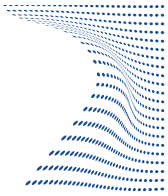
ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR



Doctoral Dissertation
Doctoral Program in Physics (33rd cycle)

Expectation Propagation Methods for Approximate Inference in Linear Estimation Problems

A statistical physics inspired approach

Submitted by:
Mirko Pieropan

Supervisors
Prof. Andrea Pagnani, Supervisor
Prof. Alfredo Braunstein, Co-supervisor

Doctoral Examination Committee:
Prof. Isaac Pérez Castillo, Referee, Universidad Autónoma Metropolitana-Iztapalapa, Mexico
Prof. Luca Leuzzi, Referee, CNR and Sapienza, Università di Roma, Italy

Politecnico di Torino
2021

I hereby declare that the contents and organisation of this dissertation constitute my own original work and do not compromise in any way the rights of third parties, including those related to the security of personal data.

Impact statement

Linear estimation problems are commonly found in a number of important applications as arising in applied optics, information processing, bioinformatics and diagnostic medical imaging. These problems often consist in ill-posed systems of linear equations, which can be solved using efficient inference techniques. This PhD thesis discusses a statistical physics based approach for the solution of sparse linear estimation problems, called expectation propagation (EP). The ideas underlying the method are rooted in the Thouless-Anderson-Palmer (TAP) approach introduced to study the physics of disordered systems and are well-suited to be applied to probabilistic modeling in general.

We focus on the Gaussian case of the EP algorithm and apply it to the compressed sensing problem and to the problem of learning a binary classification rule. Both problems can be recast as finding solutions of underconstrained systems of linear equations of the kind $\mathbf{F}\boldsymbol{x} = \boldsymbol{y}$, where $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{y} \in \mathbb{R}^M$ and $M < N$, to be solved given the knowledge of the linear transformation $\mathbf{F}$, given the set $\boldsymbol{y}$ of partial observations and given additional constraints concerning the hypothesized structure of the sought solution $\boldsymbol{x}$. In a Bayesian setting, these additional constraints can be encoded in suitable prior distributions, which, in turn, are easily incorporated in the EP approximation.

The linear transformation producing the observations often needs to fulfill several constraints dictated by the specific problem of interest, giving rise to some kind of structure in the related matrix and in the measurements produced. Such constraints may be due to the experimental setup adopted or to physical limitations associated with the measuring device used to collect the observations. As a consequence, methods that rely on the statistical independence of the entries of the matrix $\mathbf{F}$ can face serious challenges when dealing with inverse linear problems generated by structured matrices.

In this PhD thesis, the performance of Gaussian EP is compared to other message passing algorithms and is shown to be robust in several cases where correlations are introduced in the linear transformation matrix of the problem considered. Thus, the work presented in this dissertation can be relevant for applications where deterministic or random correlated linear transformations arise and can be considered a step towards the development of efficient approximate inference algorithms able to deal with structured data sets in a variety of settings.

Contents

Acknowledgements

First and foremost, I would like to thank my supervisors Andrea Pagnani and Alfredo Braunstein for giving me the opportunity to work with them on sparse linear estimation problems. I am grateful for their patience, for their scientific feedback and for sharing with me their computational expertise. I especially thank Andrea for his constant interest in my project and for giving me technical advice along the way: this has helped me improve my workflow over time. A special thanks goes to Alfredo for his scientific enthusiasm as well as for sharing with me several tricks of the trade related to Julia programming. Andrea and Alfredo have both been extremely approachable PhD supervisors, even when the COVID-19 pandemic prevented us to meet in person for many months, and they allowed me to attend several schools related to my scientific interests throughout my PhD. I consider myself very fortunate for this.

Thanks to Luca Leuzzi and Isaac Pérez Castillo for agreeing to be part of my PhD examination committee and for taking the time to review my thesis. They read my dissertation carefully and spotted some errors which I had not noticed. Any errors that are left after their revision are of course my own.

I am indebted to the SmartData interdepartmental center for data science of Politecnico di Torino for funding my PhD. I especially thank Marco Mellia for providing me and the other researchers affiliated with the center with an extremely pleasant open space located at OGR Tech. In addition, Marco agreed to buy books about machine learning, statistics and data science – many of which were suggested by me – in order to support the research activity of the center. I used several of these books while writing this PhD dissertation.

Thanks to Luca Dall'Asta, Carlo Lucibello and Marco Pretti for the scientific exchanges we had in occasion of the courses they taught or organized as well as for supporting the idea of having a Journal Club in our department. Thanks to the other members of the research group too: Jorge, Giovanni, Anna, Guido, Indaco, Andrea Gamba, Marco Zamparo, Chiara, Carla, Elsi, Elisa, Louise, Luca Sesta, Fabio, Stefano and Matteo. I am especially grateful to Jorge Fernandez-de-Cossio-Diaz, from whom I learned a great deal while working together during our time in Les Houches: that was very fun and I look forward to collaborating again in the future. During the past three years and a half, I studied my PhD at the same time and in the same research group as Giovanni Catania: thanks for our scientific discussions and for being a kind friend.

