymptotic mutual information of the spiked model with generative prior $\mathrm{P_v}$ to the mutual information between $\mathbf{v}$ taken from $\mathrm{P_v}$ and its noisy version, $\mathscr{I}(\mathbf{v};\mathbf{v}+\sqrt{\Delta/q_v}\boldsymbol{\xi})$. As mentioned before, the mutual information is intimately connected to the performance of the optimal estimator, and one can prove in particular that for the spiked-matrix model (Alaoui et al., 2018) the MMSE on the spike $\mathbf{v}^\star$ is asymptotically given by:

$$\mathrm{MMSE}_v = \rho_v - q_v^\star , \quad (216)$$

where $q_v^\star$ is the optimizer of the function $i_{\mathrm{rs}}(\Delta, q_v)$. Computing this later mutual information is itself a high-dimensional task, hard in full generality, but it can be done for a range of non-trivial $\mathrm{P_v}$. The simplest tractable case is when the prior $\mathrm{P_v}$ is separable, then it yields back exactly the formula previously known from (Krzakala et al., 2016; Barbier et al., 2016; Lelarge et al., 2019). It can also be computed for the correlated Gaussian generative model, $\mathrm{P_v}(\mathbf{v}) =$

$\mathcal{N}_{\mathbf{v}}(\mathbf{0},\mathbf{\Sigma})$, for which $\mathcal{I}(\mathbf{v};\mathbf{v}+\sqrt{\Delta/q_v}\boldsymbol{\xi}) = \mathrm{Tr}\{\log\left(\mathbf{I}_p + q_v\mathbf{\Sigma}/\Delta\right)\}/2$ is readily known.

More interestingly, the mutual information associated to the multi-layer generative prior with random weights from eq. (211), explicitly written as

$$\begin{aligned}\mathrm{P}_{\mathrm{v}}(\mathbf{v}) = \int \prod_{l=1}^{L}\prod_{\nu_l=1}^{k_l} \mathrm{d}h_{\nu_l}^{(l)}\ \mathrm{P}_{\mathrm{out}}^{(l-1)}\left(h_{\nu_l}^{(l)}\middle|\frac{1}{\sqrt{k_{l-1}}}\sum_{\nu_{l-1}=1}^{k_{l-1}} w_{\nu_l\nu_{l-1}}^{(l-1)} h_{\nu_{l-1}}\right)\\ \times\prod_{i=1}^{p}\mathrm{P}_{\mathrm{out}}^{(L)}\left(v_i\middle|\frac{1}{\sqrt{k_L}}\sum_{\nu_L=1}^{k_L} w_{i\nu_L}^{(L)} h_L\right),\end{aligned} \tag{217}$$

can also be asymptotically computed. Indeed, the corresponding single-layer formula for this mutual information has been derived and proven in (Barbier et al., 2019b). For the multi-layer case the mutual information formula has been derived in (Manoel et al., 2017; Reeves, 2017) and proven for the case of two layers in (Gabrié et al., 2018). Theorem 9.2.1 together with the results from (Barbier et al., 2019b; Manoel et al., 2017; Reeves, 2017; Gabrié et al., 2018) yields the following formula for the spiked Wigner model (209) with multi-layer generative prior (211):

$$\begin{aligned}i_{\mathrm{rs}}(\Delta,q_v) = \frac{\rho_v^2}{4\Delta} + \frac{1}{4\Delta}q_v^2\\ +\frac{1}{\alpha}\underset{\{\hat{q}_l,q_l\}_l}{\mathbf{extr}}\Bigg[\frac{1}{2}\sum_{l=1}^{L}\alpha_l\hat{q}_l q_l - \sum_{l=2}^{L}\alpha_l\Psi_{\mathrm{out}}^{(l)}\left(\hat{q}_l,q_{l-1}\right)\\ -\alpha\Psi_{\mathrm{out}}^{(L+1)}\left(\frac{q_v}{\Delta},q_L\right) - \Psi_z\left(\hat{q}_z\right)\Bigg].\end{aligned} \tag{218}$$

where $\alpha_l = k_l/k$ (note that in particular $\alpha_1 = 1$) and the functions $\Psi_z, \Psi_{\mathrm{out}}$ are defined by

$$\Psi_z(x) \equiv \mathbb{E}_{\xi}\left[\mathcal{Z}_z\left(x^{1/2}\xi,x\right)\log\left(\mathcal{Z}_z\left(x^{1/2}\xi,x\right)\right)\right], \tag{219}$$

$$\begin{aligned}\Psi_{\mathrm{out}}^{(l)}(x,y) \equiv \mathbb{E}_{\xi,\eta}\Big[\mathcal{Z}_{\mathrm{out}}^{(l)}\left(x^{1/2}\xi,x,y^{1/2}\eta,\rho_l-y\right)\\ \log\left(\mathcal{Z}_{\mathrm{out}}^{(l)}\left(x^{1/2}\xi,x,y^{1/2}\eta,\rho_l-y\right)\right)\Big],\end{aligned} \tag{220}$$

with $\xi,\eta \sim \mathcal{N}(0,1)$ i.i.d, ρ_l is the second moment of the hidden variable h_l and $\mathcal{Z}_z$, $\mathcal{Z}_{\mathrm{out}}^{(l)}$ are the normalizations of the following denoising scalar distributions:

$$\begin{aligned}\mathrm{Q}_{\mathrm{z}}\left(z;\gamma,\Lambda\right) \equiv \frac{\mathrm{P}_{\mathrm{z}}(z)}{\mathcal{Z}_z(\gamma,\Lambda)}\ e^{-\frac{\Lambda}{2}z^2+\gamma z},\\ \mathrm{Q}_{\mathrm{out}}^{(l)}\left(v,x;B,A,\omega,V\right) \equiv \frac{\mathrm{P}_{\mathrm{out}}^{(l)}(v|x)}{\mathcal{Z}_{\mathrm{out}}^{(l)}(B,A,\omega,V)}\ e^{-\frac{A}{2}v^2+Bv}\ \frac{e^{-\frac{(x-\omega)^2}{2V}}}{\sqrt{2\pi V}}.\end{aligned} \tag{221}$$

Result (218) is remarkable in that it connects the asymptotic mutual information of a high-dimensional model with a simple scalar formula that can be easily evaluated. Moreover, it fully characterize the statistical performance of the optimal estimator, allowing us to readily identify the statistical thresholds separating the region between possible and impossible inference of the spike. We now draw the consequences of eq. (218) for the most common choices of activation.

9.2.2 OPTIMAL PERFORMANCE AND STATISTICAL THRESHOLDS: PHASE DIAGRAMS

Taking the extremization over q_v and $(\hat{q}_l, q_l)_{1\leq l\leq L}$ in eq. (218), we obtain the following system of coupled fixed point equations:

$$\begin{cases} q_v = \Lambda_x\left(\frac{q_v}{\Delta}, q_L\right) \\ q_L = \Lambda_x\left(\hat{q}_L, q_{L-1}\right) \\ \qquad \vdots \\ q_l = \Lambda_x\left(\hat{q}_l, q_{l-1}\right) \\ \qquad \vdots \\ q_z = \Lambda_z\left(\hat{q}_z\right) \end{cases}, \quad \begin{cases} \hat{q}_L = \tilde{\alpha}_L \Lambda_{\text{out}}\left(\frac{q_v}{\Delta}, q_L\right) \\ \hat{q}_{L-1} = \tilde{\alpha}_{L-1} \Lambda_{\text{out}}\left(\hat{q}_L, q_{L-1}\right) \\ \qquad \vdots \\ \hat{q}_l = \tilde{\alpha}_l \Lambda_{\text{out}}\left(\hat{q}_{l+1}, q_l\right) \\ \qquad \vdots \\ \hat{q}_z = \tilde{\alpha}_1 \Lambda_{\text{out}}\left(\hat{q}_2, q_z\right) \end{cases}, \tag{222}$$

where we have defined the update functions

$$\Lambda_x(x,y) \equiv 2\partial_x \Psi_{\text{out}}(x,y)\,, \qquad \Lambda_{\text{out}}(x,y) \equiv 2\partial_y \Psi_{\text{out}}(x,y)\,,$$
$$\Lambda_z(x) \equiv 2\partial_x \Psi_z(x)\,,$$

and the layer-wise aspect ratios $\tilde{\alpha}_l = k_{l+1}/k_l = \alpha_{l+1}/\alpha_l$. As previously discussed, the fixed point of these equations provide all the information about the performance of the Bayes-optimal estimator through eq. (216).

An important first question that can be answered from eqs. (222) is when does the Bayes-optimal estimator performs better than a random guess from the prior distribution P_{v}. For instance, we intuitively expect that when the prior is not biased towards a particular direction in $\mathbb{R}^p$ and for very high noise $\Delta \gg 1$ better-than-random estimation is not possible. In terms of fixed points of eqs. (222), this situation corresponds to the existence of the *non-informative* fixed point $q_v^\star = 0$ (i.e. maximum $\mathrm{MSE}_v = \rho_v$, or zero overlap with the spike). Evaluating the right-hand side of eqs. (222) at $q_v = 0$, we can see that $q_v^\star = 0$ is a fixed point if

$$\mathbb{E}_{\mathrm{P}_z}\left[z\right] = 0 \quad \text{and} \quad \mathbb{E}_{Q_{\text{out}}^{(l),0}}\left[v\right] = 0\,, \tag{223}$$

where $Q_{\text{out}}^{(l),0}(v,x) \equiv Q_{\text{out}}^{(l)}(v,x;0,0,0,\rho_l)$ from eq. (221). Note that for multi-layer network with deterministic channels and $\varphi^{(l)} \equiv \varphi$ for all l, the second condition is equivalent to φ being an odd function.

When the condition (223) holds, $(q_v, q_L, \hat{q}_L, \ldots, \hat{q}_z, q_z) = (0,0,0,\ldots,0,0)$ is a fixed point of eq. (222). The numerical stability of this fixed point is

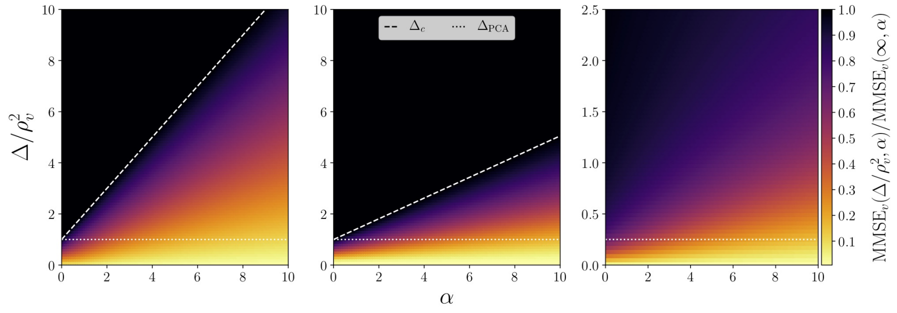


Figure 50: Spiked Wigner model: MMSE_v on the spike as a function of noise to signal ratio Δ/ρ_v^2, and single-layer generative prior with compression ratio α for (**Left**) linear $\rho_v = 1$, (**Center**) sign $\rho_v = 1$, and (**Right**) relu $\rho_v = 1/2$ activations. Dashed white lines mark the phase transitions Δ_c, matched by both the AMP and LAMP algorithms. Dotted white line marks the phase transition of canonical PCA.

determined by whether it is an attractor of the dynamics, and therefore determines a phase transition point Δ_c, defined as the noise below which the fixed point $\mathbf{0} \in \mathbb{R}^{L+1}$ becomes a repeller. The character of the fixed point can be determined by a standard linear stability analysis of the fixed point equations. The transition will then correspond to the value of Δ for which the largest eigenvalue of the Jacobian of the eqs. (222) at 0 becomes greater than one. This Jacobian is given explicitly by

$$\begin{array}{ccccccccccccccc}
q_v & \hat{q}_L & q_L & \hat{q}_{L-1} & q_{L-1} & \cdots & \hat{q}_{l+1} & q_{l+1} & \hat{q}_l & q_l & \cdots & \hat{q}_z & q_z & \\
\frac{1}{\Delta}m_{vv}^{(L)} & 0 & \frac{1}{\rho_L^2}m_{vx}^{(L)} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & q_v \\
\frac{\tilde{\alpha}_L}{\Delta}m_{vx}^{(L)} & 0 & \frac{\tilde{\alpha}_L}{\rho_L^2}m_{xx}^{(L)} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \hat{q}_L \\
0 & m_{vv}^{(L-1)} & 0 & 0 & \frac{1}{\rho_{L-1}^2}m_{vx}^{(L-1)} & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & q_L \\
0 & \tilde{\alpha}_{L-1}m_{vx}^{(L-1)} & 0 & 0 & \frac{\tilde{\alpha}_{L-1}}{\rho_{L-1}^2}m_{xx}^{(L-1)} & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \hat{q}_{L-1} \\
0 & 0 & 0 & m_{vv}^{(L-2)} & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & q_{L-1} \\
0 & 0 & 0 & \tilde{\alpha}_{L-2}m_{vx}^{(L-2)} & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \hat{q}_{L-2} \\
 & & \vdots & & & \vdots & & & \vdots & & & \vdots & & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & m_{vv}^{(l)} & 0 & 0 & \frac{1}{\rho_l^2}m_{vx}^{(l)} & \cdots & 0 & 0 & q_{l+1} \\
0 & 0 & 0 & 0 & 0 & \cdots & \tilde{\alpha}_l m_{vx}^{(l)} & 0 & 0 & \frac{\tilde{\alpha}_l}{\rho_l^2}m_{xx}^{(l)} & \cdots & 0 & 0 & \hat{q}_l \\
 & & \vdots & & & \vdots & & & \vdots & & & \vdots & & \vdots \\
0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & m_{zz} & 0 & q_z
\end{array} \tag{224}$$

where we have defined the following shorthand for the second moments of $Q_{\rm out}^{(l),0}(v,x)$:

$$\begin{aligned}
m_{vv}^{(l)} &= \left(\mathbb{E}_{Q_{\rm out}^{(l),0}}v^2\right)^2, & m_{vx}^{(l)} &= \left(\mathbb{E}_{Q_{\rm out}^{(l),0}}vx\right)^2, \\
m_{xx}^{(l)} &= \left(\mathbb{E}_{Q_{\rm out}^{(l),0}}x^2-\rho_l\right)^2, & m_{zz} &= \left(\mathbb{E}_{\mathrm{P}_z}z^2\right)^2.
\end{aligned} \tag{225}$$

This result is given in full generality, and it is instructive to compute Δ_c in specific cases.

First, consider the case of a single-layer generative prior $L=1$. Fix $\mathrm{P}_z(z) = \mathcal{N}_z(0,1)$ and $\mathrm{P}_{\rm out}^{(1)}(v|x) = \delta(v-\varphi(x))$, for $\varphi \in \{\text{linear}, \text{sign}, \text{relu}\}$. The first two choices of non-linearities are odd, and therefore in these cases we expect a transition as discussed above. It can be readily computed from the Jacobian

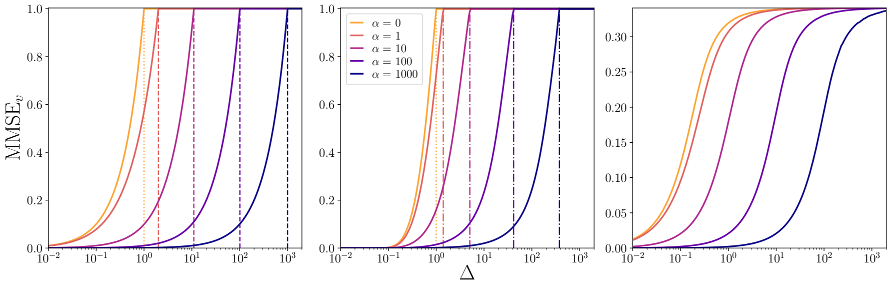


Figure 51: Spiked Wigner model: MMSE$_v$ as a function of noise Δ for $L = 1$ a wide range of compression ratios $\alpha = 0, 1, 10, 100, 1000$, for (**Left**) linear, (**Center**) sign, and (**Right**) relu activations. Unique stable fixed point of (222) is found for all these cases.

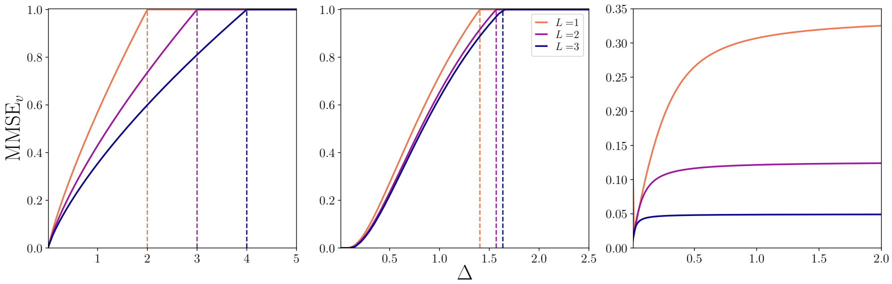


Figure 52: Spiked Wigner model: MMSE$_v$ as a function of noise Δ for $L = 1, 2, 3$ with constant compressive ratio $\alpha_1 = \alpha_2 = \alpha_3 = 1$, for (**Left**) linear, (**Center**) sign, and (**Right**) relu activations. The second moments of the variable v for $L = 1, 2, 3$ are for linear and sign activations $\rho_v^{(L)} = 1$, while for relu $\rho_v^{(L)} = 1/2^L$.

eq. (224) and yield $\Delta_c = 1 + \alpha$ for linear activation and $\Delta_c = 1 + \frac{4}{\pi^2}\alpha$ for sign activation. In both cases, since $\alpha > 0$, it is clear that knowledge of the generative prior improve reconstruction in the sense that the spike can be better reconstructed for larger amplitude of noise Δ. Moreover, the larger α (i.e. the smaller the latent dimension with respect to the signal dimension), the better the reconstruction.

Fig. 50 summarizes this discussion. We numerically solve the fixed point eqs. (222) and plot the MMSE obtained from the fixed point in a heat map, for the linear, sign and relu activations. The white dashed line marks the threshold Δ_c obtained analytically from the Jacobian in eq. (224). The property that we find the most striking is that in these three evaluated cases, for all values of Δ and α that we analyzed, we always found that eq. (222) has a unique stable fixed point. Thus we have not identified, in the physics terminology, any first order phase transition. Fig. 51 shows some examples of numerical MMSE curves for three nonlinearities discussed and different values of α. The fixed point equations were solved iteratively from uncorrelated initial condition, and from initial condition corresponding to the ground truth signal, and found that both lead to the same fixed point. This observation generalizes to deeper $L > 1$ generative priors. Consider $\mathrm{P}_z(z) = \mathcal{N}_z(0, 1)$ and layer-wise constant activation $P_{\text{out}}^{(l)}(v|x) = \delta(v - \varphi(x))$. For the previous odd activation functions discussed, we find that

> **Linear activation:** For $\varphi(x) = x$ the leading eigenvalue of the Jacobian becomes one at
>
> $$\Delta_c = 1 + \sum_{l=1}^{L} \frac{\alpha}{\alpha_l}. \tag{226}$$
>
> Note in particular that for $L = 1$ and in the limit $\alpha = 0$ we recover the phase transition $\Delta_c = 1$ known from the case with separable prior (Lesieur et al., 2017a). For $\alpha > 0$, we have $\Delta_c > 1$ meaning the spike can be estimated more efficiently when its structure is accounted for. In particular, the deeper the generative network for the spike, the easier estimation becomes.
>
> **Sign activation:** For $\varphi(x) = \text{sign}(x)$ the leading eigenvalue of the Jacobian becomes one at
>
> $$\Delta_c = 1 + \sum_{l=1}^{L} \left(\frac{4}{\pi^2}\right)^l \frac{\alpha}{\alpha_l}. \tag{227}$$
>
> For $L = 1$ and $\alpha = 0$, $P_v = \text{Bern}(1/2)$, and the transition $\Delta_c = 1$ agrees with the one found for a separable prior distribution (Lesieur et al., 2017a). As in the linear case, for $\alpha > 0$, we can estimate the spike for larger values of noise than in the separable case, and depth also improves estimation.

Note that we also did not observe first order transitions for deeper networks, at least in the first-to-come-in-mind cases that we have investigated, i.e. deterministic deep networks with $\varphi^{(l)} \equiv \varphi \in \{\text{linear}, \text{sign}, \text{relu}\}$. However, we do not expect this behavior to be completely general neither. One can engineer a situation, for instance with a very shifted relu on the last layer, and a very large intermediate layer, so that the spike $\mathbf{v}$ becomes effectively sparse with weakly correlated, almost independent, components, thus recovering the classical algorithmic gap (Lesieur et al., 2017a).

So far we have only discussed the performance of the information theoretic optimal estimator - averting the question of estimating the signal itself. In the next section we close this gap by introducing an AMP algorithm for signal reconstruction. Our algorithm has the advantaged that its performance can tracked down exactly in the thermodynamic limit, and we will show that in the cases we analyzed it exactly follows the same fixed point equations (222) as the ones derived for optimal estimator.

9.3 APPROXIMATE MESSAGE PASSING WITH GENERATIVE PRIORS

Naive sampling from the high-dimensional posterior distribution is exponentially costly, ruling this approach out from an algorithmic perspective. One should therefore appeal to algorithmically tractable approximations. AMP algorithms have proven to be particularly useful for problems defined on random graphs, and successful examples abound in the literature

In this section we derive and analyze an AMP algorithm tailored for spiked estimation with generative priors. Next, we show that the MSE of our algorithm can be tracked exactly in the thermodynamic limit, and that moreover it coincides with the optimal performance discussed in Sec. 9.2 even for large α. This result is particularly interesting when compared to the known performance of message passing algorithms for sparse-PCA, for which AMP is not able to reach optimal statistical performance in the small sparsity regime (Lesieur et al., 2017b).

AMP algorithms for spiked matrix estimation with separable priors are well known (Metzler et al., 2016; Manoel et al., 2017; Berthier et al., 2017). Our derivation draw on previous works on extending AMP to non-separable priors (Metzler et al., 2016; Manoel et al., 2017; Berthier et al., 2017) and we first focus on the more general Wishart model ($\mathbf{u}\mathbf{v}^\intercal$). After, we discuss how to get the corresponding result for the Wigner model ($\mathbf{v}\mathbf{v}^\intercal$) with a simple change.

9.3.1 DERIVATION OF THE APPROXIMATE MESSAGE PASSING ALGORITHM

AMP algorithms can be derived systematically for problems that can be written in terms of an acyclic factor graph. The standard idea is to simplify the corresponding BP equations in the limit of a large number of variables. Together with a Gaussian Ansatz for the distribution of the BP messages, the expansion of the BP yield a set of $\Theta\left(k^2\right)$ simplified equations known as rBP equations. The last step to get the corresponding AMP algorithm is to remove the target dependency of the messages that further reduces the number of iterative equations to $\Theta\left(k\right)$.

Our derivation is closely related to the derivation of AMP for a series of statistical inference problems with factorized priors, see for example (Lesieur et al., 2017a) and references therein. In the interest of the reader, instead of repeating the cumbersome steps described above, we rather describe how two known and simple AMP algorithms for independent inference problems can be combined into one for the corresponding structured problem. In particular, this is illustrated for the spiked-matrix estimation with single-layer generative model prior, which can be seen as the combination of a *rank-one matrix factorization problem* (MF) (Lesieur et al., 2017a) with a GLM

(Barbier et al., 2019b). Note that the multi-layer case follows by iterating this procedure.

9.3.1.A COMBINING FACTOR GRAPHS

Consider the factor graphs for the MF and the GLM problems with separable priors, drawn in Fig. 53. The key idea is to replace the separable prior P_v for the *structured* variable $\mathbf{v}$ in the MF model (in green) by a factorized connection channel P_out (see definition (Barbier et al., 2019b)) linking the input $\mathbf{v}$ with the output factors of the GLM (in red). The resulting factor graph for the structured Wishart model is given in Fig. 53, with the same color code.

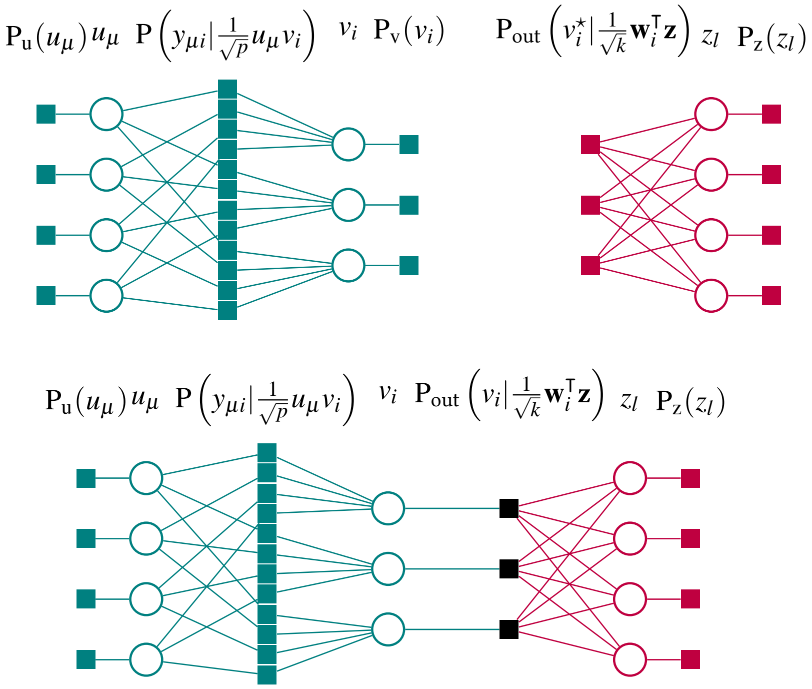

$\mathrm{P}_\mathrm{u}(u_\mu)$ u_μ $\mathrm{P}\left(y_{\mu i}|\frac{1}{\sqrt{p}}u_\mu v_i\right)$ v_i $\mathrm{P}_\mathrm{v}(v_i)$
$\mathrm{P}_\mathrm{out}\left(v_i^\star|\frac{1}{\sqrt{k}}\mathbf{w}_i^\intercal\mathbf{z}\right)$ z_l $\mathrm{P}_\mathrm{z}(z_l)$
$\mathrm{P}_\mathrm{u}(u_\mu)$ u_μ $\mathrm{P}\left(y_{\mu i}|\frac{1}{\sqrt{p}}u_\mu v_i\right)$ v_i $\mathrm{P}_\mathrm{out}\left(v_i|\frac{1}{\sqrt{k}}\mathbf{w}_i^\intercal\mathbf{z}\right)$ z_l $\mathrm{P}_\mathrm{z}(z_l)$

Figure 53: Factor graphs corresponding to a **(upper left)** low-rank matrix factorization model with separable priors $\mathrm{P}_\mathrm{u}, \mathrm{P}_\mathrm{v}$ on $\mathbf{u}, \mathbf{v}$, **(uper right)** a generalized linear model with observations $\mathbf{v}^\star$ and prior P_z on $\mathbf{z}$, and finally to **(bottom)** a low-rank matrix factorization layer (green) with a GLM prior (red) where the separable prior $\mathrm{P}_\mathrm{v}(v_i)$ is replaced by correlated factor $\mathrm{P}_\mathrm{out}(v_i|.)$.

9.3.1.B COMBINING AMP ALGORITHMS

As for the factor graphs, we start by recalling the AMP update equations in the Bayes-optimal case for the two problems in question with separable priors.

AMP equations for the Wishart MF layer (variables v and u) Consider the low-rank matrix factorization model $\mathbf{Y} = \frac{1}{\sqrt{p}}\mathbf{u}^\star\mathbf{v}^{\star\intercal} + \sqrt{\Delta}\boldsymbol{\xi}$ with separable priors P_u and P_v for the variables $\mathbf{u}$ and $\mathbf{v}$, illustrated in Fig. 53 **(up-**

per left). The corresponding Bayes-optimal AMP equations, given in (Lesieur et al., 2017a), read:

$$\begin{cases} \hat{\mathbf{u}}^{t+1} = \mathbf{f}_{\mathrm{u}}(\mathbf{b}_u^t, \mathbf{A}_u^t)\,, \\ \hat{\mathbf{C}}_u^{t+1} = \partial_{\mathbf{b}} \mathbf{f}_{\mathrm{u}}(\mathbf{b}_u^t, \mathbf{A}_u^t)\,, \\ \hat{\mathbf{v}}^{t+1} = \mathbf{f}_{\mathrm{v}}(\mathbf{b}_v^t, \mathbf{A}_v^t)\,, \\ \hat{\mathbf{C}}_v^{t+1} = \partial_{\mathbf{b}} \mathbf{f}_v(\mathbf{b}_v^t, \mathbf{A}_v^t)\,, \end{cases} \quad \text{and} \quad \begin{cases} \mathbf{b}_v^t &= \frac{1}{\Delta} \frac{\mathbf{Y}^\intercal}{\sqrt{p}} \hat{\mathbf{u}}^t - \frac{1}{\Delta} \frac{\mathbf{1}_n^\intercal \hat{\mathbf{C}}_u^t}{p} \hat{\mathbf{v}}^{t-1}\,, \\ \mathbf{A}_v^t &= \frac{1}{\Delta} \frac{\|\hat{\mathbf{u}}^t\|_2^2}{p} \mathbf{I}_p\,, \\ \mathbf{b}_u^t &= \frac{1}{\Delta} \frac{\mathbf{Y}}{\sqrt{p}} \hat{\mathbf{v}}^t - \frac{1}{\Delta} \frac{\mathbf{1}_p^\intercal \hat{\mathbf{C}}_v^t}{p} \hat{\mathbf{u}}^{t-1}\,, \\ \mathbf{A}_u^t &= \frac{1}{\Delta} \frac{\|\hat{\mathbf{v}}^t\|_2^2}{p} \mathbf{I}_n\,, \end{cases} \tag{228}$$

where the update functions f_u and f_v are respectively the means of the distributions Q_{u} and Q_{v}, defined similarly to eq. (221) as

$$\mathrm{Q}_{\mathrm{u}}(u; b, A) \equiv \frac{\mathrm{P}_{\mathrm{u}}(u)}{\mathcal{Z}_u(b, A)} e^{-\frac{1}{2}Au^2 + bu}\,, \quad \mathrm{Q}_{\mathrm{v}}(v; b, A) \equiv \frac{\mathrm{P}_{\mathrm{v}}(v)}{\mathcal{Z}_{\mathrm{v}}(b, A)} e^{-\frac{1}{2}Av^2 + bv}\,. \tag{229}$$

AMP equations for the GLM layer (variable z) On the other hand, the Bayes-optimal AMP equations for the GLM model $\mathbf{v}^\star = \varphi\left(\frac{1}{\sqrt{k}} \mathbf{W} \mathbf{z}^\star\right)$ with $z_l^\star \underset{\mathrm{iid}}{\sim} \mathrm{P}_{\mathrm{z}}$, given in (Barbier et al., 2019b) and illustrated in Fig. 53 read

$$\begin{cases} \hat{\mathbf{z}}^{t+1} = \mathbf{f}_{\mathrm{z}}(\boldsymbol{\gamma}^t, \boldsymbol{\Lambda}^t)\,, \\ \hat{\mathbf{C}}_z^{t+1} = \partial_{\boldsymbol{\gamma}} \mathbf{f}_{\mathrm{z}}(\boldsymbol{\gamma}^t, \boldsymbol{\Lambda}^t)\,, \\ \mathbf{g}^t = \mathbf{f}_{\mathrm{out}}\left(\mathbf{v}^\star, \boldsymbol{\omega}^t, \mathbf{V}^t\right)\,, \end{cases} \quad \text{and} \quad \begin{cases} \boldsymbol{\gamma}^t = \frac{1}{\sqrt{k}} \mathbf{W}^\intercal \mathbf{g}^t + \boldsymbol{\Lambda}^t \hat{\mathbf{z}}^t\,, \\ \boldsymbol{\Lambda}^t = \frac{1}{k} (\mathbf{W}^2)^\intercal (\mathbf{g}^t)^2 \mathbf{I}_k\,, \\ \boldsymbol{\omega}^t = \frac{1}{\sqrt{k}} \mathbf{W} \hat{\mathbf{z}}^t - \mathbf{V}^t \mathbf{g}^{t-1}\,, \\ \mathbf{V}^t = \frac{1}{k} (\mathbf{W}^2) \hat{\mathbf{C}}_z^t \mathbf{I}_p\,, \end{cases} \tag{230}$$

where the operation $(\cdot)^2$ is taken component-wise. f_{z} is the mean of Q_{z} defined in eq. (221) and f_{out} is the mean of $V^{-1}(x - \omega)$ with respect to

$$\mathrm{Q}_{\mathrm{out}}\left(x; v^\star, \omega, V\right) = \frac{\mathrm{P}_{\mathrm{out}}\left(v^\star | x\right)}{\mathcal{Z}_{\mathrm{out}}(v^\star, \omega, V)} \frac{e^{-\frac{1}{2} V^{-1} (x - \omega)^2}}{\sqrt{2\pi V}} \tag{231}$$

Module composition In principle, composing the AMP equations for the inference problems above is non-trivial and requires a full-blown derivation from the BP equations on the composed factor graph in Fig. 53. Surprisingly, the upshot of this cumbersome computation is rather simple: the AMP equations for the composed model are equivalent to coupling the MF eqs. (228) and the GLM eqs. (230) by replacing $\mathrm{Q}_{\mathrm{v}}(v; b, A)$ and $\mathrm{Q}_{\mathrm{out}}(x; v^\star, \omega, V)$ with the following joint distribution:

$$\mathrm{Q}_{\mathrm{out}}(v, x; b, A, \omega, V) \equiv \frac{\mathrm{P}_{\mathrm{out}}\left(v | x\right)}{\mathcal{Z}_{\mathrm{out}}(b, A, \omega, v)} e^{-\frac{1}{2}Av^2 + bv} \frac{e^{-\frac{1}{2} V^{-1} (x - \omega)^2}}{\sqrt{2\pi V}}\,. \tag{232}$$

The associated update functions f_{v}, f_{out} are thus replaced by the means of v and $V^{-1}(x - \omega)$ with respect to this new joint distribution $\mathrm{Q}_{\mathrm{out}}$. Replacing the separable distributions Q_{u} and $\mathrm{Q}_{\mathrm{out}}$ by the joint distribution eq. (232)

and corresponding update functions as described above in eq. (228)-(230), we obtain the AMP algorithm for structured model. Additionally, we note that the AMP equations above are also valid for arbitrary weight matrix $\mathbf{W} \in \mathbb{R}^{p \times k}$. In the case of interest where $w_{il} \underset{\text{iid}}{\sim} \mathcal{N}(0,1)$, using $\mathbb{E}\left[w_{il}^2\right] = 1$ we can further simplify that leads to the following algorithm Algo. 3.

The AMP algorithm for the Wigner model is very similar and can be readily obtained by imposing at each time step $\left(\hat{\mathbf{u}}^t, \hat{\mathbf{C}}_u^t\right) = \left(\hat{\mathbf{v}}^t, \hat{\mathbf{C}}_v^t\right)$ and removing the redundant equations in Algo. 3.

```
**Input:** vector $\mathbf{Y} \in \mathbf{b}R^{n \times p}$ and matrix $\mathbf{W} \in \mathbf{b}R^{p \times k}$:
*Initialize to zero:* $(\mathbf{g}, \hat{\mathbf{u}}, \hat{\mathbf{v}}, \mathbf{b}_v, \mathbf{A}_v, \mathbf{b}_u, \mathbf{A}_u)^{t=0}$
*Initialize with:* $\hat{\mathbf{u}}^{t=1} = \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$, $\hat{\mathbf{v}}^{t=1} = \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$, $\hat{\mathbf{z}}^{t=1} = \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$,
$\hat{\mathbf{C}}_u^{t=1} = \mathbf{I}_n$, $\hat{\mathbf{C}}_v^{t=1} = \mathbf{I}_p$, $\hat{\mathbf{C}}_z^{t=1} = \mathbf{I}_k$.
**repeat**
    *Spiked layer:*
    $\mathbf{b}_u^t = \frac{1}{\Delta} \frac{\mathbf{Y}}{\sqrt{p}} \hat{\mathbf{v}}^t - \frac{1}{\Delta} \frac{\mathbf{1}_p^\intercal \hat{\mathbf{C}}_v^t}{p} \mathbf{I}_n \hat{\mathbf{u}}^{t-1}$ and $\mathbf{A}_u^t = \frac{1}{\Delta} \frac{\|\hat{\mathbf{v}}^t\|_2^2}{p} \mathbf{I}_n$
    $\mathbf{b}_v^t = \frac{1}{\Delta} \frac{\mathbf{Y}^\intercal}{\sqrt{p}} \hat{\mathbf{u}}^t - \frac{1}{\Delta} \frac{\mathbf{1}_n^\intercal \hat{\mathbf{C}}_u^t}{p} \mathbf{I}_p \hat{\mathbf{v}}^{t-1}$ and $\mathbf{A}_v^t = \frac{1}{\Delta} \frac{\|\hat{\mathbf{u}}^t\|_2^2}{p} \mathbf{I}_p$
    *Generative layer:*
    $\mathbf{V}^t = \frac{1}{k} \left(\mathbf{1}_k^\intercal \hat{\mathbf{C}}_z^t\right) \mathbf{I}_p$ and $\boldsymbol{\omega}^t = \frac{1}{\sqrt{k}} \mathbf{W} \hat{\mathbf{z}}^t - \mathbf{V}^t \mathbf{g}^{t-1}$ and
    $\mathbf{g}^t = \mathbf{f}_{\text{out}}\left(\mathbf{b}_v^t, \mathbf{A}_v^t, \boldsymbol{\omega}^t, \mathbf{V}^t\right)$
    $\boldsymbol{\Lambda}^t = \frac{1}{k} \|\mathbf{g}^t\|_2^2 \mathbf{I}_k$ and $\boldsymbol{\gamma}^t = \frac{1}{\sqrt{k}} \mathbf{W}^\intercal \mathbf{g}^t + \boldsymbol{\Lambda}^t \hat{\mathbf{z}}^t$
    *Update of the estimated marginals:*
    $\hat{\mathbf{u}}^{t+1} = \mathbf{f}_{\text{u}}(\mathbf{b}_u^t, \mathbf{A}_u^t)$ and $\hat{\mathbf{C}}_u^{t+1} = \partial_{\mathbf{b}} \mathbf{f}_{\text{u}}(\mathbf{b}_u^t, \mathbf{A}_u^t)$
    $\hat{\mathbf{v}}^{t+1} = \mathbf{f}_{\text{v}}(\mathbf{b}_v^t, \mathbf{A}_v^t, \boldsymbol{\omega}^t, \mathbf{V}^t)$ and $\hat{\mathbf{C}}_v^{t+1} = \partial_{\mathbf{b}} \mathbf{f}_{\text{v}}(\mathbf{b}_v^t, \mathbf{A}_v^t, \boldsymbol{\omega}^t, \mathbf{V}^t)$
    $\hat{\mathbf{z}}^{t+1} = \mathbf{f}_{\text{z}}(\boldsymbol{\gamma}^t, \boldsymbol{\Lambda}^t)$ and $\hat{\mathbf{C}}_z^{t+1} = \partial_{\boldsymbol{\gamma}} \mathbf{f}_{\text{z}}(\boldsymbol{\gamma}^t, \boldsymbol{\Lambda}^t)$
    $t = t + 1$
**until** Convergence
**Output:** $\hat{\mathbf{u}}, \hat{\mathbf{v}}, \hat{\mathbf{z}}$
```

Algorithme 3 : Bayes-optimal AMP algorithm for the spiked Wishart model with single-layer generative prior.

9.3.2 STATE EVOLUTION EQUATIONS

Perhaps the most important virtue of AMP-type algorithms is that their asymptotic performance can be tracked exactly via a set of scalar equations called *state evolution*. The order parameters involved are the average overlap between the estimated signals and the ground truth, and are closely related to the mean square error obtained by the algorithm. This fact has been proven for a range of models including the spiked matrix models with separable priors in (Javanmard et al., 2013), and with non-separable priors in (Berthier et al., 2017). Adapting the steps of these works, we now derive the state evolution equations for our structured model. As before, we focus on the derivation for the general Wishart model $\mathbf{u}\mathbf{v}^\intercal$, from which the Bayes-optimal SE equations for the symmetric $\mathbf{v}\mathbf{v}^\intercal$ can be readily obtained.

9.3.2.a RELAXED-BELIEF PROPAGATION EQUATIONS

Note that the standard derivation starts from the rBP equations, which are roughly equivalent to AMP updates up to the Onsager terms containing messages with delayed time indices $(\cdot)^{t-1}$. We briefly recall them below where we introduced the parameters $s_{j\mu} \equiv \frac{y_{j\mu}}{\Delta}$ and $r_{j\mu} \equiv -\frac{1}{\Delta} + s_{j\mu}^2, \forall j \in \llbracket n \rrbracket \mu \in \llbracket p \rrbracket$.

Variable u

$$\begin{aligned}
\hat{u}_{j\to j\mu}^{t+1} &= f_{\mathrm{u}}\left(b_{j\to j\mu}^{u,t}, A_{j\to j\mu}^{u,t}\right), \quad \hat{C}_{j\to j\mu}^{u,t+1} = \partial_b f_{\mathrm{u}}\left(b_{j\to j\mu}^{u,t}, A_{j\to j\mu}^{u,t}\right), \\
b_{j\to j\mu}^{u,t} &= \frac{1}{\sqrt{p}} \sum_{\nu\neq\mu}^{p} s_{j\nu} \hat{v}_{\nu\to j\nu}^{t}, \\
a_{j\to j\mu}^{u,t} &= \frac{1}{p} \sum_{\nu\neq\mu}^{p} s_{j\nu}^2 (\hat{v}_{\nu\to j\nu}^{t})^2 - r_{j\nu}(\hat{C}_{\nu\to j\nu}^{v,t} + (\hat{v}_{\nu\to j\nu}^{t})^2),
\end{aligned}$$

Variable v

$$\begin{aligned}
\hat{v}_{\mu\to j\mu}^{t+1} &= f_{\mathrm{v}}(b_{\mu\to j\mu}^{v,t}, A_{\mu\to j\mu}^{v,t}, \omega_{\mu}^{t}, V_{\mu}^{t}), \\
\hat{C}_{\mu\to j\mu}^{v,t+1} &= \partial_b f_{\mathrm{v}}(b_{\mu\to j\mu}^{v,t}, A_{\mu\to j\mu}^{v,t}, \omega_{\mu}^{t}, V_{\mu}^{t}), \\
b_{\mu\to j\mu}^{v,t} &= \frac{1}{\sqrt{p}} \sum_{l\neq j}^{n} s_{l\mu} \hat{u}_{l\to l\mu}^{t}, \\
A_{\mu\to j\mu}^{v,t} &= \frac{1}{p} \sum_{l\neq j}^{n} s_{l\mu}^2 (\hat{u}_{l\to l\mu}^{t})^2 - r_{l\mu}(\hat{C}_{l\to l\mu}^{u,t} + (\hat{u}_{l\to l\mu}^{t})^2), \\
\omega_{\mu}^{t} &= \frac{1}{\sqrt{k}} \sum_{i=1}^{k} w_{\mu i} \hat{z}_{i\to\mu}^{t}, \quad V_{\mu}^{t} = \frac{1}{k} \sum_{i=1}^{k} w_{\mu i}^2 \hat{C}_{i\to\mu}^{z,t},
\end{aligned} \tag{233}$$

Variable z

$$\begin{aligned}
\hat{z}_{i\to\mu}^{t+1} &= f_{\mathrm{z}}\left(\gamma_{i\to\mu}^{t}, \Lambda_{i\to\mu}^{t}\right), \quad \hat{C}_{i\to\mu}^{z,t+1} = \partial_\gamma f_{\mathrm{z}}\left(\gamma_{i\to\mu}^{t}, \Lambda_{i\to\mu}^{t}\right), \\
\gamma_{i\to\mu}^{t} &= \sum_{\nu\neq\mu}^{p} b_{\nu\to i}^{z,t}, \quad \Lambda_{i\to\mu}^{t} = \sum_{\nu\neq\mu}^{p} A_{\nu\to i}^{z,t}, \\
b_{\nu\to i}^{z,t} &= \frac{w_{\nu i}}{\sqrt{k}} f_{\mathrm{out}}(b_{\nu}^{v,t}, A_{\nu}^{v,t}, \omega_{\nu\to i}^{t}, V_{\nu\to i}^{t}), \\
A_{\nu\to i}^{z,t} &= -\frac{w_{\nu i}^2}{k} \partial_\omega f_{\mathrm{out}}(b_{\nu}^{v,t}, A_{\nu}^{v,t}, \omega_{\nu\to i}^{t}, V_{\nu\to i}^{t}), \\
b_{\nu}^{v,t} &= \frac{1}{\sqrt{p}} \sum_{j=1}^{n} s_{j\nu} \hat{u}_{j\to j\nu}^{t}, \\
A_{\nu}^{v,t} &= \frac{1}{p} \sum_{j=1}^{n} s_{j\nu}^2 (\hat{u}_{j\to j\nu}^{t})^2 - r_{j\nu}(\hat{C}_{j\to j\nu}^{u,t} + (\hat{u}_{j\to j\nu}^{t})^2), \\
\omega_{\nu\to i}^{t} &= \frac{1}{\sqrt{k}} \sum_{l\neq i}^{k} w_{\nu l} \hat{z}_{l\to\nu}^{t}, \quad V_{\nu\to i}^{t} = \frac{1}{k} \sum_{l\neq i}^{k} w_{\nu l}^2 \hat{C}_{l\to\nu}^{z,t}.
\end{aligned} \tag{234}$$

We take this as our starting point and refer the curious reader to (Lesieur et al., 2017a) for more details. The first step is to define the overlap parameters that measure the reconstruction of our inference problem:

$$q_u^t \equiv \mathbb{E}_{\mathbf{u}^\star} \lim_{n\to\infty} \frac{(\hat{\mathbf{u}}^t)^\intercal \hat{\mathbf{u}}^t}{n} = \mathbb{E}_{\mathbf{u}^\star} \lim_{n\to\infty} \frac{(\hat{\mathbf{u}}^t)^\intercal \mathbf{u}^\star}{n} \equiv m_u^t , \tag{235}$$

$$q_v^t \equiv \mathbb{E}_{\mathbf{v}^\star} \lim_{p\to\infty} \frac{(\hat{\mathbf{v}}^t)^\intercal \hat{\mathbf{v}}^t}{p} = \mathbb{E}_{\mathbf{v}^\star} \lim_{p\to\infty} \frac{(\hat{\mathbf{v}}^t)^\intercal \mathbf{v}^\star}{p} \equiv m_v^t , \tag{236}$$

$$q_z^t \equiv \mathbb{E}_{\mathbf{z}^\star} \lim_{k\to\infty} \frac{(\hat{\mathbf{z}}^t)^\intercal \hat{\mathbf{z}}^t}{k} = \mathbb{E}_{\mathbf{z}^\star} \lim_{k\to\infty} \frac{(\hat{\mathbf{z}}^t)^\intercal \mathbf{z}^\star}{k} \equiv m_z^t ,$$

where we used the Nishimori identity see (Lesieur et al., 2017a) or Appendix. A.3 to obtain the equality between order parameters $q_x^t = m_x^t$ for $x \in \{v, u, z\}$.

9.3.2.B AVERAGE DISTRIBUTIONS

Next, to see how these order parameters come into play, we compute the distribution of the rBP messages in eqs. (233-234), taking the average over the random variables $\mathbf{W}$, $\boldsymbol{\xi}$, the planted solutions $\mathbf{v}^\star, \mathbf{u}^\star, \mathbf{z}^\star$ and taking the limit $p \to \infty$. Note that using the BP independence assumption over the messages and keeping only dominant terms in the $1/p$ expansion, the dependency in the target node disappears and yields:

- **Average over $\mathbf{b}_u, \mathbf{A}_u$**

$$\begin{aligned}
\mathbb{E}\left[\mathbf{b}_u^t\right] &= \frac{1}{\sqrt{p}\Delta}\mathbb{E}\left[\mathbf{Y}\hat{\mathbf{v}}^t\right] = \frac{1}{\sqrt{p}\Delta}\mathbb{E}\left[\left(\frac{\mathbf{u}^\star(\mathbf{v}^\star)^\intercal}{\sqrt{p}} + \sqrt{\Delta}\boldsymbol{\xi}\right)\hat{\mathbf{v}}^t\right] \\
&\xrightarrow[p\to\infty]{} \frac{q_v^t}{\Delta}\mathbf{u}^\star , \\
\mathbb{E}\left[\mathbf{b}_u^t(\mathbf{b}_u^t)^\intercal\right] &= \frac{1}{p\Delta^2}\mathbb{E}\left[\mathbf{Y}\hat{\mathbf{v}}^t(\hat{\mathbf{v}}^t)^\intercal\mathbf{Y}^\intercal\right] = \frac{1}{\Delta}\frac{1}{p}\mathbb{E}\left[\xi\hat{\mathbf{v}}^t(\hat{\mathbf{v}}^t)^\intercal\xi^\intercal\right] + o\left(1/p\right) \\
&\xrightarrow[p\to\infty]{} \frac{q_v^t}{\Delta}\mathbf{I}_n , \\
\mathbb{E}\left[\mathbf{A}_u^t\right] &= \mathbb{E}\left[\frac{1}{\Delta}\frac{\|\hat{\mathbf{v}}^t\|_2^2}{p}\mathbf{I}_n\right] \xrightarrow[p\to\infty]{} \frac{q_v^t}{\Delta}\mathbf{I}_n .
\end{aligned} \tag{237}$$

- **Average over $\mathbf{b}_v, \mathbf{A}_v$**

$$\mathbb{E}\left[\mathbf{b}_v^t\right] \xrightarrow[p\to\infty]{} \beta\frac{q_u^t}{\Delta}\mathbf{v}^\star , \quad \mathbb{E}\left[\mathbf{b}_v^t(\mathbf{b}_v^t)^\intercal\right] \xrightarrow[p\to\infty]{} \beta\frac{q_u^t}{\Delta}\mathbf{I}_p , \quad \mathbb{E}\left[\mathbf{A}_v^t\right] \xrightarrow[p\to\infty]{} \beta\frac{q_u^t}{\Delta}\mathbf{I}_p . \tag{238}$$

- **Average over $\boldsymbol{\omega}, \mathbf{V}$**

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{\omega}^t\right] = \mathbf{0}_p , \quad \mathbb{E}\left[\boldsymbol{\omega}^t(\boldsymbol{\omega}^t)^\intercal\right] = \mathbb{E}\left[\frac{1}{k}\mathbf{W}\hat{\mathbf{z}}^t(\hat{\mathbf{z}}^t)^\intercal\mathbf{W}^\intercal\right] \xrightarrow[n\to\infty]{} q_z^t\mathbf{I}_p , \\
&\mathbb{E}\left[\mathbf{V}^t\right] \xrightarrow[k\to\infty]{} (\rho_z - q_z^t)\mathbf{I}_p .
\end{aligned} \tag{239}$$

Wrapping the above equations together, we obtained the distributions of means and variances $\mathbf{b}_u, \mathbf{A}_u$, $\mathbf{b}_v, \mathbf{A}_v$ and $\boldsymbol{\omega}, \mathbf{V}$:

$$\begin{aligned} \mathbf{b}_u &\sim \frac{q_v^t}{\Delta}\mathbf{u}^\star + \sqrt{\frac{q_v^t}{\Delta}}\boldsymbol{\xi}_u , & \mathbf{A}_u^t &\sim \frac{q_v^t}{\Delta}\mathbf{I}_n , \\ \mathbf{b}_v &\sim \beta\frac{q_u^t}{\Delta}\mathbf{v}^\star + \sqrt{\beta\frac{q_u^t}{\Delta}}\boldsymbol{\xi}_v , & \mathbf{A}_v^t &\sim \beta\frac{q_u^t}{\Delta}\mathbf{I}_p , \\ \boldsymbol{\omega} &\sim \sqrt{q_z^t}\boldsymbol{\eta} , & \mathbf{V} &\sim (\rho_z - q_z^t)\mathbf{I}_p , \end{aligned} \tag{240}$$

with $\boldsymbol{\xi}_u \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n), \boldsymbol{\xi}_v \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p), \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$.

9.3.2.c STATE EVOLUTION EQUATIONS IN THE WISHART MODEL

With the averaged limiting distributions of all the messages, we can now compute the state evolution of the overlaps. Using the definition of the overlaps eq. (235) and distributions in eq. (240), we obtain:

Variable u

$$\begin{aligned} q_u^{t+1} &= \mathbb{E}_{\mathbf{u}^\star} \lim_{n\to\infty} \frac{1}{n}\mathbf{f}_{\mathrm{u}}(\mathbf{b}_u^t, \mathbf{A}_u^t)^\intercal \mathbf{f}_u(\mathbf{b}_u^t, \mathbf{A}_u^t) \\ &= \mathbb{E}_{u^\star,\xi}\left[f_{\mathrm{u}}\left(\frac{q_v^t}{\Delta}u^\star + \sqrt{\frac{q_v^t}{\Delta}}\xi, \frac{q_v^t}{\Delta} \right)^2 \right] \end{aligned} \tag{241}$$

where $u^\star \sim \mathrm{P_u}$, $\xi \sim \mathcal{N}(0,1)$.

Variable v

$$\begin{aligned} q_v^{t+1} &= \mathbb{E}_{\mathbf{v}^\star} \lim_{p\to\infty} \frac{1}{p}\mathbf{f}_{\mathrm{v}}(\mathbf{b}_v^t, \mathbf{b}v^t, \boldsymbol{\omega}^t, \mathbf{V}^t)^\intercal f_v(\mathbf{b}_v^t, \mathbf{b}v^t, \boldsymbol{\omega}^t, \mathbf{V}^t) \\ &= \mathbb{E}_{v^\star,\xi,\eta}\left[f_{\mathrm{v}}\left(\frac{\beta q_u^t}{\Delta}v^\star + \sqrt{\frac{\beta q_u^t}{\Delta}}\xi, \beta\frac{q_u^t}{\Delta}, \sqrt{q_z^t}\eta, \rho_z - q_z^t \right)^2 \right] \end{aligned} \tag{242}$$

where $v^\star \sim \mathrm{P_v}$, $\xi \sim \mathcal{N}(0,1)$.

Variable ẑ and z Even if the *hat* overlap does not have as much physical meaning as the standard overlaps that quantify the reconstruction performances, we define it as

$$\hat{q}_z^t \equiv \alpha\mathbb{E}_{v^\star,\xi,\eta}\left[f_{\mathrm{out}}\left(\frac{\beta q_u^t}{\Delta}v^\star + \sqrt{\frac{\beta q_u^t}{\Delta}}\xi, \beta\frac{q_u^t}{\Delta}\sqrt{q_z^t}\eta, \rho_z - q_z^t \right)^2 \right] , \tag{243}$$

with $v^\star \sim \mathrm{P}_{\mathrm{v}}$, $\xi, \eta \sim \mathcal{N}(0,1)$. Averages of the messages of the variable $\mathbf{z}$ are explicitly expressed as a function of the *hat* overlaps introduced just above:

$$\mathbb{E}\left[\boldsymbol{\gamma}^t\right] \xrightarrow[k\to\infty]{} \hat{q}_z^t \mathbf{z}^\star, \quad \mathbb{E}\left[\boldsymbol{\gamma}^t (\boldsymbol{\gamma}^t)^\intercal\right] \xrightarrow[k\to\infty]{} \hat{q}_z^t \mathbf{I}_k \quad \text{and} \quad \mathbb{E}\left[\boldsymbol{\Lambda}^t\right] \xrightarrow[k\to\infty]{} \hat{q}_z^t \mathbf{I}_k\,. \tag{244}$$

At the leading order, we obtain

$$\boldsymbol{\gamma}^t \sim \hat{q}_z^t \mathbf{z}^\star + \sqrt{\hat{q}_z^t}\boldsymbol{\xi}\,, \quad \boldsymbol{\Lambda}^t \sim \hat{q}_z^t \mathbf{I}_k \quad \text{with} \quad \boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{0}_k, \mathbf{I}_k\right) \tag{245}$$

and finally

$$\begin{aligned} q_z^{t+1} &\equiv \mathbb{E}_{\mathbf{z}^\star} \lim_{k\to\infty} \frac{1}{k} \mathbf{f}_z(\boldsymbol{\gamma}^t, \boldsymbol{\Lambda}^t)^\intercal \mathbf{f}_z(\boldsymbol{\gamma}^t, \boldsymbol{\Lambda}^t) \\ &= \mathbb{E}_{z^\star,\xi}\left[f_z\left(\hat{q}_z^t z^\star + \sqrt{\hat{q}_z^t}\xi, \hat{q}_z^t\right)^2\right]. \end{aligned} \tag{246}$$

As a conclusion, equations (241- 243, 246) constitute the closed set of SE equations of the Bayes-optimal AMP algorithm for the Wishart model.

9.3.2.D STATE EVOLUTION EQUATIONS IN THE WIGNER MODEL

Finally, similarly to the derivation of the AMP algorithm, the SE equations for the Wigner model ($\mathbf{v}\mathbf{v}^\intercal$) are obtained as a particular case of the above by simply restricting $q_u^t = q_v^t$ and $\beta = 1$. In the end, performing a change of variable, this leaves us with only three coupled equations:

$$\begin{aligned} q_z^{t+1} &= \mathbb{E}_\xi\left[\mathcal{Z}_z \times f_z^2\left(\sqrt{\hat{q}_z^t}\xi, \hat{q}_z^t\right)\right] = 2\partial_{\hat{q}_z}\Psi_z\left(\hat{q}_z^t\right), \\ \hat{q}_z^t &= \alpha \mathbb{E}_{\xi,\eta}\left[\mathcal{Z}_{\text{out}} \times f_{\text{out}}^2\left(\sqrt{\frac{q_v^t}{\Delta}}\xi, \frac{q_v^t}{\Delta}, \sqrt{q_z^t}\eta, \rho_z - q_z^t\right)\right] \\ &= 2\alpha\partial_{q_z}\Psi_{\text{out}}\left(\frac{q_v^t}{\Delta}, q_z^t\right), \\ q_v^{t+1} &= \mathbb{E}_{\xi,\eta}\left[\mathcal{Z}_{\text{out}} \times f_v^2\left(\sqrt{\frac{q_v^t}{\Delta}}\xi, \frac{q_v^t}{\Delta}, \sqrt{q_z^t}\eta, \rho_z - q_z^t\right)\right] \\ &= 2\partial_{q_v}\Psi_{\text{out}}\left(\frac{q_v^t}{\Delta}, q_z^t\right), \end{aligned} \tag{247}$$

with initialization $q_v^{t=0} = \varepsilon$, $q_z^{t=0} = \varepsilon$ and a small $\varepsilon > 0$. We notice immediately that (247) are the same equations as the fixed point equations related to the Bayes-optimal estimation (222) with specific time-indices and initialization, but crucially the same fixed points. Thus the analysis of fixed points in Sec. 9.2.2 applies straightforwardly here. In particular, since in all cases analyzed we found the stable fixed point of (222) to be unique, we conclude that our AMP algorithm reaches asymptotically optimal performance in these cases.

We can further check this result by numerically comparing the runs of AMP on finite size instances with the state evolution curves already presented in Fig. 51, also giving an idea of the amplitude of the finite size effects. This

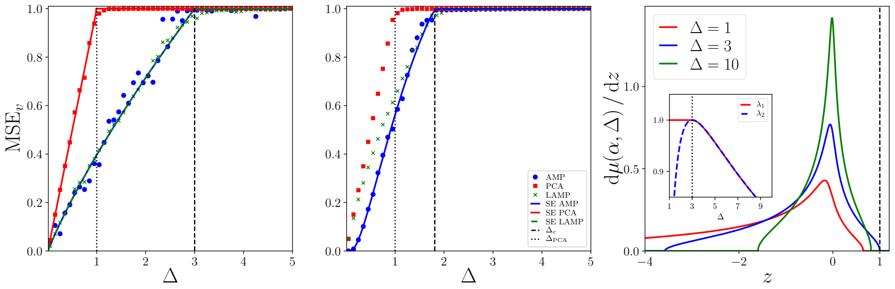


Figure 54: Comparison between PCA, LAMP and AMP for (left) the linear, (center) and sign activations, at compression ratio $\alpha = 2$. Lines correspond to the theoretical asymptotic performance of PCA (red line), LAMP (green line) and AMP (blue line). Dots correspond to simulations of PCA (red squares), LAMP (green crosses) for $k = 10^4$ and AMP (blue points) for $k = 5.10^3$, $\sigma^2 = 1$. Notice that the spectral estimators have been rescaled by a factor $(q^\star_{v,\text{AMP}})^{1/2}$ to fairly compare AMP with PCA and LAMP. (Right) Illustration of the spectral phase transition in the matrix $\mathbf{\Gamma}_p^{vv}$ eq. (257) at $\alpha = 2$ with an informative leading eigenvector with eigenvalue equal to 1 out of the bulk for $\Delta \leq 1 + \alpha$. We show the bulk spectral density $\mu(\alpha, \Delta)$. The inset shows the two leading eigenvalues.

experiment is illustrated in Fig. 54, together with a curve for PCA and for LAMP, a spectral method we derive from AMP in the next section. A code for reproducing this experiment is provided in GitHub repository (Aubin et al., 2019a).

9.4 LAMP: A SPECTRAL ALGORITHM FOR GENERATIVE PRIORS

Spectral algorithms are the most popular and simplest methods for solving the spiked matrix estimation problem. For instance, canonical PCA estimates the spike from the leading eigenvector of the matrix $\mathbf{Y}$. A classical result from Baik, Ben Arous and Péché (BBP) (Baik et al., 2005) shows that this eigenvector is correlated with the signal if and only if the signal-to-noise ratio $\rho_v^2 / \Delta > 1$. For sparse separable priors with $\rho_v^2 = \Theta(1)$, $\Delta_{\text{PCA}} = \rho_v^2$ is also the threshold for AMP and it is conjectured that no polynomial algorithm can improve upon it (Lesieur et al., 2017a). In the previous section we have shown that our structured AMP algorithm has a consistently better performance than PCA, and in particular achieve the optimal threshold for better-than-random recovery. This is not a surprise, since different from AMP, vanilla PCA doesn't take into account the information available from the prior.

Despite all its virtues, AMP is unarguably a convoluted algorithm. It would be desirable to have a simpler spectral algorithm taking into account the structured nature of the prior. In this section we design a spectral algorithm, hereafter named LAMP, matching the AMP recovery threshold. Our derivation follows the strategy pioneered in (Krzakala et al., 2013), consisting on the linearization of the AMP equations around the non-informative fixed point.

In this section, the discussion is framed on the Wigner model, the Wishart case being a straightforward generalization.

In order for the $q_v = 0$ expansion to be well-defined, we first need to insure that this is indeed a fixed point. Indeed, this condition was already discussed in Sec. 9.2 for the fixed point equations for the Bayes-optimal estimator. Not surprisingly, the same conditions can be obtained independently from the AMP equations by analyzing when $\hat{\mathbf{v}} = \mathbf{0}$ is a fixed point, and are repeated below for convenience.

$$(\hat{\mathbf{v}}, \hat{\mathbf{z}}) = (\mathbf{0}, \mathbf{0}) \quad \text{if} \quad \mathscr{C} \equiv \left\{ \mathbb{E}_{Q^0_{\text{out}}}[v] = 0 \quad \text{and} \quad \mathbb{E}_{P_{\text{v}}}[z] = 0 \right\}. \tag{248}$$

That these conditions agree exactly to the ones in eq. (223) is just a rephrasing of the fact that the AMP SE equations in eqs. (247) have the same fixed points as the Bayes-optimal estimator.

9.4.1 LINEARIZING THE AMP EQUATIONS

To lighten notations, we denote with $|_\star$ quantities that are evaluated at $(\mathbf{b}_v, \mathbf{A}_v, \boldsymbol{\omega}, \mathbf{V}, \boldsymbol{\gamma}, \boldsymbol{\Lambda}) = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \rho_z \mathbf{I}_p, \mathbf{0}, \mathbf{0})$, and we linearize the AMP equations in Algo. (3) around the uninformative fixed point

$$\begin{aligned} &(\hat{\mathbf{v}}, \hat{\mathbf{C}}_v) = (\mathbf{0}, \rho_v \mathbf{I}_p), \quad (\hat{\mathbf{z}}, \hat{\mathbf{C}}_z) = (\mathbf{0}, \rho_z \mathbf{I}_k), \quad (\mathbf{b}_v, \mathbf{A}_v) = (\mathbf{0}, \mathbf{0}), \\ &(\boldsymbol{\gamma}, \boldsymbol{\Lambda}) = (\mathbf{0}, \mathbf{0}), \quad (\boldsymbol{\omega}, \mathbf{V}, \mathbf{g}) = (\mathbf{0}, \rho_z \mathbf{I}_p, \mathbf{0}). \end{aligned} \tag{249}$$

In a scalar formulation, this yields

$$\begin{aligned} \delta \hat{v}_i^{t+1} &= \partial_b f_{\text{v}}|_\star \delta b_i^{v,t} + \partial_A f_{\text{v}}|_\star \delta A_i^{v,t} + \partial_\omega f_{\text{v}}|_\star \delta \omega_i^t + \partial_V f_{\text{v}}|_\star \delta V_i^t, \\ \delta \hat{c}_i^{v,t+1} &= \partial^2_{b,b} f_{\text{v}}|_\star \delta b_i^{v,t} + \partial^2_{A,b} f_{\text{v}}|_\star \delta A_i^{v,t} + \partial^2_{\omega,b} f_{\text{v}}|_\star \delta \omega_i^t + \partial^2_{V,b} f_{\text{v}}|_\star \delta V_i^t, \\ \delta \hat{z}_l^{t+1} &= \partial_\gamma f_{\text{z}}|_\star \delta \gamma_l^t + \partial_\Lambda f_{\text{z}}|_\star \delta \Lambda_l^t, \\ \delta \hat{c}_i^{z,t+1} &= \partial^2_{\gamma,\gamma} f_{\text{z}}|_\star \delta \gamma_l^t + \partial^2_{\Lambda,\gamma} f_{\text{z}}|_\star \delta \Lambda_l^t, \\ \delta g_i^t &= \partial_b f_{\text{out}}|_\star \delta b_i^{v,t} + \partial_A f_{\text{out}}|_\star \delta A_i^{v,t} + \partial_\omega f_{\text{out}}|_\star \delta \omega_i^t + \partial_V f_{\text{out}}|_\star \delta V_i^t, \end{aligned} \tag{250}$$

with

$$\begin{aligned} \delta b_i^{v,t} &= \frac{1}{\Delta} \sum_{j=1}^p \frac{y_{ji}}{\sqrt{p}} \delta \hat{v}_j^t - \frac{1}{\Delta} \left(\sum_{j=1}^p \frac{\hat{c}_j^{v,t}|_\star}{p} \right) \delta \hat{v}_i^{t-1} - \frac{1}{\Delta} \left(\sum_{j=1}^p \frac{\delta \hat{c}_j^{v,t}}{p} \right) \hat{v}_i^{t-1}|_\star, \\ \delta A_i^{v,t} &= \frac{2}{\Delta} \sum_{j=1}^p \frac{\hat{v}_j^t|_\star \delta \hat{v}_j^t}{p} = 0 \\ \delta \omega_i^t &= \frac{1}{\sqrt{k}} \sum_{l=1}^k w_{il} \delta \hat{z}_l^t - \delta V_i^t g_i^{t-1}|_\star - V_i^t|_\star \delta g_i^{t-1}, \quad \delta V_i^t = \frac{1}{k} \sum_{l=1}^k \delta \hat{c}_l^{z,t}, \\ \delta \gamma_l^t &= \frac{1}{\sqrt{k}} \sum_{i=1}^p W_{il} \delta \mathbf{g}_i^t + \delta \Lambda_l^t \hat{\mathbf{z}}_l^t|_\star + \Lambda_l^t|_\star \delta \hat{z}_l^t, \quad \delta \Lambda_l^t = \frac{2}{k} \sum_{i=1}^p g_i^t|_\star \delta g_i^t = 0. \end{aligned} \tag{251}$$

These equations can be simplified and closed over three vectors $\hat{\mathbf{v}} \in \mathbb{R}^p$, $\hat{\mathbf{z}} \in \mathbb{R}^k$ and $\boldsymbol{\omega} \in \mathbb{R}^p$, where we used the existence condition $\mathscr{C}$ that leads to $\partial_\omega f_{\text{out}}|_\star = \partial_V f_{\text{out}}|_\star = 0$. Finally, inserting eq. (251) in (250), rewriting the partial derivatives of f_v, f_z and f_{out} at the fixed point $|_\star$ as moments of the distributions $\mathrm{P_z}$ and $\mathrm{Q_{out}}$ and simplifying the expression using the condition $\mathscr{C}$, we finally obtain

$$\begin{aligned} \delta\hat{\mathbf{v}}^{t+1} =& \frac{1}{\Delta}\rho_v\left(\frac{\mathbf{Y}}{\sqrt{p}}\delta\hat{\mathbf{v}}^t - \rho_v\mathbf{I}_p\delta\hat{\mathbf{v}}^{t-1}\right) + \rho_z^{-1}\mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[vx]\mathbf{I}_p\delta\boldsymbol{\omega}^t \\ &+ \frac{\mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[vx^2]\mathbb{E}_{\mathrm{P_z}}\left[z^3\right]}{2\rho_z^3}\frac{\mathbf{1}_p\mathbf{1}_k^\intercal}{k}\delta\hat{\mathbf{z}}^t\,, \end{aligned} \tag{252}$$

$$\delta\hat{\mathbf{z}}^{t+1} = \frac{1}{\Delta}\mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[vx]\frac{\mathbf{W}^\intercal}{\sqrt{k}}\left[\frac{\mathbf{Y}}{\sqrt{p}}\delta\hat{\mathbf{v}}^t - \rho_v\mathbf{I}_p\delta\hat{\mathbf{v}}^{t-1}\right], \tag{253}$$

$$\begin{aligned} \delta\boldsymbol{\omega}^{t+1} =& \frac{1}{\Delta}\left(\mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[vx]\frac{\mathbf{W}\mathbf{W}^\intercal}{k}\left[\frac{\mathbf{Y}}{\sqrt{p}}\delta\hat{\mathbf{v}}^t - \rho_v\mathbf{I}_p\delta\hat{\mathbf{v}}^{t-1}\right]\right) \\ &- \mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[vx]\left[\frac{\mathbf{Y}}{\sqrt{p}}\delta\hat{\mathbf{v}}^{t-1} - \rho_v\mathbf{I}_p\delta\hat{\mathbf{v}}^{t-2}\right]. \end{aligned} \tag{254}$$

Inserting eq. (253)-(254) in (252) and dropping heuristically the time indices, we finally obtain the closed linear equation $\hat{\mathbf{v}} = \boldsymbol{\Gamma}_p^{vv}\hat{\mathbf{v}}$, where the LAMP operator $\boldsymbol{\Gamma}_p^{vv}$ is given by

$$\boldsymbol{\Gamma}_p^{vv} = \frac{1}{\Delta}\left((a-b)\mathbf{I}_p + b\frac{\mathbf{W}\mathbf{W}^\intercal}{k} + c\frac{\mathbf{1}_p\mathbf{1}_k^\intercal}{k}\frac{\mathbf{W}^\intercal}{\sqrt{k}}\right)\times\left(\frac{\mathbf{Y}}{\sqrt{p}} - a\mathbf{I}_p\right), \tag{255}$$

where the parameters are simply the moments of distributions $\mathrm{P_z}$ and $\mathrm{Q}^0_{\text{out}}$

$$\begin{aligned} &a \equiv \mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[v^2] = \rho_v\,, \quad b \equiv \rho_z^{-1}\mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[vx]^2\,, \\ &c \equiv \frac{1}{2}\rho_z^{-3}\mathbb{E}_{\mathrm{P_z}}\left[z^3\right]\mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[vx^2]\mathbb{E}_{\mathrm{Q}^0_{\text{out}}}[vx]\,. \end{aligned} \tag{256}$$

Note that in most of the cases we studied, the parameter c, taking into account the skewness of the variable $\mathbf{z}$, is zero, simplifying considerably the structured matrix. Moreover, for the specific examples already discussed in Sec. 9.2, the LAMP operator $\boldsymbol{\Gamma}_p^{vv}$ is very simple. For instance, for Gaussian z and $\mathrm{P_{out}}(v|x) = \delta(v - \text{sign}(x))$, we have $(a,b,c) = (1, 2/\pi, 0)$. Instead, for linear activation we get $(a,b,c) = (1,1,0)$. Note that in this last case, the LAMP operator can be written as

$$\boldsymbol{\Gamma}_p^{vv} = \frac{1}{\Delta}\mathbf{K}_p\left[\frac{\mathbf{Y}}{\sqrt{p}} - \mathbf{I}_p\right] \text{ with } \mathbf{K}_p = \frac{\mathbf{W}\mathbf{W}^\intercal}{k} = \boldsymbol{\Sigma} \approx \frac{1}{n}\sum_\alpha \mathbf{v}^\alpha(\mathbf{v}^\alpha)^\intercal\,, \tag{257}$$

or, in other words, $\mathbf{K}_p$ is the covariance matrix of the structured spike $\mathbf{v}$. The same observation holds for the sign activation function. Interestingly, the covariance matrix $\boldsymbol{\Sigma}$ can be empirically estimated directly from samples of spikes, without the knowledge of the generative model $(\varphi, \mathbf{W})$ itself, suggesting a simple practical implementation of LAMP. Therefore we finally use a more generic definition for LAMP as expressed in Algo. 4. From this perspective, LAMP in Algo. 4 can be interpreted as a PCA that takes into account the structure of the prior by incorporating the non-trivial correlations through $\mathbf{K}_p$ into the spectral estimation. In particular taking $\mathrm{P}_{\mathrm{v}}(\mathbf{v}) = \mathcal{N}_{\mathbf{v}}\left(\mathbf{0}, \mathbf{I}_p\right)$, we obtain $\mathbf{\Gamma}_p^{vv} = \frac{1}{\Delta}\left[\frac{\mathbf{Y}}{\sqrt{p}} - \mathbf{I}_p\right]$ and recognize the PCA operator that has been shifted. Analogously to the state evolution for AMP, the asymptotic performance of both PCA and LAMP can be evaluated in a closed-form for the spiked Wigner model with single-layer generative prior with linear activation. The corresponding expressions are derived in the next section and plotted in Fig. 54 for the three considered algorithms.

```
Input: Observed matrix $\mathbf{Y} \in \mathbb{R}^{p \times p}$, prior $\mathrm{P}_{\mathrm{v}}$ on $\mathbf{v} \in \mathbb{R}^p$
Take the leading eigenvector $\hat{\mathbf{v}} \in \mathbb{R}^p$ of
$\mathbf{\Gamma}_p^{vv} \equiv \mathbf{K}_p \left[\frac{\mathbf{Y}}{\sqrt{p}} - \mathbf{I}_p\right]$ with $\mathbf{K}_p = \mathbb{E}_{\mathrm{P}_{\mathrm{v}}}\left[\mathbf{v}\mathbf{v}^\intercal\right]$.
```

Algorithme 4 : LAMP spectral algorithm for the Wigner model.

9.4.2 STATE EVOLUTION FOR LAMP AND PCA IN THE LINEAR CASE

As we have already mentioned in Sec. 9.3.2, one of the greatest virtues of AMP is being able to track its asymptotic performance through a set of simple scalar state evolution equations. Interestingly, we can also derive state evolution equations for the LAMP algorithm in the linear case. This allows a direct comparison between the performance of LAMP and the performance of PCA.

For the noiseless linear channel $\mathrm{P}_{\mathrm{out}}(v|x) = \delta\left(v - x\right)$, the set of eqs. (252-254) are already linear. Hence the LAMP spectral method flows directly from the AMP Algo. 3. As a consequence, this means that the state evolution equations associated to the spectral method are simply dictated by the set of AMP SE equations eq. (247).

However, it is worth stressing that as the LAMP returns a normalized estimator, the LAMP MSE is not given by the AMP mean squared error. We now compute the overlaps and mean squared error performed by this spectral algorithm.

Recall that m_v and q_v are the parameters defined in eq. (235), respectively measuring the overlaps between the ground truth $\mathbf{v}^\star$ and the estimator $\hat{\mathbf{v}}$, and the norm of the estimator. In the general case, the MMSE eq. (216) becomes:

$$\mathrm{MMSE}_v = \rho_v + \mathbb{E}_{\mathbf{v}^\star} \lim_{p\to\infty} \frac{1}{p}\|\hat{\mathbf{v}}\|_2^2 - 2\mathbb{E}_{\mathbf{v}^\star} \lim_{p\to\infty} \frac{1}{p}\hat{\mathbf{v}}^\intercal \mathbf{v}^\star = \rho_v + q_v - 2m_v\,,$$

However the LAMP spectral method computes the normalized leading eigenvector of the structured matrix $\mathbf{\Gamma}_p^{vv}$. Hence the norm of the LAMP estimator is $\|\hat{\mathbf{v}}\|^2_{\mathrm{LAMP}} = q_{v,\mathrm{LAMP}} = 1$, while the Bayes-optimal AMP estimator is not normalized with $\|\hat{\mathbf{v}}\|^2_{\mathrm{AMP}} = q^\star_{v,\mathrm{AMP}} = m^\star_{v,\mathrm{AMP}} \neq 1$, solutions of eq. (247).

As the non-normalized LAMP estimator follows AMP state evolutions in the *linear case*, the overlap with the ground truth is thus given by:

$$\begin{aligned} m_{v,\mathrm{LAMP}} &\equiv \mathbb{E}_{\mathbf{v}^\star} \lim_{p\to\infty} \frac{1}{p} \hat{\mathbf{v}}_{\mathrm{LAMP}}^\intercal \mathbf{v}^\star = \mathbb{E}_{\mathbf{v}^\star} \lim_{p\to\infty} \frac{1}{p} \left(\frac{\hat{\mathbf{v}}_{\mathrm{AMP}}}{\|\hat{\mathbf{v}}\|_{\mathrm{AMP}}} \right)^\intercal \mathbf{v}^\star \\ &= \frac{m^\star_{v,\mathrm{AMP}}}{\left(q^\star_{v,\mathrm{AMP}}\right)^{1/2}} = \left(m^\star_{v,\mathrm{AMP}}\right)^{1/2} . \end{aligned}$$

Finally the mean squared error performed by the LAMP method is easily obtained from the optimal overlap reached by the AMP algorithm and yields

$$\mathrm{MSE}_{v,\mathrm{LAMP}} = \rho_v + 1 - 2\left(q^\star_{v,\mathrm{AMP}}\right)^{1/2} .$$

The respective result for PCA can be obtained from the observation that for the linear case, the $\alpha = 0$ LAMP operator reduces exactly to the matrix $\mathbf{Y}$. In other words, in this case LAMP reduces to PCA. In terms of the prior, this is clear since $\alpha = 0$ is equivalent to a separable Gaussian prior, for which the spectral algorithm derived from AMP is exactly given by PCA (Lesieur et al., 2017a). Therefore we can simply state that the mean squared error performed by PCA is computed using the optimal overlap reached by AMP at $\alpha = 0$:

$$\mathrm{MSE}_{v,\mathrm{PCA}} = \rho_v + 1 - 2\left(q^\star_{v,\mathrm{AMP}}|_{\alpha=0}\right)^{1/2} .$$

In order to fairly compare PCA, LAMP and AMP in Fig. 54, instead of showing the MSE corresponding to the *normalized* PCA and LAMP estimators, we rescale these spectral estimators by the optimal normalisation $\left(q^\star_{v,\mathrm{AMP}}\right)^{1/2}$ (obtained from AMP for instance) so that the renormalized MSE are given by

$$\mathrm{MSE}_{v,\mathrm{LAMP}} = \rho_v - m^\star_{v,\mathrm{LAMP}} , \qquad \mathrm{MSE}_{v,\mathrm{PCA}} = \rho_v - m^\star_{v,\mathrm{PCA}} .$$

Therefore in the *linear case* we simply obtain that LAMP is strictly equivalent to AMP, while PCA is sub-optimal:

$$\mathrm{MSE}_{v,\mathrm{LAMP}} = \rho_v - q^\star_{v,\mathrm{AMP}} , \qquad \mathrm{MSE}_{v,\mathrm{PCA}} = \rho_v - q^\star_{v,\mathrm{AMP}}|_{\alpha=0} .$$

Fig. 54 shows good agreement between the state evolution for LAMP and PCA with linear activation (solid lines) and the respective finite instance numerical simulations (points).

9.4.3 A RANDOM MATRIX PERSPECTIVE ON THE RECOVERY THRESHOLD

Remarkably, the performance of the spectral method based on matrix (257) can be investigated independently of AMP using random matrix theory. An analysis of the random matrix (257) shows that a spectral phase transition for generative prior with linear activation appears at $\Delta_c = 1 + \alpha$ (as for AMP). This transition is analogous to the well-known BBP transition (Baik et al., 2005), but for a non-GOE random matrix (257). For the spiked Wigner models with linear generative prior we prove two detailed theorems describing the behavior of the supremum of the bulk spectral density, the transition of the largest eigenvalue and the correlation of the corresponding eigenvector. The theorems counterparts for the linear Wishart model are very similar, and are presented in appendix. We assume in the following that $\rho_v = 1$ to simplify the analysis (without any loss of generality). Recall that we have

$$\mathbf{\Gamma}_p^{vv} \equiv \left[\frac{1}{k}\mathbf{W}\mathbf{W}^\intercal\right]\left[\frac{1}{\sqrt{\Delta p}}\boldsymbol{\xi} + \frac{1}{\Delta}\frac{\mathbf{v}\mathbf{v}^\intercal}{p} - \frac{1}{\Delta}\mathbf{I}_p\right]. \tag{258}$$

Here $\boldsymbol{\xi}/\sqrt{p}$ is a matrix from the Gaussian Orthogonal Ensemble, i.e. $\boldsymbol{\xi}$ is a real symmetric matrix with entries drawn independently from a Gaussian distribution with zero mean and variance $\mathbb{E}\,\xi_{ij}^2 = (1+\delta_{ij})$.

Theorem 9.4.1 (Bulk of the spectral density, spiked Wigner, linear activation)**.** *For any $\alpha, \Delta > 0$, as $p \to +\infty$, the spectral measure of $\mathbf{\Gamma}_p^{vv}$ converges almost surely and in the weak sense to a well-defined and compactly supported probability measure $\mu(\alpha,\Delta)$, and we denote* $\operatorname{supp}\mu$ *its support. We separate two cases:*

(i) If $\Delta \leq \frac{1}{4}$, then $\operatorname{supp}\mu \subseteq \mathbb{R}_-$.

(ii) Assume now $\Delta > \frac{1}{4}$ and denote $z_1(\Delta) \equiv -\Delta^{-1} + 2\Delta^{-1/2} > 0$. Let ρ_Δ be the probability measure on $\mathbb{R}$ with density

$$\rho_\Delta(\mathrm{d}t) = \frac{\sqrt{\Delta}}{2\pi}\sqrt{4 - \Delta\left(t + \frac{1}{\Delta}\right)^2}\,\mathbb{1}\left\{\left|t + \frac{1}{\Delta}\right| \leq \frac{2}{\sqrt{\Delta}}\right\}\mathrm{d}t. \tag{259}$$

Note that the supremum of the support of ρ_Δ is $z_1(\Delta)$. The following equation admits a unique solution for $s \in (-z_1(\Delta)^{-1}, 0)$:

$$\alpha \int \rho_\Delta(\mathrm{d}t)\left(\frac{st}{1+st}\right)^2 = 1. \tag{260}$$

We denote this solution as $s_{edge}(\alpha,\Delta)$ (or simply s_{edge}). The supremum of the support of $\mu(\alpha,\Delta)$ is denoted $\lambda_{max}(\alpha,\Delta)$ (or simply λ_{max}). It is given by:

$$\lambda_{max} = \begin{cases} -\dfrac{1}{s_{edge}} + \alpha \displaystyle\int \rho_\Delta(\mathrm{d}t)\frac{t}{1+s_{edge}t} & \text{if } \alpha \leq 1, \\ \max\left(0, -\dfrac{1}{s_{edge}} + \alpha \displaystyle\int \rho_\Delta(\mathrm{d}t)\frac{t}{1+s_{edge}t}\right) & \text{if } \alpha > 1. \end{cases} \tag{261}$$

As a function of Δ, λ_{max} has a unique global maximum, reached exactly at the point $\Delta = \Delta_c(\alpha) = 1 + \alpha$. Moreover, $\lambda_{max}(\alpha, \Delta_c(\alpha)) = 1$.

Theorem 9.4.2 (Transition of the largest eigenvalue and eigenvector, spiked Wigner, linear activation)**.** *Let $\alpha > 0$. We denote $\lambda_1 \geq \lambda_2$ the first and second eigenvalues of $\mathbf{\Gamma}_p^{vv}$.*

- *If $\Delta \geq \Delta_c(\alpha)$, then as $p \to \infty$ we have a.s. $\lambda_1 \to \lambda_{max}$ and $\lambda_2 \to \lambda_{max}$.*
- *If $\Delta \leq \Delta_c(\alpha)$, then as $p \to \infty$ we have a.s. $\lambda_1 \to 1$ and $\lambda_2 \to \lambda_{max}$.*

Further, denoting $\tilde{\mathbf{v}}$ a normalized ($\|\tilde{\mathbf{v}}\|_2^2 = p$) eigenvector of $\mathbf{\Gamma}_p^{vv}$ with eigenvalue λ_1, then $|\tilde{\mathbf{v}}^\intercal \mathbf{v}^\star|^2 / p^2 \to \varepsilon(\Delta)$ a.s., where $\varepsilon(\Delta) = 0$ for all $\Delta \geq \Delta_c(\alpha)$, $\varepsilon(\Delta) > 0$ for all $\Delta < \Delta_c(\alpha)$ and $\lim_{\Delta \to 0} \varepsilon(\Delta) = 1$.

Thm. 9.4.1 and Thm. 9.4.2 are illustrated in Fig. 54. The proof gives the value of $\varepsilon(\Delta)$, which turns out to lead to the same MSE as in Fig. 54 in the linear case. The proofs of theorems 9.4.1 and 9.4.2 are left in (Aubin et al., 2019c), along with the precise arguments used to derive the eigenvalue density, the transition of λ_1 and the computation of $\varepsilon(\Delta)$. These arguments are solely based on random matrix theory. The method of proof of Theorem 9.4.2 is very much inspired by (Benaych-Georges et al., 2011), and allows us to compute numerically the squared correlation $\varepsilon(\Delta)$. Note that while all the calculations are justified, refinements would be needed in order to be completely rigorous. These refinements would follow exactly some proofs of (Silverstein et al., 1995) and (Benaych-Georges et al., 2011), so we will refer to them when necessary. A Mathematica demonstration notebook is provided in the GitHub repository (Aubin et al., 2019a).

In the non-linear case the random matrix analysis is harder to perform. In the matrix $\mathbf{\Gamma}_p^{vv}$, the Wishart matrix $\mathbf{W}\mathbf{W}^\intercal / k$ is replaced by $a\mathbf{I} + b\mathbf{W}\mathbf{W}^\intercal$ with $a, b \geq 0$. It is thus not possible to relate the spectrum of $\mathbf{\Gamma}_p^{vv}$ to the one of a symmetric matrix of the type $\mathbf{W}\mathbf{Z}\mathbf{W}^\intercal$ with $\mathbf{W}$ a gaussian i.i.d matrix. Some techniques from free probability could make the computation nevertheless possible, but we leave this analysis for future work.

9.4.4 APPLYING LAMP TO REAL DATA

As we have already remarked, the LAMP operator in eq. (257) only depend on the generative prior through its covariance. An interesting exercise is to apply LAMP for real data by simply using the empirical covariance for n samples of the spikes, $\mathbf{v}^\alpha$, $\alpha = 1, \dots, n$.

For illustration, we perform the following experiment: the spikes $\mathbf{v}^\star$ are drawn from the standard Fashion-MNIST dataset (Xiao et al., 2017), and are used to generate the spiked matrix $\mathbf{Y}$ according to eq. (209). We then apply our LAMP algorithm to reconstruct the spikes, repeating this experiment for different values of noise Δ. In Fig. 55 we compare the reconstruction by LAMP with standard PCA over $\mathbf{Y}$. In principle, we have no theoretical guarantees about the performance of LAMP, since the Fashion-MNIST images are not drawn from the generative class studied above. Nevertheless, it is striking to observe that LAMP outperforms PCA.

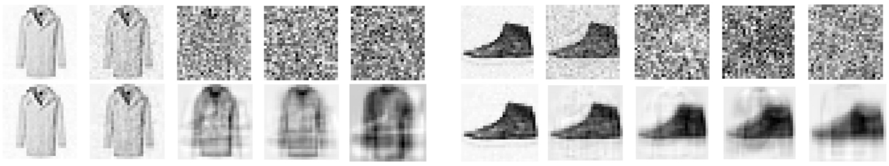


Figure 55: Illustration of canonical PCA (top line) and the LAMP (bottom line) spectral methods (257) on the spiked Wigner model. The covariance $\mathbf{\Sigma}$ is estimated empirically from the FashionMNIST database (Xiao et al., 2017). The estimation of the spike is shown for two images from FashionMNIST, with (from left to right), noise variance $\Delta = 0.01, 0.1, 1, 2, 10$.

A demonstration notebook illustrating this experiment is provided in the GitHub repository (Aubin et al., 2019a).

10

EXACT ASYMPTOTICS FOR PHASE RETRIEVAL AND COMPRESSED SENSING WITH RANDOM GENERATIVE PRIORS

Over the past decade the study of compressed sensing has lead to significant developments in the field of signal processing, with novel sub-Nyquist sampling strategies and a veritable explosion of work in sparse representation. A central observation is that sparsity allows one to measure the signal with fewer observations than its dimension (Donoho, 2006; Candes et al., 2006). The success of neural networks in the recent years suggests another powerful and generic way of representing signals with multi-layer generative priors, such as those used in generative adversarial networks GAN (Goodfellow et al., 2014) and VAE. It is therefore natural to replace sparsity by generative neural network models in compressed sensing and other inverse problems, a strategy that was successfully explored in a number of papers, e.g. (Tramel et al., 2016a; Tramel et al., 2016b; Bora et al., 2017; Manoel et al., 2017; Hand et al., 2018b; Fletcher et al., 2018; Hand et al., 2018a; Mixon et al., 2018; Aubin et al., 2019c). While this direction of research seems to have many promising applications, a systematic theory of what can be efficiently achieved still falls short of the one developed over the past decade for sparse signal processing. Our aim is therefore to dialogue with the broad program of studying how generative models can help solving inverse problems using the toolbox of statistical physics. In this chapter, we build on a line of work allowing for theoretical analysis in the case the measurement and the weight matrices of the prior are random (Manoel et al., 2017; Reeves, 2017; Fletcher et al., 2018; Gabrié et al., 2018; Aubin et al., 2019c) similarly to Chap. 9.

We employ tools originally developed in the context of statistical physics to derive precise asymptotics for the information-theoretically optimal thresholds for signal recovery and for the performance of the best known polynomial algorithm in two such inverse problems: (real-valued) phase retrieval and compressed sensing. These two problems of interest can be framed as a *generalized linear estimation*. Given a set of observations $\mathbf{y} \in \mathbb{R}^n$ generated from a fixed (but unknown) signal $\mathbf{x}^\star \in \mathbb{R}^d$ as

$$\mathbf{y} = \varphi\left(\mathbf{A}\mathbf{x}^\star\right), \tag{262}$$

the goal is to reconstruct $\mathbf{x}^\star$ from the knowledge of $\mathbf{y}$, φ and $\mathbf{A} \in \mathbb{R}^{n\times d}$. CS and Phase Retrieval (PR) are particular instances of this problem, corresponding to $\varphi(x) = x$ and $\varphi(x) = |x|$ respectively. Two key questions in these inverse problems are a) how many observations n are required for theoretically reconstructing the signal $\mathbf{x}^\star$, and b) how this can be done in practice - i. e. . to find an efficient algorithm for reconstruction. Signal structure plays an important role in the answer to both these questions, and have been the subject of intense investigation in the literature. A typical situation is to consider signals admitting a low-dimensional representation, such as sparse signals, for which $k-d$ of the d components of $\mathbf{x}^*$ are exactly zero, see e.g. (Candes et al., 2015; Netrapalli et al., 2013).

In this work, we consider instead structured signals drawn from a generative model $\mathbf{x}^\star = G(\mathbf{z})$, where $\mathbf{z} \in \mathbb{R}^k$ is a low-dimensional latent representation of $\mathbf{x}^\star$. In particular, we will focus in generative multi-layer neural networks, and in order to provide a sharp asymptotic theory, we will restrict the analysis to an ensamble of random networks with known random weights:

$$\mathbf{x}^\star = G(\mathbf{z}) = \sigma^{(L)}\left(\mathbf{W}^{(L)}\sigma^{(L-1)}\left(\mathbf{W}^{(L-1)}\cdots\sigma^{(1)}\left(\mathbf{W}^{(1)}\mathbf{z}\right)\cdots\right)\right), \quad (263)$$

where $\sigma^{(l)} : \mathbb{R} \to \mathbb{R}$, $1 \le l \le L$ are component-wise non-linearities. As aforementioned, we take $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{W}^{(l)} \in \mathbb{R}^{k_l\times k_{l-1}}$ to have i.i.d Gaussian entries with zero means and variances $1/d$ and $1/k_{l-1}$ respectively, and focus on the high-dimensional regime defined by taking $n, d, k_l \to \infty$ while keeping the measurement rate $\alpha = n/d$ and the layer-wise aspect ratios $\beta_l = k_{l+1}/k_l$ constant. We stress that in this regime the depth L is of order one when compared to the width of the generative network, which scales with the input dimension d. With this observation in mind, we adopt the standard terminology in machine learning of denoting networks with $L > 1$ as *deep*. To provide a comparison with previous results for sparse signals, it is useful to define the total compression factor $\rho = k/d$. We note, however, that the comparison between generative and sparse priors herein is not based on a quantitative comparison between the reconstruction estimation errors. Indeed, since data is generated differently in both cases, such a comparison would make little sense. Instead, we compare qualitative properties of the phase diagrams, taking as a surrogate for algorithmic hardness the size of the statistical-to-algorithmic gap in these two different reconstruction problems. Our results hold for latent variables drawn from an arbitrary separable distribution $\mathbf{z} \sim \mathrm{P}_{\mathrm{z}}$, and for arbitrary activations $\sigma^{(l)}$, although for concreteness we present results for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$ and $\sigma^{(l)} \in \{\text{linear}, \text{ReLU}\}$, as it is commonly the case in practice with GAN or VAE.

Previous results on sparsity: Sparsity is probably the most widely studied type of signal structure in linear estimation and phase retrieval. It is thus instructive to recall the main results for sparse signal reconstruction in these inverse problems in the high-dimensional regime with random measurement matrices studied in this manuscript. Optimal statistical and algorithmic thresholds have been established non-rigorously using the replica-method in a series of works (Wu et al., 2012; Krzakala et al., 2012b; Reeves et al., 2012; Zdeborová et al., 2016a). Later the information theoretic results, as well as the corresponding MMSE, has been rigorously proven in (Barbier et al., 2016; Reeves et al., 2016a; Barbier et al., 2019b). So far, the best known polynomial time algorithm in this context is the AMP algorithm, the new avatar of the mean-field approach pioneered in statistical mechanics (Mézard et al., 1987), that has been introduced in (Donoho et al., 2009; Rangan, 2011; Krzakala et al., 2012a; Schniter et al., 2014; Metzler et al., 2017) for these problems, and can be rigorously analyzed (Bayati et al., 2011b). For both (noiseless) compressed sensing and phase retrieval, the information theoretic limit for a perfect signal recovery is given by $\alpha > \alpha_{\text{IT}} = \rho_s$, with ρ_s being the fraction of non-zero components of the signal $\mathbf{x}^{\star}$.

The ability of AMP to exactly reconstruct the signal, however, is different. A non-trivial line $\alpha_{\text{alg}}^{\text{sparse}}(\rho_s) > \alpha_{\text{IT}}$ appears below which AMP fails. No polynomial algorithm achieving better performance for these problems is known. Strikingly, as discussed in (Barbier et al., 2019b), the behaviour of the sparse linear estimation and phase retrieval is drastically different: while $\alpha_{\text{alg}}^{\text{sparse}}(\rho_s)$ is going to zero as $\rho_s \to 0$ for sparse linear estimation hence allowing for compressed sensing, it is not the case for the phase retrieval, for which $\alpha_{\text{alg}}^{\text{sparse}} \to 1/2$ as $\rho_s \to 0$. As a consequence, *no efficient approach to real-valued compressed phase retrieval with small but order one ρ_s in the high-dimensional limit is known.*

Summary of results: In this work, we replace the sparse prior by the multi-layer generative model introduced in eq. (263). Our main contribution is specifying the interplay between the number of measurements needed for exact reconstruction of the signal, parametrized by α, and its latent dimension k. Of particular interest is the comparison between a sparse and separable signal (having a fraction ρ_s of non-zero components) and the structured generative model above, parametrized by $\rho = k/d$. While the number of unknown latent variables is the same in both cases if $\rho = \rho_s$, the upshot is that generative models offer algorithmic advantages over sparsity. More precisely:

1. We analyze the MMSE of the optimal Bayesian estimator for the compressed sensing and phase retrieval problems with generative priors of arbitrary depth, choice of activation and prior distribution for the latent variable. We derive sufficient conditions for the existence of an *undetectable phase* in which better-than-random estimation of $\mathbf{x}^{\star}$ is impossible, and characterize in full generality the threshold α_c beyond which partial signal recovery becomes statistically possible.

2. Fixing our attention on the natural choices of activations $\sigma \in \{$linear, ReLU$\}$, we establish the threshold α_{IT} above which perfect signal reconstruction is theoretically possible. This threshold can be intuitively understood with a simple counting argument.

3. We analyze the performance of the associated AMP algorithm (Manoel et al., 2017), conjectured to be the best known polynomial time algorithm in this setting. This allows us to establish the algorithmic threshold α_{alg} below which no known algorithm is able to perfectly reconstruct $\mathbf{x}^{\star}$.

As expected, the thresholds $\{\alpha_c, \alpha_{\text{IT}}, \alpha_{\text{alg}}\}$ are functions of the compression factor ρ, the number of layers L, the aspect ratios $\{\beta_l\}_{l=1}^{L}$ and the activation functions. In particular, for a fixed architecture we find that the algorithmic gap $\Delta_{\text{alg}} = \alpha_{\text{alg}} - \alpha_{\text{IT}}$ is drastically reduced with the depth L of the generative model, beating the algorithmic hindrance identified in (Barbier et al., 2019b) for compressive phase retrieval with sparse encoding.

10.1 INFORMATION THEORETICAL ANALYSIS

10.1.1 PERFORMANCE OF THE BAYES-OPTIMAL ESTIMATOR

In our analysis we assume that the model generating the observations $\mathbf{y} \in \mathbb{R}^n$ is known. Therefore, the optimal estimator minimizing the mean-squared-error in our setting is given by the Bayesian estimator

$$\hat{\mathbf{x}}^{\text{opt}} = \underset{\hat{\mathbf{x}}}{\text{argmin}} ||\hat{\mathbf{x}} - \mathbf{x}^{\star}||_2^2 = \mathbb{E}_{\text{P}(\mathbf{x}|\mathbf{y})}\left[\mathbf{x}\right] . \tag{264}$$

The posterior distribution of the signal given the observations is in general given by:

$$\text{P}(\mathbf{x}|\mathbf{y}) = \frac{1}{\mathcal{Z}_d(\mathbf{y})} \text{P}_{\text{x}}(\mathbf{x}) \prod_{\mu=1}^{n} \delta\left(y^{\mu} - \varphi\left(\sum_{j=1}^{d} a_j^{\mu} x_j\right)\right), \tag{265}$$

where the normalization $\mathcal{Z}_d(\mathbf{y})$ is known as the *partition function*, and φ is the nonlinearity defining the estimation problem, e.g. $\varphi(x) = |x|$ for phase retrieval and $\varphi(x) = x$ for linear estimation. We note that the presented approach generalizes straightforwardly to account for the presence of noise, but we focus in this work on the analysis of the noiseless case. For the generative model in eq. (263), the prior distribution P_{x} reads

$$\text{P}_{\text{x}}(\mathbf{x}) = \int_{\mathbb{R}^k} \text{d}\mathbf{z}\, \text{P}_{\text{z}}(\mathbf{z}) \prod_{l=1}^{L} \int_{\mathbb{R}^{k_l}} \text{d}\mathbf{h}^{(l)}\, \text{P}_{\text{out}}^{(l)}\left(\mathbf{h}^{(l+1)} \middle| \mathbf{W}^{(l)} \mathbf{h}^{(l)}\right), \tag{266}$$

where for notational convenience we denoted $\mathbf{x} \equiv \mathbf{h}^{(L+1)}$, $\mathbf{z} \equiv \mathbf{h}^{(1)}$ and defined the likelihoods $\text{P}_{\text{out}}^{(l)}$ parametrising the output distribution of each layer given its input. As before, this Bayesian treatment also accounts for stochastic activation functions, even though we focus here on deterministic ones.

Although exact sampling from the posterior is intractable in the high-dimensional regime, it is still possible to track the behavior of the minimum-mean-squared-error estimator as a function of the model parameters. Our main results are based on the line of works comparing, on one hand, the information-theoretically best possible reconstruction, analyzing the ideal Bayesian inference decoder, regardless of the computation cost, and on the other, the best reconstruction using the most efficient known polynomial algorithm - the approximate message passing.

Our analysis builds upon the statistical physics inspired multi-layer formalism introduced in (Manoel et al., 2017), who showed using the cavity and replica methods that the minimum mean-squared-error achieved by the Bayes-optimal estimator defined in eq. (264) can be written, in the limit of $n, d \to \infty$ and $\alpha = n/d = \Theta(1)$ for a generic prior distribution $\mathrm{P_x}$ as

$$\mathrm{mmse}(\alpha) = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}||\hat{\mathbf{x}}^{\mathrm{opt}} - \mathbf{x}^\star||_2^2 = \rho_x - q_x^\star \tag{267}$$

where ρ_x is the second moment of $\mathrm{P_x}$ and the scalar parameter $q_x^\star \in [0, \rho_x]$ is the solution of the following *free energy* extremisation problem

$$\Phi = -\lim_{d\to\infty}\frac{1}{d}\mathbb{E}_{\mathbf{y}} \log \mathcal{Z}_d(\mathbf{y}) = \underset{q_x, \hat{q}_x}{\mathbf{extr}} \left\{ \frac{1}{2}\hat{q}_x q_x - \alpha \Psi_y(q_x) - \Psi_x(\hat{q}_x) \right\}, \tag{268}$$

with the so-called potentials (Ψ_y, Ψ_x) given by

$$\begin{aligned} \Psi_y(t) &= \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \mathcal{Z}_y\left(y; \sqrt{t}\xi, t\right) \log \mathcal{Z}_y\left(y; \sqrt{t}\xi, t\right) \right], \\ \Psi_x(r) &= \lim_{d\to\infty}\frac{1}{d}\; \mathbb{E}_\xi \left[\mathcal{Z}_x(\sqrt{r}\xi, r) \log \mathcal{Z}_x(\sqrt{r}\xi, r) \right], \end{aligned} \tag{269}$$

where $\xi \sim \mathcal{N}(0,1)$ and $\mathcal{Z}_y$, $\mathcal{Z}_x$ are the normalizations of the auxiliary distributions

$$\begin{aligned} \mathrm{Q_y}\left(x; y, \omega, V\right) &= \frac{1}{\mathcal{Z}_y(y;\omega, V)} \frac{e^{-\frac{1}{2V}(x-\omega)^2}}{\sqrt{2\pi V}} \delta\left(y - \varphi(x)\right), \\ \mathrm{Q_x}\left(\mathbf{x}; b, A\right) &= \frac{\mathrm{P_x}(\mathbf{x})}{\mathcal{Z}_x(b, A)} e^{-\frac{A}{2}x_j^2 + b x_j}. \end{aligned} \tag{270}$$

Note that this expression is valid for arbitrary distribution $\mathrm{P_x}$, as long as the limit in Ψ_x is well-defined. In particular, it reduces to the known result in (Krzakala et al., 2012b; Barbier et al., 2019b) when $\mathrm{P_x}$ factorizes. In principle, for correlated P_x such as in the generative model of eq. (266) computing Ψ_x is itself a hard problem. However, we can see eq. (266) as a chain of generalized linear models. In the limit where $k_l \to \infty$ with $\rho = k/d = \Theta(1)$, $L = \Theta(1)$ and $\beta_l = k_{l+1}/k_l = \Theta(1)$ we can apply the observation above iteratively, layer-wise, up to the input layer for which $\mathrm{P_z}$ factorizes - and is easy to compute. This yields (Manoel et al., 2017)

$$\begin{aligned}\Phi = \underset{q_x,\hat{q}_x,\{q_l,\hat{q}_l\}}{\textbf{extr}} \Bigg\{ & -\frac{1}{2}\hat{q}_x q_x - \frac{\rho}{2}\sum_{l=1}^{L}\beta_l q_l \hat{q}_l + \alpha\Psi_y\left(q_x\right) \\ & +\rho\sum_{l=2}^{L}\beta_l \Psi^{(l)}_{\text{out}}\left(\hat{q}_l, q_{l-1}\right) + \Psi^{(L+1)}_{\text{out}}\left(\hat{q}_x, q_L\right) + \rho\Psi_z\left(\hat{q}_z\right)\Bigg\}, \end{aligned} \tag{271}$$

where we have introduced the additional potentials $(\Psi_{\text{out}}, \Psi_z)$

$$\begin{aligned}\Psi^{(l)}_{\text{out}}(r,s) &= \mathbb{E}_{\xi,\eta}\Big[\mathcal{Z}^{(l)}_{\text{out}}(\sqrt{r}\xi, r, \sqrt{s}\xi, \rho_{l-1}-s) \\ &\qquad \log \mathcal{Z}^{(l)}_{\text{out}}(\sqrt{r}\xi, r, \sqrt{s}\xi, \rho_{l-1}-s)\Big], \\ \Psi_z(t) &= \mathbb{E}_{\xi}\left[\mathcal{Z}_z(\sqrt{t}\xi, t)\log \mathcal{Z}_z(\sqrt{t}\xi, t)\right], \end{aligned} \tag{272}$$

defined in terms of the following auxiliary distributions

$$\begin{aligned} \mathrm{Q}^{(l)}_{\text{out}}(x,z;b,A,\omega,V) &= \frac{e^{-\frac{A}{2}x^2+bx}}{\mathcal{Z}_{\text{out}}(b,A,\omega,V)}\frac{e^{-\frac{1}{2V}(z-\omega)^2}}{\sqrt{2\pi V}}\mathrm{P}^{(l)}_{\text{out}}(x|z), \\ Q_{\mathrm{z}}\left(z;b,A\right) &= \frac{e^{-\frac{A}{2}z^2+bz}}{\mathcal{Z}_z(b,A)}\,\mathrm{P_z}(z), \end{aligned} \tag{273}$$

and with ρ_l the second moment of the hidden variable $\mathbf{h}^{(l)}$.

These predictions, that have also been derived with different heuristics in (Reeves, 2017), were rigorously proven for two-layers in (Gabrié et al., 2018), while deeper architectures requires additional assumptions on the concentration of the free energies to be under a rigorous control. Eq. (271) thus reduces the asymptotics of the high-dimensional estimation problem to a low-dimensional extremisation problem over the $2(L+1)$ variables $(q_x, \hat{q}_x, \{q_l, \hat{q}_l\}_{l=1}^{L})$, allowing for a mathematically sound and rigorous investigation. These parameters are also known as the *overlaps*, since they parametrize the overlap between the Bayes-optimal estimator and ground-truth signal at each layer. Solving eq. (268) provides two important statistical thresholds: the *weak recovery* threshold α_c above which better-than-random (i.e. mmse $< \rho_x$) reconstruction becomes theoretically possible and the *perfect reconstruction* threshold, above which perfect signal recovery (i.e. when mmse $= 0$) becomes possible.

Interestingly, the free energy eq. (271) also provides information about the algorithmic hardness of the problem. The above extremisation problem is closely related the state evolution of the AMP algorithm for this problem, as derived in (Manoel et al., 2017), and generalized in (Fletcher et al., 2018). It is conjectured to provide the best polynomial time algorithm for the estimation of $\mathbf{x}^{\star}$ in our considered setting. Specifically, the algorithm reaches a mean-squared error that corresponds to the local extremiser reached by gradient descent in the function (271) starting with uninformative initial conditions.

While so far we summarized results that follow from previous works, these results were up to our knowledge not systematically evaluated and analyzed for the linear estimation and phase retrieval with generative priors. This analysis and its consequences is the object of the rest of this work and constitutes the original contributions of this work.