I would also like to thank Alberto Batista Tomas for helping me find accommodation in Havana before I started my secondment there. I am grateful to Alberto, Jorge, Lidice Vaillant, Roberto Mulet, Alejandro Lage Castellanos, Amanda, Orlando, Ernesto, Carlos and JJ for making my stay in Cuba more pleasant and for the time we shared in Torino.

It would be unfair not to acknowledge ADI (*Associazione Dottorandi e dottori di ricerca in Italia*) for their dedication to the cause of better recruitment and working conditions for PhD students, postdocs and fixed-term researchers in Italy. I am especially indebted to ADI Torino for the countless hours they spent in order to obtain the extension of PhD employment for all doctoral candidates in Torino during the COVID-19 outbreak. As a result of their hard work, I was granted a five-month extension of my PhD fellowship. I really cannot thank them enough for this.

Finally, I would like to thank my flatmates Vittorio and Giorgio for brightening up my days with their happy-go-lucky approach to life, my friends and colleagues of the Master course in Physics of Complex Systems and the office mates with whom I shared parts of my PhD journey in the past few years – Davide, Serena, Angelo, Andrea Richaud, Lorenzo Rossi, Elisa, Dena, Francesca, Alessandro, Francesco, Eliana, Marilisa, Vittorio and many others. Last but not least, I would like to take this opportunity to acknowledge two people who have motivated me to pursue physics. Luigi Cariolato was my mathematics and physics teacher during high school. His astonishing dedication inspired me to study physics at university. Giovanni Garberoglio was my computational physics instructor during my bachelor degree in Physics in Trento and my first research supervisor. Working with him on improving the Poisson-Boltzmann theory of electrolites by taking into account the effects of finite ion size gave me a first taste of what doing research in theoretical and computational physics really means. Thanks to both of you for keeping the bar high back then, and I hope I have made good use of what you taught me.

To all current and future PhD students. May they be supported in their professional growth and career choices by their supervisors and academic institutions.

Chapter 1

Introduction

Linear estimation problems arise in many fields of science and engineering and consist in finding solutions to a system of linear equations based on the knowledge of a set of observations and of the linear operator which generates them. In most interesting scenarios, such observations are only partial, resulting in an underdetermined system of equations, for which infinitely many solutions exist[1]. Noise corrupting the observations is an additional challenge to be tackled and requires faithfully modeling its statistics. The development of techniques able to solve ill-posed problems of this kind is an exciting interdisciplinary field at the intersection between high dimensional statistics, Bayesian inference, statistical physics, signal processing, statistical learning theory and optimization and is of utmost importance for many real-world applications.

Among all fields involved in developing novel and effective methods to successfully deal with linear inverse problems, statistical physics has proven to be a very effective tool in order to design methods able to solve linear estimation problems efficiently, thanks to a series of improved mean field techniques originally devised to approximate complex probability distributions in the context of disordered systems. Indeed, techniques such as approximate message passing (AMP) and adaptive Thouless-Anderson-Palmer (TAP) methods have been shown to be effective for a variety of real-world relevant problems, especially in optical and medical imaging. Important examples are magnetic resonance imaging [1] and tomography [2, 3], to name a few.

The development of reconstruction methods requires that the properties of the solution to be retrieved, of the sensing matrix generating the linear projections and of the statistic of any noise affecting the measurement process are properly taken into account. Concerning the structure of the solution sought, sparsity is one key property that needs to be incorporated among the assumptions from a modeling standpoint, as

[1] By Rouché-Capelli theorem, assuming that the linear operator has maximal rank, these solutions span an affine space of dimension $N - M$, where N is the number of unknowns and M is the number of equations.

signals are often approximately sparse in some domain. Additional structural properties of the signal may involve short and long-range correlations. For instance, natural images are sparse in the wavelet basis and tend to appear smooth over localized as well as extended regions, resulting in local and long-range correlations among pixel values. Reconstruction methods need to be flexible enough to take into account any useful prior knowledge about these statistical properties. In a Bayesian framework, this translates into incorporating extra constraints in the form of highly non linear and non convex priors. The resulting posterior distribution is often *intractable*, meaning that its normalization and marginals cannot be computed analytically nor they can be computed numerically in polynomial time as a function of the size of the system. This intractability of the posterior distribution makes it necessary to resort to approximate inference schemes in order to extract useful information for the problem considered.

A particularly powerful and flexible method inspired by statistical physics able to incorporate non convex prior information is expectation propagation (EP) (see, e.g., [3]). Its Gaussian formulation will be the computational framework used in this PhD thesis in the context of linear estimation problems.