10.1.2 WEAK RECOVERY THRESHOLD

Solutions for the extremisation in eq. (271) can be found by solving the fixed point equations, obtained by taking the gradient of eq. (271) with respect of the parameters $(q_x, \hat{q}_x, \{q_l, \hat{q}_l\}_{l=1}^L)$:

$$\begin{cases} \hat{q}_x = \alpha \Lambda_y(q_x) \\ \hat{q}_L = \beta_L \Lambda_{\text{out}}(\hat{q}_x, q_L) \\ \hat{q}_{L-1} = \beta_{L-1} \Lambda_{\text{out}}(\hat{q}_L, q_{L-1}) \\ \qquad \vdots \\ \hat{q}_l = \beta_l \Lambda_{\text{out}}(\hat{q}_{l+1}, q_l) \\ \qquad \vdots \\ \hat{q}_z = \beta_1 \Lambda_{\text{out}}(\hat{q}_2, q_z) \end{cases}, \quad \begin{cases} q_x = \Lambda_x(\hat{q}_x, q_L) \\ q_L = \Lambda_x(\hat{q}_L, q_{L-1}) \\ \qquad \vdots \\ q_l = \Lambda_x(\hat{q}_l, q_{l-1}) \\ \qquad \vdots \\ q_z = \Lambda_z(\hat{q}_z) \end{cases}, \tag{274}$$

where $\Lambda_y(t) = 2\,\partial_t \Psi_y(t)$, $\Lambda_z(t) = 2\,\partial_t \Psi_z(t)$, $\Lambda_x(t) = 2\,\partial_r \Psi_{\text{out}}(r,s)$, $\Lambda_{\text{out}}(t) = 2\,\partial_s \Psi_{\text{out}}(r,s)$. The weak recovery threshold α_c is defined as the value above which one can estimate $\mathbf{x}^\star$ better than a random draw from the prior P_x. In terms of the MMSE it is defined as

$$\alpha_c = \underset{\alpha \geq 0}{\text{argmax}}\{\text{mmse}(\alpha) = \rho_x\}. \tag{275}$$

From eq. (267), it is clear that an uninformative solution mmse $= \rho_x$ of eq. (271) corresponds to a fixed point $q_x = 0$. For both the phase retrieval and linear estimation, evaluating the right-hand side of eqs. (274) at $q_x = 0$ we can see that $\hat{q}_x^\star = 0$ is a fixed point if σ is an odd function and if

$$\mathbb{E}_{P_z}[z] = 0, \qquad \text{and} \qquad \mathbb{E}_{Q_{\text{out}}^{(l),0}}[x] = 0, \tag{276}$$

where $Q_{\text{out}}^{(l),0}(x,z) = Q_{\text{out}}^{(l)}(x,z;0,0,0,\rho_{l-1})$. These conditions reflect the intuition that if the prior P_z or the likelihoods $P_{\text{out}}^{(l)}$ are biased towards certain values, this knowledge helps the statistician estimating better than a random guess. If these conditions are satisfied, then α_c can be obtained as the point for which the fixed point $q_x = 0$ becomes unstable. The stability condition is determined by the eigenvalues of the Jacobian of eqs. (274) around the fixed point $(q_x^\star, \hat{q}_x^\star, \{q_l^\star, \hat{q}_l^\star\}_{l=1}^L) = 0$. More precisely, the fixed point becomes unstable as soon as one eigenvalue of the Jacobian is bigger than one. Expanding the update functions around the fixed point and using the conditions in eq. (276),

$$\begin{aligned}
\Lambda_y(t) &\underset{t\ll 1}{=} \frac{1}{\rho_x^2}\int \mathrm{d}y\ \mathscr{Z}_y(y;0,\rho_x)\left(\mathbb{E}_{\mathrm{Q}_y^0}[\rho_x - x^2]\right)^2 t + \Theta\left(t^{3/2}\right),\\
\Lambda_x^{(l)}(r,s) &\underset{r,s\ll 1}{=} \left(\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}^{(l),0}}[x^2]\right)^2 r + \frac{1}{\rho_{l-1}^2}\left(\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}^{(l),0}}[xz]\right)^2 s + \Theta\left(r^{3/2}, s^{3/2}\right),\\
\Lambda_{\mathrm{out}}^{(l)}(r,s) &\underset{r,s\ll 1}{=} \left(\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}^{(l),0}}[xz]\right)^2 r + \frac{1}{\rho_{l-1}^2}\left(\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}^{(l),0}}[z^2] - \rho_{l-1}\right)^2 s\\
&\qquad + \Theta\left(r^{3/2}, s^{3/2}\right),\\
\Lambda_z(t) &\underset{t\ll 1}{=} \left(\mathbb{E}_{\mathrm{P}_z}[z^2]\right)^2 t + \Theta\left(t^{3/2}\right).
\end{aligned} \tag{277}$$

For a generative prior with depth L, the Jacobian is a cumbersome sparse $(L+1)\times(L+1)$ matrix, with all the entries given by the six partial derivatives above. For the sake of conciseness we only write it here for $L=1$:

$$\begin{pmatrix}
0 & \left(\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}^0}\left[x^2\right]\right)^2 & \frac{1}{\rho_z^2}\left(\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}^0}\left[xz\right]\right)^2 & 0\\
\frac{\alpha}{\rho_x^2}\int \mathrm{d}y\ \mathscr{Z}_y^0\left(\mathbb{E}_{\mathrm{Q}_y^0}\left[\rho_x - x^2\right]\right)^2 & 0 & 0 & 0\\
0 & 0 & 0 & \left(\mathbb{E}_{\mathrm{P}_z}\left[z^2\right]\right)^2\\
0 & \beta\left(\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}^0}\left[xz\right]\right)^2 & \frac{\beta}{\rho_z^2}\left(\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}^0}\left[z^2\right] - \rho_z\right)^2 & 0
\end{pmatrix}. \tag{278}$$

Note that this holds for any choice of $\mathrm{P}_{\mathrm{out}}^{(l)}$ and latent space distribution P_z, as long as conditions eq. (276) hold. For the phase retrieval with a linear generative model for instance $\mathrm{P}^{(l)}(x|z) = \delta(x-z)$, we find $\alpha_c = \frac{1}{2}\frac{1}{1+\rho^{-1}}$. For a linear network of depth L this generalizes to

$$\alpha_c = \frac{1}{2}\left(1 + \sum_{l=1}^{L}\prod_{k=0}^{l-1}\beta_{L-k}\right)^{-1}. \tag{279}$$

The linear estimation problem has exactly the same threshold, but without the global $1/2$ factor. Since $\rho, \beta_l \geq 0$, it is clear that α_c is decreasing in the depth L of the network. This analytical formula is verified by numerically solving eqs. (274), see Figs. 58 and 59. For other choices of activation satisfying condition (276) (e.g. the sign function), we always find that depth helps in the weak recovery of the signal.

10.1.3 PERFECT RECOVERY THRESHOLD

We now turn our attention to the perfect recovery threshold, above which perfect signal reconstruction becomes statistically possible. Formally, it can be defined as

$$\alpha_{\mathrm{IT}} = \underset{\alpha \geq 0}{\mathrm{argmin}} \{\mathrm{mmse}(\alpha) = 0\}, \tag{280}$$

and corresponds to the global minimum of the free energy in eq. (271). Numerically, the perfect recovery threshold is found by solving the fixed point equations (274) from an informed initialization $q_x \approx \rho_x$, corresponding to mmse ≈ 0 according to eq. (267). The resulting fixed point is then checked to be a minimizer of the free energy eq. (271). Different from α_c, it cannot be computed analytically for an arbitrary architecture. However, for the compressed sensing and phase retrieval problems with $\sigma \in \{\mathrm{linear}, \mathrm{ReLU}\}$ generative priors, α_{IT} can be analytically computed by generalizing a simple argument based on the invertibility of the linear system of equations at each layer, originally used in the usual compressive sensing (Candes et al., 2006; Tao, 2009).

First, consider the linear estimation problem with a deep linear generative prior, i.e. $\mathbf{y} = \mathbf{A}\mathbf{x}^\star \in \mathbb{R}^n$ with $\mathbf{x}^\star = \mathbf{W}^{(L)} \dots \mathbf{W}^{(1)}\mathbf{z} \in \mathbb{R}^d$ and $\mathbf{A}, \{\mathbf{W}^{(l)}\}_{l=1}^L$ i.i.d Gaussian matrices, that are full rank with high probability. For $n > d$, the system $\mathbf{y} = \mathbf{A}\mathbf{x}^\star$ is overdetermined as there are more equations than unknowns. Hence the information theoretical threshold has to verify $\alpha_{\mathrm{IT}} = \frac{n_{\mathrm{IT}}}{d} \leq 1$. For $L = 0$ (i.e. $\mathbf{x}^\star$ is Gaussian i.i.d), we have exactly $\alpha_{\mathrm{IT}}^{(0)} = 1$ as the prior does not give any additional information for solving the linear system. For $L \geq 1$ though, at each level $l \in [\![L]\!]$, we need to solve successively $\mathbf{h}^{(l)} \in \mathbb{R}^{k_l}$ in the linear system $\mathbf{y} = \mathbf{A}\mathbf{W}^{(L)} \cdots \mathbf{W}^{(l)}\mathbf{h}^{(l)}$. Again as $\mathbf{A}\mathbf{W}^{(L)} \cdots \mathbf{W}^{(l)} \in \mathbb{R}^{n \times k_l}$, if $n > k_l$ the system is over-constrained. Hence the information theoretical threshold for this equation is such that $\forall l \in [\![L]\!], n_{\mathrm{IT}}^{(l)} \leq k_l \Leftrightarrow \alpha_{\mathrm{IT}}^{(l)} \leq \prod_{k=1}^{l} \frac{1}{\beta_{L-k+1}}$. And note that $\rho \equiv \prod_{k=1}^{L} \frac{1}{\beta_{L-k+1}}$. Hence, the information theoretical threshold is obtained by taking the smallest of the above values $\alpha_{\mathrm{IT}}^{(l)}$:

$$\alpha_{\mathrm{IT}} = \min_{l \in [0:L]} \alpha_{\mathrm{IT}}^{(l)} = \min \left(1, \left\{ \prod_{k=1}^{l} \frac{1}{\beta_{L-k+1}} \right\}_{l=1}^{L-1}, \rho \right) . \tag{281}$$

This result generalizes to the real-valued phase retrieval problem. First, we note that by the data processing inequality taking $\mathbf{y} = |\mathbf{A}\mathbf{x}^\star|$ cannot increase the information about $\mathbf{x}^\star$, and therefore the transition in phase retrieval cannot be *better* than for compressed sensing. Secondly, an inefficient algorithm exists that achieve the same performance as compressed sensing for the real valued phase retrieval: one just needs to try all the possible 2^n assignments for the sign, and then solve the corresponding compressed sensing problem. This strategy that will work as soon as the compressed sensing problem is solvable. Eq. (281) is thus valid for the real phase retrieval problem as well.

One can finally generalize this analysis for a non-linear generative prior with ReLU activation at each layer, i.e. $\mathbf{x}^\star = \mathrm{relu}\left(\mathbf{W}^{(L)}\mathrm{relu}\left(\cdots \mathbf{W}^{(1)}\mathbf{z}\right)\cdots\right)$. Noting that on average $\mathbf{x}$ has half of zero entries and half of i.i.d Gaussian entries, the system can be reorganized and simplified $\mathbf{y} = \tilde{\mathbf{A}}\tilde{\mathbf{x}}$, with $\tilde{\mathbf{x}} \in \mathbb{R}^{d/2}$ the extracted vector of $\mathbf{x}$ with on average $d/2$ strictly positive entries and the corresponding reduced matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times d/2}$, is over-constrained for $n > d/2$ and hence the information theoretical threshold verifies $\alpha_{\mathrm{IT}} = \frac{n_{\mathrm{IT}}}{d} \leq \frac{1}{2}$. Noting that this observation remains valid for generative layers, we will have on average at each layer an input vector $\mathbf{h}^{(l)}$ with half of zero entries and half of Gaussian distributed entries - except at the very first layer for which the input $\mathbf{z} \in \mathbb{R}^k$ is dense. Repeating the above arguments yields the following perfect recovery threshold

$$\alpha_{\mathrm{IT}} = \min\left(\frac{1}{2}, \left\{\frac{1}{2}\prod_{k=1}^{l}\frac{1}{\beta_{L-k+1}}\right\}_{l=1}^{L-1}, \rho\right). \tag{282}$$

for both the linear estimation and phase retrieval problems. Both these results are consistent with the solution of the saddle-point eqs. (274) with a informed initialisation, see Figs. 59-61.

10.1.4 ALGORITHMIC THRESHOLD

The discussion so far focused on the statistical limitations for signal recovery, regardless of the cost of the reconstruction procedure. In practice, however, one is concerned with the algorithmic costs for reconstruction. In the high-dimensional regime we are interested, where the number of observations scale with the number of parameters in the model, only (low)-polynomial time algorithms are manageable in practice. Remarkably, the formula in eq. (271) also provides useful information about the algorithmic hindrances for the inverse problems under consideration. Indeed, with a corresponding choice of iteration schedule and initialization, the fixed point equations eq. (271) are identical to the state evolution describing the asymptotic performance of an associated AMP algorithm (Manoel et al., 2017; Fletcher et al., 2018). Moreover, the AMP aforementioned is the *best known* polynomial time algorithm for the estimation problem under consideration, and it is conjectured to be the optimal polynomial algorithm in this setting.

The AMP *state evolution* corresponds to initializing the overlap parameters $(q_x, q_l) \approx 0$ and updating, at each time step t the hat variables $\hat{q}_x^t = \alpha\Lambda_y(q_x^t)$ before the overlaps $q_x^{t+1} = \Lambda_x(\hat{q}_x^t, q_L^t)$, etc. In Fig. 56 we illustrate this equivalence by comparing the MSE obtained by iterating eqs. (274) with the averaged MSE obtained by actually running the AMP algorithm from (Manoel et al., 2017) for a specific architecture and implemented with the tramp python package (Baker et al., 2020). In particular even though the AMP state evolution is not yet rigorously proven, we see a strong agreement of our analytical results with AMP simulations.

Note that, by construction, the performance of the Bayes-optimal estimator corresponds to the global minimum of the scalar potential in eq. (271). If this potential is convex, eqs. (274) will converge to the global minimum, and the asymptotic performance of the associated AMP algorithm will be optimal. However, if the potential has also a local minimum, initializing the fixed point equations will converge to the different minima depending on the initialization. In this case, the MSE associated to the AMP algorithm (corresponding to the local minimum) differs from the Bayes-optimal one (by construction the global minimum). In the later setting, we define the *algorithmic threshold* as the threshold above which AMP is able to perfectly reconstruct the signal - or equivalently for which mmse $= 0$ when eqs. (274) are iterated from $q_x^{t=0} = q_l^{t=0} = \varepsilon \ll 1$. Note that by definition $\alpha_{\rm IT} < \alpha_{\rm alg}$, and we refer to $\Delta_{\rm alg} = \alpha_{\rm alg} - \alpha_{\rm IT}$ as the algorithmic gap. See Fig. 57 for an illustration of the evolution of the free energy landscape for increasing α.

Studying the existence of an algorithmic gap for the linear estimation and phase retrieval problems, and how it depends on the architecture and depth of the generative prior, is the subject of the next section.

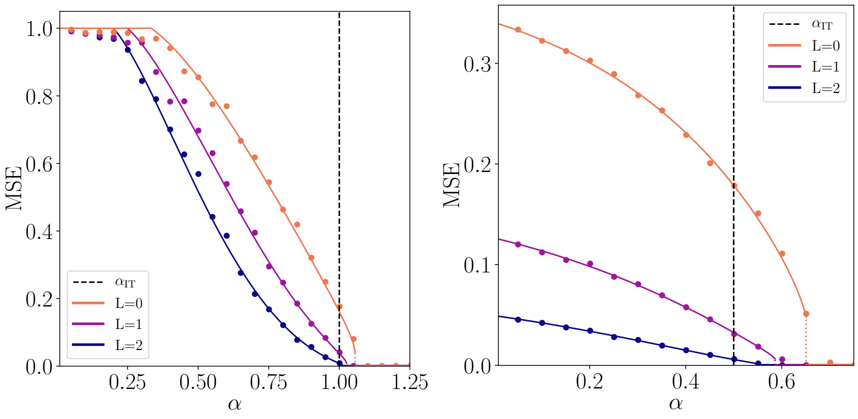


Figure 56: Mean squared error obtained by running the AMP algorithm (dots) from (Manoel et al., 2017) and implemented with the tramp package (Baker et al., 2020), for $d = 2.10^3$ averaged on 10 samples, compared to the MSE obtained from the state evolution eqs. (274) with uninformative initialization $q_x = q_l \approx 0$ (solid line) for the phase retrieval problem with linear (**Left**) and relu (**Right**) generative prior networks. Different curves correspond to different depths L, with fixed $\rho = 2$ and layer-wise aspect ratios $\beta_l = 1$. The dashed vertical line corresponds to $\alpha_{\rm IT}$. To illustrate for instance in the linear case (**Left**), $(\alpha_c^{L=0}, \alpha_c^{L=1}, \alpha_c^{L=2}) = (1/3, 1/4, 1/5)$, $\alpha_{\rm IT} = 1$ and $(\alpha_{\rm alg}^{L=0}, \alpha_{\rm alg}^{L=1}, \alpha_{\rm alg}^{L=2}) = (1.056, 1.026, 1.011)$.

10.2 PHASE DIAGRAMS

In this section we summarize the previous discussions in plots in the (ρ, α)-plane, hereafter named *phase diagrams*. Phase diagrams quantify the quality of signal reconstruction for a fixed architecture $(\beta_1, \ldots, \beta_{L-1})$ [1] as a function of the compression ρ. Moreover, it allows a direct visual comparison between the phase diagram for a sparse Gaussian prior and the multi-layer generative prior. For both the phase retrieval and compressed sensing problems we distinguish the following regions of parameters limited by the thresholds of Sec. 10.1:

- *Undetectable* region where the best achievable error is as bad as a random guess from the prior as if no measurement $\mathbf{y}$ were available. Corresponds to $\alpha < \alpha_c$.
- *Weak recovery* region where the optimal reconstruction error is better than the one of a random guess from the prior, but exact reconstruction cannot be achieved. Corresponds to $\alpha_c < \alpha < \alpha_{\mathrm{IT}}$.
- *Hard* region where exact reconstruction can be achieved information-theoretically, but no efficient algorithm achieving it is known. Corresponds to $\alpha_{\mathrm{IT}} < \alpha < \alpha_{\mathrm{alg}}$
- The so-called *easy* region where the aforementioned AMP algorithm for this problem achieves exact reconstruction of the signal. Corresponds to $\alpha > \alpha_{\mathrm{alg}}$.

As already explained, we locate the corresponding phase transitions in the following manner: for the weak recovery threshold α_c, we notice that the fixed point corresponding to an error as bad as a random guess corresponds to the values of the order parameters $q_x, q_l = 0$. This is an extremiser of the free energy (268) when the prior P_z has zero mean and the non-linearity φ is an even function. This condition is satisfied for both the linear estimation and the phase retrieval problem with linear generative priors that leads to zero-mean distributions on the components of the signal, but is not achieved for a generative prior with ReLU activation, since it biases estimation. In case this uninformative fixed point exists, we investigate its stability under the state evolution of the AMP algorithm, thus defining the threshold α_c. For $\alpha < \alpha_c$ the fixed point is stable, implying the algorithm is not able to find an estimator better than random guess. In contrast, for $\alpha > \alpha_c$ the AMP algorithm provides an estimator better than random guess. For phase retrieval with linear generative model in the setting of the present paper, this analysis leads to the threshold derived in (279). If there exists a region where the performance of the AMP algorithm and the information-theoretic one do not agree we call it the *hard* region. The hard region is delimited by threshold α_{IT} and α_{alg}.

The statistical and algorithmic thresholds defined above admit an alternative and instructive description in terms of free energy landscape, see Fig. 57. Consider a fixed ρ: for small α the free energy (271) has a single global minimum with small overlap (high MSE) with the ground truth solution $\mathbf{x}^{\star}$, referred as the *uninformative* fixed point. At a value α_{sp}, known as the *first*

[1] Note that β_L is fixed from the knowledge of $(\rho, \beta_1, \ldots, \beta_{L-1})$.

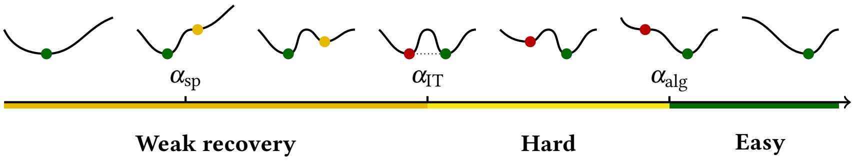


Figure 57: Illustration of the free energy landscape as a function of the overlap with the ground truth solution, when one increases α. For small $\alpha < \alpha_{\rm sp}$, there exists a unique global minimum, whose overlap with the solution is small (high MSE). At $\alpha = \alpha_{\rm sp}$, a *local* minimum (orange dot) with higher overlap (small MSE) appears. By definition, the global minimum corresponds to the MMSE of the problem, which is the MSE attained by the Bayes-optimal estimator (green dot). For $\alpha < \alpha_{\rm IT}$ the accessible solution, i.e the global minimum (green dot) has a high MSE while a better solution exists but has a higher free energy (weak recovery phase). At $\alpha = \alpha_{\rm IT}$ the two minima are global and have the same free energy. Between $\alpha_{\rm IT} < \alpha < \alpha_{\rm alg}$ (hard phase), the local minimum with higher MSE corresponds to the performance of the AMP estimator (red dot). Above $\alpha_{\rm alg}$ only the small MSE minima survive and the AMP estimator is able to achieve the Bayes-optimal performance (easy phase).

spinodal transition, a second local minimum appears with higher overlap (smaller MSE) with the ground truth, referred as *informative* fixed point. The later fixed point becomes a global minimum of the free energy at $\alpha_{\rm IT} > \alpha_{\rm sp}$, while the uninformative fixed point becomes a local minimum. A second spinodal transition occurs at $\alpha_{\rm alg}$ when the informed fixed point becomes unstable. Numerically, the informed and uninformative fixed points can be reached by iterating the saddle-point equations from different initial conditions. When the two are present, the informed fixed point can be reached by iterating from $q_x \approx \rho_x$, which corresponds to a minimum overlap with the ground truth $\mathbf{x}^\star$, and the uninformative fixed point from $q_x \approx 0$, corresponding to no initial overlap with the signal. In the noiseless linear estimation and phase retrieval studied here we observe $\alpha_{\rm IT} = \alpha_{\rm sp}$.

10.2.1 SINGLE-LAYER GENERATIVE PRIOR

First, we consider the case where the signal is generated from a single-layer generative prior, $\mathbf{x}^\star = \sigma(\mathbf{W}\mathbf{z})$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$. We analyze both the compressed sensing and the phase retrieval problem, for $\sigma \in \{\text{linear}, \text{ReLU}\}$. In this case the only free parameters of the model are (ρ, α), and therefore the phase diagram fully characterizes the recovery in these inverse problems. The aim is to compare with the phase diagram of a sparse prior with density $\rho_s = \rho$ of nonzero components.

Fig. 58 depicts the compressed sensing problem with linear (**Left**) and ReLU (**Right**) generative priors. We depict the phase transitions defined above. On the left hand side we compare to the algorithmic phase transition known from (Krzakala et al., 2012a) for sparse separable prior with fraction $1-\rho$ of zero entries and ρ of Gaussian entries of zero mean presenting an algorithmically hard phase for $\rho < \alpha < \alpha_{\rm alg}^{sparse}(\rho)$.

In the case of compressed sensing with linear generative prior we do not observe any hard phase and exact recovery is possible for $\alpha \geq \min(\rho, 1)$ due to invertibility (or the lack of there-of) of the matrix product $\mathbf{AW}$. With ReLU generative prior we have $\alpha_{\text{IT}} = \min(\rho, 1/2)$ and the hard phase exists and has interesting properties: The $\rho \to \infty$ limit corresponds to the separable prior, and thus in this limit $\alpha_{\text{alg}}(\rho \to \infty) = \alpha_{\text{alg}}^{\textit{sparse}}(\rho_s = 1/2)$. Curiously we observe $\alpha_{\text{alg}} > \alpha_{\text{IT}}$ for all $\rho \in (0, \infty)$ except at $\rho = 1/2$. Moreover the size of the hard phase is very small for $\rho < 1/2$ when compared to the one for compressed sensing with separable priors, suggesting that exploring structure in terms of generative models might be algorithmically advantageous over sparsity.

Fig. 59 depicts the phase diagram for the phase retrieval problem with linear (**Left**) and ReLU (**Right**) generative priors. The information-theoretic transition is the same as the one for compressed sensing, while numerical inspection shows that $\alpha_{\text{alg}}^{\text{PR}} > \alpha_{\text{alg}}^{\text{CS}}$ for all $\rho \neq 0, 1/2, 1$. In the left hand side we depict also the algorithmic transition corresponding to the sparse separable prior with non-zero components being Gaussian of zero mean, $\alpha_{\text{alg}}^{\textit{sparse}}(\rho_s)$, as taken from (Barbier et al., 2019b). Crucially, in that case the algorithmic transition to exact recovery does not fall bellow $\alpha = 1/2$ even for very small (yet finite) ρ_s, thus effectively disabling the possibility to sense compressively. In contrast, with both the linear and ReLU generative priors we observe $\alpha_{\text{alg}}(\rho \to 0) \to 0$. More specifically, the theory for the linear prior implies that $\alpha_{\text{alg}}/\rho(\rho \to 0) \to \alpha_{\text{alg}}^{\textit{sparse}}(\rho_s = 1) \approx 1.128$ with the hard phase being largely reduced. Again the hard phase disappears entirely for $\rho = 1$ for the linear model and $\rho = 1/2$ for ReLU.

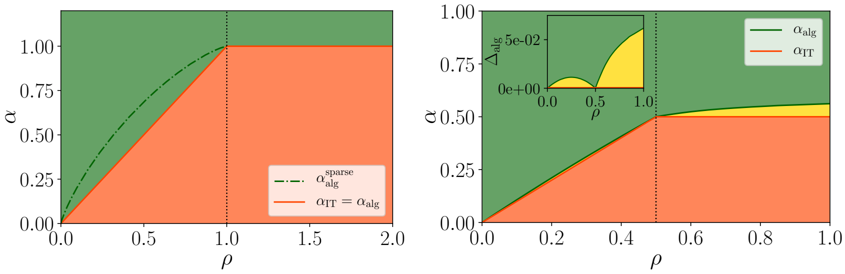


Figure 58: Phase diagrams for the compressed sensing problem with (**Left**) linear generative prior and (**Right**) ReLU generative prior, in the plane (ρ, α). The α_{IT} (red line) represents the information theoretic transition for perfect reconstruction and α_{alg} (green line) the algorithmic transition to perfect reconstruction. In the left part we depict for comparison the algorithmic phase transition for sparse separable prior $\alpha_{\text{alg}}^{\textit{sparse}}$ (dashed-dotted green line). The inset in the right part depicts the difference $\Delta_{\text{alg}} = \alpha_{\text{alg}} - \alpha_{\text{IT}}$. Colored areas correspond respectively to the *weak recovery* (orange), *hard* (yellow) and *easy* (green) phases. The behavior of the free energy landscape for increasing α and fixed ρ is illustrated in Fig. 57.

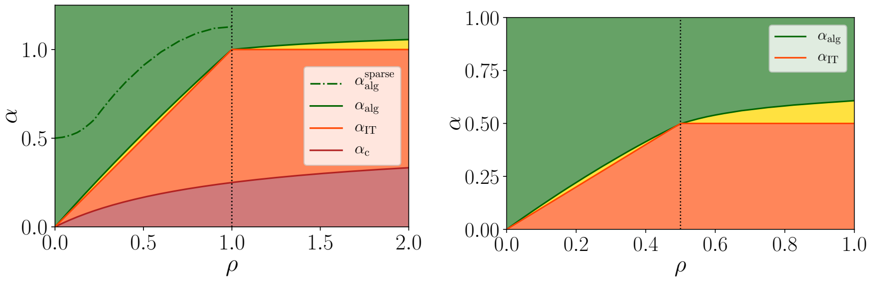


Figure 59: The same as Fig. 58 for the phase retrieval problem with (**Left**) linear generative prior and (**Right**) ReLU generative prior. A major result is that while with sparse separable priors (green dashed-dotted line) compressed phase retrieval is algorithmically hard for $\alpha < 1/2$, with generative priors compressed phase retrieval is tractable down to vanishing α (green line). In the left part we depict additionally the *weak recovery* transition $\alpha_c = \rho/[2(1+\rho)]$ (dark red line). It splits the *no-exact-recovery* phase into the *undetectable* (dark red) and the *weak-recovery* region (orange).

10.2.2 MULTI-LAYER GENERATIVE PRIOR

From the discussion above, we conclude that generative priors are algorithmically advantageous over sparse priors, allowing compressive sensing for the phase retrieval problem. We now investigate how the role of depth of the prior in this discussion. As before, we analyze both the linear estimation and phase retrieval problems, fixing $\sigma^{(l)} \equiv \sigma \in \{\text{linear}, \text{ReLU}\}$ at every layer $1 \leq l \leq L$. Different from the $L = 1$ case discussed above, for $L > 1$ we have other $L - 1$ free parameters characterizing the layer-wise compression factors $(\beta_1, \dots, \beta_{L-1})$.

First, we fix β_l and investigate the role played by depth. Fig. 60 depicts the phase diagrams for compressed sensing (**Left**) and phase retrieval (**Right**) with ReLU activation with varying depth, and a fixed architecture $\beta_l = 3$ for $1 \leq l \leq L$ and note that all these curves share the same $\alpha_{\rm IT} = \min(0.5, \rho)$. It is clear that depth improves even more the small gap already observed for a single-layer generative prior. The algorithmic advantage of multi-layer generative priors in the phase retrieval problem has been previously observed in a similar setting in (Hand et al., 2018a).

Next, we investigate the role played by the layer-wise compression factor β_l. Fig. 61 depicts the phase diagrams for the compressed sensing (**Left**) and phase retrieval (**Right**) with ReLU activation for fixed depth $L = 2$, and varying $\beta \equiv \beta_1$. According to the result in (281), we have $\alpha_{\rm IT} = \min\left(1/2, \rho, 1/2\beta\right)$. It is interesting to note that there is a trade-off between compression $\beta < 2$ and the algorithmic gap, in the following sense. For $\rho < 0.5$ fixed, $\alpha_{\rm IT}$ decreases with decreasing $\beta \ll 1$: compression helps perfect recovery. However, the algorithmic gap $\Delta_{\rm alg}$ becomes wider for fixed $\rho < 0.5$ and decreasing $\beta \ll 1$.

These observations also hold for a linear generative model. In Fig. 62 we have a closer look by plotting the algorithmic gap $\Delta_{\rm alg} \equiv \alpha_{\rm alg} - \alpha_{\rm IT}$ in the phase retrieval problem. On the left, we fix $L = 4$ and plot the gap for increasing values of $\beta \equiv \beta_l$, leading to increasing Δ_{alg}. On the right, we fix $\beta = 2$ and vary the depth, observing a monotonically decreasing Δ_{alg}.

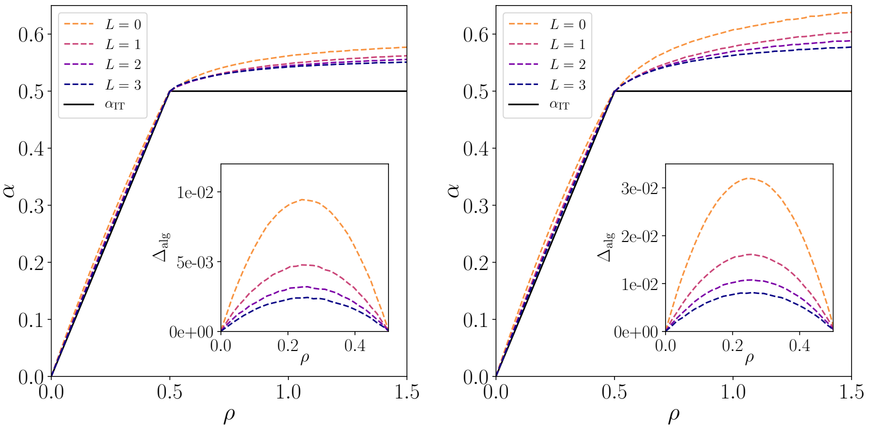


Figure 60: Phase diagrams for the compressed sensing (**Left**) and phase retrieval (**Right**) problems for different depths of the prior, with ReLU activation and fixed layer-wise compression $\beta_l = 3$. Dashed lines represent the algorithmic threshold α_{alg} and solid lines the perfect recovery threshold α_{IT}. We note that the algorithmic gap Δ_{alg} (shown in insets) decreases with the network depth L.

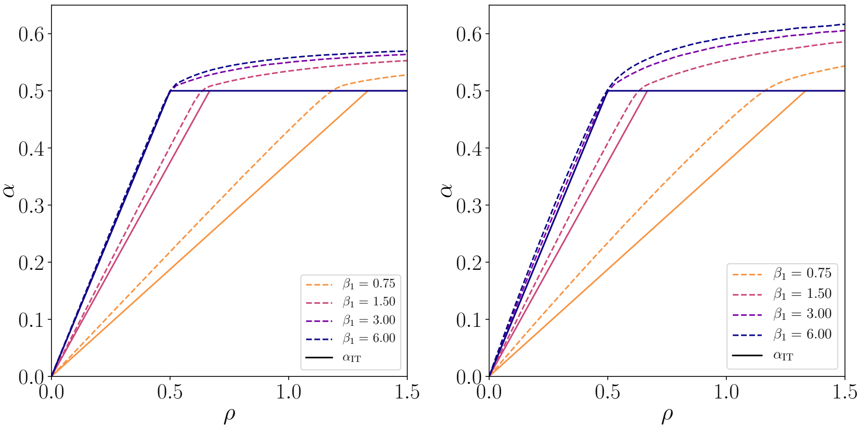


Figure 61: Phase diagrams for the compressed sensing (**Left**) and phase retrieval (**Right**) problems with $L = 2$ and ReLU activation for different values of the layer-wise compression factor β_1. Dashed lines represent the algorithmic threshold α_{alg} and solid lines the perfect recovery threshold α_{IT}. We note that for a given $\rho < 0.5$, α_{IT} is decreasing with $\beta \ll 1$. However, the algorithmic gap Δ_{alg} (shown in the inset) grows for decreasing β. Note that for $\beta_1 \geq 2$ the hard phase is hardly visible at $\rho = 0.5$, even though it disappears only in the large width limit, for both compressed sensing and phase retrieval settings.

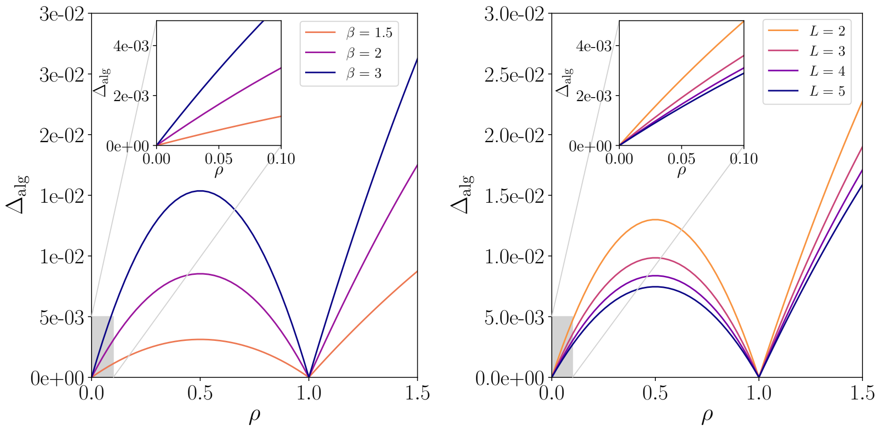


Figure 62: Algorithmic gap $\Delta_{\rm alg} = \alpha_{\rm alg} - \alpha_{\rm IT}$ for small ρ and linear activation, as a function of (**Left**) the compression $\beta \equiv \beta_l$ for fixed depth $L = 4$ and of (**Right**) depth for a fixed compression $\beta = 2$.

CONCLUSION AND PERSPECTIVES

In this chapter, we analyzed how generative priors from an ensemble of random multi-layer neural networks impact signal reconstruction in the high-dimensional limit of two important inverse problems: real-valued phase retrieval and linear estimation. More specifically, we characterized the phase diagrams describing the interplay between number of measurements needed at a given signal compression ρ, for a range of shallow and multi-layer architectures for the generative prior. We observed that although present, the algorithmic gap significantly decreases with depth in the studied architectures. This is particularly striking when compared with sparse priors at $\rho \ll 1$, for which the algorithmic gap is considerably wider. In practice, this means generative models given by random multi-layer neural networks allow for efficient compressive sensing in these problems.

In this work we have only considered independent random weight matrices for both the estimation layer and for the generative model. Ideally, one would like to introduce correlations in a setting closer to reality to show that the smaller computation-to-statistical gap also appears in real-life tasks. The hurdle is that in those cases one does not know what is the theoretically optimal performance nor what are the optimal polynomial algorithms, so that one cannot evaluate the computation-to-statistical empirically in those cases. Yet another tractable case is the study of random rotationally invariant or unitary sensing matrices, as in (Kabashima, 2008; Fletcher et al., 2018; Barbier et al., 2018b; Dudeja et al., 2019). In a different direction, it would be interesting to observe the phenomenology from this work in an experimental setting, for instance using a generative model, such as GAN or VAE, trained on a real dataset to improve the performance of AMP algorithms in a practical task. This is the purpose of the next section.

10.3 ESTIMATION WITH NON I.I.D GENERATIVE PRIORS

Instead reproducing the *plug-in* approach illustrated in the context of the spiked matrix model with generative prior in Chap. 9 Sec. 9.3.1 to derive the AMP algorithm for each structured model, we developed a python package tramp, standing for *TRee Approximate Message Passing*, that automatically build the corresponding AMP algorithm from the sub-models.

Moreover, the package provides an implementation of EP for modular compositional inference in high-dimensional tree-structured models, which is more robust than the classical AMP. In particular, while the classical AMP, discussed in the previous section, is restricted to i.i.d weights, EP implemented in tramp is able to handle non-i.i.d weights such as the ones obtained after training of a GAN or VAE. More details on the implementation can be found in (Baker et al., 2020) and the source code publicly available at *https://github.com/sphinxteam/tramp*.

To illustrate the performances of EP on structured models with correlated weights, we consider a signal $\mathbf{x} \in \mathbb{R}^d$ (with $d = 784$) drawn from the MNIST data set (LeCun et al., 2010). We want to reconstruct the original image from a corrupted observation $\mathbf{y} = \varphi(\mathbf{x}) \in \mathbb{R}^d$, where $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ represents a noisy channel. In the following the noisy channel represents either a Gaussian additive channel or an inpainting channel, that erases some pixels of the input image. In order to reconstruct correctly the MNIST image, we investigate the possibility of using a generative prior such as a VAE along the lines of (Bora et al., 2017; Fletcher et al., 2018). Note that information theoretical and approximate message passing properties of reconstruction of a low rank or GLM channel, using a dense feed-forward neural network generative prior with i.i.d weights has been studied in particular in (Aubin et al., 2019c; Aubin et al., 2020c). However, neither information theoretical or algorithmic perspective was investigated to handle a *trained* generative prior with non-i.i.d weights, such as the ones we consider in this section.

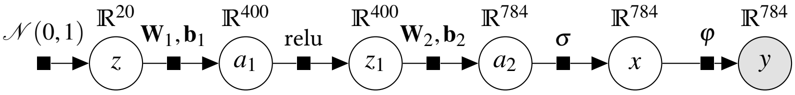


Figure 63: Denoising/inpainting of a MNIST image with a VAE prior. The weights $\mathbf{W}_1, \mathbf{W}_2$ and biases $\mathbf{b}_1, \mathbf{b}_2$ are learned beforehand on the MNIST data set and fixed during the reconstruction.

Following (Fletcher et al., 2018), we use a structured prior coming from a VAE trained itself on the MNIST data set beforehand. The VAE architecture is summarized in Fig. 63 and the training procedure follows closely the canonical one detailed in (Keras-VAE, 2020). We consider two common inference tasks: denoising and inpainting, which are simpler than the one considered in the previous section.

Denoising: In that case, the corrupted channel $\varphi_{\mathrm{den},\Delta}$ adds a Gaussian noise and corresponds to the noisy component-wise channel

$$\varphi_{\mathrm{den},\Delta}(\mathbf{x}) = \mathbf{x} + \boldsymbol{\xi} \text{ with } \xi_i \sim \mathcal{N}(0,\Delta)\,.$$

Inpaiting: The corrupted channel erases a few pixels of the input image and corresponds formally to

$$\varphi_{\mathrm{inp},\mathrm{I}_\alpha}(\mathbf{x}) = \mathbf{x} - m(\mathbf{x})\,,$$

where m represents a mask applied component-wise. Let $\alpha \in [0;1]$, I_α denotes the set of erased indexes of size $\lfloor \alpha d \rfloor$ and the masks acts according to $m(x_i) = \mathbb{1}\left[x_i \in \mathrm{I}_\alpha\right]$. As an illustration, we consider two different manner of generating the erased interval I_α:

1. A central horizontal band of width $\lfloor \alpha d \rfloor$: $\mathrm{I}_\alpha^{\mathrm{band}} = [\lfloor \frac{d}{2}(1-\alpha)\rfloor ; \lfloor \frac{d}{2}(1+\alpha)\rfloor]$
2. $\lfloor \alpha d \rfloor$ indices drawn uniformly at random: $\mathrm{I}_\alpha^{\mathrm{uni}} \sim \mathcal{U}\left([1,d];\lfloor \alpha d \rfloor\right)$

Solving these inference tasks in tramp is straightforward: first declare the structured model Fig. 63 and then run EP. A few MNIST samples $\mathbf{x}^* \in \mathbb{R}^{784}$ in the test set, which were not used to train the VAE, compared to the noisy observations $\mathbf{y} \in \mathbb{R}^{784}$ and tramp reconstructions $\hat{\mathbf{x}}$ are presented in Fig. 64. It suggests that the EP implementation of tramp is able to use the trained VAE prior information to either denoise very noisy observations or reconstruct missing pixels of the MNIST images.

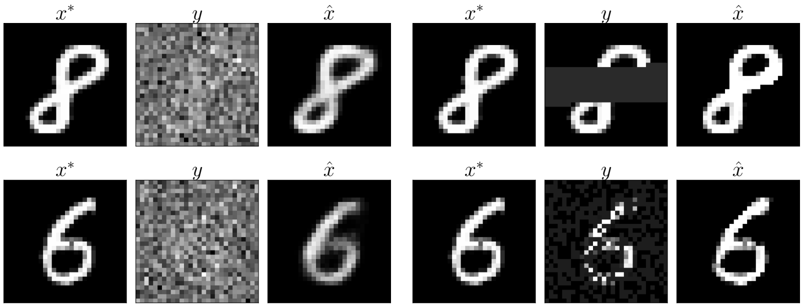


Figure 64: Illustration of the tramp prediction $\hat{x}$ using a VAE prior from observation $y = \varphi(x^*)$ with x^* a MNIST sample. **(Left)** Denoising $\varphi = \varphi_{\mathrm{den},\Delta}$ with $\Delta = 4$. **(Right-upper)** Band-inpainting $\varphi_{\mathrm{inp},I_\alpha^{\mathrm{band}}}$ with $\alpha = 0.3$ **(Right-lower)** Uniform-inpainting $\varphi_{\mathrm{inp},I_\alpha^{\mathrm{uni}}}$ with $\alpha = 0.5$.

However, analyzing the SE of EP or VAMP for such complex prior distribution encoded in the correlated weights of the VAE is still an ongoing line of research. As a conclusion, this direction shall be pushed further to provide a more accurate theoretical comparison between generative priors and separable sparse priors, and finally conclude on their respective performances on real data.

Part III

APPENDICES

A

DEFINITIONS AND MATHEMATICAL IDENTITIES

A.1 GAUSSIAN DISTRIBUTION AND MULTIVARIATE CENTRAL LIMIT THEOREM

Consider $\mathbf{m} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ a symmetric positive definite matrix. For $\mathbf{x} \in \mathbb{R}^d$, the Gaussian probability distribution is defined by

$$\mathcal{N}_{\mathbf{x}}\left(\mathbf{m}, \Sigma\right) \equiv e^{-\frac{(\mathbf{x}-\mathbf{m})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})}{2}} / \sqrt{\det\left(2\pi\boldsymbol{\Sigma}\right)}\,. \tag{283}$$

The Gaussian vector $\mathbf{x}$ has mean $\mathbb{E}\left[\mathbf{x}\right] = \mathbf{m}$ and variance $\mathrm{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$. The Gaussian distribution is crucial as it turns out to be the fixed point distribution of the sum of i.i.d random variables as stated by the CLT:

Proposition A.1.1 (Multivariate Central Limit Theorem). *Let $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ a sequence of i.i.d random vectors in $\mathbb{R}^d$ such that $\mathbb{E}\left[\mathbf{x}\right] = \mathbf{m}$ and covariance matrix $\boldsymbol{\Sigma}$. Defining $\mathbf{s}_n = \frac{1}{n}\sum_{\mu=1}^n \mathbf{x}_\mu$, then as n approaches infinity, the sum converges in distribution to a Gaussian law:*

$$\sqrt{n}\left(\mathbf{s}_n - \mathbf{m}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}_{\mathbf{x}}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)\,. \tag{284}$$

A.2 HUBBARD-STRATONOVICH TRANSFORMATION

The Hubbard-Stratonovich transformation is a simple Gaussian identity based on the fact that:

Proposition A.2.1 (Hubbard-Stratonovich transformation). *For $\boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right)$ and a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, for all $\mathbf{x} \in \mathbb{R}^d$*

$$\mathbb{E}_{\boldsymbol{\xi}} \exp\left(\boldsymbol{\xi}^\intercal \mathbf{A}^{1/2}\mathbf{x}\right) = (2\pi)^{-d/2}\int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\xi}\, e^{-\frac{1}{2}\mathbf{x}^\intercal\mathbf{x} + \boldsymbol{\xi}^\intercal\mathbf{A}^{1/2}\mathbf{x}} = e^{\frac{1}{2}\mathbf{x}^\intercal\mathbf{A}\mathbf{x}}\,. \tag{285}$$

A.3 NISHIMORI IDENTITY

We recall the Nishimori identity from (Nishimori, 1980; Nishimori, 1981; Nishimori, 2001; Zdeborová et al., 2016a; Lesieur et al., 2017a):

Proposition A.3.1 (Nishimori identity)**.** *Let* (X,Y) *a couple of random variables. Let* $\{\mathbf{x}_\mu\}_{\mu=1}^n$ $n \geq 1$ *samples drawn* i.i.d *from* $\mathrm{P}(\mathrm{X}|\mathrm{Y})$*. Let us denote* $\langle . \rangle$ *the expectation over the posterior distribution* $\mathrm{P}(\mathrm{X}|\mathrm{Y})$ *and* $\mathbb{E}$ *the expectation with respect to* (X,Y)*. For all continuous bounded function* f*:*

$$\mathbb{E}\left[\langle f(\mathrm{Y},\mathbf{x}_1,\cdots,\mathbf{x}_{n-1},\mathbf{x}_n\rangle\right] = \mathbb{E}\left[\langle f(\mathrm{Y},\mathbf{x}_1,\cdots,\mathbf{x}_{n-1},\mathrm{X}\rangle\right] .$$

Proof. This is a simple consequence of the Bayes formula. It is equivalent to sample the couple (X,Y) according to its joint distribution $\mathrm{P}(\mathrm{X},\mathrm{Y})$ or to sample first Y according to its marginal distribution $\mathrm{P}(\mathrm{Y})$ and then to sample X conditionally to Y from its conditional distribution $\mathrm{P}(\mathrm{X}|\mathrm{Y})$. Thus the $(n+1)$-tuple $(\mathrm{Y},\mathbf{x}_1,\cdots,\mathbf{x}_{n-1},\mathbf{x}_n)$ is equal in law to $(\mathrm{Y},\mathbf{x}_1,\cdots,\mathbf{x}_{n-1},\mathrm{X})$. □

A.4 DENOISING DISTRIBUTIONS, UPDATES AND FREE ENTROPY TERMS

In this section, we introduce the K-dimensional probability distributions involved in the replica free entropies and from which the AMP update equations are derived in the context of *committee machines*. The multivariate formulation can be simplify to scalar expressions for $K=1$ in the case of GLM.

A.4.1 MMSE ESTIMATION WITH COMMITTEE MACHINES

Analyzing the joint distribution $\mathrm{P}(\mathbf{y},\mathbf{X})$ for MMSE estimation in the high-dimensional regime boils down to introducing the denoising distributions $\mathrm{Q}_{\mathrm{w}},\mathrm{Q}_{\mathrm{out}}$ on $\mathbf{w}\in\mathbb{R}^K$ and $\mathbf{z}\in\mathbb{R}^K$ and their respective normalizations $\mathcal{Z}_{\mathrm{w}}$, $\mathcal{Z}_{\mathrm{out}}$ in Sec. A.4.1.a. We define as well the denoising functions $\mathbf{f}_{\mathrm{w}}, \partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}$, $\mathbf{f}_{\mathrm{out}}, \partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}$ in Sec. A.4.1.b, that play a central role in Bayesian inference. Note in particular that they correspond to the *updates* of the GAMP algorithm in (Rangan, 2011) that we recall in Algo. 5 for the committee machine hypothesis class. They are simply defined as the derivatives of $\log\mathcal{Z}_{\mathrm{w}}$ and $\log\mathcal{Z}_{\mathrm{out}}$. Finally the free entropy can be expressed as a function of simple free entropy terms $\Psi_{\mathrm{w}},\Psi_{\mathrm{out}}$ defined in Sec. A.4.1.c.

Consider $y\in\mathbb{R}$, $\boldsymbol{\gamma},\boldsymbol{\omega}\in\mathbb{R}^K$, $\boldsymbol{\Lambda},\mathbf{V}\in\mathcal{S}_K^+$, the ensemble of symmetric positive matrices of size $K\times K$, and vectors to infer $\mathbf{w},\mathbf{z}\in\mathbb{R}^K$, with prior distributions P_{w}, $\mathrm{P}_{\mathrm{out}}$.

A.4.1.A DENOISING DISTRIBUTIONS

$$\begin{aligned}
\mathrm{Q}_{\mathrm{w}}(\mathbf{w};\boldsymbol{\gamma},\boldsymbol{\Lambda}) &\equiv \frac{\mathrm{P}_{\mathrm{w}}(\mathbf{w})}{\mathcal{Z}_{\mathrm{w}}(\boldsymbol{\gamma},\boldsymbol{\Lambda})} e^{-\frac{1}{2}\mathbf{w}^{\intercal}\boldsymbol{\Lambda}\mathbf{w}+\boldsymbol{\gamma}^{\intercal}\mathbf{w}}, && (286)\\
\mathcal{Z}_{\mathrm{w}}(\boldsymbol{\gamma},\boldsymbol{\Lambda}) &\equiv \mathbb{E}_{\mathbf{w}\sim\mathrm{P}_{\mathrm{w}}}\left[e^{-\frac{1}{2}\mathbf{w}^{\intercal}\boldsymbol{\Lambda}\mathbf{w}+\boldsymbol{\gamma}^{\intercal}\mathbf{w}}\right] && (287)\\
&= \int_{\mathbb{R}^K} \mathrm{d}\mathbf{w}\ \mathrm{p}_{\mathrm{w}}(\mathbf{w}) e^{-\frac{1}{2}\mathbf{w}^{\intercal}\boldsymbol{\Lambda}\mathbf{w}+\boldsymbol{\gamma}^{\intercal}\mathbf{w}}, \\
\mathrm{Q}_{\mathrm{out}}(\mathbf{z};y,\boldsymbol{\omega},\mathbf{V}) &\equiv \frac{\mathrm{P}_{\mathrm{out}}(y|\mathbf{z})}{\mathcal{Z}_{\mathrm{out}}(y,\boldsymbol{\omega},\mathbf{V})} \frac{e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^{\intercal}\mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}}{\sqrt{\det(2\pi\mathbf{V})}}, && (288)\\
\mathcal{Z}_{\mathrm{out}}(y,\boldsymbol{\omega},\mathbf{V}) &\equiv \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_K)}\left[\mathrm{P}_{\mathrm{out}}\left(y|\mathbf{V}^{1/2}\mathbf{z}+\boldsymbol{\omega}\right)\right] && (289)\\
&= \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}\ \mathrm{p}_{\mathrm{out}}(y|\mathbf{z}) \frac{e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^{\intercal}\mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}}{\sqrt{\det(2\pi\mathbf{V})}}.
\end{aligned}$$

A.4.1.B DENOISING UPDATES

$$\begin{aligned}
\mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma},\boldsymbol{\Lambda}) &\equiv \partial_{\boldsymbol{\gamma}}\log\left(\mathcal{Z}_{\mathrm{w}}(\boldsymbol{\gamma},\boldsymbol{\Lambda})\right) = \mathbb{E}_{\mathrm{Q}_{\mathrm{w}}}[\mathbf{w}], && (290)\\
\partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma},\boldsymbol{\Lambda}) &\equiv \mathbb{E}_{\mathrm{Q}_{\mathrm{w}}}[\mathbf{w}\mathbf{w}^{\intercal}] - \mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma},\boldsymbol{\Lambda})^{\otimes 2}, && (291)\\
\mathbf{f}_{\mathrm{out}}(y,\boldsymbol{\omega},\mathbf{V}) &\equiv \partial_{\boldsymbol{\omega}}\log\left(\mathcal{Z}_{\mathrm{out}}(y,\boldsymbol{\omega},\mathbf{V})\right) = \mathbf{V}^{-1}\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}}[\mathbf{z}-\boldsymbol{\omega}], && (292)\\
\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}(y,\boldsymbol{\omega},\mathbf{V}) &\equiv \frac{\partial\mathbf{f}_{\mathrm{out}}(y,\boldsymbol{\omega},\mathbf{V})}{\partial\boldsymbol{\omega}} && (293)\\
&= \mathbf{V}^{-1}\mathbb{E}_{\mathrm{Q}_{\mathrm{out}}}\left[(\mathbf{z}-\boldsymbol{\omega})^{\otimes 2}\right]\mathbf{V}^{-1} - \mathbf{V}^{-1} - \mathbf{f}_{\mathrm{out}}(y,\boldsymbol{\omega},\mathbf{V})^{\otimes 2}.
\end{aligned}$$

A.4.1.C FREE ENTROPY TERMS

For overlap matrices $\mathbf{Q}^{\star},\mathbf{Q}\in\mathbb{R}^{K\times K}$, and second moments $\boldsymbol{\rho}^{\star},\boldsymbol{\rho}\in\mathbb{R}^{K\times K}$,

$$\Psi_{\mathrm{w}}(\mathbf{Q}^{\star},\mathbf{Q}) \equiv \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}_{\mathrm{w}}\left((\mathbf{Q}^{\star})^{1/2}\boldsymbol{\xi},\mathbf{Q}^{\star}\right)\log\left(\mathcal{Z}_{\mathrm{w}}\left(\mathbf{Q}^{1/2}\boldsymbol{\xi},\mathbf{Q}\right)\right)\right], \tag{294}$$

$$\begin{aligned}
\Psi_{\mathrm{out}}(\mathbf{Q}^{\star},\mathbf{Q},\boldsymbol{\rho}^{\star},\boldsymbol{\rho}) \equiv \mathbb{E}_{\boldsymbol{\xi}}\Big[&\mathcal{Z}_{\mathrm{out}}\left((\mathbf{Q}^{\star})^{1/2}\boldsymbol{\xi},\boldsymbol{\rho}^{\star}-\mathbf{Q}^{\star}\right)\\
&\times\log\left(\mathcal{Z}_{\mathrm{out}}\left(\mathbf{Q}^{1/2}\boldsymbol{\xi},\boldsymbol{\rho}-\mathbf{Q}\right)\right)\Big] \qquad (295)
\end{aligned}$$

A.4.2 MAP ESTIMATION WITH GLM

Before defining similar denoising functions to analyze MAP estimation, we first recall the definition of the Moreau-Yosida regularization in the scalar case $K=1$.