One important challenge that can hamper the effectiveness of reconstruction algorithms developed for linear models is the presence of structure in the linear operator that produces the observations. In fact, while there exist many results ensuring that independent and identically distributed (i.i.d.) random measurement matrices, e.g. Gaussian or Bernoulli, allow reconstruction algorithms to achieve optimal performance, in practice, the structure of sensing matrices is often dictated by the specific physical or implementation constraints associated with the problem considered or measuring device involved [4]. Thus, although very convenient from a mathematical point of view, the assumption of i.i.d. random entries is unrealistic in most cases of interest and devising reconstruction algorithms able to deal with (or even exploit) the presence of structure in deterministic or random measurements is a very active area of research.

The choice of EP as a tool to solve linear estimation problems allows us to address the issue of dealing with statistical correlations in the measurement matrix, as these are taken into account at the level of the EP approximation, as it will be argued later in this PhD thesis. Therefore, the work presented in this dissertation can be considered an important step towards the development of approximate learning algorithms able to deal with the presence of structure in the observed data.

1.1 Problems addressed and contribution of the thesis to the current state of knowledge

The main problems addressed within the EP framework in this PhD thesis are compressed sensing reconstruction and sparse perceptron learning in the teacher-student scenario.

The main contributions of this thesis to the first application, to which Chapter 5 is

devoted, are the following ones.

- In the thermodynamic limit $N \to \infty$, where N is the length of the signal, the reconstruction error of the retrieved signal obtained by means of EP when introducing a sparsity prior of the "spike-and-slab" kind on the entries of the signal exhibits a continuous phase transition on the (ρ, α) plane, where ρ is the fraction of nonzero components of the signal and α is the measurement rate (i.e., the ratio between the number of equations and the number of unknowns), analogously to those observed in the same problem considering other message passing reconstruction algorithms.

- Given a value of the parameter ρ, the "error-free" reconstruction threshold associated with EP lies at a lower α value than the one obtained by solving the relaxed version of the problem (LASSO) [5] where, rather than the number of nonzero variables, one minimizes the L_1 norm of the unknown signal vector. Thus, by using EP one can retrieve the correct solution to the problem in the case of larger compression levels of the signal as compared to what would be possible if LASSO reconstruction were employed (Fig. 5.4). Furthermore, the critical threshold $\alpha_c(\rho)$ beyond which one obtains the correct solution by means of EP is left unchanged if one introduces correlations in the measurement matrix sampling its rows from a multivariate Gaussian as compared to the standard case of i.i.d. measurements. Our analysis shows that the same fact does not hold for the other algorithms considered in our comparisons, which are not able to accurately reconstruct the signal and often do not converge.

- Both in the presence of i.i.d. and Gaussian correlated sensing matrices, we find that the parameter ρ can be accurately estimated during the EP inference procedure using maximum likelihood, by iteratively minimizing the free energy associated with EP.

The relevant paper, where these results have been published, is:

> Alfredo Braunstein, Anna Paola Muntoni, Andrea Pagnani, and *Mirko Pieropan*. "Compressed sensing reconstruction using expectation propagation". Journal of Physics A: Mathematical and Theoretical 53.18 (Apr. 2020), p. 184001.

The other application studied in this PhD thesis focuses on learning a binary classification rule implemented by a sparse teacher perceptron using a student perceptron having diluted weights as well. By using EP to train the student perceptron in this teacher-student scenario, the following results were obtained:

- We showed that EP allows the student to efficiently estimate the set of weights of the teacher. The comparison with other message passing algorithms that can be used to solve the problem, in particular 1-bit Approximate Message Passing (1bitAMP) and Generalized Vector Approximate Message Passing (grVAMP), shows that EP is comparable to these in terms of convergence properties and of goodness

of the solution when considering statistically independent examples (e.g. with patterns extracted by a standard Gaussian distribution). However, it appears to be more robust in terms of convergence and more accurate in determining the nonzero weights as well as their values (better fixed points) in the case where the patterns presented to the student perceptron are correlated.

- One example showing that using EP is advantageous in the presence of correlated patterns (as compared with 1bitAMP and grVAMP) is given by the case of binary patterns dynamically correlated by means of an asynchronous generative process by a recurrent network of perceptrons. Indeed, while EP allows the student perceptron to successfully infer the teacher weights, 1bitAMP (in the same setup) diverges and grVAMP exhibits a high failure rate in terms of convergence as compared to EP for a large interval of training set sizes.
- We show that the sparsity level of the weights, which is not necessarily known a priori, can be accurately estimated during the EP inference procedure by minimizing the free energy associated with the algorithm, analogously to the case of compressed sensing. Likewise, if the teacher mislabels some of the examples and as long as the noise level thus introduced is small enough, EP allows one not only to approximately infer the weights of the teacher, but also to estimate the number of the correctly labeled examples in case this is not known a priori.