A.4.2.a Moreau-Yosida regularization and proximal

Let $\Sigma > 0$, $f(,z)$ a convex function in $z \in \mathbb{R}$, the Moreau-Yosida regularization $\mathscr{M}_\Sigma$ and the proximal map $\mathscr{P}_\Sigma$ are defined by

$$\mathscr{P}_\Sigma[f(,.)](x) = \text{argmin}_z \left[f(,z) + \frac{1}{2\Sigma}(z-x)^2 \right] , \tag{296}$$

$$\mathscr{M}_\Sigma[f(,.)](x) = \min_z \left[f(,z) + \frac{1}{2\Sigma}(z-x)^2 \right] , \tag{297}$$

where $(,.)$ denotes all the arguments of the function f.

A.4.2.b MAP denoising functions

The MAP denoising functions for any convex loss $l(,.)$ and convex separable regularizer $r(.)$ can be written in terms of the Moreau-Yosida regularization or the proximal map as follows

$$\begin{aligned} f_{\mathrm{w}}^{\mathrm{map},r}(\gamma,\Lambda) &\equiv \mathscr{P}_{\Lambda^{-1}}\left[r(.)\right]\left(\Lambda^{-1}\gamma\right) \qquad (298) \\ &= \Lambda^{-1}\gamma - \Lambda^{-1}\partial_{\Lambda^{-1}\gamma}\mathscr{M}_{\Lambda^{-1}}\left[r(.)\right]\left(\Lambda^{-1}\gamma\right), \\ f_{\mathrm{out}}^{\mathrm{map},l}(y,\omega,V) &\equiv -\partial_\omega \mathscr{M}_V[l(y,.)](\omega) \qquad (299) \\ &= V^{-1}\left(\mathscr{P}_V[l(y,.)](\omega) - \omega\right) . \end{aligned}$$

The derivation and the applications are detailed in Appendix. I.3 of (Aubin et al., 2020a).

B

REPLICA COMPUTATIONS

B.1 TEACHER-STUDENT - COMMITTEE MACHINE WITH I.I.D DATA

In this section, we present the heuristic derivation of the replica formula of Theorem 5.2.1 using the replica method, presented in Sec. 4.1, in the context of of the *committee machine*. This computation is necessary to properly guess the formula that we then prove using the adaptive interpolation method. The reader interested in the replica approach to neural networks and the committee machine is invited to look as well to some of the classical papers (Gardner et al., 1988; Mézard, 1989; Schwarze et al., 1992; Schwarze et al., 1993; Schwarze, 1993; Monasson et al., 1995a). In the *teacher-student* setting, the committee machine estimation problem consists of trying to estimate a *teacher* signal $\mathbf{W}^\star \in \mathbb{R}^{d\times K}$ from a set of n input-output observations $\{\mathbf{X},\mathbf{y}\} \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ generated according to

$$\mathbf{y} = \varphi_{\text{out}}\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}^\star\right) = \varphi_{\text{out}}\left(\left\{\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{w}_k^\star\right\}_{k=1}^K\right).$$

The student, within the same committee machine hypothesis class, with parameters $\mathbf{W} \in \mathbb{R}^{d\times K}$, tries to learn the teacher rule generated by the ground truth weights $\mathbf{W}^\star$. Committee machines are a simple vectorized generalization of GLM, defined in Sec. 1.2.9.a, whose estimation is performed simultaneously with $K \geq 1$ GLM. Therefore, the replica computation is shown only in the general committee machine case and final expressions for GLM will be derived as a particular case for $K = 1$.

We will assume that the matrix of data inputs $\mathbf{X} \in \mathbb{R}^{n\times d}$ is drawn i.i.d with density p_{x}. We will consider them to be i.i.d Gaussian with zero mean and unit variance: $\forall \mu \in [\![n]\!], \mathbf{x}_\mu \sim \mathcal{N}_{\mathbf{x}}(\mathbf{0},\mathbf{I}_d)$. The function $\varphi_{\text{out}} : \mathbb{R}^K \mapsto \mathbb{R}$ represents a deterministic, or stochastic function associated to a probability distribution P_{out}, applied component-wise to each sample. Notice that the factor $\frac{1}{\sqrt{d}}$ is present to insure that the variance of the input data is normalized to the unit.

B.1.1 REPLICA CALCULATION

B.1.1.A ON STATISTICAL ESTIMATION

Both MMSE and MAP estimations boil down to the analysis of the posterior distribution $\mathrm{P}(\mathbf{W}|\mathbf{y},\mathbf{X})$ expressed by the Bayes rule

$$\mathrm{P}(\mathbf{W}|\mathbf{y},\mathbf{X}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{W},\mathbf{X})\,\mathbb{P}(\mathbf{W})}{\mathrm{P}(\mathbf{y},\mathbf{X})} = \frac{\mathrm{P}_{\text{out}}(\mathbf{y}|\mathbf{W},\mathbf{X})\,\mathrm{P}_{\text{w}}(\mathbf{W})}{\mathcal{Z}_d(\{\mathbf{y},\mathbf{X}\})}\,. \tag{300}$$

The joint distribution is also called the *partition function* $\mathrm{P}(\mathbf{y},\mathbf{X}) \equiv \mathcal{Z}_d(\{\mathbf{y},\mathbf{X}\})$. To connect with the statistical physics formalism, we introduce the corresponding Hamiltonian, for separable distributions $\mathrm{P}_{\text{out}}, \mathrm{P}_{\text{w}}$ along one dimension, by

$$\begin{aligned}\mathcal{H}_d(\mathbf{W},\{\mathbf{y},\mathbf{X}\}) &= -\log \mathrm{P}_{\text{out}}(\mathbf{y}|\mathbf{W},\mathbf{X}) - \log \mathrm{P}_{\text{w}}(\mathbf{W})\,,\\ &= -\sum_{\mu=1}^{n} \log \mathrm{P}_{\text{out}}(y_\mu|\mathbf{W},\mathbf{x}_\mu) - \sum_{i=1}^{d} \mathrm{P}_{\text{w}}(\mathbf{w}_i)\,.\end{aligned}$$

The spin variables classically denoted $\boldsymbol{\sigma}$ are replaced by the weights of the model $\mathbf{W} \in \mathbb{R}^{d\times K}$ and they interact through the random dataset $\{\mathbf{y},\mathbf{X}\}$ that plays the role of the quenched exchange interactions $\mathbf{J}$. However here, the interactions are not *pairwise*, as it is often the case in the Ising-like models, but instead *fully connected*, meaning that each variable $\mathbf{w}_i \in \mathbb{R}^K$ is connected to every other spin $\{\mathbf{w}_j\}_{j\in\partial j\setminus i}$ as represented in the factor graph in Fig. 65. The partition function at inverse temperature β is therefore defined by

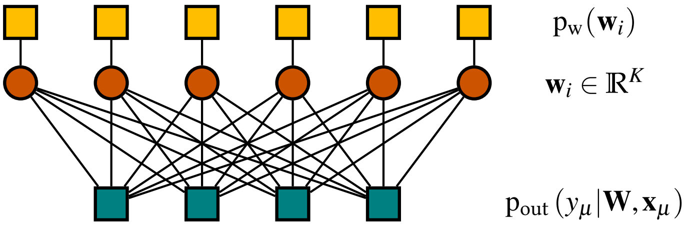

$\mathrm{p}_{\text{w}}(\mathbf{w}_i)$
$\mathbf{w}_i \in \mathbb{R}^K$
$\mathrm{p}_{\text{out}}(y_\mu|\mathbf{W},\mathbf{x}_\mu)$

Figure 65: Factor graph corresponding to the committee machines hypothesis class. The vectorial variables to infer $\mathbf{w}_i$ are fully connected through the quenched disorder $\mathbf{y} \sim \mathrm{P}_{\text{out}^\star}(.)$ and each variable follow a one-body interaction with a separable prior distribution $\mathrm{p}_{\text{w}}(\mathbf{w}_i)$.

$$\begin{aligned}\mathcal{Z}_d(\{\mathbf{y},\mathbf{X}\};\beta) \equiv \mathrm{P}(\mathbf{y},\mathbf{X}) &= \int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{W}\; e^{-\beta\mathcal{H}_d(\mathbf{W},\{\mathbf{y},\mathbf{X}\})}\\ &= \int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{w}\; e^{\beta(\log \mathrm{p}_{\text{out}}(\mathbf{y}|\mathbf{W},\mathbf{X}) + \log P_{\text{w}}(\mathbf{W}))}\\ &= \int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{w}\; \mathrm{p}_{\text{out}}(\mathbf{y}|\mathbf{W},\mathbf{X})\,\mathrm{p}_{\text{w}}(\mathbf{W})\,,\end{aligned} \tag{301}$$

and can be exactly mapped to Bayesian estimation for $\beta = 1$. In the context of ERM, MAP estimation can be analyzed by taking the limit $\beta \to \infty$ as detailed in Chap. 8. In the considered modern high-dimensional regime with $d \to \infty$,

$n \to \infty$, $\alpha = n/d = \Theta(1)$ and $K = \Theta(1)$, we are interested in computing the *free entropy* Φ (42), *averaged* over the input data $\mathbf{X}$ and teacher weights $\mathbf{W}^\star$, or equivalently over the output labels $\mathbf{y}$ generated from it, defined as

$$\Phi(\alpha) \equiv \lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\log \mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)\right] . \tag{302}$$

The heuristic replica method described in Sec. 4.1 allows to compute the above average over the random dataset $\{\mathbf{y},\mathbf{X}\}$, that plays the role of the quenched disorder in usual spin glasses. We show the computation for the more involved committee machine model class and generalization of the GLM class, only for i.i.d data. The cumbersome computation for non i.i.d data can be performed as well and lead to more complex expressions and has been performed in particular in (Kabashima, 2008) in the case of the GLM.

B.1.1.B REPLICA COMPUTATION

We present here the replica computation of the averaged free entropy $\Phi(\alpha)$ in eq. (302) for arbitrary *student* prior and channel distributions $\mathrm{P_w}, \mathrm{P_{w^\star}}$ and $\mathrm{P_{out}}, \mathrm{P_{out^\star}}$, so that the computation remains valid for both the Bayes-optimal and mismatched settings. The average in eq. (302) is intractable in general, and the computation relies on the so called *replica trick*, see Sec. 4.1.1, that consists in applying the identity

$$\mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\lim_{d\to\infty} \frac{1}{d} \log \mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)\right] = \lim_{r\to 0} \left[\lim_{d\to\infty} \frac{1}{d} \frac{\partial \log \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r\right]}{\partial r}\right] . \tag{303}$$

The replica trick has been used in a series of previous works to compute the free energy density of GLM for separable distributions (Krzakala et al., 2012b) and has been rigorous proved in this case by (Barbier et al., 2019b). Eq. (303) is interesting in the sense that it reduces the intractable average to the computation of the moments of the averaged partition function, which are easier quantities to compute. Note that for $r \in \mathbb{N}$, $\mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r = \prod_{a=1}^r \mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)$ represents the partition function of r identical non-interacting copies of the initial system, called *replicas.* Taking the quenched average over the disorder will then correlate the replicas, before taking the number of replicas $r \to 0$. Therefore, we assume there exists an analytical continuation so that $r \in \mathbb{R}$ and the limit is well defined. Finally, notice that we exchanged the order of the limits $r \to 0$ and $d \to \infty$. These technicalities are crucial points but are not rigorously justified and we will ignore them in the rest of the computation. First, in order to decouple the contributions of the channel $\mathrm{P_{out}}$ and the prior $\mathrm{P_w}$, we introduce the variable $\mathbf{Z} = \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}$ and a Dirac-delta integral:

$$\mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right) = \int_{\mathbb{R}^{n\times K}} \mathrm{d}\mathbf{z}\ \mathrm{p_{out}}\left(\mathbf{y}|\mathbf{Z}\right) \int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{w}\ \mathrm{p_w}\left(\mathbf{W}\right) \delta\left(\mathbf{Z} - \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}\right) .$$

Thus the replicated partition function for an integer $r \in \mathbb{N}$ in eq. (303) can be written as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r\right] \\
&= \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\prod_{a=1}^{r}\int_{\mathbb{R}^n} \mathrm{d}\mathbf{Z}^a \; \mathrm{p}_{\mathrm{out}^a}\left(\mathbf{y}|\mathbf{Z}^a\right)\right. \\
&\qquad\qquad \left.\times \int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{W}^a \; \mathrm{p}_{\mathrm{w}^a}\left(\mathbf{W}^a\right)\delta\left(\mathbf{Z}^a - \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}^a\right)\right] \\
&= \mathbb{E}_{\mathbf{X}}\int_{\mathbb{R}^n}\mathrm{d}\mathbf{y}\int_{\mathbb{R}^{n\times K}}\mathrm{d}\mathbf{Z}^\star \; \mathrm{p}_{\mathrm{out}^\star}\left(\mathbf{y}|\mathbf{Z}^\star\right) \\
&\qquad\qquad \times \int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{W}^\star \; \mathrm{p}_{\mathrm{w}^\star}\left(\mathbf{W}^\star\right)\delta\left(\mathbf{Z}^\star - \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}^\star\right) \\
&\times\left[\prod_{a=1}^{r}\int_{\mathbb{R}^{n\times K}} \mathrm{d}\mathbf{Z}^a \; \mathrm{p}_{\mathrm{out}^a}\left(\mathbf{y}|\mathbf{Z}^a\right)\int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{W}^a \; \mathrm{p}_{\mathrm{w}^a}\left(\mathbf{W}^a\right)\delta\left(\mathbf{Z}^a - \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}^a\right)\right] \\
&= \int_{\mathbb{R}^n}\mathrm{d}\mathbf{y}\prod_{a=0}^{r}\int_{\mathbb{R}^{n\times K}} \mathrm{d}\mathbf{Z}^a \; \mathrm{p}_{\mathrm{out}^a}\left(\mathbf{y}|\mathbf{Z}^a\right)\int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{W}^a \; \mathrm{p}_{\mathrm{w}^a}\left(\mathbf{W}^a\right) \\
&\qquad\qquad \times \underbrace{\mathbb{E}_{\mathbf{X}}\prod_{a=0}^{r}\delta\left(\mathbf{Z}^a - \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}^a\right)}_{(I)} .
\end{aligned} \tag{304}$$

Note that the average over $\mathbf{y}$ is equivalent to the one over the ground truth vector $\mathbf{W}^\star$ in the case of a *teacher-student*, which can be conveniently grouped with the other terms by just extending the replica indices and considering it as a new replica $\mathbf{W}^0$ with index $a = 0$, leading to a total of $r + 1$ replicas.

Average over the i.i.d input data X Remains to compute the average over $\mathbf{X}$ in the term (I). We suppose that inputs are drawn from an i.i.d distribution, for example a Gaussian $\mathrm{p}_{\mathrm{x}}(\mathbf{x}) = \mathcal{N}_{\mathbf{x}}\left(\mathbf{0}, \mathbf{I}_d\right)$. More precisely, for $(i,j) \in [\![d]\!]^2$, $(\mu,\nu) \in [\![n]\!]^2$, $\mathbb{E}_{\mathbf{X}}\left[x_{\mu i}x_{\nu j}\right] = \delta_{\mu\nu}\delta_{ij}$. By definition, the average in (I) defines the probability density $\mathrm{p}_{\mathrm{z}^a}(\mathbf{Z}^a)$ and as $\forall k \in [\![K]\!], \forall \mu \in [\![n]\!]$, $z^a_{\mu k} = \frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_{\mu i}w^a_{ik}$ is the sum of i.i.d random variables, the CLT insures that in the thermodynamic limit $d \to \infty$, $z^a_{\mu k}$ follows a Gaussian multivariate distribution, with first moments given by:

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[z^a_{\mu k}] &= \frac{1}{\sqrt{d}}\sum_{i=1}^{d}\mathbb{E}_{\mathbf{X}}\left[x_{\mu i}\right]w^a_{ik} = 0 \\
\mathbb{E}_{\mathbf{X}}[z^a_{\mu k}z^b_{\nu k'}] &= \frac{1}{d}\sum_{ij}\mathbb{E}_{\mathbf{X}}\left[x_{\mu i}x_{\mu j}\right]w^a_{ik}w^b_{jk'} = \frac{1}{d}\sum_{ij}\delta_{ij}w^a_{ik}w^b_{jk'}\delta_{\mu\nu} \\
&\equiv \delta_{\mu\nu}Q^{ak}_{bk'} .
\end{aligned}$$

Notice that averaging over the quenched disorder introduced correlations between replicas, which were initially independent, described by the symmetric *overlap* matrix $\{Q^{ak}_{bk'}\}_{kk'}$ of size $(r+1)K \times (r+1)K$. This matrix order parameter measures the correlations between the replicated matrices $\{\mathbf{W}^a\}_{a=0}^r$ and is formally defined by

$$\mathbf{Q}(\{\mathbf{W}^a\}_{a=0}^r) \equiv \left(\frac{1}{d}\sum_{i=1}^{d} w_{ik}^a w_{ik'}^b\right)_{k,k'=1..K}^{a,b=0..r},$$

such that $\forall(a,b) \in [\![0:r]\!]^2$, $\mathbf{Q}^{ab} \in \mathbb{R}^{K\times K}$. Therefore, again by the CLT, in the limit $d \to \infty$, the hidden variable $\mathbf{Z}^a \in \mathbb{R}^{n\times K}$ converges in distribution to the multivariate distribution

$$\mathrm{p}_{z^a}\left(\mathbf{Z}^a|\mathbf{Q}\right) = \exp\left[-\frac{1}{2}\sum_{\mu=1}^{n}\sum_{a,b=0}^{r}\sum_{k,k'=1}^{K} z_{\mu k}^a z_{\mu k'}^b (\mathbf{Q}^{-1})_{kk'}^{ab}\right] / \left(\det\left(2\pi\mathbf{Q}\right)\right)^{\frac{n}{2}}.$$

Inserting this back in the replicated partition function finally writes

$$\begin{aligned}&\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r\right] = \\ &\quad\int_{\mathbb{R}^n} \mathrm{d}\mathbf{y} \prod_{a=0}^{r}\int_{\mathbb{R}^{n\times K}} \mathrm{d}\mathbf{Z}^a\ \mathrm{p}_{\mathrm{out}^a}\left(\mathbf{y}|\mathbf{Z}^a\right)\mathrm{p}_{z^a}\left(\mathbf{Z}^a|\mathbf{Q}\right)\int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{W}^a\ \mathrm{p}_{w^a}\left(\mathbf{W}^a\right)\end{aligned}$$

Fourier representation Next we introduce the change of variable for the new order parameter $\mathbf{Q}^{ab}$ with a Dirac-δ distribution and its Fourier representation. For a variable $x \in \mathbb{R}$, the distribution $\delta(x)$ can be written as an integral over a purely imaginary parameter $\hat{x}$:

$$\delta\left(x\right) = \frac{1}{2i\pi}\int_{i\mathbb{R}} \mathrm{d}\hat{x} e^{-\hat{x}x}.$$

Applying the above identity to the change of variable, we obtain

$$\begin{aligned}1 &= \int_{\mathbb{R}^{(K\times r+1)^2}} \mathrm{d}\mathbf{Q} \prod_{0\leq a\leq b\leq r, 1\leq k,k'\leq K} \delta\left(dQ_{kk'}^{ab} - \sum_{i=1}^{d} w_{ik}^a w_{ik'}^b\right) \\ &\propto \int\int_{\mathbb{R}^{(K\times r+1)^2}} \mathrm{d}\mathbf{Q}\mathrm{d}\hat{\mathbf{Q}}\ \exp\left(-d\sum_{a=0}^{r}\sum_{k,k'}^{K} Q_{kk'}^{aa}\hat{Q}_{kk'}^{aa} - \frac{d}{2}\sum_{a\neq b}^{r}\sum_{k,k'}^{K} Q_{kk'}^{ab}\hat{Q}_{kk'}^{ab}\right) \\ &\qquad \times \exp\left(\frac{1}{2}\sum_{a=0}^{r}\sum_{k,k'}^{K}\hat{Q}_{kk'}^{aa} w_k^a w_{k'}^a + \frac{1}{2}\sum_{a\neq b}^{r}\sum_{k,k'}^{K}\hat{Q}_{kk'}^{ab} w_k^a w_{k'}^b\right),\end{aligned}$$

that involves a new ad-hoc purely imaginary matrix parameter $\hat{\mathbf{Q}} \in i\mathbb{R}^{(K\times(r+1))^2}$. Finally, multiplying the replicated partition function by 1, using the Cauchy theorem and rotating the integration, it becomes an integral over the symmetric matrices $\mathbf{Q} \in \mathbb{R}^{(K\times r+1)^2}$ and $\hat{\mathbf{Q}} \in \mathbb{R}^{(K\times r+1)^2}$

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathscr{Z}_d\left(\mathbf{y},\mathbf{X}\right)^r\right] \\
&= \int\int_{\mathbb{R}^{(K\times r+1)^2}} \mathrm{d}\mathbf{Q}\mathrm{d}\hat{\mathbf{Q}}\ \exp\left(-d\sum_{a=0}^{r}\sum_{k,k'}^{K} Q_{kk'}^{aa}\hat{Q}_{kk'}^{aa} - \frac{d}{2}\sum_{a\neq b}^{r}\sum_{k,k'}^{K} Q_{kk'}^{ab}\hat{Q}_{kk'}^{ab}\right) \\
&\qquad\qquad \times \exp\left(-\frac{1}{2}\sum_{a=0}^{r}\sum_{k,k'} \hat{Q}_{kk'}^{aa} w_k^a w_{'k}^a + \frac{1}{2}\sum_{a\neq b}\sum_{k,k'} \hat{Q}_{kk'}^{ab} w_k^a w_{k'}^b\right) \\
&\int_{\mathbb{R}^n} \mathrm{d}\mathbf{y} \prod_{a=0}^{r} \int_{\mathbb{R}^{n\times K}} \mathrm{d}\mathbf{Z}^a\ \mathrm{p}_{\mathrm{out}^a}\left(\mathbf{y}|\mathbf{Z}^a\right)\mathrm{p}_{\mathrm{Z}^a}\left(\mathbf{Z}^a|\mathbf{Q}\right) \int_{\mathbb{R}^{d\times K}} \mathrm{d}\mathbf{W}^a\ \mathrm{p}_{\mathrm{w}^a}\left(\mathbf{W}^a\right) \\
&= \int\int_{\mathbb{R}^{(K\times r+1)^2}} \mathrm{d}\mathbf{Q}\mathrm{d}\hat{\mathbf{Q}}\ \exp\left(-d\sum_{a=0}^{r}\sum_{k,k'}^{K} Q_{kk'}^{aa}\hat{Q}_{kk'}^{aa} - \frac{d}{2}\sum_{a\neq b}^{r}\sum_{k,k'}^{K} Q_{kk'}^{ab}\hat{Q}_{kk'}^{ab}\right) \\
&\qquad\qquad \times \exp\left(-\frac{1}{2}\sum_{a=0}^{r}\sum_{k,k'}^{K} \hat{Q}_{kk'}^{aa} w_k^a w_{k'}^a + \frac{1}{2}\sum_{a\neq b}^{r}\sum_{k,k'}^{K} \hat{Q}_{kk'}^{ab} w_k^a w_{k'}^b\right) \\
&\left[\int_{\mathbb{R}} \mathrm{d}y \prod_{a=0}^{r} \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}^a\ \mathrm{p}_{\mathrm{out}^a}\left(y|\mathbf{z}^a\right)\mathrm{p}_{\mathrm{z}^a}\left(\mathbf{z}^a|\mathbf{Q}\right)\right]^n \left[\prod_{a=0}^{r}\int_{\mathbb{R}^K} \mathrm{d}\mathbf{w}^a\ \mathrm{p}_{\mathrm{w}^a}\left(\mathbf{w}^a\right)\right]^d \\
&\simeq \int\int_{\mathbb{R}^{(K\times r+1)^2}} \mathrm{d}\mathbf{Q}\mathrm{d}\hat{\mathbf{Q}}\ e^{d\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}})}\,,
\end{aligned}
$$

where in the last step, we used a Laplace method (Wong, 1989) and omitted the sub-leading factors in the thermodynamic limit $d \to \infty$ to evaluate it as a function of the free entropy potential defined by

$$
\begin{aligned}
\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}}) &= -\sum_{a=0}^{r}\sum_{k,k'}^{K} Q_{kk'}^{aa}\hat{Q}_{kk'}^{aa} - \frac{1}{2}\sum_{a\neq b}^{r}\sum_{k,k'}^{K} Q_{kk'}^{ab}\hat{Q}_{kk'}^{ab} \\
&\qquad + \log\Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}}) + \alpha\log\Psi_{\mathrm{out}}^{(r)}(\mathbf{Q})\,, \\
\Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}}) &= \prod_{a=0}^{r}\int_{\mathbb{R}^K} \mathrm{d}\mathbf{w}^a\ \mathrm{p}_{\mathrm{w}^a}\left(\mathbf{w}^a\right) \\
&\qquad \times \exp\left(\sum_{a=0}^{r}\sum_{kk'} \hat{Q}_{kk'}^{aa} w_k^a w_k^a + \frac{1}{2}\sum_{a\neq b}\sum_{k,k'} \hat{Q}_{kk'}^{ab} w_k^a w_{k'}^b\right)\,, \\
\Psi_{\mathrm{out}}^{(r)}(\mathbf{Q}) &= \prod_{a=0}^{r}\int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}^a\ \mathrm{p}_{\mathrm{out}^a}\left(y|\mathbf{z}^a\right)\mathrm{p}_{\mathrm{z}^a}\left(\mathbf{z}^a|\mathbf{Q}\right)\,,
\end{aligned}
\tag{305}
$$

and where we decoupled the variable $\mathbf{Z}^a \in \mathbb{R}^{n\times K}$ and $\mathbf{W}^a \in \mathbb{R}^{d\times K}$ along the rows

$$\begin{aligned}
\mathrm{p}_{\mathrm{out}^a}\left(\mathbf{y}|\mathbf{Z}^a\right) &= \prod_{\mu=1}^{n} \mathrm{p}_{\mathrm{out}^a}\left(y_\mu|\mathbf{z}_\mu^a\right) \,, \text{ with } \mathbf{z}_\mu^a \in \mathbb{R}^K \,,\\
\mathrm{p}_{\mathrm{z}^a}(\mathbf{Z}^a|\mathbf{Q}) &= \prod_{\mu=1}^{n} \mathrm{p}(\mathbf{z}_\mu^a|\mathbf{Q}) \,,\\
\mathrm{p}_{\mathrm{w}^a}\left(\mathbf{W}^a\right) &= \prod_{i=1}^{d} \mathrm{p}_{\mathrm{w}}\left(\mathbf{w}_i^a\right) \,, \text{ with } \mathbf{w}_i^a \in \mathbb{R}^K \,,\\
\mathrm{p}_{\mathrm{z}^a}\left(\mathbf{z}^a|\mathbf{Q}\right) &= \exp\left[-\frac{1}{2}\sum_{a,b=0}^{r}\sum_{k,k'=1}^{K} z_k^a z_{k'}^b (\mathbf{Q}^{-1})_{kk'}^{ab}\right] / \left(\det\left(2\pi\mathbf{Q}\right)\right)^{\frac{1}{2}} \,.
\end{aligned}$$

Note that the averaged replicated partition function of this fully connected model can be expressed as a saddle point equation only because distributions $\mathrm{P}_{\mathrm{out}}, \mathrm{P}_{\mathrm{out}^\star}$ and $\mathrm{P}_{\mathrm{w}}, \mathrm{P}_{\mathrm{w}^\star}$ are separable so that a pre-factor scaling with the system size d dominates the exponential distribution. Finally, switching the two limits $r \to 0$ and $d \to \infty$, the quenched free entropy Φ simplifies as a saddle point equation

$$\Phi(\alpha) = \mathbf{extr}_{\mathbf{Q},\hat{\mathbf{Q}}}\left\{\lim_{r\to 0}\frac{\partial \Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}})}{\partial r}\right\} \,, \tag{306}$$

over symmetric matrices $\mathbf{Q} \in \mathbb{R}^{(K\times r+1)^2}$ and $\hat{\mathbf{Q}} \in \mathbb{R}^{(K\times r+1)^2}$. To summarize, we managed to get rid of the original high-dimensional integrals and replace them by an optimization in the space of matrices, which, in this form, is still intractable. We not only have to search in the space of $(r+1)\times(r+1)$ matrices to find the extremiser of $\Phi^{(r)}$, but we also need to compute the limit $r \to 0^+$. In the following we will assume a simple Ansatz for these matrices in order to first obtain an analytic expression in r before taking the derivative with respect to r.

B.1.1.c REPLICA SYMMETRIC FREE ENTROPY

Our goal is to express the functional $\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}})$ appearing in the free entropy as an analytical function of r, in order to perform the replica trick.

Replica symmetric ansatz To do so, we will assume that the extremum of $\Phi^{(r)}$ is attained at a point in $\mathbf{Q},\hat{\mathbf{Q}}$ space such that a *replica symmetry* property is verified. More concretely, we assume:

$$\begin{aligned}
\exists \mathbf{Q} \in \mathbb{R}^{K\times K} \text{ s.t } \quad \forall a \in [\![0:r]\!] \quad \forall (k,k') \in [\![K]\!]^2 \quad Q_{kk'}^{aa} &= Q_{kk'} \,,\\
\exists \mathbf{Q}^\star \in \mathbb{R}^{K\times K} \text{ s.t } \quad \forall (k,k') \in [\![K]\!]^2 \quad Q_{kk'}^{00} &= Q_{kk'}^\star \,,\\
\exists \mathbf{q} \in \mathbb{R}^{K\times K} \text{ s.t } \quad \forall (a<b) \in [\![0:r]\!]^2 \quad \forall (k,k') \in [\![K]\!]^2 \quad Q_{kk'}^{ab} &= q_{kk'} \,,\\
\exists \mathbf{m} \in \mathbb{R}^{K\times K} \text{ s.t } \quad \forall a \in [\![0:r]\!] \quad \forall (k,k') \in [\![K]\!]^2 \quad Q_{kk'}^{0a} &= m_{kk'} \,,
\end{aligned} \tag{307}$$

and similarly for the ad-hoc parameter

$$\begin{aligned}
\exists \hat{\mathbf{Q}} \in \mathbb{R}^{K\times K} \text{ s.t} \quad \forall a \in [\![0:r]\!] \quad \forall (k,k') \in [\![K]\!]^2 \quad \hat{Q}^{aa}_{kk'} &= -\frac{1}{2}\hat{Q}_{kk'}\,, \\
\exists \hat{\mathbf{Q}}^\star \in \mathbb{R}^{K\times K} \text{ s.t} \quad \forall (k,k') \in [\![K]\!]^2 \quad \hat{Q}^{00}_{kk'} &= \hat{Q}^\star_{kk'}\,, \\
\exists \hat{\mathbf{q}} \in \mathbb{R}^{K\times K} \text{ s.t} \quad \forall (a<b) \in [\![0:r]\!]^2 \quad \forall (k,k') \in [\![K]\!]^2 \quad \hat{Q}^{ab}_{kk'} &= \hat{q}_{kk'}\,, \\
\exists \hat{\mathbf{m}} \in \mathbb{R}^{K\times K} \text{ s.t} \quad \forall a \in [\![0:r]\!] \quad \forall (k,k') \in [\![K]\!]^2 \quad \hat{Q}^{0a}_{kk'} &= \hat{m}_{kk'}\,.
\end{aligned} \tag{308}$$

The factor $-\frac{1}{2}$ is not necessary bu useful to recover commonly used formulations. This Ansatz can be represented by symmetric RS matrices $\mathbf{Q}^{(\mathrm{rs})} \in \mathbb{R}^{(K\times r+1)^2}$ and $\hat{\mathbf{Q}}^{(\mathrm{rs})} \in \mathbb{R}^{(K\times r+1)^2}$

$$\mathbf{Q}^{(\mathrm{rs})} = \begin{pmatrix} \mathbf{Q}^\star & \mathbf{m} & \cdots & \mathbf{m} \\ \mathbf{m} & \mathbf{Q} & \mathbf{q} & \dots \\ \vdots & \mathbf{q} & \ddots & \mathbf{q} \\ \mathbf{m} & \dots & \mathbf{q} & \mathbf{Q} \end{pmatrix} \text{ and } \hat{\mathbf{Q}}^{(\mathrm{rs})} = \begin{pmatrix} -\frac{1}{2}\hat{\mathbf{Q}}^\star & \hat{\mathbf{m}} & \dots & \hat{\mathbf{m}} \\ \hat{\mathbf{m}} & -\frac{1}{2}\hat{\mathbf{Q}} & \hat{\mathbf{q}} & \dots \\ \vdots & \hat{\mathbf{q}} & \ddots & \hat{\mathbf{q}} \\ \hat{\mathbf{m}} & \dots & \hat{\mathbf{q}} & -\frac{1}{2}\hat{\mathbf{Q}} \end{pmatrix}, \tag{309}$$

where the *overlap* parameters may be reinterpreted as the scalar product between the replicas

$$\forall (a,b) \in [\![r]\!]^2,\ \mathbf{q} = \frac{1}{d}\mathbf{W}^{a\intercal}\mathbf{W}^b\,,$$

the self-overlap of each replica

$$\forall a \in [\![r]\!],\ \mathbf{Q} = \frac{1}{d}\mathbf{W}^{a\intercal}\mathbf{W}^a\,,$$

the scalar product with the ground truth

$$\forall a \in [\![r]\!],\ \mathbf{m} = \frac{1}{d}\mathbf{W}^{\star\intercal}\mathbf{W}^a\,,$$

and the second moment of the ground truth distribution

$$\mathbf{Q}^\star = \frac{1}{d}\mathbf{W}^{\star\intercal}\mathbf{W}^\star\,.$$

The above Ansatz simplifies in the scalar GLM case with $K=1$ to $q = \frac{1}{d}\mathbf{w}^a \cdot \mathbf{w}^b$ for $a \neq b$, a norm $Q = \frac{1}{d}\|\mathbf{w}^a\|_2^2$, an overlap with the ground truth $m = \frac{1}{d}\mathbf{w}^a \cdot \mathbf{w}^\star$ and a second moment $Q^\star = \frac{1}{d}\|\mathbf{w}^\star\|_2^2$.

Let's compute separately the terms involved in the functional $\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}})$ in (305) by applyting this Ansatz: the first is a trace term, the second term $\Psi_{\mathrm{w}}^{(r)}$ depends on the prior distributions P_{w}, $\mathrm{P}_{\mathrm{w}^\star}$ and finally the third term $\Psi_{\mathrm{out}}^{(r)}$ depends on the channel distributions $\mathrm{P}_{\mathrm{out}^\star}$, $\mathrm{P}_{\mathrm{out}}$.

Trace term The trace term in (305) can be easily computed at the RS fixed point and takes the following form

$$
\begin{aligned}
&-\sum_{a=0}^{r}\sum_{k,k'}^{K} Q_{kk'}^{aa}\hat{Q}_{kk'}^{aa} - \frac{1}{2}\sum_{a\neq b}^{r}\sum_{k,k'}^{K} Q_{kk'}^{ab}\hat{Q}_{kk'}^{ab}\Bigg|_{\mathrm{rs}} \\
&= \frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}^{\star}\hat{\mathbf{Q}}^{\star}\right) + \frac{1}{2}r\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right) - r\mathrm{Tr}\left(\mathbf{m}\hat{\mathbf{m}}\right) - \frac{r(r-1)}{2}\mathrm{Tr}\left(\mathbf{q}\hat{\mathbf{q}}\right),
\end{aligned}
$$

and taking the derivative and the limit $r \to 0$ we obtain

$$
\begin{aligned}
&\lim_{r\to 0}\partial_r\left(-\sum_{a=0}^{r}\sum_{k,k'}^{K} Q_{kk'}^{aa}\hat{Q}_{kk'}^{aa} - \frac{1}{2}\sum_{a\neq b}^{r}\sum_{k,k'}^{K} Q_{kk'}^{ab}\hat{Q}_{kk'}^{ab}\right)\Bigg|_{\mathrm{rs}} \\
&\qquad = \frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right) - \mathrm{Tr}\left(\mathbf{m}\hat{\mathbf{m}}\right) + \frac{1}{2}\mathrm{Tr}\left(\mathbf{q}\hat{\mathbf{q}}\right)
\end{aligned} \tag{310}
$$

Prior integral $\Psi_{\mathrm{w}}^{(r)}$ Evaluated at the RS fixed point the quadratic form reads

$$
\begin{aligned}
&\sum_{a=0}^{r}\sum_{kk'}\hat{Q}_{kk'}^{aa}w_k^a w_k^a + \frac{1}{2}\sum_{a\neq b}\sum_{k,k'}\hat{Q}_{kk'}^{ab}w_k^a w_{k'}^b \\
&= \sum_{a=1}^{r}\mathbf{w}^{\star\intercal}\hat{\mathbf{m}}\mathbf{w}^a - \frac{1}{2}\sum_{a=1}^{r}\mathbf{w}^{a\intercal}\hat{\mathbf{Q}}\mathbf{w}^a + \sum_{1\leq a<b\leq r}\mathbf{w}^{a\intercal}\hat{\mathbf{q}}\mathbf{w}^b \\
&= \sum_{a=1}^{r}\mathbf{w}^{\star\intercal}\hat{\mathbf{m}}\mathbf{w}^a - \frac{1}{2}\sum_{a=1}^{r}\mathbf{w}^{a\intercal}\left(\hat{\mathbf{Q}}+\hat{\mathbf{q}}\right)\mathbf{w}^a + \frac{1}{2}\left(\sum_{a=1}^{r}\mathbf{w}^a\right)^{\intercal}\hat{\mathbf{q}}\left(\sum_{a=1}^{r}\mathbf{w}^a\right).
\end{aligned}
$$

Using a Hubbard-Stratonovich transformation presented in Appendix. A.2, the prior integral can be further simplified

$$
\begin{aligned}
&\Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}})\Big|_{\mathrm{rs}} = \int_{\mathbb{R}^{(r+1)\times K}} \mathrm{d}\mathbf{W}\mathrm{p}_{\tilde{\mathrm{w}}}\left(\mathbf{W}\right) e^{\sum_{a=0}^{r}\sum_{k,k'}^{K}\hat{Q}_{kk'}^{aa}w_k^a w_k^a + \frac{1}{2}\sum_{a\neq b}\sum_{k,k'}^{K}\hat{Q}_{kk'}^{ab}w_k^a w_{k'}^b} \\
&= \mathbb{E}_{\boldsymbol{\xi},\mathbf{w}^{\star}\sim\mathrm{P}_{\mathrm{w}^{\star}}}\left[e^{\mathbf{w}^{\star\intercal}\hat{\mathbf{Q}}^{\star}\mathbf{w}^{\star}}\mathbb{E}_{\mathbf{w}\sim\mathrm{P}_{\mathrm{w}}}\left[e^{\left(\mathbf{w}^{\intercal}\hat{\mathbf{m}}\mathbf{w}^{\star} - \frac{1}{2}\mathbf{w}^{\intercal}(\hat{\mathbf{Q}}+\hat{\mathbf{q}})\mathbf{w} + \mathbf{w}^{\intercal}\hat{\mathbf{q}}^{1/2}\boldsymbol{\xi}\right)}\right]^{r}\right].
\end{aligned} \tag{311}
$$

Channel integral $\Psi_{\mathbf{out}}^{(r)}$ Let us focus on the matrix $\mathbf{Q}^{(\mathrm{rs})}$ involved in the expression of $\Psi_{\mathrm{out}}^{(r)}$ in (305). The elements of its inverse block matrix

$$
\left(\mathbf{Q}^{(\mathrm{rs})}\right)^{-1} = \begin{bmatrix} \tilde{\mathbf{Q}}^{\star} & \tilde{\mathbf{m}} & \cdots & \tilde{\mathbf{m}} \\ \tilde{\mathbf{m}} & \tilde{\mathbf{Q}} & \tilde{\mathbf{q}} & \cdots \\ \vdots & \tilde{\mathbf{q}} & \ddots & \tilde{\mathbf{q}} \\ \tilde{\mathbf{m}} & \cdots & \tilde{\mathbf{q}} & \tilde{\mathbf{Q}} \end{bmatrix} \tag{312}
$$

can be computed and given by

$$
\begin{aligned}
\tilde{\mathbf{Q}}^\star &= \left(\mathbf{Q}^\star - r\mathbf{m}(\mathbf{Q}+(r-1)\mathbf{q})^{-1}\mathbf{m}^\intercal\right)^{-1} \\
\tilde{\mathbf{m}} &= -\left(\mathbf{Q}^\star - r\mathbf{m}(\mathbf{Q}+(r-1)\mathbf{q})^{-1}\mathbf{m}^\intercal\right)^{-1}\mathbf{m}(\mathbf{Q}+(r-1)\mathbf{q})^{-1} \\
\tilde{\mathbf{Q}} &= (\mathbf{Q}-\mathbf{q})^{-1} - (\mathbf{Q}+(r-1)\mathbf{q})^{-1}\mathbf{q}(\mathbf{Q}-\mathbf{q})^{-1} \\
&\quad + (\mathbf{Q}+(r-1)\mathbf{q})^{-1}\mathbf{m}^\intercal \\
&\qquad \times \left(\mathbf{Q}^\star - r\mathbf{m}(\mathbf{Q}+(r-1)\mathbf{q})^{-1}\mathbf{m}^\intercal\right)^{-1}\mathbf{m}(\mathbf{Q}+(r-1)\mathbf{q})^{-1} \\
\tilde{\mathbf{q}} &= \tilde{\mathbf{Q}} - (\mathbf{Q}-\mathbf{q})^{-1}
\end{aligned}
$$

and its determinant by

$$
\begin{aligned}
\det\left(\mathbf{Q}^{(\mathrm{rs})}\right) &= \det(\mathbf{Q}-\mathbf{q})^{r-1}\det(\mathbf{Q}+(r-1)\mathbf{q}) \\
&\quad \times \det\left(\mathbf{Q}^\star - r\mathbf{m}(\mathbf{Q}+(r-1)\mathbf{q})^{-1}\mathbf{m}^\intercal\right). \qquad (313)
\end{aligned}
$$

Therefore the quadratic form in $\mathrm{p}_{z^a}(\mathbf{z}^a|\mathbf{Q}^{(\mathrm{rs})}$ reads

$$
\begin{aligned}
&-\frac{1}{2}\sum_{a,b}\sum_{k,k'} z_k^a z_{k'}^b (\mathbf{Q}^{-1})_{kk'}^{ab} \\
&= -\frac{1}{2}\mathbf{z}^{\star\intercal}\tilde{\mathbf{Q}}^\star\mathbf{z} - \sum_{a=1}^{r}\mathbf{z}^{\star\intercal}\tilde{\mathbf{m}}\mathbf{z}^a \\
&\qquad -\frac{1}{2}\sum_{a=1}^{r}\mathbf{z}^{a\intercal}\left(\tilde{\mathbf{Q}}-\tilde{\mathbf{q}}\right)\mathbf{z}^a - \frac{1}{2}\left(\sum_a^r \mathbf{z}^a\right)^\intercal \tilde{\mathbf{q}}\left(\sum_a^r \mathbf{z}^a\right),
\end{aligned}
$$

and using another Gaussian transformation, see Appendix. A.2, we finally obtain

$$
\begin{aligned}
\Psi_{\mathrm{out}}^{(r)}(\mathbf{Q})\Big|_{\mathrm{rs}} &= \int \mathrm{d}y \int_{\mathbb{R}^{(r+1)\times K}} \mathrm{d}\mathbf{Z}\, \mathrm{p}_{\mathrm{out}}(y|\mathbf{Z})\, \mathrm{p}(\mathbf{Z}|\mathbf{Q}) \\
&= \int \mathrm{d}y \int_{\mathbb{R}^{(r+1)\times K}} \mathrm{d}\mathbf{Z}\mathrm{p}_{\mathrm{out}}(y|\mathbf{Z})\, e^{-\frac{1}{2}\sum_{a,b=0}^{r}\sum_{k,k'=1}^{K} z_k^a z_{k'}^b (\mathbf{Q}^{-1})_{kk'}^{ab}} \\
&\qquad\qquad / \left(\det(2\pi\mathbf{Q})^{(\mathrm{rs})}\right)^{\frac{1}{2}} \\
&= \int \mathrm{d}y\, \mathbb{E}_{\boldsymbol{\xi}}\, \mathrm{e}^{-\frac{1}{2}\log\left(\det\left(2\pi\mathbf{Q}^{(\mathrm{rs})}\right)\right)} \times \int \mathrm{d}\mathbf{z}^\star\, \mathrm{p}_{\mathrm{out}^\star}(y|\mathbf{z}^\star)\, e^{-\frac{1}{2}\mathbf{z}^{\star\intercal}\tilde{\mathbf{Q}}^\star\mathbf{z}^\star} \qquad (314) \\
&\times \left[\int \mathrm{d}\mathbf{z}\, \mathrm{p}_{\mathrm{out}}(y|\mathbf{z}) \exp\left(-\mathbf{z}^{\star\intercal}\tilde{\mathbf{m}}\mathbf{z} - \frac{1}{2}\mathbf{z}^\intercal\left(\tilde{\mathbf{Q}}-\tilde{\mathbf{q}}\right)\mathbf{z} + \mathbf{z}^\intercal(-\tilde{\mathbf{q}})^{1/2}\boldsymbol{\xi}\right)\right]^r,
\end{aligned}
$$

with $\det\left(\mathbf{Q}^{(\mathrm{rs})}\right)$ given by (313).

B.1.1.D CONSISTENCY CONDITIONS $r \to 0$: $\Theta(1)$ TERMS

It remains to take the limit $r \to 0^+$ of the expressions for $\Psi_{\mathrm{w}}^{(r)}$ and $\Psi_{\mathrm{out}}^{(r)}$ that are now analytical in r. First, our assumptions must be consistent and thus we need to check the consistency conditions in the limit $r \to 0$. Indeed, if $\Phi^{(r)}$ is finite we could obtain divergence taking the limit $\lim_{r\to 0}\frac{1}{r}\Phi^{(r)} = \infty$. Therefore to avoid such divergence, we must at least impose that $\lim_{r\to 0}\Phi^{(r)} = 0$:

$$\lim_{r\to 0}\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}}) = -\mathrm{Tr}\left(\mathbf{Q}^{\star}\hat{\mathbf{Q}}^{\star}\right) + \log\Psi^{0}_{\mathrm{w}}(\hat{\mathbf{Q}}^{\star}) + \alpha\log\Psi^{0}_{\mathrm{out}}(\mathbf{Q}^{\star})$$

with

$$\begin{aligned}\Psi^{0}_{\mathrm{w}}(\hat{\mathbf{Q}}^{\star}) &\equiv \mathbb{E}_{\mathbf{w}^{\star}}\exp\left(\mathbf{w}^{\star\intercal}\hat{\mathbf{Q}}^{\star}\mathbf{w}^{\star}\right),\\ \Psi^{0}_{\mathrm{out}}(\mathbf{Q}^{\star}) &\equiv \int_{\mathbb{R}} \mathrm{d}y \int \mathrm{d}\mathbf{z}^{\star}\, \mathrm{p}_{\mathrm{out}^{\star}}(y|\mathbf{z}^{\star})\,\mathcal{N}_{\mathbf{z}^{\star}}(\mathbf{0},\mathbf{Q}^{\star}) = 1\,.\end{aligned}$$

Taking the saddle point equations over $\mathbf{Q}^{\star}$ and $\hat{\mathbf{Q}}^{\star}$, imposing the consistency condition $\lim_{r\to 0}\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}}) = 0$, we finally obtain

$$\hat{\mathbf{Q}}^{\star} = \mathbf{0} \quad \text{and} \quad \mathbf{Q}^{\star} = \mathbb{E}_{\mathbf{w}^{\star}}\left[\mathbf{w}^{\star\intercal}\mathbf{w}^{\star}\right]. \tag{315}$$

B.1.1.E REPLICA TRICK $r \to 0$ LIMIT: $\Theta(r)$ TERMS

Imposing the conditions (315) avoids divergence in the replica trick, and we can therefore proceed with the $\Theta(r)$ terms.

Prior integral $\Psi^{(r)}_{\mathrm{w}}$ The limit $r \to 0$ and the derivative of the logarithm of the prior integral (311) can be trivially computed

$$\begin{aligned}&\lim_{r\to 0}\partial_r \log\Psi^{(r)}_{\mathrm{w}}(\hat{\mathbf{Q}})\Big|_{\mathrm{rs}}\\ &= \mathbb{E}_{\boldsymbol{\xi},\mathbf{w}^{\star}}\\ &\times \log\left[\mathbb{E}_{\mathbf{w}}\exp\left(\left[\mathbf{w}^{\star\intercal}\hat{\mathbf{m}}\mathbf{w} - \frac{1}{2}\mathbf{w}^{\intercal}(\hat{\mathbf{Q}}+\hat{\mathbf{q}})\mathbf{w} + \boldsymbol{\xi}^{\intercal}\hat{\mathbf{q}}^{1/2}\mathbf{w}\right]\right)\right],\end{aligned}$$

with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0},\mathbf{I}_K)$ and $\mathbf{w}^{\star} \sim \mathrm{P}_{\mathrm{w}^{\star}}$. To conclude, we can symmetrize and decouple the *teacher* and *student* expectations $\mathbb{E}_{\mathbf{w}^{\star}},\mathbb{E}_{\mathbf{w}}$. By performing the change of variable $\boldsymbol{\xi} \leftarrow \boldsymbol{\xi} + \hat{\mathbf{q}}^{-1/2}\hat{\mathbf{m}}\mathbf{w}^{\star}$, we finally obtain

$$\begin{aligned}&\lim_{r\to 0}\partial_r \log\Psi^{(r)}_{\mathrm{w}}(\hat{\mathbf{Q}})\Big|_{\mathrm{rs}}\\ &= \mathbb{E}_{\boldsymbol{\xi},\mathbf{w}^{\star}}\exp\left(-\frac{1}{2}\mathbf{w}^{\star\intercal}\hat{\mathbf{m}}^{\intercal}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}}\mathbf{w}^{\star} + \boldsymbol{\xi}^{\intercal}\hat{\mathbf{q}}^{-1/2}\hat{\mathbf{m}}\mathbf{w}^{\star}\right)\\ &\qquad \times\log\left[\mathbb{E}_{\mathbf{w}}\exp\left(\left[-\frac{1}{2}\mathbf{w}^{\intercal}(\hat{\mathbf{Q}}+\hat{\mathbf{q}})\mathbf{w} + \boldsymbol{\xi}^{\intercal}\hat{\mathbf{q}}^{1/2}\mathbf{w}\right]\right)\right]\\ &\equiv \mathbb{E}_{\boldsymbol{\xi},\mathbf{w}^{\star}}\mathcal{Z}_{\mathrm{w}^{\star}}\left(\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2}\boldsymbol{\xi},\hat{\mathbf{m}}^{\intercal}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}}\right)\log\mathcal{Z}_{\mathrm{w}}\left(\hat{\mathbf{q}}^{1/2}\boldsymbol{\xi},\hat{\mathbf{Q}}+\hat{\mathbf{q}}\right),\end{aligned} \tag{316}$$

with the corresponding denoising distribution Q_{w} and functions $\mathcal{Z}_{\mathrm{w}^{\star}},\mathcal{Z}_{\mathrm{w}}$ defined in Sec. A.4.1.a respectively with distributipn $\mathrm{P}_{\mathrm{w}^{\star}}$ and P_{w}.