These results have been published in:

> Alfredo Braunstein, Thomas Gueudré, Andrea Pagnani, and *Mirko Pieropan*. "Expectation propagation on the diluted Bayesian classifier". Phys. Rev. E 103.4 (Apr. 2021), p. 043301.

1.2 Outline of the thesis

The outline of this PhD thesis is as follows. After this initial Chapter, where we state the relevance of the topics addressed and the contributions of this PhD dissertation to the current state of knowledge, Chapter 2 introduces the fields of compressed sensing and of standard and generalized linear estimation problems. The same Chapter also gives some examples of these problems as arising in physics and discusses the role of sparsity and how to model it by means of prior distributions in order to encode it in Bayesian probabilistic models. Chapter 3 presents several advanced mean field methods obtained by means of a variational principle, from the well known naive mean field method to sophisticated message passing techniques such as belief propagation, approximate message passing and vector approximate message passing, some of which were mentioned above. Chapter 4 focuses on the theoretical and computational aspects concerning EP, with a particular emphasis on EP with univariate approximating Gaussian factors, which is the particular formulation employed in this PhD thesis, including

its dynamical update rules and its fixed point equations. Furthermore, its free energy and the variational derivation of its fixed point conditions are presented, together with the relationship between EP and the methods described in Chapter 3. Chapter 5 discusses the results obtained applying EP to the compressed sensing problem, the related phase diagram, its reconstruction properties in the presence of correlated measurement matrices and compares its performance to other state-of-the-art reconstruction methods widely employed in compressed sensing. Chapter 6 studies the problem of learning a binary classification rule from labeled examples using a sparse perceptron and discusses the results obtained in a variety of regimes, both in the presence of statistically independent and of correlated patterns presented to the perceptron and both in the case where labels are noiseless and where they are partly corrupted by noise. Finally, Chapter 7 is devoted to the conclusions of the thesis and to future directions building upon the work done so far.

Chapter 2

A brief review of linear estimation problems

2.1 Standard linear estimation problems and compressed sensing

In this section, we will introduce a general family of inverse problems known as *linear estimation problems* and relate them to the interdisciplinary field of *compressed sensing*, which has drawn considerable attention and has been extensively studied in signal processing, computer science, electrical engineering, applied mathematics and statistical physics in the last twenty years.

Linear estimation problems can be formulated as systems of M linear equations in N unknowns:

$$\boldsymbol{y} = \mathbf{F}\boldsymbol{w} + \boldsymbol{n}, \tag{2.1}$$

where $\boldsymbol{y} \in \mathbb{R}^M$ is a set of observed outcomes or measurements corrupted by a noise vector $\boldsymbol{n} \in \mathbb{R}^M$, $\mathbf{F} \in \mathbb{R}^{M \times N}$ is a known linear operator and $\boldsymbol{w} \in \mathbb{R}^N$ is an unknown signal to be retrieved from the knowledge of the noisy linear projections $\boldsymbol{y}$ and of the projection operator $\mathbf{F}$. This definition includes both linear systems with $M \geq N$ noisy measurements and linear systems with $M < N$ in the noiseless or noisy regime. When $M < N$, we say that the linear system is *underconstrained*. The focus of this PhD thesis will be on underconstrained linear models, as this is the most interesting case for applications.

If the system in Eq. (2.1) is underconstrained, then it admits infinite solutions. Therefore, additional constraints are required in order to recover the unknown vector $\boldsymbol{w}$. Problems of this kind are often referred to as *standard linear estimation problems* in order to distinguish them from their *generalized* counterpart, where the output vector $\boldsymbol{y}$ is generated by a nonlinear or stochastic function of the right hand side of Eq. (2.1). Generalized linear estimation problems will be introduced and discussed in more details in section 2.2.

The problem of reconstructing sparse signals in the context of compressed sensing is a particular instance of standard linear estimation problems. In this case, the linear operator $\mathbf{F}$ is called the *measurement*, *sensing* or *data* matrix. Compressed sensing (CS) is a mathematical framework for the acquisition and the recovery of signals that admit a *sparse* representation in some basis, meaning that a signal of length N can be represented using $K \ll N$ coefficients [6]. Compressed sensing leverages sparsity in order to retrieve the signal from a significantly smaller number of measurements compared to conventional sampling. In fact, classical sampling is based on Nyquist-Shannon theorem [7], which states that any signal with finite bandwidth can be fully reconstructed by collecting samples at a rate at least equal to twice the largest frequency appearing in the signal considered (the so called *Nyquist rate*). However, depending on the specific task to be accomplished and on the physical limitations of the instrument used to take the measurements, sampling at the required frequency or storing a large number of measurements can often be difficult or costly. The practical advantage of CS is that instead of sampling data at a large rate and then performing a compression, one can directly acquire the signal in a compressed form by sampling it at a lower rate. One important example is medical imaging, where a lower number of measurements with a comparable reconstruction accuracy implies that the images can be acquired in less time.