Prior integral $\Psi_{\mathrm{out}}^{(r)}$ The limit $r \to 0$ and the derivative of the logarithm of the channel integral (314) is more tricky. First, the limit of the determinant simplifies easily and yields

$$\det\left(\mathbf{Q}^{(\mathrm{rs})}\right) \underset{r\to 0}{\longrightarrow} \det\left(\mathbf{Q}^{\star}\right)$$

and the matrix elements of $\left(\mathbf{Q}^{(\mathrm{rs})}\right)^{-1}$ in this limit become

$$\begin{aligned}
\tilde{\mathbf{Q}}^{\star} &\underset{r\to 0}{\longrightarrow} \left(\mathbf{Q}^{\star}\right)^{-1}, \\
\tilde{\mathbf{m}} &\underset{r\to 0}{\longrightarrow} -\left(\mathbf{Q}^{\star}\right)^{-1}\mathbf{m}(\mathbf{Q}-\mathbf{q})^{-1}, \\
\tilde{\mathbf{Q}} &\underset{r\to 0}{\longrightarrow} -(\mathbf{Q}-\mathbf{q})^{-1}\left(\mathbf{q}-\mathbf{m}\left(\mathbf{Q}^{\star}\right)^{-1}\mathbf{m}\right)(\mathbf{Q}-\mathbf{q})^{-1}, \\
\tilde{\mathbf{q}} &\underset{r\to 0}{\longrightarrow} \tilde{\mathbf{Q}}+(\mathbf{Q}-\mathbf{q})^{-1}.
\end{aligned}$$

By taking properly the $r \to 0$ limit and performing the change of variable $\mathbf{z}^{\star} \leftarrow \left(\mathbf{Q}^{\star}-\mathbf{m}^{\intercal}\mathbf{q}^{-1}\mathbf{m}\right)^{1/2}\mathbf{z}^{\star}+\mathbf{m}\mathbf{q}^{-1/2}\boldsymbol{\xi}$, we finally obtain

$$\begin{aligned}
&\lim_{r\to 0}\partial_r \log \Psi_{\mathrm{out}}^{(r)}(\mathbf{Q})\Big|_{\mathrm{rs}} \\
&= \int_{\mathbb{R}} \mathrm{d}y\, \mathbb{E}_{\boldsymbol{\xi}} \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}^{\star}\, \mathrm{p}_{\mathrm{out}^{\star}}\left(y|\mathbf{z}^{\star}\right) \\
&\times \exp\left(-\frac{1}{2}\left(\mathbf{z}^{\star}-\mathbf{m}\mathbf{q}^{-1/2}\boldsymbol{\xi}\right)^{\intercal}\left(\mathbf{Q}^{\star}-\mathbf{m}^{\intercal}\mathbf{q}^{-1}\mathbf{q}\right)\left(\mathbf{z}^{\star}-\mathbf{m}\mathbf{q}^{-1/2}\boldsymbol{\xi}\right)\right) \\
&\times \log\left[\int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}\, \mathrm{p}_{\mathrm{out}}\left(y|\mathbf{z}\right) e^{-\frac{1}{2}\left(\mathbf{z}-\mathbf{q}^{-1/2}\boldsymbol{\xi}\right)^{\intercal}(\mathbf{Q}-\mathbf{q})\left(\mathbf{z}-\mathbf{q}^{-1/2}\boldsymbol{\xi}\right)}\right] \qquad (317) \\
&= \int_{\mathbb{R}} \mathrm{d}y\, \mathbb{E}_{\boldsymbol{\xi}}\, \mathcal{Z}_{\mathrm{out}^{\star}}\left(\mathbf{m}\mathbf{q}^{-1/2}\boldsymbol{\xi}, \mathbf{Q}^{\star}-\mathbf{m}\mathbf{q}^{-1}\mathbf{m}\right) \log \mathcal{Z}_{\mathrm{out}}\left(\mathbf{q}^{1/2}\boldsymbol{\xi}, \mathbf{Q}-\mathbf{q}\right),
\end{aligned}$$

where the denoising distribution $\mathrm{Q}_{\mathrm{out}}$ and functions $\mathcal{Z}_{\mathrm{out}^{\star}}, \mathcal{Z}_{\mathrm{out}}$ are defined in Sec. A.4.1.a for the distributions $\mathrm{P}_{\mathrm{out}^{\star}}, \mathrm{P}_{\mathrm{out}}$.

B.1.1.f SUMMARY - MISMATCHED CASE

In the mismatched case, where the teacher and the student have not the same prior distributions, we finally obtain the replica symmetric free entropy Φ_{rs} for the committee machine hypothesis class:

$$\begin{aligned}
\Phi_{\mathrm{rs}}(\alpha) &\equiv \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\lim_{d\to\infty}\frac{1}{d}\log\left(\mathcal{Z}_d\left(\mathbf{y},\mathbf{X}\right)\right)\right] \\
&= \mathbf{extr}_{\mathbf{Q},\hat{\mathbf{Q}},\mathbf{q},\hat{\mathbf{q}},\mathbf{m},\hat{\mathbf{m}}}\left\{-\mathrm{Tr}\left(\mathbf{m}\hat{\mathbf{m}}\right)+\frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right)+\frac{1}{2}\mathrm{Tr}\left(\mathbf{q}\hat{\mathbf{q}}\right)\right. \qquad (318) \\
&\qquad \left.+\Psi_{\mathrm{w}}\left(\hat{\mathbf{Q}},\hat{\mathbf{m}},\hat{\mathbf{q}}\right)+\alpha\Psi_{\mathrm{out}}\left(\mathbf{Q},\mathbf{m},\mathbf{q};\boldsymbol{\rho}_{\mathrm{w}^{\star}}\right)\right\},
\end{aligned}$$

where $\boldsymbol{\rho}_{\mathrm{w}^\star} \equiv \lim_{d\to\infty} \mathbf{Q}^\star = \lim_{d\to\infty} \mathbb{E}_{\mathbf{W}^\star} \frac{1}{d} \mathbf{W}^{\star\intercal}\mathbf{W}^\star$ and the channel and prior integrals are defined by

$$\begin{aligned} \Psi_{\mathrm{w}}\left(\hat{\mathbf{Q}},\hat{\mathbf{m}},\hat{\mathbf{q}}\right) \equiv \mathbb{E}_{\boldsymbol{\xi}} &\left[\mathcal{Z}_{\mathrm{w}^\star}\left(\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2}\boldsymbol{\xi},\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}}\right)\right. \\ &\left.\times \log \mathcal{Z}_{\mathrm{w}}\left(\hat{\mathbf{q}}^{1/2}\boldsymbol{\xi},\hat{\mathbf{Q}}+\hat{\mathbf{q}}\right)\right], \\ \Psi_{\mathrm{out}}\left(\mathbf{Q},\mathbf{m},\mathbf{q};\boldsymbol{\rho}_{\mathrm{w}^\star}\right) \equiv \mathbb{E}_{y,\boldsymbol{\xi}} &\left[\mathcal{Z}_{\mathrm{out}^\star}\left(y,\mathbf{m}\mathbf{q}^{-1/2}\boldsymbol{\xi},\boldsymbol{\rho}_{\mathrm{w}^\star}-\mathbf{m}^\intercal\mathbf{q}^{-1}\mathbf{m}\right)\right. \\ &\left.\times \log \mathcal{Z}_{\mathrm{out}}\left(y,\mathbf{q}^{1/2}\boldsymbol{\xi},\mathbf{Q}-\mathbf{q}\right)\right], \end{aligned} \tag{319}$$

where again $\mathcal{Z}_{\mathrm{out}^\star}, \mathcal{Z}_{\mathrm{w}^\star}$ and $\mathcal{Z}_{\mathrm{out}}, \mathcal{Z}_{\mathrm{w}}$ are defined in Sec. A.4.1.a and depend respectively on channel and prior distributions of the *teacher* and *student*.

B.1.1.G SUMMARY - BAYES OPTIMAL MMSE ESTIMATION

For MMSE estimation in the Bayes-optimal setting, the student has access to the ground truth distributions of the teacher $\mathrm{P}_{\mathrm{out}}\left(\mathbf{y}|\mathbf{Z}\right) = \mathrm{P}_{\mathrm{out}^\star}\left(\mathbf{y}|\mathbf{Z}\right)$ and $\mathrm{P}_{\mathrm{w}}\left(\mathbf{W}\right) = \mathrm{P}_{\mathrm{w}^\star}(\mathbf{W})$, and therefore $\mathcal{Z}_{\mathrm{out}} = \mathcal{Z}_{\mathrm{out}^\star}$, $\mathcal{Z}_{\mathrm{w}} = \mathcal{Z}_{\mathrm{w}^\star}$. In this idealistic setting, the Nishimori conditions, recalled in Appendix. A.3, imply that

$$\mathbf{Q} = \mathbf{Q}_{\mathrm{w}^\star}, \quad \hat{\mathbf{Q}} = \mathbf{0}, \quad \mathbf{m} = \mathbf{q} \equiv \mathbf{q}_{\mathrm{b}}, \quad \hat{\mathbf{m}} = \hat{\mathbf{q}} \equiv \hat{\mathbf{q}}_{\mathrm{b}}. \tag{320}$$

Therefore the free entropy in eq. (319) simplifies as an optimization problem over the *overlaps* $\mathbf{q}_{\mathrm{b}}, \hat{\mathbf{q}}_{\mathrm{b}} \in \mathbb{R}^{K\times K}$

$$\Phi_{\mathrm{rs}}^{\mathrm{b}}(\alpha) = \mathbf{extr}_{\mathbf{q}_{\mathrm{b}},\hat{\mathbf{q}}_{\mathrm{b}}} \left\{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{q}_{\mathrm{b}}\hat{\mathbf{q}}_{\mathrm{b}}\right) + \Psi_{\mathrm{w}}^{\mathrm{b}}\left(\hat{\mathbf{q}}_{\mathrm{b}}\right) + \alpha\Psi_{\mathrm{out}}^{\mathrm{b}}\left(\mathbf{q}_{\mathrm{b}};\boldsymbol{\rho}_{\mathrm{w}^\star}\right)\right\}, \tag{321}$$

with free entropy terms $\Psi_{\mathrm{w}}^{\mathrm{b}}$ and $\Psi_{\mathrm{out}}^{\mathrm{b}}$ given by

$$\begin{aligned} \Psi_{\mathrm{w}}^{\mathrm{b}}\left(\hat{\mathbf{q}}\right) &= \mathbb{E}_{\xi}\left[\mathcal{Z}_{\mathrm{w}^\star}\left(\hat{\mathbf{q}}^{1/2}\boldsymbol{\xi},\hat{\mathbf{q}}\right)\log \mathcal{Z}_{\mathrm{w}^\star}\left(\hat{\mathbf{q}}^{1/2}\boldsymbol{\xi},\hat{\mathbf{q}}\right)\right], \\ \Psi_{\mathrm{out}}^{\mathrm{b}}\left(\mathbf{q};\boldsymbol{\rho}_{\mathrm{w}^\star}\right) &= \mathbb{E}_{y,\boldsymbol{\xi}}\left[\mathcal{Z}_{\mathrm{out}}\left(y,\mathbf{q}^{1/2}\boldsymbol{\xi},\boldsymbol{\rho}_{\mathrm{w}^\star}-\mathbf{q}\right)\right. \\ &\qquad\left.\times \log \mathcal{Z}_{\mathrm{out}}\left(y,\mathbf{q}^{1/2}\boldsymbol{\xi},\boldsymbol{\rho}_{\mathrm{w}^\star}-\mathbf{q}\right)\right]. \end{aligned} \tag{322}$$

Application to the GLM For the GLM hypothesis class, the same equations are valid if we take $K = 1$ for both the teacher and the student. As a result, we recover the replica symmetric free entropy in the Bayes-optimal setting rigorously proven in (Barbier et al., 2019b).

B.1.2 FIXED POINT EQUATIONS

The overlaps parameters, such as $\mathbf{m},\mathbf{q}$, play a crucial role since they measure the performances of the statistical estimation. Their behaviours are respectively characterized by the extremization of the free entropy (318) in the mismatched setting and (321) in the Bayes-optimal case. In this section, we give the expressions of the corresponding fixed point equations, whose derivations can be found in (Aubin et al., 2020a) Appendix. IV.4-5 for $K=1$ which can be extended to $K \geq 1$.

B.1.2.A MISMATCHED SETTING

Extremizing the free entropy eq. (318), we easily obtain the set of six fixed point equations

$$\begin{aligned}
\hat{\mathbf{Q}} &= -2\alpha \partial_{\mathbf{Q}} \Psi_{\text{out}}\left(\mathbf{Q}, \mathbf{m}, \mathbf{q}; \boldsymbol{\rho}_{\text{w}^\star}\right), & \mathbf{Q} &= -2\partial_{\hat{\mathbf{Q}}} \Psi_{\text{w}}\left(\hat{\mathbf{Q}}, \hat{\mathbf{m}}, \hat{\mathbf{q}}\right) \\
\hat{\mathbf{q}} &= -2\alpha \partial_{\mathbf{q}} \Psi_{\text{out}}\left(\mathbf{Q}, \mathbf{m}, \mathbf{q}; \boldsymbol{\rho}_{\text{w}^\star}\right), & \mathbf{q} &= -2\partial_{\hat{\mathbf{q}}} \Psi_{\text{w}}\left(\hat{\mathbf{Q}}, \hat{\mathbf{m}}, \hat{\mathbf{q}}\right), \\
\hat{\mathbf{m}} &= \alpha \partial_{\mathbf{m}} \Psi_{\text{out}}\left(\mathbf{Q}, \mathbf{m}, \mathbf{q}; \boldsymbol{\rho}_{\text{w}^\star}\right), & \mathbf{m} &= \partial_{\hat{\mathbf{m}}} \Psi_{\text{w}}\left(\hat{\mathbf{Q}}, \hat{\mathbf{m}}, \hat{\mathbf{q}}\right).
\end{aligned} \tag{323}$$

Interestingly, these equations can be reformulated as functions of $\mathcal{Z}_{\text{out}^\star}$, $\mathcal{Z}_{\text{w}^\star}$ and the denoising functions $f_{\text{out}^\star}, f_{\text{w}^\star}, f_{\text{out}}, f_{\text{w}}$ defined in (290)-(292) in Sec. A.4.1.b. Defining the natural variables $\boldsymbol{\Sigma} = \mathbf{Q} - \mathbf{q}$ and $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{Q}} + \hat{\mathbf{q}}$ they can reformulated as

$$\begin{aligned}
\hat{\mathbf{m}} &= \alpha \mathbb{E}_{y,\boldsymbol{\xi}}\left[\mathcal{Z}_{\text{out}^\star} \times \mathbf{f}_{\text{out}^\star}\left(y, \mathbf{m}\mathbf{q}^{-1/2}\boldsymbol{\xi}, \boldsymbol{\rho}_{\text{w}^\star} - \mathbf{m}^\intercal \mathbf{q}^{-1}\mathbf{m}\right)\right. \\
&\qquad\qquad\qquad\qquad\left. \times \mathbf{f}_{\text{out}}\left(y, \mathbf{q}^{1/2}\boldsymbol{\xi}, \boldsymbol{\Sigma}\right)^\intercal\right], \\
\hat{\mathbf{q}} &= \alpha \mathbb{E}_{y,\boldsymbol{\xi}}\left[\mathcal{Z}_{\text{out}^\star}\left(y, \mathbf{m}\mathbf{q}^{-1/2}\boldsymbol{\xi}, \boldsymbol{\rho}_{\text{w}^\star} - \mathbf{m}^\intercal \mathbf{q}^{-1}\mathbf{m}\right) \mathbf{f}_{\text{out}}\left(y, \mathbf{q}^{1/2}\boldsymbol{\xi}, \boldsymbol{\Sigma}\right)^{\otimes 2}\right], \\
\hat{\boldsymbol{\Sigma}} &= -\alpha \mathbb{E}_{y,\boldsymbol{\xi}}\left[\mathcal{Z}_{\text{out}^\star}\left(y, \mathbf{m}\mathbf{q}^{-1/2}\boldsymbol{\xi}, \boldsymbol{\rho}_{\text{w}^\star} - \mathbf{m}^\intercal \mathbf{q}^{-1}\mathbf{m}\right)\right. \\
&\qquad\qquad\qquad\qquad\left. \times \partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out}}\left(y, \mathbf{q}^{1/2}\boldsymbol{\xi}, \boldsymbol{\Sigma}\right)\right], \\
\mathbf{m} &= \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}_{\text{w}^\star} \times f_{\text{w}^\star}\left(\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2}\boldsymbol{\xi}, \hat{\mathbf{m}}^\intercal \hat{\mathbf{q}}^{-1}\hat{\mathbf{m}}\right) \mathbf{f}_{\text{w}}\left(\hat{\mathbf{q}}^{1/2}\boldsymbol{\xi}, \hat{\boldsymbol{\Sigma}}\right)\right], \\
\mathbf{q} &= \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}_{\text{w}^\star}\left(\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2}\boldsymbol{\xi}, \hat{\mathbf{m}}^\intercal \hat{\mathbf{q}}^{-1}\hat{\mathbf{m}}\right) \mathbf{f}_{\text{w}}\left(\hat{\mathbf{q}}^{1/2}\boldsymbol{\xi}, \hat{\boldsymbol{\Sigma}}\right)^{2}\right], \\
\boldsymbol{\Sigma} &= \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}_{\text{w}^\star}\left(\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2}\boldsymbol{\xi}, \hat{\mathbf{m}}^\intercal \hat{\mathbf{q}}^{-1}\hat{\mathbf{m}}\right) \partial_{\boldsymbol{\gamma}} \mathbf{f}_{\text{w}}\left(\hat{\mathbf{q}}^{1/2}\boldsymbol{\xi}, \hat{\boldsymbol{\Sigma}}\right)\right],
\end{aligned} \tag{324}$$

where we use the abusive notation $\mathbb{E}_y = \int_{\mathbb{R}} \mathrm{d}y$.

B.1.2.B BAYES-OPTIMAL ESTIMATION

Extremizing the Bayes-optimal free entropy eq. (321), we easily obtain the set of fixed point equations over the scalar parameters $\mathbf{q}_{\text{b}}, \hat{\mathbf{q}}_{\text{b}}$. It can be deduced from eq. (324) using the Nishimori conditions $\mathbf{f}_{\text{w}} = \mathbf{f}_{\text{w}^\star}$, $\mathbf{f}_{\text{out}} = \mathbf{f}_{\text{out}^\star}$, $\mathbf{m} = \mathbf{q}, \boldsymbol{\Sigma} = \boldsymbol{\rho}_{\text{w}^\star} - \mathbf{q}, \hat{\mathbf{m}} = \hat{\mathbf{q}}$ and $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{q}}$ that lead to

$$\begin{aligned}
\hat{\mathbf{q}}_{\text{b}} &= \alpha \mathbb{E}_{y,\boldsymbol{\xi}}\left[\mathcal{Z}_{\text{out}^\star}\left(y, \mathbf{q}_{\text{b}}^{1/2}\boldsymbol{\xi}, \boldsymbol{\rho}_{\text{w}^\star} - \mathbf{q}_{\text{b}}\right) \mathbf{f}_{\text{out}^\star}\left(y, \mathbf{q}_{\text{b}}^{1/2}\boldsymbol{\xi}, \boldsymbol{\rho}_{\text{w}^\star} - \mathbf{q}_{\text{b}}\right)^{\otimes 2}\right], \\
\mathbf{q}_{\text{b}} &= \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}_{\text{w}^\star}\left(\hat{\mathbf{q}}_{\text{b}}^{1/2}\boldsymbol{\xi}, \hat{\mathbf{q}}_{\text{b}}\right) \mathbf{f}_{\text{w}^\star}\left(\hat{\mathbf{q}}_{\text{b}}^{1/2}\boldsymbol{\xi}, \hat{\mathbf{q}}_{\text{b}}\right)^{\otimes 2}\right].
\end{aligned} \tag{325}$$

B.2 RANDOM LABELS - GLM WITH I.I.D DATA

In this section, we present the replica computation of GLM corresponding to the hypothesis class $\mathscr{F}_\varphi$ in eq. (164). We focus on data $\{\mathbf{x}_1^\intercal, \dots, \mathbf{x}_n^\intercal\} = \mathbf{X} \in \mathbb{R}^{n\times d}$, with $\alpha = n/d$, drawn i.i.d from a distribution $\mathrm{P_x}(\mathbf{x}) = \mathscr{N}_{\mathbf{x}}(\mathbf{o}, \mathbf{I}_d)$, and labels $\mathbf{y}$ drawn randomly from $\mathrm{P_y}(.)$. We consider for the moment a generic prior distribution $\mathbf{w} \sim \mathrm{P_w}(.)$ that factorizes, and a component-wise activation function $\varphi(.)$. Defining the linear transformation applied by the model $z_\mu \equiv \frac{1}{\sqrt{d}}\mathbf{w}^\intercal \mathbf{x}_\mu$, we introduce the corresponding cost function of a given sample $(\mathbf{x}_\mu, y_\mu)$ according to $V(y_\mu|z_\mu) = \mathbb{1}\left[y_\mu \neq \varphi(z_\mu)\right]$ which is 0 if the estimator classifies the example correctly (i. e. when $y_\mu = \varphi(z_\mu)$) and 1 otherwise. Finally we define the *constraint function* $\mathscr{C}$ at inverse temperature β

$$\mathscr{C}(\mathbf{y}|\mathbf{z}, \beta) \equiv \prod_{\mu=1}^{n} e^{-\beta V(y_\mu|z_\mu)} = e^{-\beta \mathscr{H}_d(\{\mathbf{y},\mathbf{X}\},\mathbf{w})}, \tag{326}$$

which, denoting the output of the estimator $f_\mathbf{w}(\mathbf{x}_\mu) = \varphi(z_\mu)$, depends explicitly on the Hamiltonian

$$\mathscr{H}_d(\{\mathbf{y},\mathbf{X}\},\mathbf{w}) \equiv \sum_{\mu=1}^{n} \mathbb{1}\left[y_\mu \neq f_\mathbf{w}(\mathbf{x}_\mu)\right]. \tag{327}$$

Notice that at zero temperature the *soft* constraint function $\mathscr{C}$ converges to a *hard* constraint function $\mathscr{C}(\mathbf{y}|\mathbf{z},\beta) \underset{\beta\to\infty}{\longrightarrow} \prod_{\mu=1}^n \mathbb{1}\left[V(y_\mu|z_\mu) = 0\right]$, which tolerates only configurations that satisfy simultaneously all the constraints. In this context, the partition function simply reads

$$\mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta) = \int_{\mathbb{R}^d} \mathrm{dP_w}(\mathbf{w})\, \mathscr{C}(\mathbf{y}|\mathbf{z},\beta). \tag{328}$$

In order to compute the quenched free entropy average, we use the replica trick, see Sec. 4.1.1, and consider the partition function of $r \in \mathbb{N}$ identical copies of the initial system. Assuming there exists an analytical continuation for $r \to 0^+$ and we can revert limits, the averaged free entropy $\Phi(\alpha,\beta) \equiv \lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathbf{y},\mathbf{X}} \log \mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)$ of the initial system becomes

$$\Phi(\alpha,\beta) = \lim_{r\to 0}\left[\lim_{d\to\infty} \frac{1}{d} \frac{\partial \log \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^r\right]}{\partial r}\right], \tag{329}$$

where the replicated partition function writes

$$\begin{aligned}\mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^r\right] &= \int_{\mathbb{R}^n} \mathrm{dP_y}(\mathbf{y}) \int_{\mathbb{R}^{n\times d}} \mathrm{dP_x}(\mathbf{X})\; \mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^r \\ &= \int_{\mathbb{R}^n} \mathrm{dP_y}(\mathbf{y}) \int_{\mathbb{R}^{n\times d}} \mathrm{dP_x}(\mathbf{X}) \\ &\quad \times \prod_{a=1}^{r} \int_{\mathbb{R}^d} \mathrm{dP_w}(\mathbf{w}^a) \prod_{\mu=1}^{n} \int \mathrm{d}z_\mu^a\; \mathscr{C}(y_\mu|z_\mu^a,\beta)\,\delta\left(z_\mu^a - \frac{1}{\sqrt{d}}\mathbf{w}^{a\intercal}\mathbf{x}_\mu\right).\end{aligned} \tag{330}$$

B.2.1 AVERAGE OVER IID INPUTS

As the data matrix is taken (Gaussian) i.i.d, for $(i,j) \in [\![d]\!]^2$, $(\mu,\nu) \in [\![n]\!]^2$, $\mathbb{E}_{\mathbf{X}}[x_{\mu i} x_{\nu j}] = \delta_{\mu\nu}\delta_{ij}$. Hence $z_\mu^a = \frac{1}{\sqrt{d}}\sum_{i=1}^d x_{\mu i} w_i^a$ is the sum of i.i.d random variables. The CLT guarantees that in the large size limit $d \to \infty$, $z_\mu^a \sim \mathcal{N}\left(\mathbb{E}_{\mathbf{X}}[z_\mu^a], \mathbb{E}_{\mathbf{X}}[z_\mu^a z_\mu^b]\right)$, with the two first moments given by

$$\begin{aligned} \mathbb{E}_{\mathbf{X}}[z_\mu^a] &= \frac{1}{\sqrt{d}} \sum_{i=1}^d \mathbb{E}_{\mathbf{X}}[x_{\mu i}] w_i^a = 0 , \\ \mathbb{E}_{\mathbf{X}}[z_\mu^a z_\nu^b] &= \frac{1}{d} \sum_{ij} \mathbb{E}_{\mathbf{X}}[x_{\mu i} x_{\nu j}] w_i^a w_j^b = \left(\frac{1}{d} \sum_{i=1}^d w_i^a w_i^b \right) \delta_{\mu\nu} . \end{aligned} \tag{331}$$

In the following, we introduce the overlap matrix $\mathbf{Q} \equiv \left(\frac{1}{d}\mathbf{w}^a \cdot \mathbf{w}^b\right)_{a,b=1..r} \in \mathbb{R}^{r\times r}$ and we define the replicated vectors $\tilde{\mathbf{z}}_\mu \in \mathbb{R}^r \equiv (z_\mu^a)_{a=1..r}$, $\tilde{\mathbf{w}}_i \equiv (w_i^a)_{a=1..r} \in \mathbb{R}^r$. From the above calculation $\tilde{\mathbf{z}}_\mu$ follows a multivariate Gaussian distribution $\tilde{\mathbf{z}}_\mu \sim \mathrm{P}_{\mathrm{z}}(\tilde{\mathbf{z}}, \mathbf{Q}) \triangleq \mathcal{N}_{\tilde{\mathbf{z}}}(\mathbf{0}_r, \mathbf{Q})$ and $\mathrm{P}_{\mathrm{w}}(\tilde{\mathbf{w}}_i) = \prod_{a=1}^r \mathrm{P}_{\mathrm{w}}(\tilde{w}_i^a)$. Introducing the change of variable and the Fourier representation of the Dirac-δ distribution that involves a new ad-hoc matrix order parameter $\hat{\mathbf{Q}} \in \mathbb{R}^{r\times r}$:

$$\begin{aligned} 1 &= \int_{\mathbb{R}^{r\times r}} \mathrm{d}\mathbf{Q} \prod_{a\leq b} \delta\left(dQ_{ab} - \sum_{i=1}^d w_i^a w_i^b \right) \\ &\propto \int_{\mathbb{R}^{r\times r}} \mathrm{d}\mathbf{Q} \int_{\mathbb{R}^{r\times r}} d\hat{\mathbf{Q}} \exp\left(-\frac{d}{2} \mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right) \right) \exp\left(\frac{1}{2} \sum_{i=1}^d \tilde{\mathbf{w}}_i^\intercal \hat{\mathbf{Q}} \tilde{\mathbf{w}}_i \right) , \end{aligned} \tag{332}$$

the replicated partition function factorizes and becomes an integral over the matrix order parameters $\mathbf{Q}$ and $\hat{\mathbf{Q}}$, that can be evaluated using a Laplace method in the $d \to \infty$ limit,

$$\begin{aligned} \mathbb{E}_{\mathbf{y},\mathbf{X}}\left[\mathcal{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha,\beta)^r\right] &\propto \int \mathrm{d}\mathbf{Q}\, \mathrm{d}\hat{\mathbf{Q}}\, e^{d\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)} \\ &\underset{d\to\infty}{\simeq} \exp\left(d \cdot \mathbf{extr}_{\mathbf{Q},\hat{\mathbf{Q}}} \left\{ \Phi^{(r)}\left(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta\right) \right\} \right) , \end{aligned} \tag{333}$$

where the replica potential is defined by

$$\begin{aligned} \Phi^{(r)}\left(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta\right) &\equiv -\frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right) + \log \Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}}) + \alpha \log \Psi_{\mathrm{out}}^{(r)}(\mathbf{Q},\beta) , \\ \Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}}) &= \int_{\mathbb{R}^r} \mathrm{dP}_{\mathrm{w}}(\tilde{\mathbf{w}})\, e^{\frac{1}{2}\tilde{\mathbf{w}}^\intercal \hat{\mathbf{Q}} \tilde{\mathbf{w}}} , \\ \Psi_{\mathrm{out}}^{(r)}(\mathbf{Q},\beta) &= \int \mathrm{dP}_{\mathrm{y}}(y) \int_{\mathbb{R}^r} \mathrm{dP}_{\mathrm{z}}(\tilde{\mathbf{z}},\mathbf{Q})\, \mathscr{C}(y|\tilde{\mathbf{z}},\beta) . \end{aligned} \tag{334}$$

Finally, using eq. (329) and switching the two limits $r \to 0$ and $d \to \infty$, the quenched free entropy Φ simplifies as an extremization problem

$$\Phi(\alpha,\beta) = \mathbf{extr}_{\mathbf{Q},\hat{\mathbf{Q}}} \left\{ \lim_{r\to 0} \frac{\partial \Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)}{\partial r} \right\} , \tag{335}$$

over general symmetric matrices $\mathbf{Q}$ and $\hat{\mathbf{Q}}$. In the following we assume simple Ansätze for these matrices that allow to obtain analytic expressions in r in order to take the derivative and the limit $r \to 0^+$.

B.2.2 ANNEALED COMPUTATION

We can use the replica calculation (334) to compute the *annealed* free entropy $\Phi^a(\alpha) = \log \mathbb{E}_{\mathbf{y},\mathbf{X}} \left[\mathscr{Z}_d(\{\mathbf{y},\mathbf{X}\},\alpha)\right]$, see Sec. 2.3.5, by assuming there exists a single replica with $r=1$, $\mathbf{Q}=q$ and $\hat{\mathbf{Q}}=\hat{q}$ (Krauth et al., 1989).

$$\Phi^a(\alpha,\beta) = \underset{q,\hat{q}}{\mathbf{extr}} \left\{ -\frac{1}{2} q\hat{q} + \log \Psi_{\mathrm{w}}^a(\hat{q}) + \alpha \log \Psi_{\mathrm{out}}^a(q,\beta) \right\}, \tag{336}$$

with

$$\begin{aligned} \Psi_{\mathrm{w}}^a(\hat{q}) &\equiv \int_{\mathbb{R}} \mathrm{dP_w}(w)\ \exp\left(\frac{1}{2}\hat{q}w^2\right), \\ \Psi_{\mathrm{out}}^a(q,\beta) &\equiv \int_{\mathbb{R}} \mathrm{dP}_y \int_{\mathbb{R}} \mathrm{d}z\ \frac{e^{-\frac{z^2}{2q}}}{\sqrt{2\pi q}} \mathscr{C}(y|z,\beta)\,. \end{aligned} \tag{337}$$

Finally, in the case of binary weights $\mathrm{P}(w) = (\delta(w-1)+\delta(w+1))$ we obtain $\Psi_{\mathrm{w}}^a(\hat{q}) = 2\exp\left(\frac{1}{2}\hat{q}\right)$. Taking the derivative of (336) with respect to $\hat{q}$ we obtain $q=1$ so that the annealed free entropy writes

$$\Phi^a(\alpha,\beta) = \log(2) + \alpha \log\left(\int_{\mathbb{R}} \mathrm{dP_y}(y) \int_{\mathbb{R}} \mathrm{D}z\, \mathscr{C}(y|z,\beta) \right). \tag{338}$$

We can compute therefore the annealed capacity α_a at zero temperature $\beta \to \infty$, such that the annealed entropy $\Phi^a(\alpha, \beta \to \infty)$ vanishes:

$$\alpha_a = \frac{-\log(2)}{\log\left(\int_{\mathbb{R}} \mathrm{dP_y}(y) \int_{\mathbb{R}} \mathrm{D}z\, \mathscr{C}(y|z)\right)}. \tag{339}$$

B.2.3 CHOOSING AN ANSATZ

Back to the quenched average computation in (335), optimizing over the space of matrices is intractable. Therefore, one needs to assume simple Ansätze about the matrices structure to push the computation further, see Sec. **??**, such as the so-called

- Replica Symmetry (RS) Ansatz: $\mathbf{Q}^{(\mathrm{rs})} = (Q-q_0)\mathbf{I}_r + q_0\mathbf{J}_r$
- 1-Step Replica Symmetry Breaking (1RSB) Ansatz: $\mathbf{Q}^{(\mathrm{1rsb})} = (Q-q_1)\mathbf{I}_r + (q_1-q_0)\mathbf{I}_{r/x_0} \otimes \mathbf{J}_{x_0} + q_0\mathbf{J}_r$,
- 2-Step Replica Symmetry Breaking (2RSB) Ansatz:

$$\mathbf{Q}^{(\mathrm{2rsb})} = (Q-q_2)\,\mathbf{I}_r + (q_2-q_1)\,\mathbf{I}_{r/x_1} \otimes \mathbf{J}_{x_1} + (q_1-q_0)\,\mathbf{I}_{r/x_0} \otimes \mathbf{J}_{x_0} + q_0\mathbf{J}_r$$

where $\mathbf{I}_k$ is the identity matrix of size k, and $\mathbf{J}_k$ is the matrix of size k full of ones. Plugging these Ansätze, taking the derivative and the $r \to 0^+$ limit, extremizing over the space of matrices boils down to much simpler optimization problems over a few scalar order parameters, as illustrated in the next sections.

B.2.4 RS FREE ENTROPY FOR I.I.D DATA

Let us compute the free entropy potential $\Phi^{(r)}(\mathbf{Q}, \hat{\mathbf{Q}}, \alpha, \beta)$ in (334) in the RS Ansatz. The latter assumes that all replicas remain equivalent with a common overlap $q_0 = \frac{1}{d}\sum_{i=1}^d w_i^a w_i^b$ for $a \neq b$ and a norm $Q = \frac{1}{d}\sum_{i=1}^d w_i^a w_i^a$, leading to the following expressions for matrices $\mathbf{Q}$ and $\hat{\mathbf{Q}} \in \mathbb{R}^{r\times r}$:

$$\mathbf{Q}^{(\mathrm{rs})} = \begin{pmatrix} Q & q_0 & \cdots & q_0 \\ q_0 & Q & \ddots & \vdots \\ \vdots & \ddots & \ddots & q_0 \\ q_0 & \cdots & q_0 & Q \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{Q}}^{(\mathrm{rs})} = \begin{pmatrix} \hat{Q} & \hat{q}_0 & \cdots & \hat{q}_0 \\ \hat{q}_0 & \hat{Q} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \hat{q}_0 \\ \hat{q}_0 & \cdots & \hat{q}_0 & \hat{Q} \end{pmatrix}. \tag{340}$$

Let us compute separately the terms involved in the functional $\Phi^{(r)}(\mathbf{Q}, \hat{\mathbf{Q}}, \alpha, \beta)$ eq. (334): the first is a trace term, the second a term $\Psi_{\mathrm{w}}^{(r)}$ depends on the prior distribution P_{w} and finally the third $\Psi_{\mathrm{out}}^{(r)}$ on the constraint $\mathscr{C}(y|z)$ in (326).

Trace The trace term can be easily computed as

$$\frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right)\Big|_{\mathrm{rs}} = \frac{1}{2}\left(rQ\hat{Q} + r(r-1)q_0\hat{q}_0\right). \tag{341}$$

Prior integral Evaluated at the RS fixed point, and using a Hubbard-Stratonovich transformation, see A.2, the prior integral can be further simplified

$$\begin{aligned} \Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}})\Big|_{\mathrm{rs}} &= \int \mathrm{dP}_{\mathrm{w}}(\tilde{\mathbf{w}})\, e^{\frac{1}{2}\tilde{\mathbf{w}}^{\intercal}\hat{\mathbf{Q}}^{(\mathrm{rs})}\tilde{\mathbf{w}}} = \int \mathrm{dP}_{\mathrm{w}}(\tilde{\mathbf{w}}) \\ &\times \exp\left(\frac{(\hat{Q}-\hat{q}_0)}{2}\sum_{a=1}^{r}(w^a)^2\right)\exp\left(\hat{q}_0\left(\sum_{a=1}^{r} w^a\right)^2\right) \\ &= \int \mathrm{D}\xi_0 \left[\int \mathrm{dP}_{\mathrm{w}}(w)\, \exp\left(\left(\frac{(\hat{Q}-\hat{q}_0)}{2}w^2 + \xi_0\sqrt{\hat{q}_0}w\right)\right)\right]^r. \end{aligned} \tag{342}$$

Constraint integral Recall the vector $\tilde{\mathbf{z}} \sim \mathrm{P}_{\mathrm{z}} \triangleq \mathcal{N}(\mathbf{0}, \mathbf{Q})$ follows a Gaussian distribution with zero mean and covariance matrix $\mathbf{Q}$. In the RS Ansatz, the covariance can be rewritten as a linear combination of the identity $\mathbf{I}_r$ and $\mathbf{J}_r$: $\mathbf{Q}|_{\mathrm{rs}} = (Q-q_0)\mathbf{I}_r + q_0\mathbf{J}_r$, that allows to split the variable $z^a =$

$\sqrt{Q-q_0}u^a + \sqrt{q_0}\xi_0$ with $\xi_0 \sim \mathcal{N}(0,1)$ and $\forall a \in [\![r]\!]$, $u_a \sim \mathcal{N}(0,1)$. The constraint integral finally reads

$$\begin{aligned}
\Psi_{\text{out}}^{(r)}(\mathbf{Q},\beta)\Big|_{\text{rs}} &= \int \mathrm{dP}_y(y) \int_{\mathbb{R}^r} \mathrm{dP}_{\text{z}}(\tilde{\mathbf{z}})\, \mathscr{C}(y|\tilde{\mathbf{z}},\beta) \\
&= \int \mathrm{dP}_y(y) \int \mathrm{d}\xi_0 \int \prod_{a=1}^{r} \mathrm{d}u^a\, \mathscr{C}\left(y|\sqrt{Q-q_0}u^a + \sqrt{q_0}\xi_0,\beta\right) \\
&= \int \mathrm{dP}_y(y) \int \mathrm{d}\xi_0 \left[\int \mathrm{d}z\, \mathscr{C}\left(y|\sqrt{Q-q_0}z + \sqrt{q_0}\xi_0,\beta\right)\right]^r .
\end{aligned} \tag{343}$$

Finally, putting pieces together, the functional $\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)$ taken at the RS fixed point has an explicit formula and dependency in r:

$$\begin{aligned}
&\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)\Big|_{\text{rs}} \underset{r\to 0}{\simeq} -\frac{1}{2}\left(rQ\hat{Q} + r(r-1)q_0\hat{q}_0\right) \\
&+ r\int \mathrm{d}\xi_0 \log\left(\int \mathrm{dP}_{\text{w}}(w)\, \exp\left\{\left(\frac{(\hat{Q}-\hat{q}_0)}{2}w^2 + \xi_0\sqrt{\hat{q}_0}w\right)\right\}\right) \\
&+ r\alpha \int \mathrm{dP}_y(y) \int \mathrm{d}\xi_0 \log\left(\int \mathrm{d}z\, \mathscr{C}\left(y|\sqrt{Q-q_0}z + \sqrt{q_0}\xi_0,\beta\right)\right).
\end{aligned} \tag{344}$$

B.2.4.A SUMMARY OF RS FREE ENTROPY - GENERAL CASE

Taking the derivative with respect to r and the $r \to 0^+$ limit, the RS free entropy has a simple expression

$$\begin{aligned}
\Phi^{(\text{rs})}(\alpha,\beta) &= \underset{q_0,\hat{q}_0}{\mathbf{extr}}\left\{-\frac{1}{2}Q\hat{Q} + \frac{1}{2}q_0\hat{q}_0 + \Psi_{\text{w}}^{(\text{rs})}(\hat{q}_0) + \alpha\Psi_{\text{out}}^{(\text{rs})}(q_0,\beta)\right\}, \\
\Psi_{\text{w}}^{\text{rs}}(\hat{q}_0) &\equiv \mathbb{E}_{\xi_0}\log\mathbb{E}_w\left[\exp\left(\frac{(\hat{Q}-\hat{q}_0)}{2}w^2 + \xi_0\sqrt{\hat{q}_0}w\right)\right], \\
\Psi_{\text{out}}^{\text{rs}}(q_0,\beta) &\equiv \mathbb{E}_y\,\mathbb{E}_{\xi_0}\log\mathbb{E}_z\left[\mathscr{C}\left(y|\sqrt{Q-q_0}z + \sqrt{q_0}\xi_0,\beta\right)\right],
\end{aligned} \tag{345}$$

where $\xi_0, z \sim \mathcal{N}(0,1)$, $w \sim \mathrm{P}_{\text{w}}(.)$, $y \sim \mathrm{P}_y(.)$ and $Q = \hat{Q} = 1$.

B.2.4.B SUMMARY OF RS FREE ENTROPY - SPHERICAL CASE

In the spherical (or equivalently in the Gaussian case with a correctly defined variance) such that the weights verify $\|\tilde{\mathbf{w}}\|_2^2 = d$, $\Psi_{\text{w}}^{(r)}(\hat{\mathbf{Q}})$ in eq. (334) can be directly integrated

$$\Psi_{\text{w}}^{(r)}(\hat{\mathbf{Q}}) = \int_{\|\tilde{\mathbf{w}}\|_2^2=d} \mathrm{d}\tilde{\mathbf{w}}\, \exp\left(\frac{1}{2}\tilde{\mathbf{w}}^\intercal\hat{\mathbf{Q}}\tilde{\mathbf{w}}\right) = -\frac{1}{2}\log\det\left(2\pi(\mathbf{I}_r + \hat{\mathbf{Q}})\right). \tag{346}$$

Besides, taking the derivative of eq. (334) with respect to $\hat{\mathbf{Q}}$ we obtain $\mathbf{Q}^{-1} = (\mathbf{I}_r + \hat{\mathbf{Q}})$. Injecting it, we can get rid of $\hat{\mathbf{Q}}$ and obtain

$$\Phi^{(r)}\left(\mathbf{Q}, \alpha, \beta\right) \equiv \frac{1}{2} \log\det\left(2\pi\mathbf{Q}\right) + \alpha \log \Psi_{\text{out}}^{(r)}(\mathbf{Q}, \beta)\,. \tag{347}$$

Determinant The above determinant reads in the RS Ansatz

$$\frac{1}{2}\det\left(\mathbf{Q}\right)\Big|_{\text{rs}} \simeq \frac{r}{2}\left(\log(1-q_0) + \frac{q_0}{1+(r-1)q_0} + \dots\right), \tag{348}$$

so that it leads to the RS free entropy

$$\Phi^{(\text{rs})}(\alpha, \beta) = \mathbf{extr}_{q_0}\left\{\frac{1}{2(1-q_0)} + \frac{1}{2}\log(2\pi) + \frac{1}{2}\log(1-q_0)\right. \\ \left. + \alpha\Psi_{\text{out}}^{\text{rs}}(q_0, \beta)\right\}, \tag{349}$$

with $\Psi_{\text{out}}^{\text{rs}}$ defined in eq. (345).

B.2.5 RS STABILITY

B.2.5.A DE ALMEIDA THOULESS RS STABILITY

The stability of a given saddle point Ansatz is related to the positivity the Hessian of the functional $-\Phi^{(r)}$. Following (Almeida et al., 1978; Gardner et al., 1988; Engel et al., 2001), the stability analysis leads to computing the first unstable eigenvalues of the Hessian, the so-called *replicons eigenvalues*. In the context of the RS Ansatz λ_3^A and λ_3^B can be expressed as functions of $\{g_i^w, f_i^z\}_{i=0}^2$ defined in Chap. 6 - eq. (136):

$$\begin{aligned} \lambda_3^A(q_0) &= \frac{1}{(Q-q_0)^2}\mathbb{E}_{\xi_0}\left[\frac{\left(f_0^z(f_0^z - f_2^z) + (f_1^z)^2\right)^2}{(f_0^z)^4}(\xi_0, q_0)\right], \\ \lambda_3^B(\hat{q}_0) &= \mathbb{E}_{\xi_0}\left[\frac{\left(g_0^w g_2^w - (g_1^w)^2\right)^2}{(g_0^w)^4}(\xi_0, \hat{q}_0)\right], \end{aligned} \tag{350}$$

for $\xi_0 \sim \mathcal{N}(0,1)$. The instability dAT-line is defined when the determinant of the Hessian vanishes, i. e. when the first negative eigenvalues appear. This translates as an implicit equation over α, where $q_0, \hat{q}_0$ are solution of the saddle point equations eq. (134) at $\alpha = \alpha_{\text{at}}$:

$$\frac{1}{\alpha_{\text{at}}} = \lambda_3^A\left(q_0(\alpha_{\text{at}}), \beta\right) \times \lambda_3^B\left(\hat{q}_0(\alpha_{\text{at}})\right). \tag{351}$$

However for $\alpha < \alpha_{\text{at}}$, $(q_0, \hat{q}_0) = (0,0)$ is the only solution. Defining for $z \sim \mathcal{N}(0,1)$ and $w \sim \text{P}_\text{w}$

$$\tilde{f}_i^z \equiv \mathbb{E}_z\left[z^i \varphi(z)\right], \qquad \tilde{g}_i^w \equiv \mathbb{E}_w\left[w^i \exp\left(w^2/2\right)\right], \tag{352}$$

, this expression can be simplified in the case where the prior distribution $\mathrm{P_w}$ and the activation φ are symmetric. In fact the symmetry imposes $\tilde{f}_1^z = 0$ and $\tilde{g}_1^w = 0$ and the condition simplifies to

$$\frac{1}{\alpha_{\mathrm{at}}} = \left(\frac{\tilde{f}_2^z - \tilde{f}_0^z}{\tilde{f}_0^z} \right)^2 \left(\frac{\tilde{g}_2^w}{\tilde{g}_0^w} \right)^2 . \tag{353}$$

B.2.5.B EXISTENCE AND STABILITY OF THE RS FIXED POINT

We provide an alternative approach to get the instability condition of the RS solution for symmetric prior $\mathrm{P_w}$ and activation φ. In this symmetric case, the stability can be derived from the existence and stability of the symmetric fixed point $(q_0, \hat{q}_0) = (0,0)$. Let us define

$$\begin{aligned} F(q_0) &\equiv \alpha \mathbb{E}_{\xi_0} \left[\frac{(f_1^z)^2 - 2\xi_0 \sqrt{q_0} f_0^z f_1^z + q_0 \xi_0^2 (f_0^z)^2}{(1-q_0)^2 (f_0^z)^2} (\xi_0, q_0) \right] , \\ G(\hat{q}_0) &\equiv \mathbb{E}_{\xi_0} \left[\frac{g_2^w - \xi_0 \hat{q}_0^{-1/2} g_1^w}{g_0^w} (\xi_0, \hat{q}_0) \right] . \end{aligned} \tag{354}$$

In fact the saddle point equations at the RS fixed point eq. (345) can be written using the functions F, G, and can be reduced to a single fixed point equation over q_0

$$\begin{cases} q_0 = G(\hat{q}_0) \, , \\ \hat{q}_0 = F(q_0) \, , \end{cases} \qquad \Rightarrow \qquad q_0 = G \circ F(q_0) \equiv H(q_0) \, . \tag{355}$$

The RS stability of the fixed point $(q_0, \hat{q}_0) = (0,0)$ can be analyzed from the above fixed point equation eq. (355). Computing F, F', G, G' in the limit

$(q_0, \hat{q}_0) \to (0,0)$, expanding $\{f_i^z, g_i^w\}_i$ as functions of $\{\tilde{f}_i^z, \tilde{g}_i^w\}_i$ and finally using the symmetry conditions $\tilde{f}_1^z = 0$ and $\tilde{g}_1^w = 0$, we finally obtain

$$\begin{aligned}
F(q_0) \underset{q_0 \to 0}{=} \alpha & \left[\left(\frac{\tilde{f}_1^z}{\tilde{f}_0^z} \right)^2 + q_0 \left(\frac{(\tilde{f}_2^z - \tilde{f}_0^z)^2}{(\tilde{f}_0^z)^2} + 3 \frac{(\tilde{f}_1^z)^4}{(\tilde{f}_0^z)^4} \right. \right. \\
& \left. \left. - 4 \frac{(\tilde{f}_1^z)^2 (\tilde{f}_2^z - \tilde{f}_0^z)}{(\tilde{f}_0^z)^3} \right) + \Theta(q_0^2) \right] \sim \alpha q_0 \left(\frac{\tilde{f}_2^z - \tilde{f}_0^z}{\tilde{f}_0^z} \right)^2 \underset{q_0 \to 0}{\longrightarrow} 0 \,, \\
\partial_{q_0} F(q_0) \underset{q_0 \to 0}{=} \alpha & \left[\left(\frac{\tilde{f}_2^z - \tilde{f}_0^z}{\tilde{f}_0^z} \right)^2 + \left(\frac{\tilde{f}_1^z}{\tilde{f}_0^z} \right)^2 \left(3 \frac{(\tilde{f}_1^z)^2}{(\tilde{f}_0^z)^2} - 4 \frac{(\tilde{f}_2^z - \tilde{f}_0^z)}{\tilde{f}_0^z} \right) \right. \\
& \left. + \Theta(q_0) \right] \underset{q_0 \to 0}{\longrightarrow} \alpha \left(\frac{\tilde{f}_2^z - \tilde{f}_0^z}{\tilde{f}_0^z} \right)^2 \,, \\
G(\hat{q}_0) \underset{\hat{q}_0 \to 0}{=} & \left(\frac{\tilde{g}_1^w}{\tilde{g}_0^w} \right)^2 + \hat{q}_0 \left(\left(\frac{\tilde{g}_2^w}{\tilde{g}_0^w} \right)^2 + \frac{\tilde{g}_1^w}{\tilde{g}_0^w} \left(3 \left(\frac{\tilde{g}_1^w}{\tilde{g}_0^w} \right)^3 - 4 \frac{\tilde{g}_1^w \tilde{g}_2^w}{(\tilde{g}_0^w)^2} \right) \right) \\
& + \Theta(\hat{q}_0^{3/2}) \underset{\hat{q}_0 \to 0}{\longrightarrow} 0 \,, \\
\partial_{\hat{q}_0} G(\hat{q}_0) \underset{\hat{q}_0 \to 0}{=} & \left(\frac{\tilde{g}_2^w}{\tilde{g}_0^w} \right)^2 + \frac{\tilde{g}_1^w}{\tilde{g}_0^w} \left(3 \left(\frac{\tilde{g}_1^w}{\tilde{g}_0^w} \right)^3 - 4 \frac{\tilde{g}_1^w \tilde{g}_2^w}{(\tilde{g}_0^w)^2} \right) + \Theta(\sqrt{\hat{q}_0}) \\
& \underset{\hat{q}_0 \to 0}{\longrightarrow} \left(\frac{\tilde{g}_2^w}{\tilde{g}_0^w} \right)^2 \,.
\end{aligned}$$

Finally, the existence and stability conditions of the fixed point $(q_0, \hat{q}_0) = (0,0)$ translate as an explicit condition over α that implicitly defines $\alpha_{\rm at}$

$$\begin{cases} H(q_0) = G \circ F(q_0) \underset{q_0 \to 0}{\to} 0 \\ \left. \frac{\partial H}{\partial q_0} \right|_{q_0 = 0} = \frac{\partial G}{\partial \hat{q}_0} \times \left. \frac{\partial F}{\partial q_0} \right|_{q_0 = 0} \leq 1 \,, \end{cases} \Rightarrow \alpha \leq \left[\left(\frac{\tilde{f}_2^z - \tilde{f}_0^z}{\tilde{f}_0^z} \right)^2 \left(\frac{\tilde{g}_2^w}{\tilde{g}_0^w} \right)^2 \right]^{-1} \equiv \alpha_{\rm at} \,. \tag{356}$$

B.2.6 1RSB FREE ENTROPY FOR I.I.D DATA

The free entropy potential eq. (334) can also be evaluated at the simplest non trivial fixed point: the one-step Replica Symmetry Breaking Ansatz (1RSB), see Sec. ??. Instead of assuming that replicas are equivalent, it states that the symmetry between the replicas is broken and that the replicas are clustered in different *states*, with inner-overlap q_1 and outer-overlap q_0. Translating this analytically, the matrices can be expressed as function of the Parisi parameter x_0, which controls the size of the clusters:

$$\begin{aligned}
\mathbf{Q}^{(\rm 1rsb)} &= q_0 \mathbf{J}_r + (q_1 - q_0) \mathbf{I}_{\frac{r}{x_0}} \otimes \mathbf{J}_{x_0} + (Q - q_1) \mathbf{I}_r \,, \\
\hat{\mathbf{Q}}^{(\rm 1rsb)} &= \hat{q}_0 \mathbf{J}_r + (\hat{q}_1 - \hat{q}_0) \mathbf{I}_{\frac{r}{x_0}} \otimes \mathbf{J}_{x_0} + (\hat{Q} - \hat{q}_1) \mathbf{I}_r \,.
\end{aligned} \tag{357}$$

Trace term Again, the trace term can be easily computed

$$\frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}\hat{\mathbf{Q}}\right)\Big|_{\text{1rsb}} = \frac{1}{2}\left(rQ\hat{Q} + r(x_0-1)q_1\hat{q}_1 + r(r-x_0)q_0\hat{q}_0\right). \tag{358}$$

Prior integral To decouple the replicas with different overlaps q_0, q_1, and using Hubbard-Stratonovich transformations in Appendix. A.2, the prior integral can be written

$$\begin{aligned}
\Psi_{\mathrm{w}}^{(r)}(\hat{\mathbf{Q}})\Big|_{\text{1rsb}} &= \int_{\mathbb{R}^r} \mathrm{dP_w}(\tilde{\mathbf{w}}) \exp\left(\frac{(\hat{Q}-\hat{q}_1)}{2}\sum_{a=1}^{r}(w^a)^2\right.\\
&\quad \left.+\frac{(\hat{q}_1-\hat{q}_0)}{2}\sum_{k=1}^{\frac{r}{x_0}}\sum_{a,b=(k-1)x_0+1}^{kx_0} w^a w^b + \frac{\hat{q}_0}{2}\left(\sum_{a=1}^{r} w^a\right)^2\right)\\
&= \int \mathrm{D}\xi_0 \left[\int \mathrm{D}\xi_1\right.\\
&\quad \left.\left[\int \mathrm{dP_w}(w) \exp\left(\frac{(\hat{Q}-\hat{q}_1)}{2}w^2 + \left(\sqrt{\hat{q}_0}\xi_0 + \sqrt{\hat{q}_1-\hat{q}_0}\xi_1\right)w\right)\right]^{x_0}\right]^{\frac{r}{x_0}},
\end{aligned} \tag{359}$$

with $\xi_0, \xi_1 \sim \mathcal{N}(0,1)$.

Constraint integral The replicated vector $\tilde{\mathbf{z}} \sim \mathrm{P_z}(.) \triangleq \mathcal{N}_{\mathbf{z}}\left(\mathbf{0}, \mathbf{Q}^{(\text{1rsb})}\right)$ follows a Gaussian vector with zero mean and covariance matrix $\mathbf{Q}^{(\text{1rsb})}$ that can be decomposed in a sum of normal Gaussian vectors $\xi_0 \sim \mathcal{N}(0,1)$, $\forall k \in [\![1; \frac{r}{x_0}]\!]$, $\xi_k \sim \mathcal{N}(0,1)$ and $\forall a \in [\![(k-1)x+1; kx]\!]$, $u_a \sim \mathcal{N}(0,1)$:

$$z^a = \sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_k + \sqrt{Q-q_1}u_a.$$

Finally, the constraint integral reads

$$\begin{aligned}
&\Psi_{\text{out}}^{(r)}(Q,\beta)\Big|_{\text{1rsb}}\\
&= \int \mathrm{dP}_y(y) \int \mathrm{D}\xi_0 \int \prod_{k=1}^{\frac{r}{x_0}} \mathrm{D}\xi_k\\
&\quad \times \int \prod_{a=(k-1)x+1}^{kx} \mathrm{D}u_a\, \mathscr{C}\left(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_k + \sqrt{Q-q_1}u_a, \beta\right)\\
&= \int \mathrm{dP}_y(y) \int \mathrm{D}\xi_0\\
&\quad \times \left[\int \mathrm{D}\xi_1 \left[\int \mathrm{D}z\, \mathscr{C}\left(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_1 + \sqrt{Q-q_1}z, \beta\right)\right]^{x}\right]^{\frac{r}{x_0}}.
\end{aligned} \tag{360}$$

B.2.6.a Summary of the 1RSB free entropy - general case

Gathering the previous computations eq. (358, 359, 360), the functional $\Phi^{(r)}$ evaluated at the 1RSB fixed point reads:

$$\begin{aligned}&\Phi^{(r)}(\mathbf{Q},\hat{\mathbf{Q}},\alpha,\beta)\Big|_{\text{1rsb}}\\ &\underset{r\to 0}{\simeq} -\frac{1}{2}\left(rQ\hat{Q}+r(x_0-1)q_1\hat{q}_1+r(r-x_0)q_0\hat{q}_0\right)\\ &\qquad\qquad +r\Psi_{\text{w}}^{(\text{1rsb})}(\hat{\mathbf{q}})+r\alpha\Psi_{\text{out}}^{(\text{1rsb})}(\mathbf{q},\beta)\end{aligned}$$

with

$$\begin{aligned}&\Psi_{\text{w}}^{(\text{1rsb})}(\hat{\mathbf{q}},x_0)\equiv\frac{1}{x_0}\mathbb{E}_{\xi_0}\\ &\log\left(\mathbb{E}_{\xi_1}\mathbb{E}_w\left[\exp\left(\frac{(\hat{Q}-\hat{q}_1)}{2}w^2+\left(\sqrt{\hat{q}_0}\xi_0+\sqrt{\hat{q}_1-\hat{q}_0}\xi_1\right)w\right)\right]^{x_0}\right),\\ &\Psi_{\text{out}}^{(\text{1rsb})}(\mathbf{q},x_0,\beta)\equiv\frac{1}{x_0}\mathbb{E}_y\mathbb{E}_{\xi_0} \qquad (361)\\ &\qquad\log\left(\mathbb{E}_{\xi_1}\mathbb{E}_z\left[\mathscr{C}(y|\sqrt{q_0}\xi_0+\sqrt{q_1-q_0}\xi_1+\sqrt{Q-q_1}z,\beta)\right]_0^x\right),\end{aligned}$$

where $\mathbf{q}=(q_0,q_1)$, $\hat{\mathbf{q}}=(\hat{q}_0,\hat{q}_1)$, $\xi_0,\xi_1,z\sim\mathcal{N}(0,1)$, $w\sim \mathrm{P}_{\text{w}}(.)$, $y\sim\mathrm{P}_y(.)$ and $Q=\hat{Q}=1$. Finally taking the derivative with respect to r and the limit $r\to 0^+$, we obtain the 1RSB free entropy

$$\begin{aligned}\Phi^{(\text{1rsb})}(\alpha,\beta)=\underset{\mathbf{q},\hat{\mathbf{q}},x_0}{\textbf{extr}}\Bigg\{&\frac{1}{2}\left(q_1\hat{q}_1-Q\hat{Q}\right)+\frac{x_0}{2}\left(q_0\hat{q}_0-q_1\hat{q}_1\right)\\ &+\Psi_{\text{w}}^{(\text{1rsb})}(\hat{\mathbf{q}},x_0)+\alpha\Psi_{\text{out}}^{(\text{1rsb})}(\mathbf{q},x_0,\beta)\Bigg\}. \qquad (362)\end{aligned}$$

B.2.6.b Summary of the 1RSB free entropy - spherical case

In the 1RSB, the simplification eq. (347) remains valid. Therefore, we can simply compute the determinant in the 1RSB Ansatz.

Determinant

$$\begin{aligned}\det(\mathbf{Q})|_{\text{1rsb}}&=(rq_0+x_0(q_1-q_0)+(1-q_1))\\ &\qquad\times(1-q_1)^{r-r/x_0}\times(x_0(q_1-q_0)+(1-q_1))^{r/x_0-1},\end{aligned}$$

so that

$$\begin{aligned}\log\det(\mathbf{Q})|_{\text{1rsb}}\simeq r\Bigg(&\frac{x_0-1}{x_0}\log(1-q_1)+\\ &\frac{1}{x_0}\log\left(x_0(q_1-q_0)+(1-q_1)\right)+\frac{q_0}{x_0(q_1-q_0)+(1-q_1)}\Bigg).\end{aligned}$$

Using the above expression for the determinant and the simplified replica potential in eq. (347) we obtain

$$\Phi^{(1\text{rsb})}(\alpha,\beta) = \textbf{extr}_{q_0,q_1,x}\left\{\frac{1}{2}\log(2\pi) + \frac{x-1}{2x}\log(1-q_1) + \right.$$
$$+\frac{1}{2x}\log\left(x(q_1-q_0)+(1-q_1)\right) + \frac{q_0}{2\left(x(q_1-q_0)+(1-q_1)\right)} \tag{363}$$
$$\left. +\alpha\Psi_{\text{out}}^{(1\text{rsb})}(\mathbf{q},\beta)\right\}.$$

B.2.7 GROUND STATE ENERGIES

We focus on the particular case of the spherical perceptron with parameters $\mathbf{w} \in \mathbb{R}^d$ lying on the sphere and verifying $\|\mathbf{w}\|_2^2 = d$.

RS capacity In the case of the step-perceptron activation function $\varphi(z) = \theta(z-\kappa)$ for $\kappa \geq 0$, we can compute the capacity α_c taking the extremization over q_0 in eq. (349):

$$q_0 = -2\alpha(1-q_0)^2\partial_{q_0}\Psi_{\text{out}}^{(\text{rs})}(q_0) \simeq 2\alpha(1-q_0)^2\int_{-\infty}^{\kappa}\text{d}t\,\frac{(\kappa-t)^2}{(1-q_0)^2}\,. \tag{364}$$

At the critical capacity, we have $q_0 \to 1$, which leads to the expression

$$\alpha_c = \left(\int_{-\infty}^{\kappa}\text{d}t\,(\kappa-t)^2\right)^{-1}. \tag{365}$$

Notice that for $\kappa = 0$, this approach performed in (Gardner et al., 1988) naturally leads to Cover's result $\alpha_c = 2$ (Cover, 1965). Above this capacity α_c, the constraints cannot be satisfied simultaneously and the ground state energy is necessarily positive.

B.2.7.A RS GROUND STATE ENERGY

To compute the ground state energy, we first need to take both limits $q_0 \to 1$ and $\beta \to \infty$, keeping the product $\chi = \beta(Q-q_0)$ finite (Majer et al., 1993;

Erichsen et al., 1993; Whyte et al., 1996). Recall eq. (345), we obtain using the definition of $\mathscr{C}$ in (326)

$$\begin{aligned}\Psi_{\text{out}}^{\text{rs}}(q_0,\beta) &\equiv \mathbb{E}_y\, \mathbb{E}_{\xi_0} \log \mathbb{E}_z \left[\mathscr{C}\left(y\middle|\sqrt{Q-q_0}z+\sqrt{q_0}\xi_0,\beta\right)\right] \\ &= \int \mathrm{dP_y}(y) \int \mathrm{D}\xi_0 \, \log\left(\int \mathrm{d}z\, \mathscr{N}_z\left(\sqrt{q}_0\xi_0, Q-q_0\right) e^{-\beta V(y|z)}\right) \\ &\underset{(q_0,\beta)\to(1,\infty)}{\simeq} -\frac{1}{2}\log(2\pi(Q-q_0)) \\ &\qquad -\beta \int \mathrm{dP_y}(y) \int \mathrm{D}\xi_0 \, \min_z \left[V(y|z) + \frac{(z-\xi_0)^2}{2\chi}\right].\end{aligned} \tag{366}$$

Taking the limits $q_0 \to 1$, $\beta \to \infty$ in eq. (349), we obtain to the RS ground state energy of the spherical perceptron

$$e_{\text{gs}}^{\text{rs}} = \textbf{extr}_\chi \left\{-\frac{1}{2\chi} + \alpha \mathbb{E}_{y,\xi_0} \min_z \left[V(y|z) + \frac{(z-\xi_0)^2}{2\chi}\right]\right\} \tag{367}$$

Application to the step-perceptron For the step function $V(y|z) = \theta(\kappa - z)$ with $\mathrm{P_y}(y) = \delta(y-1)$ and $\kappa \geq 0$, it leads to the Gardner expression (Gardner et al., 1988)

$$e_{\text{gs}}^{\text{rs}} = \textbf{extr}_\chi \left\{-\frac{1}{2\chi} + \alpha\left(\int_{-\infty}^{\kappa-\sqrt{2\chi}} \mathrm{D}\xi_0 + \int_{\kappa-\sqrt{2\chi}}^{\kappa} \mathrm{D}\xi_0 \, \frac{(\xi_0-\kappa)^2}{2\chi}\right)\right\} \tag{368}$$

B.2.7.B 1RSB GROUND STATE ENERGY

To compute the ground state energy in the 1RSB Ansatz, we take similarly the limits $q_1 \to 1$, $\beta \to \infty$ and $x_0 \to 0$, keeping the products $\chi \equiv \beta(Q-q_1)$ and $\omega_0 \equiv x_0\beta$ finite (Whyte et al., 1996), with $\Delta q = 1 - q_0$

$$\begin{aligned}&\Psi_{\text{out}}^{(\text{1rsb})}(\mathbf{q},\beta) \equiv \\ &\frac{1}{x_0}\mathbb{E}_y\, \mathbb{E}_{\xi_0} \log\left(\mathbb{E}_{\xi_1}\mathbb{E}_z\left[\mathscr{C}(y|\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_1 + \sqrt{Q-q_1}z,\beta)\right]^{x_0}\right) \\ &= \frac{1}{x_0}\int \mathrm{dP_y}(y) \int \mathrm{D}\xi_0 \, \log \\ &\qquad \int \mathrm{D}\xi_1 \, \left(\int \mathrm{d}z\, \mathscr{N}_z(\sqrt{q_0}\xi_0 + \sqrt{q_1-q_0}\xi_1, 1-q_1)\, e^{-\beta V(y|z)}\right)^{x_0} \\ &\simeq \frac{1}{x_0}\int \mathrm{dP_y}(y) \int \mathrm{D}\xi_0 \, \log \\ &\qquad \int \mathrm{D}\xi_1 \, e^{-x_0\beta \min_z \left[V(y|z) + \frac{1}{2\beta(1-q_1)}\left(z-\sqrt{q_0}\xi_0-\sqrt{q_1-q_0}\xi_1\right)^2\right]}.\end{aligned}$$

Finally, taking $q_1 \to 1$ with $\beta \to \infty$ and $x \to 0$ in eq. (364), defining $\Omega_0 \equiv \frac{\omega_0}{\chi}$, we obtain the 1RSB ground state energy

$$e_{\rm gs}^{(\rm 1rsb)} = \textbf{extr}_{\chi,\Omega_0,q_0} \left\{ \frac{1}{2\Omega_0\chi}\log\left(1+\Omega_0\Delta q\right) + \frac{q_0}{2\chi\left(1+\Omega_0\Delta q\right)} \right. \\ \left. + \frac{\alpha}{\chi\Omega_0}\mathbb{E}_{\xi_0}\log\mathbb{E}_{\xi_1}e^{-\Omega_0\chi\min_z\left[V(y|z)+\frac{1}{2\chi}\left(z-\sqrt{q_0}\xi_0-\sqrt{\Delta q}\xi_1\right)^2\right]} \right\}. \tag{369}$$

B.2.7.C 2RSB GROUND STATE ENERGY $e_{\rm gs}$

Similarly taking $q_2 \to 1$ with $\beta \to \infty$, we define $\Omega_0 \equiv \frac{x_0\beta}{\chi}, \Omega_1 \equiv \frac{x_1\beta}{\chi}$ and we obtain similarly the 2RSB ground state energy of the spherical perceptron (Whyte et al., 1996)

$$e_{\rm gs}^{(\rm 2rsb)} = \textbf{extr}_{\chi,\Omega_1,\Omega_0,q_1,q_0,} \left\{ \frac{q_0}{2\chi(1+\Omega_1(1-q_1)+\Omega_0(q_1-q_0))} \right. \\ + \frac{1}{2\Omega_1\chi}\log(1+\Omega_1(1-q_1)) \\ + \frac{1}{2\Omega_0\chi}\log\left(1+\frac{\Omega_0(q_1-q_0)}{1+\Omega_1(1-q_1)}\right) + \frac{\alpha}{\chi\Omega_0}\mathbb{E}_{\xi_0} \\ + \log\mathbb{E}_{\xi_1}\left[\mathbb{E}_{\xi_2}\exp\left(-\Omega_1\chi\min_z\left[V(y|z)\right.\right.\right. \\ \left.\left.\left.\left. + \frac{1}{2\chi}\left(z-\sqrt{q_0}\xi_0-\sqrt{q_1-q_0}\xi_1-\sqrt{1-q_1}\xi_2\right)^2\right]\right)\right]^{\Omega_0/\Omega_1}\right\} \tag{370}$$

and notice that taking $q_1 = q_0, x_0 = x_1$ we recover the 1RSB expression.

C

AMP DERIVATION - COMMITTEE MACHINE

The AMP algorithm can be seen as Taylor expansion of the loopy BP approach (Mézard et al., 1987; Mézard et al., 2009; Wainwright et al., 2008), similar to the so-called TAP equation in spin glass theory (Thouless et al., 1977). While the behavior of AMP can be rigorously studied (Bayati et al., 2011b; Javanmard et al., 2013; Bayati et al., 2015), it is useful and instructive to see how the derivation can be performed in the framework of BP and the cavity method, as was pioneered in (Mézard, 1989) for the single layer problem. The derivation uses the GAMP notations of (Rangan, 2011) and follows closely the one of (Zdeborová et al., 2016a). The computation is presented for the committee machine hypothesis class, which is the vectorized version of the GLM, with $K \geq 1$ vectorial parameters $\mathbf{W} = \{\mathbf{w}_k\}_{k=1}^{K} \in \mathbb{R}^{d \times K}$.

C.1 FACTOR GRAPH AND BP EQUATIONS

As a central illustration, we present the instructive derivation of the rBP equations starting with the BP equations in the context of committee machines, already discussed in Appendix. B.1.1. We recall the JPD

$$\mathrm{P}_d\left(\mathbf{W}|\mathbf{y},\mathbf{X}\right) = \frac{\mathrm{P}_{\mathrm{out}}(\mathbf{y}|\mathbf{Z})\mathrm{P}_{\mathrm{w}}(\mathbf{W})}{\mathcal{Z}_d(\mathbf{y},\mathbf{X})} = \frac{\prod_{\mu=1}^{n} \mathrm{P}_{\mathrm{out}}(y_\mu|\mathbf{z}_\mu)\prod_{i=1}^{d} \mathrm{P}_{\mathrm{w}}(\mathbf{w}_i)}{\mathcal{Z}_d(\mathbf{y},\mathbf{X})}\,, \tag{371}$$

where we defined $\mathbf{Z} = \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{Z} \in \mathbb{R}^{n\times K}$ and we assume that the channel and prior distributions factorize over factors $\mathrm{P}_{\mathrm{out}}(y_\mu|\mathbf{z}_\mu)$ and variables $\mathrm{P}_{\mathrm{w}}(\mathbf{w}_i)$.

C.1.1 FACTOR GRAPH

The posterior distribution may be represented by the following bipartite factor graph in Fig. 66. In the following, we attach a set of *messages* $\{m_{i\to\mu}, \tilde{m}_{\mu\to i}\}_{i=1..n}^{\mu=1..m}$ to the edges of this bipartite factor graph. These messages correspond to the marginal probabilities of $\mathbf{w}_i \in \mathbb{R}^K$ if we remove the edges $(i \to \mu)$ or $(\mu \to i)$. We define the auxiliary variable $\mathbf{z}_\mu = \frac{1}{\sqrt{d}}\mathbf{x}_\mu^\intercal \mathbf{W} \in \mathbb{R}^K$ which is $\Theta(1)$ thanks to the pre-factor rescaling $1/\sqrt{d}$. This scaling is crucial as it allows the BP equations to hold true even though the factor graph is not tree-like and is instead fully connected with short loops.

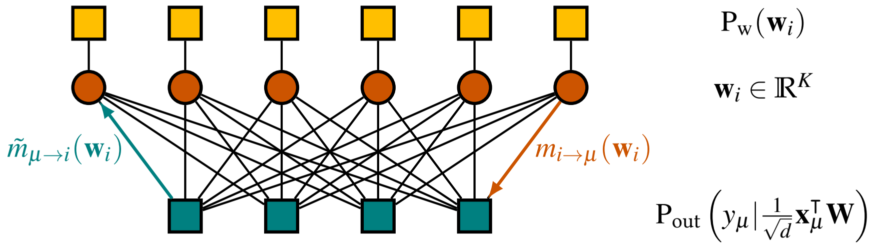

$\mathrm{P_w}(\mathbf{w}_i)$
$\mathbf{w}_i \in \mathbb{R}^K$
$\tilde{m}_{\mu\to i}(\mathbf{w}_i)$
$m_{i\to\mu}(\mathbf{w}_i)$
$\mathrm{P_{out}}\left(y_\mu|\frac{1}{\sqrt{d}}\mathbf{x}_\mu^\intercal\mathbf{W}\right)$

Figure 66: Factor graph representation of the joint distribution for committee machines.

C.1.2 BP EQUATIONS

The BP equations (also called the sum-product equations) for $\mathbf{w}_i = (w_{ik})_{k=1..K} \in \mathbb{R}^K$ on the factor graph Fig. 66 can be formulated, see Sec. 4.3, as:

$$m_{i\to\mu}^{t+1}(\mathbf{w}_i) = \frac{1}{\mathcal{Z}_{i\to\mu}}\mathrm{P_w}(\mathbf{w}_i)\prod_{\nu\neq\mu}^{n}\tilde{m}_{\nu\to i}^t(\mathbf{w}_i) \tag{372}$$

$$\tilde{m}_{\mu\to i}^t(\mathbf{w}_i) = \frac{1}{\mathcal{Z}_{\mu\to i}}\int_{\mathbb{R}^K}\prod_{j\neq i}^{d}\mathrm{d}\mathbf{w}_j\,\mathrm{P_{out}}\left(y_\mu|\frac{1}{\sqrt{d}}\sum_{j=1}^{d}x_{\mu j}\mathbf{w}_j\right)m_{j\to\mu}^t(\mathbf{w}_j)\,,$$

C.2 RELAXED BP EQUATIONS

The idea of the relaxed BP equations is to simply expand in the limit $d\to\infty$ the set of $\Theta(d^2)$ messages $\tilde{m}$ of the BP equations in (372) before plugging them in m. Truncating the expansion and keeping only terms of order $\Theta(1/d)$, messages become *Gaussian*. Hence messages are therefore parametrized only by the mean $\hat{\mathbf{w}}_{i\to\mu}^t$ and the covariance matrix $\hat{\mathbf{C}}_{i\to\mu}^t$ of the marginal distribution at time t:

$$\begin{aligned}\hat{\mathbf{w}}_{i\to\mu}^t &\equiv \int_{\mathbb{R}^K}\mathrm{d}\mathbf{w}_i\, m_{i\to\mu}^t(\mathbf{w}_i)\,\mathbf{w}_i\\ \hat{\mathbf{C}}_{i\to\mu}^t &\equiv \int_{\mathbb{R}^K}\mathrm{d}\mathbf{w}_i\, m_{i\to\mu}^t(\mathbf{w}_i)\,\mathbf{w}_i\mathbf{w}_i^\intercal - \hat{\mathbf{w}}_{i\to\mu}^t(\hat{\mathbf{w}}_{i\to\mu}^t)^\intercal\end{aligned} \tag{373}$$

To decouple the argument of $\mathrm{P_{out}}$, we first by introducing its Fourier transform $\hat{\mathrm{P}}_{\mathrm{out}}$ according to

$$\mathrm{P_{out}}\left(y_\mu|\frac{1}{\sqrt{d}}\sum_{j=1}^{d}x_{\mu j}\mathbf{w}_j\right) = \frac{1}{(2\pi)^{K/2}} \times\int_{\mathbb{R}^K}\mathrm{d}\boldsymbol{\xi}\exp\left(i\boldsymbol{\xi}^\intercal\left(\frac{1}{\sqrt{d}}\sum_{j=1}^{d}x_{\mu j}\mathbf{w}_j\right)\hat{\mathrm{P}}_{\mathrm{out}}(y_\mu,\boldsymbol{\xi})\right).$$

Injecting this representation in the BP equations, (372) becomes:

$$\tilde{m}^t_{\mu\to i}(\mathbf{w}_i) = \frac{1}{(2\pi)^{K/2}\mathscr{Z}_{\mu\to i}} \int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\xi}\ \hat{\mathrm{P}}_{\mathrm{out}}(y_\mu, \boldsymbol{\xi})\ \exp\left(i\boldsymbol{\xi}^\intercal \frac{1}{\sqrt{d}} x_{\mu i}\mathbf{w}_i\right)$$
$$\times \prod_{j\neq i}^{d} \underbrace{\int_{\mathbb{R}^K} \mathrm{d}\mathbf{w}_j\ m^t_{j\to\mu}(\mathbf{w}_j)\ \exp\left(i\boldsymbol{\xi}^\intercal \frac{1}{\sqrt{d}} x_{\mu j}\mathbf{w}_j\right)}_{\equiv I_j} \qquad (374)$$

In the limit $d\to\infty$ the term I_j can be easily expanded and expressed using $\hat{\mathbf{w}}$ and $\hat{\mathbf{C}}$ in (373):

$$I_j = \int_{\mathbb{R}^K} \mathrm{d}\mathbf{w}_j\ m^t_{j\to\mu}(\mathbf{w}_j)\ \exp\left(i\frac{x_{\mu j}}{\sqrt{d}}\boldsymbol{\xi}^\intercal \mathbf{w}_j\right)$$
$$\simeq \exp\left(i\frac{x_{\mu j}}{\sqrt{d}}\boldsymbol{\xi}^\intercal \hat{\mathbf{w}}^t_{j\to\mu} - \frac{1}{2}\frac{x^2_{\mu j}}{d}\boldsymbol{\xi}^\intercal \hat{\mathbf{C}}^t_{j\to\mu}\boldsymbol{\xi}\right).$$

Finally using the inverse Fourier transform:

$$\tilde{m}^t_{\mu\to i}(\mathbf{w}_i) = \frac{1}{(2\pi)^{K/2}\mathscr{Z}_{\mu\to i}} \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}\mathrm{P}_{\mathrm{out}}(y_\mu|\mathbf{z}) \int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\xi} e^{-i\boldsymbol{\xi}^\intercal\mathbf{z}} e^{ix_{\mu i}\boldsymbol{\xi}^\intercal\mathbf{w}_i}$$
$$\times \prod_{j\neq i}^{d} \exp\left(i\frac{x_{\mu j}}{\sqrt{d}}\boldsymbol{\xi}^\intercal \hat{\mathbf{w}}^t_{j\to\mu} - \frac{1}{2}\frac{x^2_{\mu j}}{d}\boldsymbol{\xi}^\intercal \hat{\mathbf{C}}^t_{j\to\mu}\boldsymbol{\xi}\right)$$
$$= \frac{1}{(2\pi)^{K}\mathscr{Z}_{\mu\to i}} \int_{\mathbb{R}} \mathrm{d}\mathbf{z}\ \mathrm{P}_{\mathrm{out}}(y_\mu|\mathbf{z})$$
$$\int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\xi}\ e^{-i\boldsymbol{\xi}^\intercal\mathbf{z}} e^{ix_{\mu i}\boldsymbol{\xi}^\intercal\mathbf{w}_i} e^{i\sum_{j\neq i}^{d}\frac{x_{\mu j}}{\sqrt{d}}\boldsymbol{\xi}^\intercal\hat{\mathbf{w}}^t_{j\to\mu}} e^{-\frac{1}{2}\sum_{j\neq i}^{d}\frac{x^2_{\mu j}}{d}\boldsymbol{\xi}^\intercal\hat{\mathbf{C}}^t_{j\to\mu}\boldsymbol{\xi}}$$
$$= \frac{1}{(2\pi)^{K}\mathscr{Z}_{\mu\to i}} \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}\ \mathrm{P}_{\mathrm{out}}(y_\mu|\mathbf{z})$$
$$\times \sqrt{\frac{(2\pi)^K}{\det\left(V^t_{\mu\to i}\right)}}\ \underbrace{e^{-\frac{1}{2}\left(\mathbf{z}-\frac{x_{\mu i}}{\sqrt{d}}\mathbf{w}_i-\boldsymbol{\omega}^t_{\mu\to i}\right)^\intercal(\mathbf{V}^t_{\mu\to i})^{-1}\left(\mathbf{z}-\frac{x_{\mu i}}{\sqrt{d}}\mathbf{w}_i-\boldsymbol{\omega}^t_{\mu\to i}\right)}}_{\equiv H_{\mu\to i}},$$

where we defined the mean and variance, depending on the node i:

$$\boldsymbol{\omega}^t_{\mu\to i} \equiv \frac{1}{\sqrt{d}}\sum_{j\neq i}^{d} x_{\mu j}\hat{\mathbf{w}}^t_{j\to\mu}, \qquad \mathbf{V}^t_{\mu\to i} \equiv \frac{1}{d}\sum_{j\neq i}^{d} x^2_{\mu j}\hat{\mathbf{C}}^t_{j\to\mu}.$$

Again, in the limit $d\to\infty$, the term $H_{\mu\to i}$ can be expanded as

$$H_{\mu\to i} \simeq e^{-\frac{1}{2}\left(\mathbf{z}-\boldsymbol{\omega}^t_{\mu\to i}\right)^\intercal(\mathbf{V}^t_{\mu\to i})^{-1}\left(\mathbf{z}-\boldsymbol{\omega}^t_{\mu\to i}\right)}$$
$$\times\left(1+\frac{x_{\mu i}}{\sqrt{d}}\mathbf{w}_i^\intercal(\mathbf{V}^t_{\mu\to i})^{-1}(\mathbf{z}-\boldsymbol{\omega}^t_{\mu\to i}) - \frac{1}{2}\frac{x^2_{\mu i}}{d}\mathbf{w}_i^\intercal(\mathbf{V}^t_{\mu\to i})^{-1}\mathbf{w}_i\right.$$
$$\left.+\frac{1}{2}\frac{x^2_{\mu i}}{d}\mathbf{w}_i^\intercal(\mathbf{V}^t_{\mu\to i})^{-1}(\mathbf{z}-\boldsymbol{\omega}^t_{\mu\to i})(\mathbf{z}-\boldsymbol{\omega}^t_{\mu\to i})^\intercal(\mathbf{V}^t_{\mu\to i})^{-1}\mathbf{w}_i\right).$$

Putting all pieces together, the message $\tilde{m}_{\mu\to i}$ can be expressed using definitions of $\mathbf{f}_{\text{out}}$ and $\partial_{\omega}\mathbf{f}_{\text{out}}$ in Appendix. A.4.1.b. We finally obtain

$$
\begin{aligned}
\tilde{m}^t_{\mu\to i}(\mathbf{w}_i) &\sim \frac{1}{\mathcal{Z}_{\mu\to i}}\left\{1+\frac{x_{\mu i}}{\sqrt{d}}\mathbf{w}_i^\intercal \mathbf{f}_{\text{out}}(y_\mu,\boldsymbol{\omega}^t_{\mu\to i},\mathbf{V}^t_{\mu\to i})\right.\\
&\qquad +\frac{1}{2}\frac{x^2_{\mu i}}{d}\mathbf{w}_i^\intercal \mathbf{f}_{\text{out}}\mathbf{f}_{\text{out}}^\intercal(y_\mu,\boldsymbol{\omega}^t_{\mu\to i},\mathbf{V}^t_{\mu\to i})\mathbf{w}_i\\
&\qquad \left.+\frac{1}{2}\frac{x^2_{\mu i}}{d}\mathbf{w}_i^\intercal \partial_{\boldsymbol{\omega}}\mathbf{f}_{\text{out}}(y_\mu,\boldsymbol{\omega}^t_{\mu\to i},\mathbf{V}^t_{\mu\to i})\mathbf{w}_i\right\}\\
&= \frac{1}{\mathcal{Z}_{\mu\to i}}\left\{1+\mathbf{w}_i^\intercal\mathbf{b}^t_{\mu\to i}+\frac{1}{2}\mathbf{w}_i^\intercal\mathbf{b}^t_{\mu\to i}(\mathbf{b}^t_{\mu\to i})^\intercal(\mathbf{w}_i)-\frac{1}{2}\mathbf{w}_i^\intercal\mathbf{A}^t_{\mu\to i}w_i\right\}\\
&= \sqrt{\frac{\det\left(\mathbf{A}^t_{\mu\to i}\right)}{(2\pi)^K}}e^{-\frac{1}{2}\left(\mathbf{w}_i^\intercal-(\mathbf{A}^t_{\mu\to i})^{-1}\mathbf{b}^t_{\mu\to i}\right)^\intercal\mathbf{A}^t_{\mu\to i}\left(\mathbf{w}_i^\intercal-(\mathbf{A}^t_{\mu\to i})^{-1}\mathbf{b}^t_{\mu\to i}\right)}
\end{aligned}
$$

with the following definitions of $\mathbf{A}_{\mu\to i}$ and $\mathbf{b}_{\mu\to i}$

$$
\begin{aligned}
\mathbf{b}^t_{\mu\to i} &\equiv \frac{x_{\mu i}}{\sqrt{d}}\mathbf{f}_{\text{out}}(y_\mu,\boldsymbol{\omega}^t_{\mu\to i},\mathbf{V}^t_{\mu\to i}),\\
\mathbf{A}^t_{\mu\to i} &\equiv -\frac{x^2_{\mu i}}{d}\partial_{\boldsymbol{\omega}}\mathbf{f}_{\text{out}}(y_\mu,\boldsymbol{\omega}^t_{\mu\to i},\mathbf{V}^t_{\mu\to i}).
\end{aligned}
$$

The set of BP equations can finally be closed over the Gaussian messages $\{m_{i\to\mu}\}_{i=1..d}^{\mu=1..n}$ according to

$$
\begin{aligned}
m^{t+1}_{i\to\mu}(\mathbf{w}_i) = \frac{1}{\mathcal{Z}_{i\to\mu}}\mathrm{P}_{\mathrm{w}}(\mathbf{w}_i)\prod_{\nu\neq\mu}^{n}\sqrt{\frac{\det(\mathbf{A}^t_{\nu\to i})}{(2\pi)^K}}\\
\times e^{-\frac{1}{2}\left(\mathbf{w}_i-(\mathbf{A}^t_{\nu\to i})^{-1}\mathbf{b}^t_{\nu\to i}\right)^\intercal\mathbf{A}^t_{\nu\to i}\left(w_i-(\mathbf{A}^t_{\nu\to i})^{-1}\mathbf{b}^t_{\nu\to i}\right)}.
\end{aligned}
$$

In the end, computing the mean and variance of the product of Gaussians, the messages are updated using $\mathbf{f}_{\mathrm{w}}$ and $\partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}$, defined in Appendix. A.4.1.b, according to

$$
\hat{\mathbf{w}}^{t+1}_{i\to\mu} = \mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma}^t_{\mu\to i},\boldsymbol{\Lambda}^t_{\mu\to i}), \qquad \hat{\mathbf{C}}^{t+1}_{i\to\mu} = \partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma}^t_{\mu\to i},\boldsymbol{\Lambda}^t_{\mu\to i}),
$$

with

$$
\boldsymbol{\gamma}^t_{\mu\to i} = \sum_{\nu\neq\mu}^{n}\mathbf{b}^t_{\nu\to i}, \qquad \boldsymbol{\Lambda}^t_{\mu\to i} = \sum_{\nu\neq\mu}^{n}\mathbf{A}^t_{\nu\to i}.
$$

Summary of the rBP equations In the end, the rBP equations are simply the following set of equations:

$$\begin{aligned}
\hat{\mathbf{w}}^{t+1}_{i\to\mu} &= \mathbf{f}_{\mathrm{w}}\left(\boldsymbol{\gamma}^t_{\mu\to i}, \boldsymbol{\Lambda}^t_{\mu\to i}\right), & \hat{\mathbf{C}}^{t+1}_{i\to\mu} &= \partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}\left(\boldsymbol{\gamma}^t_{\mu\to i}, \boldsymbol{\Lambda}^t_{\mu\to i}\right) \\
\boldsymbol{\gamma}^t_{\mu\to i} &= \sum_{\nu\neq\mu}^{n} \mathbf{b}^t_{\nu\to i}, & \boldsymbol{\Lambda}^t_{\mu\to i} &= \sum_{\nu\neq\mu}^{n} \mathbf{A}^t_{\nu\to i} \\
\mathbf{b}^t_{\mu\to i} &= \frac{x_{\mu i}}{\sqrt{d}}\mathbf{f}_{\mathrm{out}}\left(y_\mu, \boldsymbol{\omega}^t_{\mu\to i}, \mathbf{V}^t_{\mu\to i}\right), & & \\
\mathbf{A}^t_{\mu\to i} &= -\frac{x^2_{\mu i}}{d}\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}\left(y_\mu, \boldsymbol{\omega}^t_{\mu\to i}, \mathbf{V}^t_{\mu\to i}\right) & & \\
\boldsymbol{\omega}^t_{\mu\to i} &= \sum_{j\neq i}^{d} \frac{x_{\mu j}}{\sqrt{d}}\hat{\mathbf{w}}^t_{j\to\mu}, & \mathbf{V}^t_{\mu\to i} &= \sum_{j\neq i}^{d} \frac{x^2_{\mu j}}{d}\hat{\mathbf{C}}^t_{j\to\mu}.
\end{aligned} \tag{375}$$

C.3 AMP ALGORITHM

The rBP equations eq. (375) contains $\Theta(d^2)$ messages. However all the messages depend weakly on the target node. The missing message is negligible in the limit $d\to\infty$, that allows us to expand the rBP around the *full* messages:

$$\begin{aligned}
\boldsymbol{\omega}^t_{\mu} &\equiv \sum_{j=1}^{d} \frac{x_{\mu j}}{\sqrt{d}}\hat{\mathbf{w}}^t_{j\to\mu}, & \mathbf{V}^t_{\mu} &\equiv \sum_{j=1}^{d} \frac{x^2_{\mu j}}{d}\hat{\mathbf{C}}^t_{j\to\mu} \\
\boldsymbol{\gamma}^t_{i} &\equiv \sum_{\mu=1}^{n} \mathbf{b}^t_{\mu\to i}, & \boldsymbol{\Lambda}^t_{i} &\equiv \sum_{\mu=1}^{n} \mathbf{A}^t_{\mu\to i}.
\end{aligned} \tag{376}$$

By completing the sum, we naturally remove the target node dependence and reduce the set of messages to $\Theta(d)$. Let us now perform the expansion of the rBP messages.

Partial covariance $\mathbf{f}_{\mathrm{w}}$: $\boldsymbol{\Lambda}^t_{\mu\to i}$

$$\begin{aligned}
\boldsymbol{\Lambda}^t_{\mu\to i} &= \sum_{\nu\neq\mu}^{n} \mathbf{A}^t_{\nu\to i} = \sum_{\nu=1}^{n} \mathbf{A}^t_{\nu\to i} - \mathbf{A}^t_{\mu\to i} \\
&= \boldsymbol{\Lambda}^t_i - \mathbf{A}^t_{\mu\to i} = \boldsymbol{\Lambda}^t_i + \Theta\left(\frac{1}{d}\right).
\end{aligned}$$

Partial mean $\mathbf{f}_{\mathrm{w}}$: $\boldsymbol{\gamma}^t_{\mu\to i}$

$$\boldsymbol{\gamma}^t_{\mu\to i} = \sum_{\nu\neq\mu}^{n} \mathbf{b}^t_{\nu\to i} = \sum_{\nu=1}^{n} \mathbf{b}^t_{\nu\to i} - \mathbf{b}^t_{\mu\to i} = \boldsymbol{\gamma}^t_i - \mathbf{b}^t_{\mu\to i} + \Theta\left(\frac{1}{d}\right).$$

Mean $\hat{\mathbf{w}}_{i\to\mu}^{t+1}$ update

$$\begin{aligned}
\hat{\mathbf{w}}_{i\to\mu}^{t+1} &= \mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma}_{\mu\to i}^t, \boldsymbol{\Lambda}_{\mu\to i}^t) = \mathbf{f}_{\mathrm{w}}\left(\boldsymbol{\gamma}_i^t - \mathbf{b}_{\mu\to i}^t, \boldsymbol{\Lambda}_i^t\right) + \Theta\left(\frac{1}{d}\right) \\
&= \mathbf{f}_{\mathrm{w}}\left(\boldsymbol{\gamma}_i^t, \boldsymbol{\Lambda}_i^t\right) - \partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}\left(\boldsymbol{\gamma}_i^t, \boldsymbol{\Lambda}_i^t\right)\mathbf{b}_{\mu\to i}^t + \Theta\left(\frac{1}{d}\right) \\
&= \hat{\mathbf{w}}_i^{t+1} - \hat{\mathbf{C}}_i^{t+1}\mathbf{b}_{\mu\to i}^t + \Theta\left(\frac{1}{d}\right) \\
&= \hat{\mathbf{w}}_i^{t+1} - \frac{x_{\mu i}}{\sqrt{d}}\hat{\mathbf{C}}_i^{t+1}\mathbf{f}_{\mathrm{out}}(y_\mu, \boldsymbol{\omega}_\mu^t, \mathbf{V}_\mu^t) + \Theta\left(\frac{1}{d}\right).
\end{aligned}$$

where we defined the prior updates

$$\hat{\mathbf{w}}_i^{t+1} \equiv \mathbf{f}_{\mathrm{w}}\left(\boldsymbol{\gamma}_i^t, \boldsymbol{\Lambda}_i^t\right), \qquad \hat{\mathbf{C}}_i^{t+1} \equiv \partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}\left(\boldsymbol{\gamma}_i^t, \boldsymbol{\Lambda}_i^t\right),$$

and used the fact that $\mathbf{b}_{\mu\to i}^t \simeq \frac{x_{\mu i}}{\sqrt{d}}\hat{\mathbf{C}}_i^{t+1}\mathbf{f}_{\mathrm{out}}(y_\mu, \boldsymbol{\omega}_\mu^t, \mathbf{V}_\mu^t)$ by expanding the equation over $\mathbf{b}_{\mu\to i}^t$ in (375).

Covariance $\hat{\mathbf{C}}_{i\to\mu}^{t+1}$ update

$$\begin{aligned}
\hat{\mathbf{C}}_{i\to\mu}^{t+1} &= \partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma}_{\mu\to i}^t, \boldsymbol{\Lambda}_{\mu\to i}^t) \\
&\simeq \partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}(\boldsymbol{\gamma}_i^t, \boldsymbol{\Lambda}_i^t) + \Theta\left(\frac{1}{\sqrt{d}}\right) = \hat{\mathbf{C}}_i^{t+1} + \Theta\left(\frac{1}{\sqrt{d}}\right).
\end{aligned}$$

Channel update function $\mathbf{f}_{\mathbf{out}}(y_\mu, \boldsymbol{\omega}_{\mu\to i}^t, \mathbf{V}_{\mu\to i}^t)$

$$\begin{aligned}
&\mathbf{f}_{\mathrm{out}}(y_\mu, \boldsymbol{\omega}_{\mu\to i}^t, \mathbf{V}_{\mu\to i}^t) = \mathbf{f}_{\mathrm{out}}\left(y_\mu, \boldsymbol{\omega}_\mu^t - \frac{x_{\mu i}}{\sqrt{d}}\hat{\mathbf{w}}_{i\to\mu}^t, \mathbf{V}_\mu^t - \frac{x_{\mu i}^2}{d}\hat{\mathbf{C}}_{i\to l}^t\right) \\
&= \mathbf{f}_{\mathrm{out}}\left(y_\mu, \boldsymbol{\omega}_\mu^t, \mathbf{V}_\mu^t\right) - \frac{x_{\mu i}}{\sqrt{d}}\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}\left(y_\mu, \boldsymbol{\omega}_\mu^t, \mathbf{V}_\mu^t\right) \underbrace{\hat{\mathbf{w}}_{i\to\mu}^t}_{=\hat{\mathbf{w}}_i^t + \Theta\left(\frac{1}{\sqrt{d}}\right)} + \Theta\left(\frac{1}{d}\right) \\
&= \mathbf{f}_{\mathrm{out}}\left(y_\mu, \boldsymbol{\omega}_\mu^t, \mathbf{V}_\mu^t\right) - \frac{x_{\mu i}}{\sqrt{d}}\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}\left(y_\mu, \boldsymbol{\omega}_\mu^t, \mathbf{V}_\mu^t\right)\hat{\mathbf{w}}_i^t + \Theta\left(\frac{1}{d}\right).
\end{aligned}$$

Covariance $\mathbf{f}_{\mathbf{out}}$: $\mathbf{V}_\mu^t$

$$\mathbf{V}_\mu^t \equiv \sum_{j=1}^{d}\frac{x_{\mu j}^2}{d}\hat{\mathbf{C}}_{j\to\mu}^t = \sum_{j=1}^{d}\frac{x_{\mu j}^2}{d}\hat{\mathbf{C}}_{j\to\mu}^t + \Theta\left(\frac{1}{d^{3/2}}\right).$$

Mean $\mathbf{f}_{\text{out}}$: $\boldsymbol{\omega}^t_\mu$

$$
\begin{aligned}
\boldsymbol{\omega}^t_\mu &= \sum_{i=1}^{d} \frac{x_{\mu i}}{\sqrt{d}} \hat{\mathbf{w}}^t_{i\to\mu} \\
&= \sum_{i=1}^{d} \frac{x_{\mu i}}{\sqrt{d}} \left(\hat{\mathbf{w}}^t_i - x_{\mu i} \hat{\mathbf{C}}^t_i \mathbf{f}_{\text{out}}(y_\mu, \boldsymbol{\omega}^{t-1}_\mu, \mathbf{V}^{t-1}_\mu) + \Theta\left(\frac{1}{d}\right) \right) \\
&= \sum_{i=1}^{d} \frac{x_{\mu i}}{\sqrt{d}} \hat{\mathbf{w}}^t_i - \sum_{i=1}^{d} \frac{x^2_{\mu i}}{d} \hat{\mathbf{C}}^t_i \mathbf{f}_{\text{out}}(y_\mu, \boldsymbol{\omega}^{t-1}_\mu, \mathbf{V}^{t-1}_\mu) + \Theta\left(\frac{1}{d^{3/2}}\right) .
\end{aligned}
$$

Covariance $\mathbf{f}_{\text{w}}$: $\boldsymbol{\Lambda}^t_i$

$$
\begin{aligned}
\boldsymbol{\Lambda}^t_i \equiv \sum_{\mu=1}^{n} \mathbf{A}^t_{\mu\to i} &= \sum_{\nu=1}^{n} -\frac{x^2_{\mu i}}{d} \partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out}}(y_\mu, \boldsymbol{\omega}^t_{\mu\to i}, \mathbf{V}^t_{\mu\to i}) \\
&= \sum_{\mu=1}^{n} -\frac{x^2_{\mu i}}{d} \partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out}}(y_\mu, \boldsymbol{\omega}^t_\mu, \mathbf{V}^t_\mu) + \Theta\left(\frac{1}{d^{3/2}}\right) .
\end{aligned}
$$

Mean $\mathbf{f}_{\text{w}}$: $\boldsymbol{\gamma}^t_i$

$$
\begin{aligned}
\boldsymbol{\gamma}^t_i = \sum_{\mu=1}^{n} \mathbf{b}^t_{\mu\to i} &= \sum_{\mu=1}^{n} \frac{x_{\mu i}}{\sqrt{d}} \mathbf{f}_{\text{out}}(y_\mu, \boldsymbol{\omega}^t_{\mu\to i}, \mathbf{V}^t_{\mu\to i}) \\
&= \sum_{\mu=1}^{n} \frac{x_{\mu i}}{\sqrt{d}} \mathbf{f}_{\text{out}}(y_\mu, \boldsymbol{\omega}^t_\mu, \mathbf{V}^t_\mu) \\
&\quad - \frac{x^2_{\mu i}}{d} \partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out}}(y_\mu, \boldsymbol{\omega}^t_\mu, \mathbf{V}^t_\mu) \hat{\mathbf{w}}^t_i + \Theta\left(\frac{1}{d^{3/2}}\right) .
\end{aligned}
$$

SUMMARY - AMP ALGORITHM

We finally obtain the AMP algorithm as a reduced set of $\Theta(d)$ messages in Algo. 5.

C.4 STATE EVOLUTION EQUATIONS OF AMP

In this section we derive the behavior of the AMP algorithm in Algo. 5 in the thermodynamic limit $d \to \infty$. This average asymptotic behavior can be tracked with some overlap parameters at time t, $\mathbf{m}^t$, $\mathbf{q}^t$, $\boldsymbol{\Sigma}^t$, that respectively measure the correlation of the AMP estimator with the ground truth, the norms of student and teacher weights, the estimator variance and the second moment of the teacher network $\boldsymbol{\rho}_{\text{w}^\star}$, defined by

$$
\begin{aligned}
\mathbf{m}^t &\equiv \mathbb{E} \lim_{d\to\infty} \frac{1}{d} \hat{\mathbf{W}}^{t\intercal} \hat{\mathbf{W}}^\star , & \mathbf{q}^t &\equiv \mathbb{E} \lim_{d\to\infty} \frac{1}{d} \hat{\mathbf{W}}^{t\intercal} \hat{\mathbf{W}}^t , \\
\boldsymbol{\Sigma}^t &\equiv \mathbb{E} \lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \hat{\mathbf{C}}^t_i , & \boldsymbol{\rho}_{\text{w}^\star} &\equiv \mathbb{E} \lim_{d\to\infty} \frac{1}{d} \mathbf{W}^{\star\intercal} \mathbf{W}^\star ,
\end{aligned} \tag{377}
$$

```
Input: vector $\mathbf{y} \in \mathbb{R}^n$ and matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$:
Initialize: $\hat{\mathbf{w}}_i$, $\mathbf{f}_{\text{out},\mu} \in \mathbb{R}^K$ and $\hat{\mathbf{V}}_i$, $\partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out},\mu} \in \mathbb{R}^{K \times K}$ for $1 \leq i \leq d$ and $1 \leq \mu \leq n$ at $t = 0$.
repeat
  Channel: Update the mean $\omega_\mu \in \mathbb{R}^K$ and variance $V_\mu \in \mathbb{R}^{K \times K}$:
  $\mathbf{V}_\mu^t = \sum_{i=1}^{d} \frac{x_{\mu j}^2}{d} \hat{\mathbf{C}}_i^t$
  $\boldsymbol{\omega}_\mu^t = \sum_{i=1}^{d} \frac{x_{\mu i}}{\sqrt{d}} \hat{\mathbf{w}}_i^t - \mathbf{V}_\mu^t \mathbf{f}_{\text{out},\mu}^{t-1}$ ,
  Update $\mathbf{f}_{\text{out},\mu}$ and $\partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out},\mu}$:
  $\mathbf{f}_{\text{out},\mu}^t = \mathbf{f}_{\text{out}} \left( y_\mu, \boldsymbol{\omega}_\mu^t, \mathbf{V}_\mu^t \right)$ , $\partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out},\mu}^t = \partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out}} \left( y_\mu, \boldsymbol{\omega}_\mu^t, \mathbf{V}_\mu^t \right)$
  Prior: Update the mean $\boldsymbol{\gamma}_i \in \mathbb{R}^K$ and variance $\boldsymbol{\Lambda}_i \in \mathbb{R}^{K \times K}$:
  $\boldsymbol{\Lambda}_i^t = \sum_{\mu=1}^{n} -\frac{x_{\mu i}^2}{d} \partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out},\mu}$
  $\boldsymbol{\gamma}_i^t = \sum_{\mu=1}^{n} \frac{x_{\mu i}}{\sqrt{d}} \mathbf{f}_{\text{out},\mu} + \boldsymbol{\Lambda}_i^t \hat{\mathbf{w}}_i^t$ ,
  Update the estimated marginals $\hat{\mathbf{w}}_i \in \mathbb{R}$ and $\hat{\mathbf{C}}_i \in \mathbb{R}^+$:
  $\hat{\mathbf{w}}_i^{t+1} = \mathbf{f}_{\text{w}} \left( \boldsymbol{\gamma}_i^t, \boldsymbol{\Lambda}_i^t \right)$ , $\hat{\mathbf{C}}_i^{t+1} = \partial_{\boldsymbol{\gamma}} \mathbf{f}_{\text{w}} \left( \boldsymbol{\gamma}_i^t, \boldsymbol{\Lambda}_i^t \right)$
  $t \leftarrow t + 1$
until Convergence on $\hat{\mathbf{w}}_i$, $\hat{\mathbf{C}}_i$.
Output: $\{\hat{\mathbf{w}}_i\}_{i=1}^d$ and $\{\hat{\mathbf{C}}_i\}_{i=1}^d$.
```

Algorithme 5 : Approximate Message Passing algorithm for committee machines.

where the expectation is over ground truth signals $\mathbf{W}^\star$ and input data $\mathbf{X}$. The aim is to derive the asymptotic behavior of these overlap parameters, called SE. The idea is simply to compute the overlap distributions starting with the set of rBP equations in (375).

C.4.1 MESSAGES DISTRIBUTION

In order to get the asymptotic behavior of the overlap parameters, we first need to compute the distribution of $\mathbf{W}^{t+1}$ and, as a result, of the mean $\boldsymbol{\gamma}^t_{\mu\to i}$ and covariance $\boldsymbol{\Lambda}^t_{\mu\to i}$. Recalling that under the BP assumption incoming messages are independent, the messages $\boldsymbol{\omega}^t_{\mu\to i}$ and $\mathbf{z}_\mu$ are the sum of independent variables and follow Gaussian distributions. However, these two variables are correlated and we need to compute correctly the covariance matrix.

To compute it, we will make use of different ingredients. First, we recall that in the T-S scenario, the output has been generated by a teacher such that $\forall \mu \in [\![n]\!]$, $y_\mu = \varphi_{\text{out}^\star}\left(\frac{1}{\sqrt{d}}\mathbf{x}_\mu^\intercal \mathbf{W}^\star\right)$. By convenience, we define $\mathbf{z}_\mu \equiv \frac{1}{\sqrt{d}}\mathbf{x}_\mu^\intercal \mathbf{W}^\star = \frac{1}{\sqrt{d}}\sum_{i=1}^d x_{\mu i}\mathbf{w}_i^\star$ and $z_{\mu\to i} \equiv \frac{1}{\sqrt{d}}\sum_{j\neq i}^d x_{\mu j}\mathbf{w}_j^\star$. Second, in the case the input data are i.i.d Gaussian, we have $\mathbb{E}_{\mathbf{X}}[x_{\mu i}] = 0$ and $\mathbb{E}_{\mathbf{X}}[x_{\mu i}^2] = 1$.

Partial mean $\mathbf{f}_{\text{out}}$: $\boldsymbol{\omega}^t_{\mu\to i}$ Let's compute the first two moments, using expansions of the rBP equations (375):

$$\mathbb{E}\left[\boldsymbol{\omega}^t_{\mu\to i}\right] = \frac{1}{\sqrt{d}}\sum_{j\neq i}^d \mathbb{E}_{\mathbf{X}}\left[x_{\mu j}\right]\mathbb{E}\left[\hat{\mathbf{w}}^t_{j\to\mu}\right] = \mathbf{0},$$

$$\mathbb{E}\left[\boldsymbol{\omega}^t_{\mu\to i}\left(\boldsymbol{\omega}^t_{\mu\to i}\right)^\intercal\right] = \frac{1}{d}\sum_{j\neq i}^d \mathbb{E}_{\mathbf{X}}\left[x_{\mu j}^2\right]\mathbb{E}\left[\hat{\mathbf{w}}^t_{j\to\mu}\left(\hat{\mathbf{w}}^t_{j\to\mu}\right)^\intercal\right]$$

$$= \frac{1}{d}\sum_{i=1}^d \mathbb{E}_{\mathbf{X}}\left[x_{\mu j}^2\right]\mathbb{E}\left[\hat{\mathbf{w}}^t_i\left(\hat{\mathbf{w}}_i\right)^\intercal\right] + \Theta\left(d^{-3/2}\right) \underset{d\to\infty}{\longrightarrow} \mathbf{q}^t .$$

Hidden variable $\mathbf{z}_\mu$ Let us compute the first moments of the hidden variable $\mathbf{z}_\mu$:

$$\mathbb{E}\left[\mathbf{z}_\mu\right] = \frac{1}{\sqrt{d}}\sum_{i=1}^d \mathbb{E}_{\mathbf{X}}\left[x_{\mu i}\right]\mathbb{E}_{\mathbf{W}^\star}\left[\mathbf{w}_i^\star\right] = \mathbf{0},$$

$$\mathbb{E}\left[\mathbf{z}_\mu\mathbf{z}_\mu^\intercal\right] = \frac{1}{d}\sum_{i=1}^d \mathbb{E}_{\mathbf{X}}\left[x_{\mu i}^2\right]\mathbb{E}_{\mathbf{W}^\star}\left[\mathbf{w}_i^\star(\mathbf{w}_i^\star)^\intercal\right] \underset{d\to\infty}{\longrightarrow} \boldsymbol{\rho}_{\mathbf{w}^\star} .$$

Correlation between $\mathbf{z}_\mu$ and $\boldsymbol{\omega}^t_{\mu\to i}$ The cross correlation is given by

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\omega}^t_{\mu\to i}\mathbf{z}_\mu^\intercal\right] &= \frac{1}{d}\sum_{j\neq i,k=1}^{d}\mathbb{E}_{\mathbf{X}}\left[x_{\mu j}x_{\mu k}\right]\mathbb{E}_{\mathbf{W}^\star}\left[\hat{\mathbf{w}}^t_{j\to\mu}(\mathbf{w}^\star_k)^\intercal\right] \\
&= \frac{1}{d}\sum_{j\neq i}^{d}\mathbb{E}_{\mathbf{W}^\star}\left[\hat{\mathbf{w}}^t_{j\to\mu}(\mathbf{w}^\star_j)^\intercal\right] = \frac{1}{d}\sum_{i}^{d}\mathbb{E}_{\mathbf{W}^\star}\left[\hat{\mathbf{w}}^t_i(\mathbf{w}^\star_i)^\intercal\right] + \Theta\left(d^{-3/2}\right) \\
&\underset{d\to\infty}{\longrightarrow} \mathbf{m}^t\,.
\end{aligned}$$

Hence asymptotically the random vector $(\mathbf{z}_\mu,\, \boldsymbol{\omega}^t_{\mu\to i})$ follow a multivariate Gaussian distribution with covariance matrix $\mathbf{Q}^t = \begin{bmatrix}\boldsymbol{\rho}_{\mathrm{w}^\star} & \mathbf{m}^t \\ \mathbf{m}^t & \mathbf{q}^t\end{bmatrix} \in \mathbb{R}^{(2K)\times(2K)}$.