The compressed sensing problem can be phrased as:

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0, \quad \text{subject to } \mathbf{F}\boldsymbol{w} = \boldsymbol{y}, \tag{2.2}$$

if the measurements are noiseless, and as:

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_0, \quad \text{subject to } \|\mathbf{F}\boldsymbol{w} - \boldsymbol{y}\|^2 \leq \epsilon, \tag{2.3}$$

if the measurements are noisy, where the notation $\|\boldsymbol{w}\|_0$ denotes the cardinality of the support $\text{supp}(\boldsymbol{w}) := \{i | w_i \neq 0\}$ of the signal, which is commonly referred to as the L_0 *"norm"*, where the quotation marks are a reminder that this is not a proper norm[1]. However, finding the sparsest solution satisfying the linear constraints leads to a non-convex problem and is NP-hard, as it requires an exhaustive search over all the possible supports of the signal to be reconstructed and the number of these supports scales exponentially with the dimension (or length) N of the signal of interest [4]. Therefore, several approaches have been proposed where other L_p norms are minimized, where, for $1 \leq p \leq \infty$, one defines:

$$\|\boldsymbol{w}\|_p := \begin{cases} \left(\sum_{i=1}^{N} |w_i|^p\right)^{\frac{1}{p}}, & 1 \leq p < \infty; \\ \max\limits_{i=1,\ldots,N} |w_i|, & p = \infty. \end{cases} \tag{2.4}$$

[1] In the sequel, with some abuse of terminology, we will refer to the L_0 "norm" without quotation marks, as customary in the signal processing and information theory literature.

The latter definition (2.4) is often extended to $0 < p < 1$, in which case the resulting "norms" are rather "quasinorms" [6], in that they fail to fulfill the triangle inequality. For the special case $p = 0$, which is not a quasinorm either, the notation is justified by the fact that [6]:

$$\|\boldsymbol{w}\|_0 = \lim_{p\to 0} \|\boldsymbol{w}\|_p^p. \tag{2.5}$$

A graphical representation of various L_p norms for different values of p is shown in Fig. 2.1: as it can be seen from the isosurfaces plotted, L_p norms with $p \geq 1$ are convex, whereas L_p norms with $p < 1$ are concave. Furthermore, when attempting to find a solution with minimal norm, L_p norms with $p \leq 1$ enforce sparsity: intuitively, this can be realized from Fig. 2.1, where minimizing L_p norms with $p = \frac{1}{2}$ and $p = 1$ leads to solutions that are found on one of the axes, contrary to the case $p = 2$ and $p = \infty$, where the solution is not sparse, although large values of the norm are indeed penalized.

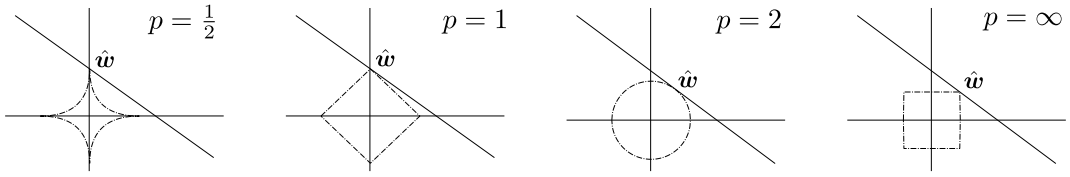


Figure 2.1: Best approximation of a vector constrained to lie on a linear subspace embedded in $\mathbb{R}^2$ and with minimal L_p norm, for $p \in \left\{\frac{1}{2}, 1, 2, \infty\right\}$.

Among the reconstruction methods proposed, one of the most successful ones is the convex relaxation of the L_0 minimization problem where the L_0 norm is replaced with the L_1 norm, leading to the *basis pursuit* problem (BP) in the noiseless case [8]:

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_1, \quad \text{subject to } \mathbf{F}\boldsymbol{w} = \boldsymbol{y}. \tag{2.6}$$

and to the *basis pursuit denoising* problem (BPDN) if the measurements are affected by noise [8]:

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_1, \quad \text{subject to } \|\mathbf{F}\boldsymbol{w} - \boldsymbol{y}\|_2^2 \leq \epsilon. \tag{2.7}$$

The analogous unconstrained problem, known as the *Least Absolute Shrinkage and Selection Operator* (LASSO) problem [9], is often considered:

$$\min_{\boldsymbol{w}} \left(\|\mathbf{F}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \tau\|\boldsymbol{w}\|_1\right) \tag{2.8}$$

where τ is a regularization parameter that penalizes large values of the L_1 norm and that is related to the parameter ϵ in Eq. (2.7). For suitable choices of the regularization parameter, the LASSO and the BPDN solutions coincide.

The results contained in a series of papers [10–14] show that, under certain conditions[2], the basis pursuit problem is equivalent to L_0 minimization and can thus be solved using convex programming reconstruction techniques [16, 17].