Partial variance $\mathbf{f}_{\mathbf{out}}$: $\mathbf{V}_{\mu\to i}$ $\mathbf{V}_{\mu\to i}$ concentrates around its mean:

$$\mathbb{E}\left[\mathbf{V}^t_{\mu\to i}\right] = \frac{1}{d}\sum_{j\neq i}^{d}\mathbb{E}_{\mathbf{X}}\left[x^2_{\mu j}\right]\hat{\mathbf{C}}^t_{j\to\mu} = \frac{1}{d}\sum_{i}^{d}\hat{\mathbf{C}}^t_i + \Theta(d^{-3/2}) \underset{d\to\infty}{\longrightarrow} \boldsymbol{\Sigma}^t\,.$$

Ad-hoc overlaps Let us define some other ad-hoc order parameters, that will appear in the following:

$$\begin{aligned}
\hat{\mathbf{q}}^t &\equiv \alpha\mathbb{E}_{\boldsymbol{\omega},\mathbf{z}}\left[\mathbf{f}_{\mathrm{out}}(\varphi_{\mathrm{out}^\star}(\mathbf{z}),\boldsymbol{\omega},\boldsymbol{\Sigma}^t)^{\otimes 2}\right]\,, \\
\hat{\mathbf{m}}^t &\equiv \alpha\mathbb{E}_{\boldsymbol{\omega},\mathbf{z}}\left[\partial_{\mathbf{z}}\mathbf{f}_{\mathrm{out}}(\varphi_{\mathrm{out}^\star}(\mathbf{z}),\boldsymbol{\omega},\boldsymbol{\Sigma}^t)\right]\,, \\
\hat{\boldsymbol{\chi}}^t &\equiv \alpha\mathbb{E}_{\boldsymbol{\omega},\mathbf{z}}\left[-\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}(\varphi_{\mathrm{out}^\star}(\mathbf{z}),\boldsymbol{\omega},\boldsymbol{\Sigma}^t)\right]\,.
\end{aligned} \tag{378}$$

Partial mean $\mathbf{f}_{\mathrm{w}}$: $\boldsymbol{\gamma}^t_{\mu\to i}$ Using the expression $y_\nu = \varphi_{\mathrm{out}^\star}\left(\mathbf{z}_{\nu\to i} + \frac{1}{\sqrt{d}}x_{\nu i}\mathbf{w}^\star_i\right)$ and expanding $\boldsymbol{\gamma}^t_{\mu\to i}$, we obtain

$$\begin{aligned}
\boldsymbol{\gamma}^t_{\mu\to i} &= \sum_{\nu\neq\mu}^{n}\mathbf{b}^t_{\nu\to i} = \sum_{\nu\neq\mu}^{n}\frac{x_{\nu i}}{\sqrt{d}}\mathbf{f}_{\mathrm{out}}\left(y_\nu, \boldsymbol{\omega}^t_{\nu\to i}, \mathbf{V}^t_{\nu\to i}\right) \\
&= \frac{1}{\sqrt{d}}\sum_{\nu\neq\mu}^{n}x_{\nu i}\mathbf{f}_{\mathrm{out}}\left(\varphi_{\mathrm{out}^\star}\left(\mathbf{z}_{\nu\to i}\right), \boldsymbol{\omega}^t_{\nu\to i}, \mathbf{V}^t_{\nu\to i}\right) \\
&+ \frac{1}{d}\sum_{\nu\neq\mu}^{n}x^2_{\nu i}\partial_{\mathbf{z}}\mathbf{f}_{\mathrm{out}}\left(\varphi_{\mathrm{out}^\star}\left(\mathbf{z}_{\nu\to i}\right), \boldsymbol{\omega}^t_{\nu\to i}, \mathbf{V}^t_{\nu\to i}\right)\mathbf{w}^\star_i\,.
\end{aligned}$$

Thus, taking the average

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\gamma}^t_{\mu\to i}\right] &= \mathbf{0} + \frac{1}{d}\sum_{\nu\neq\mu}^{n}\mathbb{E}_{\mathbf{z},\boldsymbol{\omega}}\left[\partial_{\mathbf{z}}\mathbf{f}_{\mathrm{out}}\left(\varphi_{\mathrm{out}^\star}\left(\mathbf{z}_{\nu\to i}\right), \boldsymbol{\omega}^t_{\nu\to i}, \mathbf{V}^t_{\nu\to i}\right)\right]\mathbf{w}^\star_i \\
&\underset{d\to\infty}{\longrightarrow} \hat{\mathbf{m}}^t\mathbf{w}^\star_i\,, \\
\mathbb{E}\left[\left(\boldsymbol{\gamma}^t_{\mu\to i}\right)^{\otimes 2}\right] &= \frac{1}{d}\sum_{\nu\neq\mu}^{n}\mathbb{E}_{\mathbf{z},\boldsymbol{\omega}}\left[\mathbf{f}_{\mathrm{out}}\left(\varphi_{\mathrm{out}^\star}\left(\mathbf{z}_{\nu\to i}\right), \boldsymbol{\omega}^t_{\nu\to i}, \mathbf{V}^t_{\nu\to i}\right)^{\otimes 2}\right] \\
&\underset{d\to\infty}{\longrightarrow} \hat{\mathbf{q}}^t\,.
\end{aligned}$$

Hence $\boldsymbol{\gamma}^t_{\mu\to i} \sim \hat{\mathbf{m}}^t \mathbf{w}^\star_i + (\hat{\mathbf{q}}^t)^{1/2}\boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$.

Partial covariance $\mathbf{f}_{\mathrm{w}}$: $\boldsymbol{\Lambda}^t_{\mu\to i}$

$$\begin{aligned}\boldsymbol{\Lambda}^t_{\mu\to i} &= \sum_{\nu\neq\mu}^{n} \mathbf{A}^t_{\nu\to i} = -\frac{1}{d}\sum_{\nu\neq\mu}^{n} x^2_{\mu i}\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}(y_\nu, \boldsymbol{\omega}^t_{\nu\to i}, \mathbf{V}^t_{\nu\to i})\\ &= -\frac{1}{d}\sum_{\nu\neq\mu}^{n} x^2_{\mu i}\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}(\varphi_{\mathrm{out}^\star}(\mathbf{z}_{\nu\to i}), \boldsymbol{\omega}^t_{\nu\to i}, \mathbf{V}^t_{\nu\to i}) + \Theta\left(d^{-3/2}\right)\end{aligned}$$

and taking the average

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\Lambda}^t_{\mu\to i}\right] &= -\frac{1}{d}\sum_{\nu\neq\mu}^{n} \mathbb{E}_{\mathbf{z},\boldsymbol{\omega}}\left[\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}(\varphi_{\mathrm{out}^\star}(\mathbf{z}_{\nu\to i}), \boldsymbol{\omega}^t_{\nu\to i}, \mathbf{V}^t_{\nu\to i})\right]\\ &\underset{d\to\infty}{\longrightarrow} \hat{\boldsymbol{\chi}}^t ,\end{aligned}$$

so that in the thermodynamic limit $\boldsymbol{\Lambda}^t_{\mu\to i} \sim \hat{\boldsymbol{\chi}}^t$.

C.4.2 SUMMARY OF THE SE - MISMATCHED SETTING

Using the definition of the overlaps in (377) at time $t+1$ and the message distributions, we finally obtain the set of SE equations of the AMP algorithm in Algo. 5 in the mismatched setting:

$$\begin{aligned}\mathbf{m}^{t+1} &\equiv \mathbb{E}\lim_{d\to\infty}\frac{1}{d}\hat{\mathbf{W}}^{t+1\intercal}\hat{\mathbf{W}}^\star = \mathbb{E}_{\mathbf{w}^\star,\boldsymbol{\xi}}\left[\mathbf{f}_{\mathrm{w}}\left(\hat{\mathbf{m}}^t\mathbf{w}^\star + (\hat{\mathbf{q}}^t)^{1/2}\boldsymbol{\xi}, \hat{\boldsymbol{\chi}}^t\right)\mathbf{w}^{\star\intercal}\right],\\ \mathbf{q}^{t+1} &\equiv \mathbb{E}\lim_{d\to\infty}\frac{1}{d}\hat{\mathbf{W}}^{t+1\intercal}\hat{\mathbf{W}}^{t+1} = \mathbb{E}_{\mathbf{w}^\star,\boldsymbol{\xi}}\left[\mathbf{f}_{\mathrm{w}}\left(\hat{\mathbf{m}}^t\mathbf{w}^\star + (\hat{\mathbf{q}}^t)^{1/2}\boldsymbol{\xi}, \hat{\boldsymbol{\chi}}^t\right)^{\otimes 2}\right],\\ \boldsymbol{\Sigma}^{t+1} &\equiv \mathbb{E}\lim_{d\to\infty}\frac{1}{d}\sum_{i=1}^{d}\hat{\mathbf{C}}^{t+1}_i = \mathbb{E}_{\mathbf{w}^\star,\boldsymbol{\xi}}\left[\partial_{\boldsymbol{\gamma}}\mathbf{f}_{\mathrm{w}}\left(\hat{\mathbf{m}}^t\mathbf{w}^\star + (\hat{\mathbf{q}}^t)^{1/2}\boldsymbol{\xi}, \hat{\boldsymbol{\chi}}^t\right)\right],\end{aligned} \tag{379}$$

and

$$\begin{aligned}\hat{\mathbf{q}}^t &= \alpha\int_{\mathbb{R}^K}\int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\omega}\ \mathrm{d}\mathbf{z}\ \mathcal{N}_{(\mathbf{z},\boldsymbol{\omega})}\left(\mathbf{0}_{2K}, \mathbf{Q}^t\right)\mathbf{f}_{\mathrm{out}}(\varphi_{\mathrm{out}^\star}(\mathbf{z}), \boldsymbol{\omega}, \boldsymbol{\Sigma}^t)^{\otimes 2}\\ \hat{\mathbf{m}}^t &= \alpha\int_{\mathbb{R}^K}\int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\omega}\ \mathrm{d}\mathbf{z}\ \mathcal{N}_{(\mathbf{z},\boldsymbol{\omega})}\left(\mathbf{0}_{2K}, \mathbf{Q}^t\right))\partial_{\mathbf{z}}\mathbf{f}_{\mathrm{out}}(\varphi_{\mathrm{out}^\star}(\mathbf{z}), \boldsymbol{\omega}, \boldsymbol{\Sigma}^t),\\ \hat{\boldsymbol{\chi}}^t &= -\alpha\int_{\mathbb{R}^K}\int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\omega}\ \mathrm{d}\mathbf{z}\ \mathcal{N}_{(\mathbf{z},\boldsymbol{\omega})}\left(\mathbf{0}_{2K}, \mathbf{Q}^t\right)\partial_{\boldsymbol{\omega}}\mathbf{f}_{\mathrm{out}}(\varphi_{\mathrm{out}^\star}(\mathbf{z}), \boldsymbol{\omega}, \boldsymbol{\Sigma}^t).\end{aligned} \tag{380}$$

with $\mathbf{Q}^t = \begin{bmatrix}\boldsymbol{\rho}_{\mathrm{w}^\star} & \mathbf{m}^t\\ \mathbf{m}^t & \mathbf{q}^t\end{bmatrix} \in \mathbb{R}^{(2K)\times(2K)}$.

C.4.3 SUMMARY OF THE SE - BAYES-OPTIMAL SETTING

In the Bayes-optimal setting, the student $\mathrm{P_w} = \mathrm{P_{w^\star}}$ and $\mathrm{P_{out}} = \mathrm{P_{out^\star}}$, so that we have $\mathbf{f_w} = \mathbf{f_{w^\star}}$ and $\mathbf{f_{out}} = \mathbf{f_{out^\star}}$. Moreover, the Nishimori conditions, recalled in Appendix. A.3, imply that

$$\mathbf{m}^t = \mathbf{q}^t \equiv \mathbf{q}_\mathrm{b}^t\,, \qquad \hat{\mathbf{q}}^t = \hat{\mathbf{m}}^t = \hat{\boldsymbol{\chi}}^t \equiv \hat{\mathbf{q}}_\mathrm{b}^t\,, \qquad \boldsymbol{\Sigma}^t = \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}^t\,.$$

Therefore the set of SE equations simplify and reduce to

$$\mathbf{q}_\mathrm{b}^{t+1} = \mathbb{E}_{\mathbf{w}^\star,\boldsymbol{\xi}}\left[\mathbf{f}_{\mathrm{w}^\star}\left(\hat{\mathbf{q}}_\mathrm{b}^t\mathbf{w}^\star + (\hat{\mathbf{q}}_\mathrm{b}^t)^{1/2}\boldsymbol{\xi}, \hat{\mathbf{q}}_\mathrm{b}^t\right)^{\otimes 2}\right] \tag{381}$$

$$\hat{\mathbf{q}}_\mathrm{b}^t = \alpha\int_{\mathbb{R}^K}\int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\mathbf{z}\,\mathcal{N}_{(\mathbf{z},\boldsymbol{\omega})}\left(\mathbf{0}_{2K}, \mathbf{Q}_\mathrm{b}^t\right)\mathbf{f}_{\mathrm{out}^\star}\left(\varphi_{\mathrm{out}^\star}(\mathbf{z}), \boldsymbol{\omega}, \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_\mathrm{b}^t\right)^{\otimes 2}$$

with the simplified covariance matrix $\mathbf{Q}_\mathrm{b}^t = \begin{bmatrix} \boldsymbol{\rho}_{\mathrm{w}^\star} & \mathbf{q}_\mathrm{b}^t \\ \mathbf{q}_\mathrm{b}^t & \mathbf{q}_\mathrm{b}^t \end{bmatrix}$.

C.4.4 CONSISTENCE WITH THE REPLICA COMPUTATION

Very surprisingly, the SE of the AMP algorithm can be obtained in a convoluted and more rapid way. It turns out that in the Bayes-optimal setting, AMP performs a gradient ascent on the RS free entropy in (321). Meaning that at convergence, and under good initialization, the AMP overlaps are given by the saddle point equations of the RS free entropy $\Phi^{(\mathrm{rs})}$. To see this, we shall start performing the change of variable $\boldsymbol{\xi} \leftarrow \boldsymbol{\xi} + (\hat{\mathbf{q}}_\mathrm{b}^t)^{1/2}\mathbf{w}^\star$ in (381) so that we directly obtain the first equation of (325) with the corresponding time indices

$$\mathbf{q}_\mathrm{b}^{t+1} = \mathbb{E}_{\mathbf{w}^\star,\boldsymbol{\xi}}\left[\mathscr{Z}_{\mathrm{w}^\star}\left((\hat{\mathbf{q}}_\mathrm{b}^t)^{1/2}\boldsymbol{\xi}, \hat{\mathbf{q}}_\mathrm{b}^t\right)\mathbf{f}_{\mathrm{w}^\star}\left((\hat{\mathbf{q}}_\mathrm{b}^t)^{1/2}\boldsymbol{\xi}, \hat{\mathbf{q}}_\mathrm{b}^t\right)^{\otimes 2}\right]. \tag{382}$$

Moreover in this setting, we notice that variables $\boldsymbol{\omega}_{\mu\to i}^t$ and $\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}^t$ become independent since

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\omega}_{\mu\to i}^t\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}^t\right)^\intercal\right] &\underset{d\to\infty}{\longrightarrow} \mathbf{m}^t - \mathbf{q}^t = \mathbf{q}_\mathrm{b}^t - \mathbf{q}_\mathrm{b}^t = \mathbf{0}\,,\\
\mathbb{E}\left[\boldsymbol{\omega}_{\mu\to i}^t(\boldsymbol{\omega}_{\mu\to i}^t)^\intercal\right] &\underset{d\to\infty}{\longrightarrow} \mathbf{q}_\mathrm{b}^t\,,\\
\mathbb{E}\left[\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}^t\right)\left(\mathbf{z}_\mu - \boldsymbol{\omega}_{\mu\to i}^t\right)^\intercal\right] &\underset{d\to\infty}{\longrightarrow} \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_\mathrm{b}^t\,,
\end{aligned}$$

so that the multivariate Gaussian distribution factorize to

$$\mathcal{N}_{(\mathbf{z},\boldsymbol{\omega})}\left(\mathbf{0}, \mathbf{Q}_\mathrm{b}^t\right) = \mathcal{N}_{\boldsymbol{\omega}}\left(\mathbf{0}_K, \mathbf{q}_\mathrm{b}^t\right)\mathcal{N}_{\mathbf{z}}\left(\boldsymbol{\omega}, \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_\mathrm{b}^t\right).$$

Using $\mathrm{P}_{\mathrm{out}^\star}(y|\mathbf{z}) = \delta\left(y - \varphi_{\mathrm{out}^\star}(\mathbf{z})\right)$ the second equation of (381) becomes

$$\begin{aligned}
\hat{\mathbf{q}}^t &= \alpha \int_{\mathbb{R}^K} \int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\omega}\, \mathrm{d}\mathbf{z}\, \mathcal{N}_{(\mathbf{z},\boldsymbol{\omega})}\left(\mathbf{0}_{2K}, \mathbf{Q}_{\mathrm{b}}^t\right) \mathbf{f}_{\mathrm{out}^\star}\left(\varphi_{\mathrm{out}^\star}(\mathbf{z}), \boldsymbol{\omega}, \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_{\mathrm{b}}^t\right)^{\otimes 2} \\
&= \alpha \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\omega}\, \mathcal{N}_{\boldsymbol{\omega}}\left(\mathbf{0}_K, \mathbf{q}_{\mathrm{b}}^t\right) \\
&\qquad \times \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}\, \mathrm{p}_{\mathrm{out}^\star}(y|\mathbf{z}) \mathcal{N}_{\mathbf{z}}\left(\boldsymbol{\omega}; \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_{\mathrm{b}}^t\right) \mathbf{f}_{\mathrm{out}^\star}\left(y, \boldsymbol{\omega}, \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_{\mathrm{b}}^t\right)^{\otimes 2} \\
&= \alpha \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^K} \mathrm{d}\boldsymbol{\xi}\, \mathcal{N}_{\boldsymbol{\xi}}\left(\mathbf{0}; \mathbf{I}_K\right) \int_{\mathbb{R}^K} \mathrm{d}\mathbf{z}\, \mathrm{p}_{\mathrm{out}^\star}(y|\mathbf{z}) \\
&\qquad\qquad \times \mathcal{N}_{\mathbf{z}}\left((\mathbf{q}_{\mathrm{b}}^t)^{1/2} \boldsymbol{\xi}; \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_{\mathrm{b}}^t \right) \mathbf{f}_{\mathrm{out}^\star}\left(y, (\mathbf{q}_{\mathrm{b}}^t)^{1/2} \boldsymbol{\xi}, \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_{\mathrm{b}}^t\right)^{\otimes 2} \\
&\qquad\qquad\qquad\qquad \text{(Change of variable } \boldsymbol{\xi} \leftarrow (\mathbf{q}_{\mathrm{b}}^t)^{-1/2} \boldsymbol{\omega}^t\text{)} \\
&= \alpha \int_{\mathbb{R}} \mathrm{d}y\, \mathbb{E}_{\boldsymbol{\xi}} \mathcal{Z}_{\mathrm{out}^\star}\left(y, (\mathbf{q}_{\mathrm{b}}^t)^{1/2} \boldsymbol{\xi}, \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_{\mathrm{b}}^t \right) \\
&\qquad\qquad\qquad \times \mathbf{f}_{\mathrm{out}^\star}\left(y, (\mathbf{q}_{\mathrm{b}}^t)^{1/2} \boldsymbol{\xi}, \boldsymbol{\rho}_{\mathrm{w}^\star} - \mathbf{q}_{\mathrm{b}}^t \right),
\end{aligned}$$

which is exactly the second fixed point equation of the RS free entropy (381).

Part IV

BIBLIOGRAPHY

BIBLIOGRAPHY

Abbara, Alia, Benjamin Aubin, Florent Krzakala, and Lenka Zdeborová (2020). 'Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning.' In: *Mathematical and Scientific Machine Learning*. PMLR, pp. 27–54. eprint: *https://arxiv.org/abs/1912.02729*.

Achlioptas, Dimitris and Amin Coja-Oghlan (2008). 'Algorithmic barriers from phase transitions.' In: *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*. IEEE, pp. 793–802.

Achlioptas, Dimitris, Amin Coja-Oghlan, and Federico Ricci-Tersenghi (2011). 'On the solution-space geometry of random constraint satisfaction problems.' In: *Random Structures & Algorithms* 38.3, pp. 251–268.

Achlioptas, Dimitris and Cristopher Moore (2002). 'The asymptotic order of the random k-SAT threshold.' In: *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*. IEEE, pp. 779–788.

Ackley, David H, Geoffrey E Hinton, and Terrence J Sejnowski (1985). 'A learning algorithm for Boltzmann machines.' In: *Cognitive science* 9.1, pp. 147–169.

Advani, Madhu S and Andrew M Saxe (2017). 'High-dimensional dynamics of generalization error in neural networks.' In: *arXiv preprint arXiv:1710.03667*.

Advani, Madhu and Surya Ganguli (2016a). 'An equivalence between high dimensional Bayes optimal inference and M-estimation.' In: *Advances in Neural Information Processing Systems* 1, pp. 3386–3394.

– (2016b). 'Statistical mechanics of optimal convex inference in high dimensions.' In: *Physical Review X* 6.3, pp. 1–16.

Agoritsas, Elisabeth, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi (2018). 'Out-of-equilibrium dynamical mean-field equations for the perceptron model.' In: *Journal of Physics A: Mathematical and Theoretical* 51.8, p. 085002.

Aizerman, Mark A (1964). 'Theoretical foundations of the potential function method in pattern recognition learning.' In: *Automation and remote control* 25, pp. 821–837.

Aji, S. M. and R. J. McEliece (2000). 'The generalized distributive law.' In: *IEEE Transactions on Information Theory* 46.2, pp. 325–343.

Alaoui, Ahmed El and Florent Krzakala (2018). 'Estimation in the Spiked Wigner Model: A Short Proof of the Replica Formula.' In: *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1874–1878.

Aldrich, John et al. (2008). 'RA Fisher on Bayes and Bayes' theorem.' In: *Bayesian Analysis* 3.1, pp. 161–170.

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song (2019). 'A convergence theory for deep learning via over-parameterization.' In: *International Conference on Machine Learning*. PMLR, pp. 242–252.

Allison, Rupert and Joanna Dunkley (Dec. 2013). 'Comparison of sampling techniques for Bayesian parameter estimation.' In: *Monthly Notices of the Royal Astronomical Society* 437.4, pp. 3918–3928.

Almeida, J R L de and D J Thouless (1978). 'Stability of the Sherrington-Kirkpatrick solution of a spin glass model.' In: *Journal of Physics A: Mathematical and General* 11.5, pp. 983–990.

Amini, Arash A and Martin J Wainwright (2009). 'High-Dimensional Analysis of Semidefinite Relaxations for Sparse Principal Components.' In: *The Annals of Statistics*, pp. 2877–2921.

Amit, Daniel J, Hanoch Gutfreund, and Haim Sompolinsky (1985a). 'Spin-glass models of neural networks.' In: *Physical Review A* 32.2, p. 1007.

– (1985b). 'Storing infinite numbers of patterns in a spin-glass model of neural networks.' In: *Physical Review Letters* 55.14, p. 1530.

– (1987). 'Statistical mechanics of neural networks near saturation.' In: *Annals of physics* 173.1, pp. 30–67.

Anderson, P. W. (1972). 'More Is Different.' In: *Science* 177.4047, pp. 393–396.

Andrieu, Christophe, Nando De Freitas, Arnaud Doucet, and Michael I Jordan (2003). 'An introduction to MCMC for machine learning.' In: *Machine learning* 50.1-2, pp. 5–43.

Apt, Krzysztof (2003). *Principles of constraint programming*. Cambridge university press.

Arora, Sanjeev and Boaz Barak (2009). *Computational complexity: a modern approach.* Cambridge University Press.

Arora, Sanjeev, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang (2019). 'On exact computation with an infinitely wide neural net.' In: *Advances in Neural Information Processing Systems*, pp. 8141–8150.

Arora, Sanjeev, Rong Ge, Behnam Neyshabur, and Yi Zhang (2018a). 'Stronger generalization bounds for deep nets via a compression approach.' In: *arXiv preprint arXiv:1802.05296*.

Arora, Sanjeev, Andrej Risteski, and Yi Zhang (2018b). 'Do GANs learn the distribution? some theory and empirics.' In: *International Conference on Learning Representations*.

Aubin, Benjamin, Florent Krzakala, Yue M Lu, and Lenka Zdeborová (2020a). 'Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization.' In: *Advances in Neural Information Processing Systems 33*. eprint: *https://arxiv.org/abs/2006.06560*.

Aubin, Benjamin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová (2020b). 'Exact asymptotics for phase retrieval and compressed sensing with random generative priors.' In: *Mathematical and Scientific Machine Learning*. PMLR, pp. 55–73. eprint: *https://arxiv.org/abs/1912.02008*.

– (2020c). 'Exact asymptotics for phase retrieval and compressed sensing with random generative priors.' In: *Mathematical and Scientific Machine Learning*. PMLR, pp. 55–73. eprint: *https://arxiv.org/abs/1912.02008*.

Aubin, Benjamin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová (2019a). *Demonstration codes - The spiked matrix model with generative priors*. *https://github.com/benjaminaubin/StructuredPrior_demo*.

Aubin, Benjamin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová (2019b). 'The spiked matrix model with generative priors.' In: *Advances in Neural Information Processing Systems 32*, pp. 8366–8377. eprint: *https://arxiv.org/abs/1905.12385*.

– (2019c). 'The spiked matrix model with generative priors.' In: *Advances in Neural Information Processing Systems 32*, pp. 8366–8377. eprint: *https://arxiv.org/abs/1905.12385*.

Aubin, Benjamin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová (2018a). *AMP implementation of the committee machine*. *https://github.com/benjaminaubin/TheCommitteeMachine*.

Aubin, Benjamin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová (2018b). 'The committee machine: Computational to statistical gaps in learning a two-layers neural network.' In: *Advances in Neural Information Processing Systems 31*, pp. 3223–3234. eprint: *https://arxiv.org/abs/1806.05451*.

Aubin, Benjamin, Will Perkins, and Lenka Zdeborová (2019d). 'Storage capacity in symmetric binary perceptrons.' In: *Journal of Physics A: Mathematical and Theoretical* 52.29, p. 294003. eprint: *https://arxiv.org/abs/1901.00314*.

– (2019e). 'Storage capacity in symmetric binary perceptrons.' In: *Journal of Physics A: Mathematical and Theoretical* 52.29, p. 294003. eprint: *https://arxiv.org/abs/1901.00314*.

Baik, Jinho, Gérard Ben Arous, Sandrine Péché, et al. (2005). 'Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices.' In: *The Annals of Probability* 33.5, pp. 1643–1697.

Baity-Jest, Marco et al. (Jan. 2018). 'Comparing Dynamics: Deep Neural Networks versus Glassy Systems.' English (US). In: *35th International Conference on Machine Learning, ICML 2018*. Vol. 1. International Machine Learning Society (IMLS), pp. 526–535.

Baker, Antoine, Benjamin Aubin, Florent Krzakala, and Lenka Zdeborová (2020). 'TRAMP: Compositional Inference with TRee Approximate Message Passing.' In: *arXiv preprint arXiv:2004.01571. Submitted to Journal of Machine Learning Research*. eprint: *https://arxiv.org/abs/2004.01571*.

Baldassi, Carlo, Alfredo Braunstein, Nicolas Brunel, and Riccardo Zecchina (2007). 'Efficient supervised learning in networks with binary synapses.' In: *Proceedings of the National Academy of Sciences* 104.26, pp. 11079–11084.

Baldassi, Carlo, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina (2015). 'Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses.' In: *Physical review letters* 115.12, p. 128101.

Baldassi, Carlo et al. (2016). 'Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes.' In: *Proceedings of the National Academy of Sciences* 113.48, E7655–E7662.

Baldi, Pierre F and Kurt Hornik (1995). 'Learning in linear neural networks: A survey.' In: *IEEE Transactions on neural networks* 6.4, pp. 837–858.

Baldi, Pierre and Yves Chauvin (1991). 'Temporal evolution of generalization during learning in linear networks.' In: *Neural Computation* 3.4, pp. 589–603.

Balian, Roger, Édouard Brézin, and Jean-Claude Tolédano (1986). *Physique statistique*. École polytechnique, Département de physique.

Bandeira, Afonso, Amelia Perry, and Alexander S. Wein (Jan. 2018). 'Notes on computational-to-statistical gaps: Predictions using statistical physics.' English (US). In: *Portugaliae Mathematica* 75.2, pp. 159–186.

Bansal, Nikhil and Joel H. Spencer (2019). 'On-Line Balancing of Random Inputs.' In: *arXiv preprint arXiv:1903.06898*.

Barber, David (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barbier, J, CL Chan, and N Macris (2019a). 'Concentration of multi-overlaps for random ferromagnetic spin models.' In: *arXiv preprint arXiv:1901.06521*.

Barbier, Jean, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová (2016). 'Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula.' In: *Advances in Neural Information Processing Systems*, pp. 424–432.

Barbier, Jean, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová (2019b). 'Optimal errors and phase transitions in high-dimensional generalized linear models.' In: *Proceedings of the National Academy of Sciences* 116.12, pp. 5451–5460.

Barbier, Jean and Nicolas Macris (2018a). 'The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference.' In: *Probability Theory and Related Fields*, pp. 1–53.

Barbier, Jean, Nicolas Macris, Mohamad Dia, and Florent Krzakala (2017). 'Mutual Information and Optimality of Approximate Message-Passing in Random Linear Estimation.' In: *arXiv preprint arXiv:1701.05823*.

Barbier, Jean, Nicolas Macris, Antoine Maillard, and Florent Krzakala (2018b). 'The mutual information in random linear estimation beyond iid matrices.' In: *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 1390–1394.

Bartlett, Peter L and Wolfgang Maass (2003). 'Vapnik-Chervonenkis dimension of neural nets.' In: *The handbook of brain theory and neural networks*, pp. 1188–1192.

Bartlett, Peter L and Shahar Mendelson (2002). 'Rademacher and Gaussian complexities: Risk bounds and structural results.' In: *Journal of Machine Learning Research* 3.Nov, pp. 463–482.

Bartlett, Peter and John Shawe-taylor (1998). *Generalization Performance of Support Vector Machines and Other Pattern Classifiers*.

Bayati, Mohsen, Marc Lelarge, Andrea Montanari, et al. (2015). 'Universality in polytope phase transitions and message passing algorithms.' In: *The Annals of Applied Probability* 25.2, pp. 753–822.

Bayati, Mohsen and Andrea Montanari (2011a). 'The LASSO risk for Gaussian matrices.' In: *IEEE Transactions on Information Theory* 58.4, pp. 1997–2017.

– (2011b). 'The dynamics of message passing on dense graphs, with applications to compressed sensing.' In: *IEEE Transactions on Information Theory* 57.2, pp. 764–785.

Bayes, Thomas (1763). 'LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S.' In: *Philosophical transactions of the Royal Society of London* 53, pp. 370–418.

Bean, Derek, Peter J. Bickel, Noureddine El Karoui, and Bin Yu (2013). 'Optimal M-estimation in high-dimensional regression.' In: *Proceedings of the National Academy of Sciences of the United States of America* 110.36, pp. 14563–14568.

Belanger, DP and AP Young (1991). 'The random field Ising model.' In: *Journal of magnetism and magnetic materials* 100.1-3, pp. 272–291.

Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal (2019a). 'Reconciling modern machine-learning practice and the classical bias–variance trade-off.' In: *Proceedings of the National Academy of Sciences* 116.32, pp. 15849–15854.

Belkin, Mikhail, Daniel Hsu, and Ji Xu (2019b). 'Two models of double descent for weak features.' In: *arXiv preprint arXiv:1903.07571*.

Ben Arous, Gerard, Amir Dembo, and Alice Guionnet (2006). 'Cugliandolo-Kurchan equations for dynamics of spin-glasses.' In: *Probability theory and related fields* 136.4, pp. 619–660.

Benaych-Georges, Florent and Raj Rao Nadakuditi (2011). 'The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices.' In: *Advances in Mathematics* 227.1, pp. 494–521.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). 'Representation learning: A review and new perspectives.' In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828.

Bengio, Yoshua, Yann LeCun, et al. (2007). 'Scaling learning algorithms towards AI.' In: *Large-scale kernel machines* 34.5, pp. 1–41.

Berthet, Quentin and Philippe Rigollet (2013). 'Computational lower bounds for sparse PCA.' In: *arXiv preprint arXiv:1304.0828.*

Berthier, Ludovic and Giulio Biroli (2011). 'Theoretical perspective on the glass transition and amorphous materials.' In: *Rev. Mod. Phys.* 83 (2), pp. 587–645.

Berthier, Raphael, Andrea Montanari, and Phan-Minh Nguyen (2017). 'State evolution for approximate message passing with non-separable functions.' In: *Information and Inference: A Journal of the IMA.* preprint arXiv:1708.03950.

Bethe, Hans A (1935). 'Statistical theory of superlattices.' In: *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* 150.871, pp. 552–575.

Bex, Geert Jan, Roger Serneels, and Christian Van den Broeck (1995). 'Storage capacity and generalization error for the reversed-wedge Ising perceptron.' In: *Physical Review E* 51.6, p. 6309.

Biehl, Michael, Nestor Caticha, Manfred Opper, and Thomas Villmann (2019). 'Statistical physics of learning and inference.' In: *ESANN.*

Binder, Kurt and A Peter Young (1986). 'Spin glasses: Experimental facts, theoretical concepts, and open questions.' In: *Reviews of Modern physics* 58.4, p. 801.

Biroli, Giulio and Jorge Kurchan (2001). 'Metastable states in glassy systems.' In: *Physical Review E* 64.1, p. 016101.

Bishop, Christopher M. (2006). *Pattern recognition and machine learning.* Springer, p. 738.

Blandin, André, Marc Gabay, and Thomas Garel (1980). 'On the mean-field theory of spin glasses.' In: *Journal of Physics C: Solid State Physics* 13.3, p. 403.

Blandin, Annie (1978). 'Theories versus experiments in the spin glass systems.' In: *Le Journal de Physique Colloques* 39.C6, pp. C6–1499.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). 'Variational inference: A review for statisticians.' In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Blum, Avrim L and Ronald L Rivest (1992). 'Training a 3-node neural network is NP-complete.' In: *Neural Networks* 5.1, pp. 117–127.

Blumer, Anselm, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth (1989). 'Learnability and the Vapnik-Chervonenkis dimension.' In: *Journal of the ACM (JACM)* 36.4, pp. 929–965.

Bobrow, Daniel G (1964). 'Natural language input for a computer problem solving system.' In:

Bolthausen, Erwin (2014). 'An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model.' In: *Communications in Mathematical Physics* 325.1, pp. 333–366.

Bolthausen, Erwin and Anton Bovier (2007). *Spin glasses.* Springer.

Bora, Ashish, Ajil Jalal, Eric Price, and Alexandros G Dimakis (2017). 'Compressed sensing using generative models.' In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR. org, pp. 537–546.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik (1992). 'A training algorithm for optimal margin classifiers.' In: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152.

Bottou, Léon (2010). 'Large-Scale Machine Learning with Stochastic Gradient Descent.' In: *Proceedings of the 19th International Conference on Computational Statistics.* Paris, France: Springer, pp. 177–187.

Bottou, Léon (2012). 'Stochastic gradient descent tricks.' In: *Neural networks: Tricks of the trade.* Springer, pp. 421–436.

Bouchaud, Jean-Philippe, Leticia F Cugliandolo, Jorge Kurchan, and Marc Mezard (1998). 'Out of equilibrium dynamics in spin-glasses and other glassy systems.' In: *Spin glasses and random fields*, pp. 161–223.

Bouchaud, Jean-Philippe and Marc Potters (2003). *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management.* 2nd ed. Cambridge University Press.

Boutros, Joseph and Giuseppe Caire (2002). 'Iterative multiuser joint decoding: Unified framework and asymptotic analysis.' In: *IEEE Transactions on Information Theory* 48.7, pp. 1772–1793.

Boyd, Stephen, Stephen P Boyd, and Lieven Vandenberghe (2004). *Convex optimization.* Cambridge university press.

Braunstein, A., M. Mézard, and R. Zecchina (2005). 'Survey propagation: An algorithm for satisfiability.' In: *Random Structures & Algorithms* 27.2, pp. 201–226.

Braunstein, Alfredo and Riccardo Zecchina (2006). 'Learning by message passing in networks of discrete synapses.' In: *Physical review letters* 96.3, p. 030201.

Bray, A J and M A Moore (1980). 'Broken replica symmetry and metastable states in spin glasses.' In: *Journal of Physics C: Solid State Physics* 13.31, pp. L907–L912.

Breiman, Leo (1995). 'Reflections after refereeing papers for NIPS.' In: *The Mathematics of Generalization*, pp. 11–15.

Brémaud, Pierre (2017). *Discrete probability models and methods.* Vol. 78. Springer.

Burges, Christopher JC (1998). 'A tutorial on support vector machines for pattern recognition.' In: *Data mining and knowledge discovery* 2.2, pp. 121–167.

C., De Dominicis and Garel T. (1979). In: *J. Phys. Lett.* 40.L575.

Cakmak, Burak, Ole Winther, and Bernard H Fleury (2014). 'S-AMP: Approximate message passing for general matrix ensembles.' In: *2014 IEEE Information Theory Workshop (ITW 2014).* IEEE, pp. 192–196.

Campbell, Murray, A Joseph Hoane Jr, and Feng-hsiung Hsu (2002). 'Deep blue.' In: *Artificial intelligence* 134.1-2, pp. 57–83.

Candes, Emmanuel J, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski (2015). 'Phase retrieval via matrix completion.' In: *SIAM review* 57.2, pp. 225–251.

Candès, Emmanuel J, Pragya Sur, et al. (2020). *The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression.*

Candes, Emmanuel J and Terence Tao (2006). 'Near-optimal signal recovery from random projections: Universal encoding strategies?' In: *IEEE transactions on information theory* 52.12, pp. 5406–5425.

Carleo, Giuseppe et al. (2019). 'Machine learning and the physical sciences.' In: *Reviews of Modern Physics* 91.4.

Caruana, Rich (1997). 'Multitask learning.' In: *Machine learning* 28.1, pp. 41–75.

Castellani, Tommaso and Andrea Cavagna (2005). 'Spin-glass theory for pedestrians.' In: *Journal of Statistical Mechanics: Theory and Experiment* 5, pp. 215–266.

Cauchy, Augustin (1847). 'Méthode générale pour la résolution des systemes d'équations simultanées.' In: *Comp. Rend. Sci. Paris* 25.1847, pp. 536–538.

Charbonneau, Patrick, Jorge Kurchan, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi (2014). 'Fractal free energy landscapes in structural glasses.' In: *Nature communications* 5.1, pp. 1–6.

– (2017). 'Glass and jamming transitions: From exact results to finite-dimensional descriptions.' In: *Annual Review of Condensed Matter Physics* 8, pp. 265–288.

Chaudhari, P et al. (2017). 'Entropy-SGD: Biasing Gradient Descent Into Wide Valleys.' In: *International Conference on Learning Representations (ICLR).*

Cherkassky, Vladimir and Filip M Mulier (2007). *Learning from data: concepts, theory, and methods.* John Wiley & Sons.

Chizat, Lenaic and Francis Bach (2018). 'On the global convergence of gradient descent for over-parameterized models using optimal transport.' In: *Advances in neural information processing systems*, pp. 3036–3046.

Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman (2017). 'Lip reading sentences in the wild.' In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, pp. 3444–3453.

Coja-Oghlan, Amin, Florent Krzakala, Will Perkins, and Lenka Zdeborová (2018). 'Information-theoretic thresholds from the cavity method.' In: *Advances in Mathematics* 333, pp. 694–795.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). 'Natural language processing (almost) from scratch.' In: *Journal of machine learning research* 12.ARTICLE, pp. 2493–2537.

Cortes, Corinna and Vladimir Vapnik (1995). 'Support-vector networks.' In: *Machine learning* 20.3, pp. 273–297.

Couillet, Romain and Merouane Debbah (2011). *Random matrix methods for wireless communications.* Cambridge University Press.

Cover, Thomas M (1965). 'Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition.' In: *IEEE transactions on electronic computers* 3, pp. 326–334.

Cover, Thomas M and Joy A Thomas (2012). *Elements of information theory.* John Wiley & Sons.

Craiu, Radu V. and Jeffrey S. Rosenthal (2014). 'Bayesian Computation Via Markov Chain Monte Carlo.' In: *Annual Review of Statistics and Its Application* 1.1, pp. 179–201.

Crisanti, A. and H. J. Sommers (1992). 'The spherical p-spin interaction spin glass model: the statics.' In: *Zeitschrift für Physik B Condensed Matter* 87.3, pp. 341–354.

Crisanti, Andrea and H-J Sommers (1995). 'Thouless-Anderson-Palmer approach to the spherical p-spin spin glass model.' In: *Journal de Physique I* 5.7, pp. 805–813.

Cugliandolo, Leticia F. (2002). *Dynamics of glassy systems.*

Cugliandolo, Leticia F and Jorge Kurchan (1993). 'Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model.' In: *Physical Review Letters* 71.1, p. 173.

– (1994). 'On the out-of-equilibrium relaxation of the Sherrington-Kirkpatrick model.' In: *Journal of Physics A: Mathematical and General* 27.17, p. 5749.

Curie, Pierre (1895). *Propriétés magnétiques des corps a diverses températures.* 4. Gauthier-Villars et fils.

Cybenko, G. (1989). 'Approximation by superpositions of a sigmoidal function.' In: *Mathematics of Control, Signals and Systems* 2.4, pp. 303–314.

De Dominicis, Cirano and Irene Giardina (2006). *Random Fields and Spin Glasses: A Field Theory Approach.* Cambridge University Press.

Decelle, Aurelien, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová (2011). 'Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications.' In: *Physical Review E* 84.6, p. 066106.

Dechter, Rina and Judea Pearl (1988). 'Network-based heuristics for constraint-satisfaction problems.' In: *Search in artificial intelligence.* Springer, pp. 370–425.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). 'Imagenet: A large-scale hierarchical image database.' In: pp. 248–255.

Deng, Zeyu, Abla Kammoun, and Christos Thrampoulidis (2019). 'A Model of Double Descent for High-dimensional Binary Linear Classification.' In: *arXiv preprint arXiv:1911.05822.*

Derrida, Bernard (1981). 'Random-energy model: An exactly solvable model of disordered systems.' In: *Phys. Rev. B* 24 (5), pp. 2613–2626.

Derrida, Bernard, Elizabeth Gardner, and Anne Zippelius (1987). 'An exactly solvable asymmetric neural network model.' In: *EPL (Europhysics Letters)* 4.2, p. 167.

Deshpande, Yash and Andrea Montanari (2014a). 'Information-theoretically optimal sparse PCA.' In: *2014 IEEE International Symposium on Information Theory.* IEEE, pp. 2197–2201.

– (2014b). 'Sparse PCA via covariance thresholding.' In: *Advances in Neural Information Processing Systems*, pp. 334–342.

– (2015). 'Finding Hidden Cliques of Size \sqrt {N/e} N/e in Nearly Linear Time.' In: *Foundations of Computational Mathematics* 15.4, pp. 1069–1128.

Ding, Jian, Allan Sly, and Nike Sun (2015). 'Proof of the satisfiability conjecture for large k.' In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing.* ACM, pp. 59–68.

Ding, Jian and Nike Sun (2019). 'Capacity lower bound for the Ising perceptron.' In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing.* ACM, pp. 816–827.

Diu, Bernard, Bernard Roulet, Claudine Guthmann, and Danielle Lederer (1989). *Eléments de physique statistique.* Hermann.

Doersch, Carl (2016). 'Tutorial on variational autoencoders.' In: *arXiv:1606.05908.*

Donoho, David L (2006). 'Compressed sensing.' In: *IEEE Transactions on information theory* 52.4, pp. 1289–1306.

Donoho, David L, Adel Javanmard, and Andrea Montanari (2013a). 'Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing.' In: *IEEE transactions on information theory* 59.11, pp. 7434–7464.

Donoho, David L, Iain Johnstone, and Andrea Montanari (2013b). 'Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising.' In: *IEEE transactions on information theory* 59.6, pp. 3396–3433.

Donoho, David L, Arian Maleki, and Andrea Montanari (2009). 'Message-passing algorithms for compressed sensing.' In: *Proceedings of the National Academy of Sciences* 106.45, pp. 18914–18919.

Donoho, David and Andrea Montanari (2016). 'High dimensional robust M-estimation: asymptotic variance via approximate message passing.' In: *Probability Theory and Related Fields* 166.3-4, pp. 935–969.

Dreyfus, Hubert L (1965). *Alchemy and artificial intelligence.* Tech. rep.

Dreyfus, Hubert L, Rose-Marie Vassallo-Villaneau, Daniel Andler, Jacques Perriault, Jacques Arsac, and Mario Borillo (1984). *Intelligence artificielle: mythes et limites.* Vol. 198. Flammarion Paris.

Du, Simon S, Xiyu Zhai, Barnabas Poczos, and Aarti Singh (2018). 'Gradient descent provably optimizes over-parameterized neural networks.' In: *arXiv preprint arXiv:1810.02054*.

Duchi, John, Elad Hazan, and Yoram Singer (2011). 'Adaptive subgradient methods for online learning and stochastic optimization.' In: *Journal of machine learning research* 12.7.

Dudeja, Rishabh, Junjie Ma, and Arian Maleki (2019). 'Information Theoretic Limits for Phase Retrieval with Subsampled Haar Sensing Matrices.' In: *arXiv preprint arXiv:1910.11849*.

Dudley, Richard M (1967). 'The sizes of compact subsets of Hilbert space and continuity of Gaussian processes.' In: *Journal of Functional Analysis* 1.3, pp. 290–330.

Dunmur, AP and DJ Wallace (1993). 'Learning and generalization in a linear perceptron stochastically trained with noisy data.' In: *Journal of Physics A: Mathematical and General* 26.21, p. 5767.

Edwards, S F and P W Anderson (1975). 'Theory of spin glasses.' In: *Journal of Physics F: Metal Physics* 5.5, pp. 965–974.

El Alaoui, Ahmed, Aaditya Ramdas, Florent Krzakala, Lenka Zdeborová, and Michael I Jordan (2017). 'Decoding from pooled data: Phase transitions of message passing.' In: *Information Theory (ISIT), 2017 IEEE International Symposium on.* IEEE, pp. 2780–2784.

El Alaoui, Ahmed, Aaditya Ramdas, Florent Krzakala, Lenka Zdeborova, and Michael Jordan (Nov. 2016). 'Decoding from Pooled Data: Sharp Information-Theoretic Bounds.' In: *SIAM Journal on Mathematics of Data Science* 1.

El Karoui, Noureddine, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu (2013). 'On robust regression with high-dimensional predictors.' In: *Proceedings of the National Academy of Sciences* 110.36, pp. 14557–14562.

Ellis, Richard S et al. (1984). 'Large deviations for a general class of random vectors.' In: *The Annals of Probability* 12.1, pp. 1–12.

Engel, A and W Fink (1993). 'Statistical mechanics calculation of Vapnik-Chervonenkis bounds for perceptrons.' In: *Journal of Physics A: Mathematical and General* 26.23, p. 6893.

Engel, Andreas and Christian Van den Broeck (2001). *Statistical mechanics of learning.* Cambridge University Press.

Erichsen, R and W K Thuemann (1993). 'Optimal storage of a neural network model: a replica symmetry-breaking solution.' In: *Journal of Physics A: Mathematical and General* 26.2, pp. L61–L68.

Evans, Thomas G (1964). 'A heuristic program to solve geometric-analogy problems.' In: *Proceedings of the April 21-23, 1964, spring joint computer conference*, pp. 327–338.

Fischer, Konrad H and John A Hertz (1993). *Spin glasses.* Vol. 1. Cambridge university press.

Fisher, R.A. (1925). *Statistical methods for research workers.* American Psychological Association.

Fletcher, Alyson K., Sundeep Rangan, and Philip Schniter (2018). 'Inference in Deep Networks in High Dimensions.' In: *IEEE International Symposium on Information Theory - Proceedings* 2018-June, pp. 1884–1888.

Franco, John and Marvin Paull (1983). 'Probabilistic analysis of the Davis Putnam procedure for solving the satisfiability problem.' In: *Discrete Applied Mathematics* 5.1, pp. 77–87.

Franz, S., G. Parisi, M. Sevelev, P. Urbani, and F. Zamponi (2017). 'Universality of the SAT-UNSAT (jamming) threshold in non-convex continuous constraint satisfaction problems.' In: *SciPost Phys.*

Franz, Silvio and Giorgio Parisi (1997). 'Phase diagram of coupled glassy systems: A mean-field study.' In: *Physical review letters* 79.13, p. 2486.

– (2016). 'The simplest model of jamming.' In: *Journal of Physics A: Mathematical and Theoretical* 49.14, p. 145001.

Franz, Silvio, Guilhem Semerjian, et al. (2011). 'Analytical approaches to time-and length scales in models of glasses.' In: *Dynamical Heterogeneities in Glasses, Colloids, and Granular Media* 407.

Friedgut, Ehud (1999). 'Sharp thresholds of graph properties, and the k-SAT problem.' In: *Journal of the American mathematical Society* 12.4, pp. 1017–1054.

Gabrié, Marylou (2020). 'Mean-field inference methods for neural networks.' In: *Journal of Physics A: Mathematical and Theoretical* 53.22, p. 223002.

Gabrié, Marylou, Varsha Dani, Guilhem Semerjian, and Lenka Zdeborová (2017). 'Phase transitions in the q-coloring of random hypergraphs.' In: *Journal of Physics A: Mathematical and Theoretical* 50.50.

Gabrié, Marylou et al. (2018). 'Entropy and mutual information in models of deep neural networks.' In: *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., pp. 1821–1831.

Gallager, Robert (1962). 'Low-density parity-check codes.' In: *IRE Transactions on information theory* 8.1, pp. 21–28.

Ganascia, Jean-Gabriel (1993). *L'intelligence artificielle*. Flammarion Paris.

Gardner, Elisabeth (1985). 'Spin glasses with p-spin interactions.' In: *Nuclear Physics B* 257, pp. 747–765.

Gardner, Elizabeth (1987). 'Maximum storage capacity in neural networks.' In: *EPL (Europhysics Letters)* 4.4, p. 481.

– (1988). 'The space of interactions in neural network models.' In: *Journal of physics A: Mathematical and general* 21.1, p. 257.

Gardner, Elizabeth and Bernard Derrida (1988). 'Optimal storage properties of neural network models.' In: *Journal of Physics A: Mathematical and general* 21.1, p. 271.

– (1989). 'Three unfinished works on the optimal storage capacity of networks.' In: *Journal of Physics A: Mathematical and General* 22.12, p. 1983.

Gärtner, Jürgen (1977). 'On large deviations from the invariant measure.' In: *Theory of Probability & Its Applications* 22.1, pp. 24–39.

Gaspari, George and Joseph Rudnick (1986). 'n-vector model in the limit n→0 and the statistics of linear polymer systems: A Ginzburg-Landau theory.' In: *Phys. Rev. B* 33 (5), pp. 3295–3305.

Gauvin, Laetitia, Jean Vannimenus, and J-P Nadal (2009). 'Phase diagram of a Schelling segregation model.' In: *The European Physical Journal B* 70.2, pp. 293–304.

Geiger, Mario et al. (2019). 'Jamming transition as a paradigm to understand the loss landscape of deep neural networks.' In: *Physical Review E* 100.1.

Geiger, Mario et al. (2020). 'Scaling description of generalization with number of parameters in deep learning.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2020.2, p. 023401.

Gennes, P.G. de (1972). 'Exponents for the excluded volume problem as derived by the Wilson method.' In: *Physics Letters A* 38.5, pp. 339 –340.

Georges, Antoine, David Hansel, Pierre Le Doussal, and J-P Bouchaud (1985). 'Exact properties of spin glasses. II. Nishimori's line: new results and physical implications.' In: *Journal de Physique* 46.11, pp. 1827–1836.

Georges, Antoine and Marc Mézard (2004). *Introduction à la théorie statistique des champs: majeure de physique: Promotion 2001, Année 3, Majeure 2, PHY557B*. École polytechnique, Département de Physique.

Georges, Antoine and Jonathan S Yedidia (1991). 'How to expand around mean-field theory using high-temperature expansions.' In: *Journal of Physics A: Mathematical and General* 24, pp. 2173–2192.

Gerace, Federica, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová (2020). 'Generalisation error in learning with random features and the hidden manifold model.' In: *arXiv preprint arXiv:2002.09339*.

Gerbelot, Cédric, Alia Abbara, and Florent Krzakala (2020). 'Asymptotic errors for convex penalized linear regression beyond Gaussian matrices.' In: *arXiv preprint arXiv:2002.04372*.

Goldt, Sebastian, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová (2019a). 'Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup.' In: *Advances in Neural Information Processing Systems*, pp. 6981–6991.

Goldt, Sebastian, Marc Mézard, Florent Krzakala, and Lenka Zdeborová (2019b). 'Modelling the influence of data structure on learning in neural networks.' In: *arXiv preprint arXiv:1909.11500*.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. The MIT Press.