The sensing and reconstruction processes in the context of a typical compressed sensing problem are summarized in Fig. 2.2.

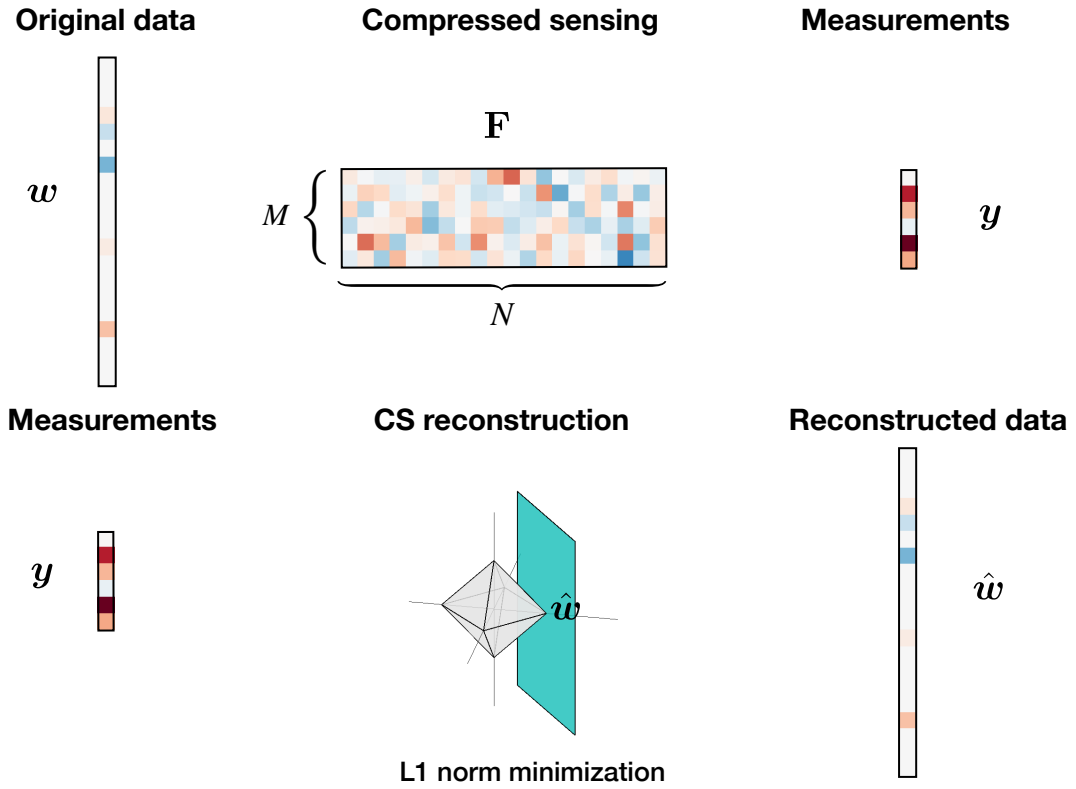


Figure 2.2: Example of a compressed sensing setup.

2.2 Generalized linear estimation problems

Generalized linear estimation problems [18] (GLM) are models where the observed output vector $\boldsymbol{y} \in \mathbb{R}^M$ is related to a linear transformation of the input data $\boldsymbol{w} \in \mathbb{R}^N$ by means of a nonlinear or stochastic function. Denoting by:

$$\boldsymbol{z} = \mathbf{F}\boldsymbol{w} \tag{2.9}$$

[2]These conditions include the incoherence of the linear projections and the approximate preservation of distances between any pair of K-sparse signals when projecting by means of the linear transformation $\mathbf{F}$. The latter condition corresponds to the so called restricted isometry property (RIP) [15].

the hidden result of the linear transformation, where $\mathbf{F} \in \mathbb{R}^{M \times N}$ is assumed to be known, the measurements are sampled from a so called "measurement channel" given by:

$$p(\boldsymbol{y}|\boldsymbol{w}) = p_{y|z}(\boldsymbol{y}|\mathbf{F}\boldsymbol{w}). \tag{2.10}$$

If the channel is deterministic, then the observations are obtained from a nonlinear output function $\boldsymbol{g}$:

$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{z}). \tag{2.11}$$

Standard linear problems can then be viewed as a particular case of this formulation, where the likelihood function is given by $p_{y|z} = \delta(\boldsymbol{y} - \mathbf{F}\boldsymbol{z})$ or, equivalently, the function $\boldsymbol{g}$ is simply the identity transformation.

While we refer to Sec. 2.5 for a more detailed description and for more examples, we shall briefly mention here a few instances of generalized linear models commonly arising in physics and engineering:

- **Quantized compressed sensing** [19] consists in considering a generalized linear model of the kind:

  $$y_m = Q(z_m + n_m), \quad m = 1, \dots, M, \tag{2.12}$$

  where $Q : \mathbb{R} \to \{0,1\}^r$ is a component-wise quantizer that returns the binary representation of each component of its argument using a predefined number r of bits and n_m for $m = 1, \dots, M$ is i.i.d. noise. This generalization of the compressed sensing problem is motivated by the fact that measurements need to be quantized to a finite number of bits before they can be stored and processed in hardware implementations [4].

- **Binary classification** [20] consists in finding a rule to assign each member of a set of objects to one of two classes by associating a binary label with it. This process can be viewed as the extreme case $Q : \mathbb{R} \to \{0,1\}$ of a quantized compressed sensing procedure, where only one bit is retained as a result of the measurement process. The corresponding GLM is given by:

  $$y_m = \operatorname{sign}(z_m + n_m), \quad m = 1, \dots, M, \tag{2.13}$$

  where sign denotes the sign function and n_m for $m = 1, \dots, M$ is i.i.d. statistical noise. The resulting model is known as probit when the noise is Gaussian and as logistic when the distribution of the noise is logistic. The problem of learning a binary classification rule will be dealt with in Chapter 6.

- **Phase retrieval** is a classic problem in crystallography, optics and astronomy (see, e.g., [21]). It arises when one needs to retrieve a signal from a set of measurement magnitudes, which are typically obtained in some transform domain (e.g., the Fourier domain):

  $$y_m = |z_m + n_m|, \quad m = 1, \dots, M, \tag{2.14}$$

  where $z_m, n_m \in \mathbb{C}$.

- The **ReLU perceptron** is a generalized linear model widely adopted in computer vision and speech recognition applications involving multilayer artificial neural networks. It is expressed as:

$$y_m = \max(0, z_m), \quad m = 1, \dots, M, \tag{2.15}$$

  where the function appearing on the right hand side is called rectified linear unit (ReLU) activation function.

- **Poisson noise linear models** [22] arise, for example, in the context of low-light acquisition and are formulated in terms of measurements drawn from a Poisson distribution:

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}) = \prod_{j=1}^{N} \frac{\lambda_j^{y_j}}{y_j!} e^{-\lambda_j}, \tag{2.16}$$

  where $\boldsymbol{\lambda}$ is a vector of latent intensities with components given by the linear model $\boldsymbol{\lambda} = \mathbf{F}\boldsymbol{w}$, $\boldsymbol{w}$ is the image of interest and $\mathbf{F}$ is a linear distortion operator that models image acquisition and is positive in order to impose non-negativity of the mean photon count.

In this thesis, the generalized linear estimation problem of binary classification of sparse input vectors will be studied using the expectation propagation algorithm in section 6.

2.3 Essential concepts of probabilistic modeling

2.3.1 Bayesian inference

Quoting from Wasserman [23], *"Statistical inference, or "learning" as it is called in computer science, is the process of using data to infer the distribution that generated the data"* or some properties of this distribution. In general [24], we have a model:

$$\boldsymbol{y} = f(\omega; \text{noise}), \tag{2.17}$$

where ω is an unknown object that we would like to estimate, $\boldsymbol{y}$ is a set of observations and $f(\cdot; \text{noise})$ a statistical model or probability distribution to which the variability of the generated observations is ascribed. Typically, in a *parametric* setting, the assumptions about the generative model that produced the data are encoded in the choice of a family of probability distributions with some given functional form and a finite dimensional set of parameters ω to be determined. In a *non-parametric* framework, instead, ω is a high dimensional or infinite-dimensional object (a function, for example) in order to make as few assumptions as possible on the statistical model. The general goal of inference is to estimate the unknown object ω, where the accuracy of the estimator

used can be evaluated under some metric. We shall give some examples of estimators in section 2.3.2.

There exist two different approaches to statistical inference. In the *frequentist* approach, ω is an unknown constant, whereas in the *Bayesian* approach, ω is considered a latent (or hidden) random variable and its degree of uncertainty is captured by a *prior distribution* encoding any a priori information or assumptions ('beliefs') about the typical realizations of ω. In this thesis, we will mostly use a Bayesian framework for the inference problems that will be addressed. By Bayes rule, we can write the posterior distribution of the hidden variable as:

$$p(\omega|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\omega)p(\omega)}{p(\boldsymbol{y})}, \tag{2.18}$$

where $p(\boldsymbol{y}|\omega)$ is called *likelihood*, $p(\omega)$ is the prior distribution over ω and the normalization constant $p(\boldsymbol{y})$ in the right-hand side of Eq. (2.18) is known in statistics as the model *evidence* and is used to perform model selection [25].