Goodfellow, Ian et al. (2014). 'Generative adversarial nets.' In: *Advances in neural information processing systems*, pp. 2672–2680.

Gordon, Yehoram (1985). 'Some inequalities for Gaussian processes and applications.' In: *Israel Journal of Mathematics* 50.4, pp. 265–289.

Grassberger, Peter and Jean-Pierre Nadal (2012). *From statistical physics to statistical inference and back*. Vol. 428. Springer Science & Business Media.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). 'Speech recognition with deep recurrent neural networks.' In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 6645–6649.

Gribonval, Rémi (2011). 'Should penalized least squares regression be interpreted as maximum a posteriori estimation?' In: *IEEE Transactions on Signal Processing* 59.5, pp. 2405–2410.

Gribonval, Remi and Pierre Machart (2013). 'Reconciling "priors" &amp; "priors" without prejudice?' In: *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pp. 2193–2201.

Gribonval, Rémi and Mila Nikolova (2018). 'A characterization of proximity operators.' In: *arXiv preprint arXiv:1807.04014*.

– (2019). 'On bayesian estimation and proximity operators.' In: *Applied and Computational Harmonic Analysis*.

Griewank, Andreas (1992). 'Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation.' In: *Optimization Methods and software* 1.1, pp. 35–54.

Grossberg, Stephen (1976). 'Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors.' In: *Biological cybernetics* 23.3, pp. 121–134.

Guerra, Francesco (2003). 'Broken Replica Symmetry Bounds in the Mean Field Spin Glass Model.' In: *Communications in Mathematical Physics* 233.1, pp. 1–12.

Guerra, Francesco and Fabio Lucio Toninelli (2002a). 'Central limit theorem for fluctuations in the high temperature region of the Sherrington–Kirkpatrick spin glass model.' In: *Journal of Mathematical Physics* 43.12, pp. 6224–6237.

Guerra, Francesco and Fabio Lucio Toninelli (2002b). 'The thermodynamic limit in mean field spin glass models.' In: *Communications in Mathematical Physics* 230.1, pp. 71–79.

Guo, Dongning, S. Shamai, and S. Verdú (2005a). 'Mutual information and minimum mean-square error in Gaussian channels.' In: *IEEE Transactions on Information Theory* 51.4, pp. 1261–1282.

Guo, Dongning and Sergio Verdú (2005b). 'Randomly spread CDMA: Asymptotics via statistical physics.' In: *IEEE Transactions on Information Theory* 51.6, pp. 1983–2010.

Györgyi, G. (2001). 'Techniques of replica symmetry breaking and the storage problem of the McCulloch-Pitts neuron.' In: *Physics Report* 342.4-5, pp. 263–392.

Györgyi, Géza (1990). 'First-order transition to perfect generalization in a neural network with binary synapses.' In: *Physical Review A* 41.12, p. 7097.

H. Huang, K.Y.M Wong & Y. Kabashima (2013). 'Entropy landscape of solutions in the binary perceptron problem.' In: *Journal of Physics A: Mathematical and Theoretical*.

Hammersley, John M and Peter Clifford (1971). 'Markov fields on finite graphs and lattices.' In: *Unpublished manuscript* 46.

Hand, Paul, Oscar Leong, and Vlad Voroninski (2018a). 'Phase retrieval under a generative prior.' In: *Advances in Neural Information Processing Systems*, pp. 9136–9146.

Hand, Paul and Vladislav Voroninski (2018b). 'Global Guarantees for Enforcing Deep Generative Priors by Empirical Risk.' In: *Conference On Learning Theory*, pp. 970–978.

Hansel, David and Haim Sompolinsky (1990). 'Learning from examples in a single-layer neural network.' In: *EPL (Europhysics Letters)* 11.7, p. 687.

Hartman, P. (1982). *Ordinary Differential Equations: Second Edition*. Classics in Applied Mathematics.

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani (2019). *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Haussler, David, Michael Kearns, H Sebastian Seung, and Naftali Tishby (1996). 'Rigorous learning curve bounds from statistical mechanics.' In: *Machine Learning* 25.2-3, pp. 195–236.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). 'Deep residual learning for image recognition.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hebb, Donald Olding (1962). *The organization of behavior: a neuropsychological theory*. Science Editions.

Hecht-Nielsen, Robert (1989). 'Neurocomputer applications.' In: *Neural computers*. Springer, pp. 445–453.

Hertz, John and Holm Schwarze (1993). 'Generalization in large committee machines.' In: *Physica A: Statistical Mechanics and its Applications* 200.1-4, pp. 563–569.

Heskes, Tom, Manfred Opper, Wim Wiegerinck, Ole Winther, and Onno Zoeter (2005). 'Approximate inference techniques with expectation constraints.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.11, P11015.

Heskes, Tom and Onno Zoeter (2012). 'Expectation Propogation for approximate inference in dynamic Bayesian networks.' In: *arXiv preprint arXiv:1301.0572*.

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh (2006). 'A fast learning algorithm for deep belief nets.' In: *Neural computation* 18.7, pp. 1527–1554.

Hinton, Geoffrey E, Terrence J Sejnowski, et al. (1986). 'Learning and relearning in Boltzmann machines.' In: *Parallel distributed processing: Explorations in the microstructure of cognition* 1.282-317, p. 2.

Hiriart-Urruty, J.-B. and C. Lemarechal (1993). *Convex analysis and minimization algorithms II*. Springer-Verlag.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). 'Long short-term memory.' In: *Neural computation* 9.8, pp. 1735–1780.

Hoeffding, Wassily (1994). 'Probability inequalities for sums of bounded random variables.' In: *The Collected Works of Wassily Hoeffding*. Springer, pp. 409–426.

Hoerl, Arthur E and Robert W Kennard (1970). 'Ridge regression: Biased estimation for nonorthogonal problems.' In: *Technometrics* 12.1, pp. 55–67.

Hopfield, John J (1982). 'Neural networks and physical systems with emergent collective computational abilities.' In: *Proceedings of the national academy of sciences* 79.8, pp. 2554–2558.

Hornik, Kurt (1991). 'Approximation capabilities of multilayer feedforward networks.' In: *Neural Networks* 4.2, pp. 251 –257.

Hosaka, Tadaaki, Yoshiyuki Kabashima, and Hidetoshi Nishimori (2002). 'Statistical mechanics of lossy data compression using a nonmonotonic perceptron.' In: *Physical Review E* 66.6, p. 066126.

Huang, Haiping and Yoshiyuki Kabashima (2014). 'Origin of the computational hardness for learning with binary synapses.' In: *Physical Review E* 90.5, p. 052813.

Hutchins, W John (2001). 'Machine translation over fifty years.' In: *Histoire epistémologie langage* 23.1, pp. 7–31.

Iba, Yukito (1999). 'The Nishimori line and Bayesian statistics.' In: *Journal of Physics A: Mathematical and General* 32.21, p. 3875.

Imbrie, John Z (1984). 'Lower critical dimension of the random-field Ising model.' In: *Physical review letters* 53.18, p. 1747.

Ioffe, Sergey and Christian Szegedy (2015). 'Batch normalization: Accelerating deep network training by reducing internal covariate shift.' In: *arXiv preprint arXiv:1502.03167*.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros (2017). 'Image-to-image translation with conditional adversarial networks.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.

Jaakkola, Tommi S and Michael I Jordan (2000). 'Bayesian parameter estimation via variational methods.' In: *Statistics and Computing* 10.1, pp. 25–37.

Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018). 'Neural tangent kernel: Convergence and generalization in neural networks.' In: *Advances in neural information processing systems*, pp. 8571–8580.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Javanmard, Adel and Andrea Montanari (2013). 'State evolution for general approximate message passing algorithms, with applications to spatial coupling.' In: *Information and Inference: A Journal of the IMA* 2.2, pp. 115–144.

Jaynes, E. T. (1957). 'Information Theory and Statistical Mechanics.' In: *Phys. Rev.* 106 (4), pp. 620–630.

– (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Jenatton, Rodolphe, Guillaume Obozinski, and Francis Bach (2010). 'Structured sparse principal component analysis.' In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 366–373.

Jensen, Tommy R and Bjarne Toft (2011). *Graph coloring problems*. Vol. 39. John Wiley & Sons.

Jolliffe, Ian T (1986). 'Principal components in regression analysis.' In: *Principal component analysis*. Springer, pp. 129–155.

Jordan, Michael I, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul (1999). 'An introduction to variational methods for graphical models.' In: *Machine learning* 37.2, pp. 183–233.

Jordan, Michael I et al. (2004). 'Graphical models.' In: *Statistical science* 19.1, pp. 140–155.

Kabashima, Yoshiyuki (2003). 'A CDMA multiuser detection algorithm on the basis of belief propagation.' In: *Journal of Physics A: Mathematical and General* 36.43, p. 11111.

– (2008). 'Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels.' In: *Journal of Physics: Conference Series*. Vol. 95. 1. IOP Publishing, p. 012001.

Kabashima, Yoshiyuki and David Saad (1998). 'Belief propagation vs. TAP for decoding corrupted messages.' In: *EPL (Europhysics Letters)* 44.5, p. 668.

Kabashima, Yoshiyuki and Shinsuke Uda (2004). 'A BP-based algorithm for performing Bayesian inference in large perceptron-type networks.' In: *International Conference on Algorithmic Learning Theory*. Springer, pp. 479–493.

Kac (1968). 'Broken ergodicity.' In: *Trondheim Theoretical Physics Seminar, Nordita Publ* 286.

Kamilov, Ulugbek, Sundeep Rangan, Michael Unser, and Alyson K Fletcher (2012). 'Approximate message passing with consistent parameter estimation and applications to sparse learning.' In: *Advances in Neural Information Processing Systems*, pp. 2438–2446.

Kardar, Mehran (2007). *Statistical physics of particles*. Cambridge University Press.

Kardar, Mehran, Giorgio Parisi, and Yi-Cheng Zhang (1986). 'Dynamic scaling of growing interfaces.' In: *Physical Review Letters* 56.9, p. 889.

Kaufman, Leonard and Peter J Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

Keras-VAE (2020). *Example of VAE on MNIST dataset using MLP*. *https://keras.io/examples/generative/vae/*.

Kim, Jeong Han and James R Roche (1998). 'Covering cubes by random half cubes, with applications to binary neural networks.' In: *Journal of Computer and System Sciences* 56.2, pp. 223–252.

Kingma, Diederik P and Jimmy Ba (2014). 'Adam: A method for stochastic optimization.' In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). 'Auto-encoding variational bayes.' In: *arXiv preprint arXiv:1312.6114*.

– (2019). 'An introduction to variational autoencoders.' In: *arXiv preprint arXiv:1906.02691*.

Kini, Ganesh and Christos Thrampoulidis (2020). 'Analytic Study of Double Descent in Binary Classification: The Impact of Loss.' In: *arXiv preprint arXiv:2001.11572*.

Kinouchi, Osame and Nestor Caticha (1996). 'Learning algorithm that gives the bayes generalization limit for perceptrons.' In: *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 54.1, R54–R57.

Kirkpatrick, TR and PG Wolynes (1987a). 'Stable and metastable states in mean-field Potts and structural glasses.' In: *Physical Review B* 36.16, p. 8552.

Kirkpatrick, Theodore R and Devarajan Thirumalai (1987b). 'Dynamics of the structural glass transition and the p-spin—interaction spin-glass model.' In: *Physical review letters* 58.20, p. 2091.

– (1987c). 'p-spin-interaction spin-glass models: Connections with the structural glass problem.' In: *Physical Review B* 36.10, p. 5388.

Kohonen, T, P Lehtiö, J Rovamo, J Hyvärinen, K Bry, and L Vainio (1977). 'A principle of neural associative memory.' In: *Neuroscience* 2.6, pp. 1065–1076.

Koller, Daphne and Nir Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Krauth, Werner and Marc Mézard (1987). 'Learning algorithms with optimal stability in neural networks.' In: *Journal of Physics A: Mathematical and General* 20.11, p. L745.

– (1989). 'Storage capacity of memory networks with binary couplings.' In: *Journal de Physique* 50.20, pp. 3057–3066.

Krizhevsky, Alex, Vinod Nair, and Geoffrey Hinton (2010). 'CIFAR-10 (Canadian Institute for Advanced Research).' In:

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). 'Imagenet classification with deep convolutional neural networks.' In: *Advances in neural information processing systems*, pp. 1097–1105.

Krogh, Anders and John A Hertz (1992). 'A simple weight decay can improve generalization.' In: *Advances in neural information processing systems*, pp. 950–957.

Krzakala, F., M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová (May 2012a). 'Statistical-Physics-Based Reconstruction in Compressed Sensing.' en. In: *Physical Review X* 2.2.

Krzakala, F. et al. (Dec. 2013). 'Spectral redemption in clustering sparse networks.' en. In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20935–20940.

Krzakala, Florent, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová (2012b). 'Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.08, P08009.

Krzakała, Florent, Andrea Montanari, Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová (2007). 'Gibbs states and the set of solutions of random constraint satisfaction problems.' In: *Proceedings of the National Academy of Sciences* 104.25, pp. 10318–10323.

Krzakala, Florent, Jiaming Xu, and Lenka Zdeborová (Sept. 2016). 'Mutual Information in Rank-One Matrix Estimation.' In: *2016 IEEE Information Theory Workshop (ITW)*. arXiv: 1603.08447, pp. 71–75.

Krzakala, Florent and Lenka Zdeborová (2009). 'Hiding quiet solutions in random constraint satisfaction problems.' In: *Physical review letters* 102.23, p. 238701.

Lazard, Emmanuel and Pierre-Eric Mounier-Kuhn (2016). *Histoire illustrée de l'infor matique*. EDP sciences.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). 'Deep learning.' In: *Nature* 521.7553, pp. 436–444.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). 'Gradient-based learning applied to document recognition.' In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

LeCun, Yann and Corinna Cortes (2010). 'MNIST handwritten digit database.' In:

LeCun, Yann, Patrick Haffner, Léon Bottou, and Yoshua Bengio (1999). 'Object recognition with gradient-based learning.' In: *Shape, contour and grouping in computer vision*. Springer, pp. 319–345.

LeCun, Yann et al. (1989). 'Backpropagation applied to handwritten zip code recognition.' In: *Neural computation* 1.4, pp. 541–551.

Ledoux, Michel and Michel Talagrand (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lee, Jaehoon et al. (2019). 'Wide neural networks of any depth evolve as linear models under gradient descent.' In: *Advances in neural information processing systems*, pp. 8572–8583.

Lelarge, Marc and Léo Miolane (2019). 'Fundamental limits of symmetric low-rank matrix estimation.' In: *Probability Theory and Related Fields* 173.3-4, pp. 859–929.

Lesieur, Thibault, Florent Krzakala, and Lenka Zdeborová (2015). 'Phase transitions in sparse PCA.' In: *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 1635–1639.

– (2017a). 'Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2017.7, p. 073403.

Lesieur, Thibault, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová (June 2017b). 'Statistical and computational phase transitions in spiked tensor estimation.' In: *2017 IEEE International Symposium on Information Theory (ISIT)*. arXiv: 1701.08010, pp. 511–515.

Lighthill, James (1973). 'Artificial intelligence: A general survey.' In: *Artificial Intelligence: a paper symposium*. Science Research Council London, pp. 1–21.

Louart, Cosme, Zhenyu Liao, Romain Couillet, et al. (2018). 'A random matrix approach to neural networks.' In: *The Annals of Applied Probability* 28.2, pp. 1190–1248.

Lubchenko, Vassiliy and Peter G Wolynes (2007). 'Theory of structural glasses and supercooled liquids.' In: *Annu. Rev. Phys. Chem.* 58, pp. 235–266.

Ma, Shang-Keng (2018). *Modern theory of critical phenomena*. Routledge.

MacKay, David JC and David JC Mac Kay (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

MacKay, David JC and Radford M Neal (1996). 'Near Shannon limit performance of low density parity check codes.' In: *Electronics letters* 32.18, pp. 1645–1646.

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu (2017). 'Towards deep learning models resistant to adversarial attacks.' In: *arXiv preprint arXiv:1706.06083*.

Majer, P., A. Engel, and A. Zippelius (1993). 'Perceptrons above saturation.' In: *Journal of Physics A: Mathematical and General* 26.24, pp. 7405–7416.

Maleki, Arian (2011). 'Approximate message passing algorithms for compressed sensing.' In: *a degree of Doctor of Philosophy, Stanford University*.

Mannelli, Stefano Sarao, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová (2020). 'Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference.' In: *Physical Review X* 10.1, p. 011057.

Manoel, Andre, Florent Krzakala, Marc Mézard, and Lenka Zdeborová (2017). 'Multi-layer generalized linear estimation.' In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 2098–2102.

Mantegna, Rosario and H. Stanley (Dec. 2000). *An Introduction to Econophysics: Correlations and Complexity in Finance*. Vol. 53.

Martin, Charles H and Michael W Mahoney (2017). 'Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior.' In: *arXiv preprint arXiv:1710.09553*.

Martin, OC, M Mézard, and O Rivoire (2004). 'Frozen glass phase in the multi-index matching problem.' In: *Physical review letters* 93.21, p. 217205.

Massart, Pascal (2000). 'Some applications of concentration inequalities to statistics.' In: *Annales de la Faculté des sciences de Toulouse: Mathématiques*. Vol. 9, pp. 245–303.

Mato, German and Nestor Parga (1992). 'Generalization properties of multilayered neural networks.' In: *Journal of Physics A: Mathematical and General* 25.19, p. 5047.

McClelland, James L, David E Rumelhart, and Geoffrey E Hinton (1986). 'The appeal of parallel distributed processing.' In: *MIT Press, Cambridge MA*, pp. 3–44.

McCulloch, Warren S and Walter Pitts (1943). 'A logical calculus of the ideas immanent in nervous activity.' In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133.

Mehta, Pankaj and David J Schwab (2014). 'An exact mapping between the variational renormalization group and deep learning.' In: *arXiv preprint arXiv:1410.3831*.

Mehta, Pankaj et al. (2019). 'A high-bias, low-variance introduction to Machine Learning for physicists.' In: *Physics Reports* 810, pp. 1–124.

Mei, Song and Andrea Montanari (2019). 'The generalization error of random features regression: Precise asymptotics and double descent curve.' In: *arXiv preprint arXiv:1908.05355*.

Mei, Song, Andrea Montanari, and Phan-Minh Nguyen (2018). 'A mean field view of the landscape of two-layer neural networks.' In: *Proceedings of the National Academy of Sciences* 115.33, E7665–E7671.

Metzler, Christopher A, Arian Maleki, and Richard G Baraniuk (2016). 'From denoising to compressed sensing.' In: *IEEE Transactions on Information Theory* 62.9, pp. 5117–5144.

Metzler, Christopher A, Manoj K Sharma, Sudarshan Nagesh, Richard G Baraniuk, Oliver Cossairt, and Ashok Veeraraghavan (2017). 'Coherent inverse scattering via transmission matrices: Efficient phase retrieval algorithms and a public dataset.' In: *2017 IEEE International Conference on Computational Photography (ICCP)*. IEEE, pp. 1–16.

Mézard, M., G. Parisi, and M. A. Virasoro (1986). 'SK model: The replica solution without replicas.' In: *Epl* 1.2, pp. 77–82.

Mézard, Marc (1989). 'The space of interactions in neural networks: Gardner's computation with the cavity method.' In: *Journal of Physics A: Mathematical and General* 22.12, p. 2181.

Mézard, Marc and Andrea Montanari (2009). *Information, physics, and computation*. Oxford University Press.

Mézard, Marc, Thierry Mora, and Riccardo Zecchina (2005). 'Clustering of solutions in the random satisfiability problem.' In: *Physical Review Letters* 94.19, p. 197205.

Mézard, Marc and Giorgio Parisi (1986a). 'A replica analysis of the travelling salesman problem.' In: *Journal de Physique* 47.8, pp. 1285–1296.

– (1986b). 'Mean-field equations for the matching and the travelling salesman problems.' In: *EPL (Europhysics Letters)* 2.12, p. 913.

– (2000). 'Statistical physics of structural glasses.' In: *Journal of Physics: Condensed Matter* 12.29, p. 6655.

– (2001). 'The Bethe lattice spin glass revisited.' In: *The European Physical Journal B-Condensed Matter and Complex Systems* 20.2, pp. 217–233.

– (2003). 'The cavity method at zero temperature.' In: *Journal of Statistical Physics* 111.1-2, pp. 1–34.

Mézard, Marc, Giorgio Parisi, and Miguel Virasoro (1987). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Publishing Company.

Mézard, Marc and Allan Peter Young (1992). 'Replica symmetry breaking in the random field Ising model.' In: *EPL (Europhysics Letters)* 18.7, p. 653.

Mézard, Marc and Riccardo Zecchina (2002). 'Random k-satisfiability problem: From an analytic solution to an efficient algorithm.' In: *Physical Review E* 66.5, p. 056126.

Mignacco, Francesca, Florent Krzakala, Yue M Lu, and Lenka Zdeborová (2020a). 'The role of regularization in classification of high-dimensional noisy Gaussian mixture.' In: *arXiv preprint arXiv:2002.11544*.

Mignacco, Francesca, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová (2020b). 'Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification.' In: *arXiv preprint arXiv:2006.06098*.

Minka, Thomas P. (2001a). 'Expectation Propagation for approximate Bayesian inference.' In: pp. 362–369.

Minka, Thomas Peter (2001b). 'A family of algorithms for approximate Bayesian inference.' PhD thesis. Massachusetts Institute of Technology.

Minsky, Marvin and Seymour A Papert (1969). *Perceptrons: An introduction to computational geometry*. MIT press.

Miolane, Léo (2017). 'Fundamental limits of low-rank matrix estimation: the non-symmetric case.' In: *arXiv preprint arXiv:1702.00473*.

Mitchell, Thomas M. (1997). *Machine Learning*. 1st ed. USA: McGraw-Hill, Inc.

Mitchison, G. J. and R. M. Durbin (1989). 'Bounds on the learning capacity of some multi-layer networks.' In: *Biological Cybernetics* 60.5, pp. 345–365.

Mitra, Partha P (2019). *Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation*.

Mixon, Dustin G and Soledad Villar (2018). 'SUNLayer: Stable denoising with generative networks.' In: *arXiv preprint arXiv:1803.09319*.

Mnih, Volodymyr et al. (2013). 'Playing atari with deep reinforcement learning.' In: *arXiv preprint arXiv:1312.5602*.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2012). *Foundations of Machine Learning*. The MIT Press.

Monasson, Rémi and Riccardo Zecchina (1995a). 'Learning and generalization theories of large committee-machines.' In: *Modern Physics Letters B* 9.30, pp. 1887–1897.

– (1995b). 'Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks.' In: *Physical review letters* 75.12, p. 2432.

Monasson, Rémi, Riccardo Zecchina, Scott Kirkpatrick, Bart Selman, and Lidror Troyansky (1999). 'Determining computational complexity from characteristic 'phase transitions'.' In: *Nature* 400.6740, pp. 133–137.

Montanari, Andrea (2012). *Graphical models concepts in compressed sensing*.

Montanari, Andrea, Feng Ruan, Youngtak Sohn, and Jun Yan (2019). *The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime*.

Montanari, Andrea and David Tse (2006). 'Analysis of belief propagation for non-linear problems: The example of CDMA (or: How to prove Tanaka's formula).' In: *2006 IEEE Information Theory Workshop-ITW'06 Punta del Este*. IEEE, pp. 160–164.

Moore, Cristopher and Stephan Mertens (2011). *The nature of computation*. OUP Oxford.

Mossel, Elchanan, Joe Neeman, and Allan Sly (2015). 'Reconstruction and estimation in the planted partition model.' In: *Probability Theory and Related Fields* 162.3-4, pp. 431–461.

Mulet, Roberto, Andrea Pagnani, Martin Weigt, and Riccardo Zecchina (2002). 'Coloring random graphs.' In: *Physical review letters* 89.26, p. 268701.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press.

Netrapalli, Praneeth, Prateek Jain, and Sujay Sanghavi (2013). 'Phase retrieval using alternating minimization.' In: *Advances in Neural Information Processing Systems*, pp. 2796–2804.

Neyshabur, Behnam, Srinadh Bhojanapalli, David McAllester, and Nati Srebro (2017). 'Exploring generalization in deep learning.' In: *Advances in neural information processing systems*, pp. 5947–5956.

Nishimori, Hidetoshi (1980). 'Exact results and critical properties of the Ising model with competing interactions.' In: *Journal of Physics C: Solid State Physics* 13.21, p. 4071.

– (1981). 'Internal energy, specific heat and correlation function of the bond-random ising model.' In: *Progress of Theoretical Physics* 66.4, pp. 1169–1181.

– (2001). *Statistical physics of spin glasses and information processing: an introduction*. 111. Clarendon Press.

Onsager, Lars (1944). 'Crystal statistics. I. A two-dimensional model with an order-disorder transition.' In: *Physical Review* 65.3-4, p. 117.

Oono, Yoshitsugu (June 1989). 'Large Deviation and Statistical Physics.' In: *Progress of Theoretical Physics Supplement* 99, pp. 165–205.

Opper, M. and W. Kinzel (1996a). *Models of neural networks III.* Springer. Chap. Statistical mechanics of generalization, pp. 151–209.

Opper, M., W. Kinzel, J. Kleinz, and R. Nehl (1990). 'On the ability of the optimal perceptron to generalise.' In: *Journal of Physics A: General Physics* 23.11.

Opper, Manfred (1995). 'Statistical physics estimates for the complexity of feedforward neural networks.' In: *Physical Review E* 51.4, p. 3613.

Opper, Manfred and David Haussler (1991a). 'Calculation of the learning curve of Bayes optimal classification algorithm for learning a perceptron with noise.' In: Citeseer.

– (1991b). 'Generalization performance of Bayes optimal classification algorithm for learning a perceptron.' In: *Physical Review Letters* 66.20, pp. 2677–2680.

Opper, Manfred and David Saad (2001a). *Advanced mean field methods: Theory and practice.* MIT press.

Opper, Manfred and Ole Winther (1996b). 'Mean field approach to Bayes learning in feed-forward neural networks.' In: *Physical review letters* 76.11, p. 1964.

– (2001b). 'Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling.' In: *Physical Review E* 64.5, p. 056131.

– (2001c). 'Tractable approximations for probabilistic models: The adaptive Thouless-Anderson-Palmer mean field approach.' In: *Physical Review Letters* 86.17, p. 3695.

– (2005). 'Expectation consistent free energies for approximate inference.' In: *Advances in Neural Information Processing Systems*, pp. 1001–1008.

Orlandini, E, M C Tesi, and S G Whittington (2002). 'Self-averaging in the statistical mechanics of some lattice models.' In: *Journal of Physics A: Mathematical and General* 35.19, pp. 4219–4227.

Palmer, R.G.xs (1982). 'Broken ergodicity.' In: *Advances in Physics* 31.6, pp. 669–735.

Pan, Sinno Jialin and Qiang Yang (2010). 'A survey on transfer learning.' In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.

Panchenko, Dmitry (2013). *The Sherrington-Kirkpatrick model.* Springer Science & Business Media.

– (2014). 'The Parisi formula for mixed *p*-spin models.' In: *The Annals of Probability* 42.3, pp. 946–958.

Panchenko, Dmitry and Michel Talagrand (2004). 'Bounds for diluted mean-fields spin glass models.' In: *Probability Theory and Related Fields* 130.3, pp. 319–336.

Panchenko, Dmitry et al. (2018). 'Free energy in the Potts spin glass.' In: *The Annals of Probability* 46.2, pp. 829–864.

Parisi, G (1980a). 'A sequence of approximated solutions to the S-K model for spin glasses.' In: *Journal of Physics A: Mathematical and General* 13.4, pp. L115–L121.

– (1980b). 'Magnetic properties of spin glasses in a new mean field theory.' In: *Journal of Physics A: Mathematical and General* 13.5, pp. 1887–1895.

Parisi, Giorgio (1979). 'Infinite number of order parameters for spin-glasses.' In: *Physical Review Letters* 43.23, p. 1754.

– (1980c). 'A sequence of approximated solutions to the SK model for spin glasses.' In: *Journal of Physics A: Mathematical and General* 13.4, p. L115.

– (1980d). 'The order parameter for spin glasses: a function on the interval 0-1.' In: *Journal of Physics A: Mathematical and General* 13.3, p. 1101.

– (1983). 'Order Parameter for Spin-Glasses.' In: *Phys. Rev. Lett.* 50 (24), pp. 1946–1948.

Parisi, Giorgio and Marc Potters (1995). 'Mean-field equations for spin models with orthogonal interaction matrices.' In: *Journal of Physics A: Mathematical and General* 28.18, p. 5267.

Parisi, Giorgio, Pierfrancesco Urbani, and Francesco Zamponi (2020). *Theory of Simple Glasses: Exact Solutions in Infinite Dimensions.* Cambridge University Press.

Parisi, Giorgio and Francesco Zamponi (2010). 'Mean-field theory of hard sphere glasses and jamming.' In: *Reviews of Modern Physics* 82.1, p. 789.

Parker, Jason T, Philip Schniter, and Volkan Cevher (2014). 'Bilinear generalized approximate message passing—Part I: Derivation.' In: *IEEE Transactions on Signal Processing* 62.22, pp. 5839–5853.

Parkhi, Omkar M, Andrea Vedaldi, and Andrew Zisserman (2015). 'Deep face recognition.' In:

Patarnello, S and P Carnevali (1987). 'Learning Networks of Neurons with Boolean Logic.' In: *Europhysics Letters (EPL)* 4.4, pp. 503–508.

Pearl, Judea (1982). *Reverend Bayes on inference engines: A distributed hierarchical approach.*

– (1986). 'Fusion, propagation, and structuring in belief networks.' In: *Artificial intelligence* 29.3, pp. 241–288.

Pedregosa, F. et al. (2011). 'Scikit-learn: Machine Learning in Python.' In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peierls, Rudolf (1936). 'Statistical theory of superlattices with unequal concentrations of the components.' In: *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* 154.881, pp. 207–222.

Percus, Allon, Gabriel Istrate, and Cristopher Moore (2006). *Computational complexity and statistical physics.* OUP USA.

Perry, Amelia, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra (2016). 'Optimality and sub-optimality of PCA for spiked random matrices and synchronization.' In: *arXiv preprint arXiv:1609.05573.*

Plefka, Timm (1982). 'Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model.' In: *Journal of Physics A: Mathematical and general* 15.6, p. 1971.

Quillian, M Ross (1969). 'The teachable language comprehender: A simulation program and theory of language.' In: *Communications of the ACM* 12.8, pp. 459–476.

Rangan, Sundeep (2011). 'Generalized approximate message passing for estimation with random linear mixing.' In: *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on.* IEEE, pp. 2168–2172.

Rangan, Sundeep and Alyson K Fletcher (2012). 'Iterative estimation of constrained rank-one matrices in noise.' In: *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on.* IEEE, pp. 1246–1250.

Rangan, Sundeep, Vivek Goyal, and Alyson K Fletcher (2009). 'Asymptotic analysis of map estimation via the replica method and compressed sensing.' In: *Advances in Neural Information Processing Systems*, pp. 1545–1553.

Rangan, Sundeep, Philip Schniter, and Alyson K Fletcher (2019a). 'Vector approximate message passing.' In: *IEEE Transactions on Information Theory* 65.10, pp. 6664–6684.

Rangan, Sundeep, Philip Schniter, Alyson K Fletcher, and Subrata Sarkar (2019b). 'On the convergence of approximate message passing with arbitrary matrices.' In: *IEEE Transactions on Information Theory* 65.9, pp. 5339–5351.

Reed, Scott, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee (2016). 'Generative adversarial text to image synthesis.' In: *arXiv preprint arXiv:1605.05396.*

Reeves, G. and H. D. Pfister (2016a). 'The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact.' In: *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 665–669.

Reeves, Galen (2017). 'Additivity of information in multilayer networks via additive Gaussian noise transforms.' In: *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton).* IEEE, pp. 1064–1070.

Reeves, Galen and Michael Gastpar (2012). 'Compressed sensing phase transitions: Rigorous bounds versus replica predictions.' In: *2012 46th Annual Conference on Information Sciences and Systems (CISS).* IEEE, pp. 1–6.

Reeves, Galen and Henry D. Pfister (2016b). 'The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact.' In: *IEEE International Symposium on Information Theory - Proceedings.*

Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra (2014). 'Stochastic backpropagation and approximate inference in deep generative models.' In: *arXiv preprint arXiv:1401.4082.*

Ricci-Tersenghi, Federico, Martin Weigt, and Riccardo Zecchina (2001). 'Simplest random k-satisfiability problem.' In: *Physical Review E* 63.2, p. 026702.

Rieger, H. (1992). 'The number of solutions of the Thouless-Anderson-Palmer equations for p-spin-interaction spin glasses.' In: *Phys. Rev. B* 46 (22), pp. 14655–14661.

Robbins, Herbert E. (1951). 'A Stochastic Approximation Method.' In: *Annals of Mathematical Statistics* 22, pp. 400–407.

Rosenblatt, Frank (1958). 'The perceptron: a probabilistic model for information storage and organization in the brain.' In: *Psychological review* 65.6, p. 386.

Rosset, Saharon, Ji Zhu, and Trevor J. Hastie (2004). 'Margin Maximizing Loss Functions.' In: *Advances in Neural Information Processing Systems 16.* MIT Press, pp. 1237–1244.

Rotskoff, Grant M and Eric Vanden-Eijnden (2018). 'Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error.' In: *stat* 1050, p. 22.

Ruder, Sebastian (2016). 'An overview of gradient descent optimization algorithms.' In: *arXiv preprint arXiv:1609.04747.*

Rumelhart, David E, Geoffrey E Hinton, James L McClelland, et al. (1986a). 'A general framework for parallel distributed processing.' In: *Parallel distributed processing: Explorations in the microstructure of cognition* 1.45-76, p. 26.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986b). 'Learning representations by back-propagating errors.' In: *nature* 323.6088, pp. 533–536.

Saad, David and Sara A. Solla (1995a). 'Exact solution for on-line learning in multilayer neural networks.' In: *Physical Review Letters* 74.21, pp. 4337–4340.

Saad, David and Sara A Solla (1995b). 'On-line learning in soft committee machines.' In: *Physical Review E* 52.4, p. 4225.

Safran, Itay and Ohad Shamir (2018). 'Spurious Local Minima are Common in Two-Layer ReLU Neural Networks.' In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, pp. 4433–4441.

Santoro, Adam et al. (2017). 'A simple neural network module for relational reasoning.' In: *Advances in neural information processing systems*, pp. 4967–4976.

Sauer, Norbert (1972). 'On the density of families of sets.' In: *Journal of Combinatorial Theory, Series A* 13.1, pp. 145–147.

Schmidhuber, Jürgen (2015). 'Deep learning in neural networks: An overview.' In: *Neural networks* 61, pp. 85–117.

Schniter, Philip and Sundeep Rangan (2014). 'Compressive phase retrieval via generalized approximate message passing.' In: *IEEE Transactions on Signal Processing* 63.4, pp. 1043–1055.

Scholkopf, Bernhard et al. (1999). 'Input space versus feature space in kernel-based methods.' In: *IEEE transactions on neural networks* 10.5, pp. 1000–1017.

Schülke, Christophe (2016). *Statistical physics of linear and bilinear inference problems.*

Schwarze, Henry (1993). 'Learning a rule in a multilayer neural network.' In: *Journal of Physics A: Mathematical and General* 26.21, p. 5781.

Schwarze, Henry and John Hertz (1992). 'Generalization in a large committee machine.' In: *EPL (Europhysics Letters)* 20.4, p. 375.

– (1993). 'Generalization in fully connected committee machines.' In: *EPL (Europhysics Letters)* 21.7, p. 785.

Sejnowski, Terrence J (2018). *The deep learning revolution.* Mit Press.

Semerjian, Guilhem (2008). 'On the freezing of variables in random constraint satisfaction problems.' In: *Journal of Statistical Physics* 130.2, pp. 251–293.

Sethna, James et al. (2006). *Statistical mechanics: entropy, order parameters, and complexity.* Vol. 14. Oxford University Press.

Seung, Sebastian, Haim Sompolinsky, and Naftali Tishby (1992). 'Statistical mechanics of learning from examples.' In: *Physical Review A* 45.8, p. 6056.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

Shannon, Claude E (1948). 'A mathematical theory of communication.' In: *The Bell system technical journal* 27.3, pp. 379–423.

Shcherbina, Mariya and Brunello Tirozzi (2003). 'Rigorous solution of the Gardner problem.' In: *Communications in mathematical physics* 234.3, pp. 383–422.

Shelah, Saharon (1972). 'A combinatorial problem; stability and order for models and theories in infinitary languages.' In: *Pacific Journal of Mathematics* 41.1, pp. 247–261.

Sherrington, D and S Kirkpatrick (1975). *Solvable Model of a Spin Glass.*

Silver, David et al. (2016). 'Mastering the game of Go with deep neural networks and tree search.' In: *nature* 529.7587, pp. 484–489.

Silverstein, Jack W and ZD Bai (1995). 'On the empirical distribution of eigenvalues of a class of large dimensional random matrices.' In: *Journal of Multivariate analysis* 54.2, pp. 175–192.

Skansi, Sandro (2018). *Introduction to Deep Learning: from logical calculus to artificial intelligence.* Springer.

Sommers, H. J. (1978). 'Solution of the long-range gaussian-random Ising model.' In: *Zeitschrift für Physik B Condensed Matter* 31.3, pp. 301–307.

Sommers, H-J (1979). 'The Sherrington-Kirkpatrick spin glass model: results of a new theory.' In: *Zeitschrift für Physik B Condensed Matter* 33.2, pp. 173–180.

Sompolinsky, H., N. Tishby, and H. S. Seung (1990). 'Learning from examples in large neural networks.' In: *Physical Review Letters* 65.13, pp. 1683–1686.

Sompolinsky, Haim and Annette Zippelius (1982). 'Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses.' In: *Physical Review B* 25.11, p. 6860.

Sourlas, Nicolas (1989). 'Spin-glass models as error-correcting codes.' In: *Nature* 339.6227, pp. 693–695.

Stanley, H. E. (1968). 'Dependence of Critical Properties on Dimensionality of Spins.' In: *Phys. Rev. Lett.* 20 (12), pp. 589–592.

Stojnic, Mihailo (2013a). 'Another look at the Gardner problem.' In: *arXiv preprint arXiv:1306.3979*.

– (2013b). 'Discrete perceptrons.' In: *arXiv preprint arXiv:1306.4375*.

– (2013c). 'Negative spherical perceptron.' In: *arXiv preprint arXiv:1306.3980*.

Sutskever, Ilya, James Martens, George Dahl, and Geoffrey Hinton (2013). 'On the importance of initialization and momentum in deep learning.' In: *International conference on machine learning*, pp. 1139–1147.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). 'Sequence to sequence learning with neural networks.' In: *Advances in neural information processing systems*, pp. 3104–3112.

Sutton, Richard S, Andrew G Barto, et al. (1998). *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge.

Sutton, Richard S, David A McAllester, Satinder P Singh, and Yishay Mansour (2000). 'Policy gradient methods for reinforcement learning with function approximation.' In: *Advances in neural information processing systems*, pp. 1057–1063.

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf (2014). 'Deepface: Closing the gap to human-level performance in face verification.' In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708.

Talagrand, Michel (1998). 'The Sherrington–Kirkpatrick model: A challenge for mathematicians.' In: *Probability theory and related fields* 110.2, pp. 109–176.

– (2003). *Spin glasses: a challenge for mathematicians: cavity and mean field models*. Vol. 46. Springer Science & Business Media.

– (2006a). 'Free energy of the spherical mean field model.' In: *Probability Theory and Related Fields* 134.3, pp. 339–382.

– (2006b). 'The Parisi formula.' In: *Annals of mathematics*, pp. 221–263.

Tanaka, Toshiyuki (2002). 'A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors.' In: *IEEE Transactions on Information theory* 48.11, pp. 2888–2910.

Tao, Terence (2009). *Compressed sensing, or: the equation $Ax = b$, revisited*. Clay-Mahler Lecture Series.

Thouless, David J, Philip W Anderson, and Robert G Palmer (1977). 'Solution of'solvable model of a spin glass'.' In: *Philosophical Magazine* 35.3, pp. 593–601.

Thrampoulidis, Christos, Ehsan Abbasi, and Babak Hassibi (2018). 'Precise Error Analysis of Regularized M-Estimators in High Dimensions.' In: *IEEE Transactions on Information Theory* 64.8, pp. 5592–5628.

Thrampoulidis, Christos, Samet Oymak, and Babak Hassibi (2015). 'Regularized Linear Regression: A Precise Analysis of the Estimation Error.' In: *Proceedings of The 28th Conference on Learning Theory*. Vol. 40. Proceedings of Machine Learning Research. Paris, France: PMLR, pp. 1683–1709.

Tibshirani, Robert (1996). 'Regression shrinkage and selection via the lasso.' In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tishby, Naftali, Esther Levin, and Sara A. Solla (Dec. 1989). 'Consistent inference of probabilities in layered networks: Predictions and generalization.' English (US). In: *IJCNN Int Jt Conf Neural Network*. Ed. by Anon. IJCNN International Joint Conference on Neural Networks ; Conference date: 18-06-1989 Through 22-06-1989. Publ by IEEE, pp. 403–409.

Tishby, Naftali, Fernando C Pereira, and William Bialek (2000). 'The information bottleneck method.' In: *arXiv preprint physics/0004057*.

Tishby, Naftali and Noga Zaslavsky (2015). 'Deep learning and the information bottleneck principle.' In: *2015 IEEE Information Theory Workshop (ITW)*. IEEE, pp. 1–5.

Toledano, Pierre and Jean-claude Toledano (1987). *Landau Theory Of Phase Transitions, The: Application To Structural, Incommensurate, Magnetic And Liquid Crystal Systems*. Vol. 3. World Scientific Publishing Company.

Touchette, Hugo (Apr. 2008). 'Touchette, H.: The large deviation approach to statistical mechanics. Phys. Rep. 478, 1-69.' In: *Physics Reports* 478.

Toulouse, G et al. (1987). 'Theory of the frustration effect in spin glasses: I.' In: *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* 9, p. 99.

Tramel, Eric W, Angélique Drémeau, and Florent Krzakala (2016a). 'Approximate message passing with restricted Boltzmann machine priors.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2016.7, p. 073401.

Tramel, Eric W, Andre Manoel, Francesco Caltagirone, Marylou Gabrié, and Florent Krzakala (2016b). 'Inferring sparsity: Compressed sensing using generalized restricted Boltzmann machines.' In: *2016 IEEE Information Theory Workshop (ITW)*. IEEE, pp. 265–269.

Tsang, Edward (2014). *Foundations of constraint satisfaction: the classic text*. BoD–Books on Demand.

Tsybakov, Alexandre B (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Tulino, Antonia M, Giuseppe Caire, Sergio Verdu, and Shlomo Shamai (2013). 'Support recovery with sparsely sampled free random matrices.' In: *IEEE Transactions on Information Theory* 59.7, pp. 4243–4271.

Turing, Alan M (2009). 'Computing machinery and intelligence.' In: *Parsing the turing test*. Springer, pp. 23–65.

Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky (2018). 'Deep image prior.' In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454.

Urbanczik, R (1997). 'Storage capacity of the fully-connected committee machine.' In: *Journal of Physics A: Mathematical and General* 30.11, p. L387.

Valiant, Leslie G (1984). 'A theory of the learnable.' In: *Communications of the ACM* 27.11, pp. 1134–1142.

Vapnik, Vladimir N and A Ya Chervonenkis (2015). 'On the uniform convergence of relative frequencies of events to their probabilities.' In: *Measures of complexity*. Springer, pp. 11–30.

Vapnik, Vladimir (1998). *Statistical learning theory. 1998*. Wiley, New York.

– (2006). *Estimation of dependences based on empirical data*. Springer Science & Business Media.

– (2013). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, Vladimir, Esther Levin, and Yann Le Cun (1994). 'Measuring the VC-dimension of a learning machine.' In: *Neural computation* 6.5, pp. 851–876.

Varadhan, S. R. S. (Mar. 2008). 'Large deviations.' In: *Ann. Probab.* 36.2, pp. 397–419.

Vaswani, Ashish et al. (2017). 'Attention is all you need.' In: *Advances in neural information processing systems*, pp. 5998–6008.

Vila, Jeremy, Philip Schniter, Sundeep Rangan, Florent Krzakala, and Lenka Zdeborová (2015). 'Adaptive damping and mean removal for the generalized approximate message passing algorithm.' In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2021–2025.

Villar, Soledad (2018). 'Generative models are the new sparsity?'

Vincent, Eric (2007). 'Ageing, rejuvenation and memory: the example of spin-glasses.' In: *Ageing and the glass transition*. Springer, pp. 7–60.

Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou (2010). 'Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.' In: *Journal of machine learning research* 11.12.

Voit, Johannes (2013). *The statistical mechanics of financial markets*. Springer Science & Business Media.

Wainwright, Martin J, Michael I Jordan, et al. (2008). 'Graphical models, exponential families, and variational inference.' In: *Foundations and Trends® in Machine Learning* 1.1–2, pp. 1–305.

Watkin, Timothy LH, Albrecht Rau, and Michael Biehl (1993). 'The statistical mechanics of learning a rule.' In: *Reviews of Modern Physics* 65.2, p. 499.

Weiss, Pierre (1907). 'L'hypothèse du champ moléculaire et la propriété ferromagnétique.' In:

Wendel, James G (1962). 'A problem in geometric probability.' In: *Math. Scand* 11, pp. 109–111.

Whyte, W. and D. Sherrington (1996). 'Replica-symmetry breaking in perceptrons.' In: *Journal of Physics A: Mathematical and General* 29.12, pp. 3063–3073.

Widrow, Bernard and Marcian E Hoff (1960). *Adaptive switching circuits*. Tech. rep. Stanford Univ Ca Stanford Electronics Labs.

Williams, Christopher KI and Carl Edward Rasmussen (1996). 'Gaussian processes for regression.' In: *Advances in neural information processing systems*, pp. 514–520.

Wilson, Kenneth G (1983). 'The renormalization group and critical phenomena.' In: *Reviews of Modern Physics* 55.3, p. 583.

Wong, R. (1989). *Asymptotic Approximations of Integrals Computer Science and Scientific Computing*.

Wu, F. Y. (1982). 'The Potts model.' In: *Rev. Mod. Phys.* 54 (1), pp. 235–268.

Wu, Yihong and Sergio Verdú (2012). 'Optimal phase transitions in compressed sensing.' In: *IEEE Transactions on Information Theory* 58.10, pp. 6241–6263.

Wu, Yonghui et al. (2016). 'Google's neural machine translation system: Bridging the gap between human and machine translation.' In: *arXiv preprint arXiv:1609.08144*.

Xiao, Han, Kashif Rasul, and Roland Vollgraf (2017). *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms.*

Xiong, Yuansheng, Chulan Kwon, and Jong-Hoon Oh (1998). 'The storage capacity of a fully-connected committee machine.' In: *Advances in Neural Information Processing Systems*, pp. 378–384.

Yedidia, JS and WT Freeman (2001a). 'Understanding belief propagation and its generalizations.' In: *Exploring artificial intelligence in the new millennium*, pp. 239 –269.

Yedidia, Jonathan S, William T Freeman, and Yair Weiss (2001b). 'Generalized belief propagation.' In: *Advances in neural information processing systems*, pp. 689–695.

– (2002). 'Bethe free energy Kikuchi approximations and BP algorithms.' In: TR2002-35.

Yedidia, Jonathan S., William T. Freeman, and Yair Weiss (2005). 'Constructing free-energy approximations and generalized belief propagation algorithms.' In: *IEEE Transactions on Information Theory* 51.7, pp. 2282–2312.

Yuille, Alan (2001). 'A double-loop algorithm to minimize the bethe free energy.' In: *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition.* Springer, pp. 3–18.

Zamponi, Francesco (2010). 'Mean field theory of spin glasses.' In:

Zdeborová, Lenka (2017). 'Machine learning: New tool in the box.' In: *Nature Physics* 13.5, pp. 420–421.

– (2020). 'Understanding deep learning is also a job for physicists.' In: *Nature Physics* 16.6, pp. 602–604.

Zdeborová, Lenka and Florent Krząkała (2007). 'Phase transitions in the coloring of random graphs.' In: *Physical Review E* 76.3, p. 031131.

Zdeborová, Lenka and Florent Krzakala (2011). 'Quiet planting in the locked constraint satisfaction problems.' In: *SIAM Journal on Discrete Mathematics* 25.2, pp. 750–770.

Zdeborová, Lenka and Florent Krzakala (2016a). 'Statistical physics of inference: Thresholds and algorithms.' In: *Advances in Physics* 65.5, pp. 453–552.

– (Sept. 2016b). 'Statistical physics of inference: thresholds and algorithms.' In: *Advances in Physics* 65.5, pp. 453–552.

Zdeborová, Lenka and Marc Mézard (2008a). 'Constraint satisfaction problems with isolated solutions are hard.' In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.12, P12004.

– (2008b). 'Locked constraint satisfaction problems.' In: *Physical review letters* 101.7, p. 078702.

Zeiler, Matthew D (2012). 'Adadelta: an adaptive learning rate method.' In: *arXiv preprint arXiv:1212.5701.*

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2016). 'Understanding deep learning requires rethinking generalization.' In: *ICLR 2017, preprint arXiv:1611.03530.*

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros (2017a). 'Unpaired image-to-image translation using cycle-consistent adversarial networks.' In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Zhu, Junan, Dror Baron, and Florent Krzakala (2017b). 'Performance limits for noisy multimeasurement vector problems.' In: *IEEE Transactions on Signal Processing* 65.9, pp. 2444–2454.

Zia, Royce KP, Edward F Redish, and Susan R McKay (2009). 'Making sense of the Legendre transform.' In: *American Journal of Physics* 77.7, pp. 614–622.

Zou, Hui and Trevor Hastie (2005). 'Regularization and variable selection via the elastic net.' In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.

d'Ascoli, Stéphane, Maria Refinetti, Giulio Biroli, and Florent Krzakala (2020). 'Double trouble in double descent: Bias and variance (s) in the lazy regime.' In: *arXiv preprint arXiv:2003.01054.*

Ph.D dissertation of Benjamin Aubin
Mean-field methods and algorithmic perspectives for high-dimensional machine learning © December 16, 2020

SUPERVISORS:
Lenka Zdeborová and Florent Krzakala

LOCATION:
Institut de Physique Théorique
CEA & Université Paris-Saclay, Saclay, France

TIME FRAME:
October 1, 2017 — December 16, 2020

université
PARIS-SACLAY

ÉCOLE DOCTORALE
Physique en
l'Île-de-France (PIF)

Titre: Méthodes à champ moyen et perspectives algorithmiques pour l'apprentissage automatique en haute dimension.

Mots clés: Physique statistique, apprentissage automatique, réseaux de neurones, estimation statistique, algorithmes de passage de messages, méthode des répliques.

Résumé: À une époque où l'utilisation des données a atteint un niveau sans précédent, l'apprentissage machine, et plus particulièrement l'apprentissage profond basé sur des réseaux de neurones artificiels, a été responsable de très importants progrès pratiques. Leur utilisation est désormais omniprésente dans de nombreux domaines d'application, de la classification d'images à la reconnaissance vocale en passant par la prédiction de séries temporelles et l'analyse de texte. Pourtant, la compréhension de nombreux algorithmes utilisés en pratique est principalement empirique et leur comportement reste difficile à analyser. Ces lacunes théoriques soulèvent de nombreuses questions sur leur efficacité et leurs potentiels risques. Etablir des fondements théoriques sur lesquels asseoir les observations numériques est devenu l'un des défis majeurs de la communauté scientifique. La principale difficulté qui se pose lors de l'analyse de la plupart des algorithmes d'apprentissage automatique est de traiter analytiquement et numériquement un grand nombre de variables aléatoires en interaction. Dans ce manuscrit, nous revisitons une approche basée sur les outils de la physique statistique des systèmes désordonnés. Développés au long d'une riche littérature, ils ont été précisément conçus pour décrire le comportement macroscopique d'un grand nombre de particules, à partir de leurs interactions microscopiques. Au cœur de ce travail, nous mettons fortement à profit le lien profond entre la méthode des répliques et les algorithmes de passage de messages pour mettre en lumière les diagrammes de phase de divers modèles théoriques, en portant l'accent sur les potentiels écarts entre seuils statistiques et algorithmiques. Nous nous concentrons essentiellement sur des tâches et données synthétiques générées dans le paradigme enseignant-élève. En particulier, nous appliquons ces méthodes à champ moyen à l'analyse Bayes-optimale des machines à comité, à l'analyse des bornes de généralisation de Rademacher pour les perceptrons, et à la minimisation du risque empirique dans le contexte des modèles linéaires généralisés. Enfin, nous développons un cadre pour analyser des modèles d'estimation avec des informations *à priori* structurées, produites par exemple par des réseaux de neurones génératifs avec des poids aléatoires.

Title: Mean-field methods and algorithmic perspectives for high-dimensional machine learning

Keywords: Statistical physics, machine learning, neural networks, statistical estimation, message-passing algorithms, replica method

Abstract: At a time when the use of data has reached an unprecedented level, machine learning, and more specifically deep learning based on artificial neural networks, has been responsible for very important practical advances. Their use is now ubiquitous in many fields of application, from image classification, text mining to speech recognition, including time series prediction and text analysis. However, the understanding of many algorithms used in practice is mainly empirical and their behavior remains difficult to analyze. These theoretical gaps raise many questions about their effectiveness and potential risks. Establishing theoretical foundations on which to base numerical observations has become one of the fundamental challenges of the scientific community. The main difficulty that arises in the analysis of most machine learning algorithms is to handle, analytically and numerically, a large number of interacting random variables. In this manuscript, we revisit an approach based on the tools of statistical physics of disordered systems. Developed through a rich literature, they have been precisely designed to infer the macroscopic behavior of a large number of particles from their microscopic interactions. At the heart of this work, we strongly capitalize on the deep connection between the replica method and message passing algorithms in order to shed light on the phase diagrams of various theoretical models, with an emphasis on the potential differences between statistical and algorithmic thresholds. We essentially focus on synthetic tasks and data generated in the teacher-student paradigm. In particular, we apply these mean-field methods to the Bayes-optimal analysis of committee machines, to the worst-case analysis of Rademacher generalization bounds for perceptrons, and to empirical risk minimization in the context of generalized linear models. Finally, we develop a framework to analyze estimation models with structured prior informations, produced for instance by deep neural networks based generative models with random weights.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France