In the problems that will be addressed in this thesis, ω is typically an high dimensional vector, which we shall denote by $\boldsymbol{w}$, and the posterior distribution of interest is given by $p(\boldsymbol{w}|\boldsymbol{y})$. We will be typically interested in the following inference tasks [26, 27]:

- obtaining the marginals (marginalization problem):

$$p_i(w_i|\boldsymbol{y}) := \int \mathrm{d}\boldsymbol{w}_{\setminus i} p(\boldsymbol{w}|\boldsymbol{y}), \tag{2.19}$$

  which is needed in order to compute the moments of the latent variables. Here $\boldsymbol{w}_{\setminus i}$ denotes the set of all components of $\boldsymbol{w}$ except w_i;

- computing the model evidence (normalization problem):

$$Z := p(\boldsymbol{y}) = \int \mathrm{d}\boldsymbol{w} p(\boldsymbol{w}, \boldsymbol{y}), \tag{2.20}$$

  which plays the role of a partition function in statistical physics;

- computing the mode of the posterior:

$$\arg\max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{y}), \tag{2.21}$$

  which gives the most probable value of $\boldsymbol{w}$ given a specific observation $\boldsymbol{y}$;

- computing the likelihood of the parameters of the model (those appearing in the prior distribution if their values are not assumed to be known a priori, for instance) given the set of observations. This will be developed in more detail in the context of EP based approximate inference in section 4.10.

As will be made clear in Sec. 2.3.2, solving these inference problems allows one to compute suitable estimators for the unknown variables $\boldsymbol{w}$. Each of these estimators is optimal with respect to a specific criterion that quantifies the accuracy of the associated estimate.

2.3.2 Estimators

Maximum a posteriori estimator (MAP). The maximum a posteriori estimator selects the value of $\boldsymbol{x}$ at which the maximum of the posterior distribution $P(\boldsymbol{x}|\boldsymbol{y})$ is attained:

$$\hat{\boldsymbol{x}}_{MAP} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{y}). \tag{2.22}$$

While this is a reasonable estimator in many cases, it can be a bad estimator in some situations, for example when $P(\boldsymbol{x}|\boldsymbol{y})$ takes large values over extended regions but its maximum coincides with a very narrow peak which is located elsewhere in the support.

Maximum likelihood estimator (MLE). The maximum likelihood estimator returns the value of $\boldsymbol{x}$ which maximizes the likelihood $L(\boldsymbol{x}) = P(\boldsymbol{y}|\boldsymbol{x})$ of the signal given the measurements:

$$\hat{\boldsymbol{x}}_{ML} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{y}|\boldsymbol{x}). \tag{2.23}$$

It coincides with the MAP estimate if all priors are uniform over the *whole* space where the signal $\boldsymbol{x}$ is defined.

Minimal mean squared error estimator (MMSE). The minimal mean squared error estimate is given by the conditional expectation of $\boldsymbol{x}$ given $\boldsymbol{y}$:

$$\hat{\boldsymbol{x}}_{MMSE} = \int \boldsymbol{x} P(\boldsymbol{x}|\boldsymbol{y}) \mathrm{d}\boldsymbol{x}, \tag{2.24}$$

and corresponds to taking the component-wise mean of the marginal posterior distribution. The MMSE estimate has the property that the average mean squared error between the actual signal $\boldsymbol{s}$ and the predicted signal $\boldsymbol{x}$:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - s_i)^2 \tag{2.25}$$

is minimal, as can be easily checked by imposing stationary conditions for $\langle MSE \rangle_{P(\boldsymbol{x}|\boldsymbol{y})}$ with respect to the signal estimate $\boldsymbol{x}$.

Minimum mean absolute error estimator. The minimum mean absolute error estimator minimizes the mean absolute error (MAE), which is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - s_i| \tag{2.26}$$

and is given by the component-wise median of the marginal posterior distribution.

2.3.3 Factor graphs

Systems composed of many interdependent variables can often be described in terms of 'local' constraints involving smaller subsets of variables. In a statistical physics perspective, this can be seen as a consequence of the fact that, in many relevant cases, physical interactions have a limited range, resulting in local interactions. The constraints (or dependencies or interactions) to which a given subset of variables is subject can be represented as a *factor* in the joint probability distribution of the configurations of the system. The mutual dependencies among the variables composing the system and their joint distribution can then be represented using *factor graphs* [28].

Given a probability distribution of N variables $\boldsymbol{x}$ written in the factorized form:

$$P(\boldsymbol{x}) = \prod_{a=1}^{M} \psi_a(\boldsymbol{x}_a), \tag{2.27}$$

where $\boldsymbol{x}_a$ denotes the subset of variables that participates in the factor ψ_a, a factor graph is an undirected bipartite graph $G = (V, F, E)$, where V is an index set for the variables, F is an index set for the factors and E is a set of edges that only connect a variable node and a factor node. In particular, an edge $(i, a) \in E$ if and only if $i \in V$, $a \in F$ and x_i is one of the arguments of the factor ψ_a. In a factor graph, variable nodes $i \in V$ are represented by circles and factor nodes $a \in F$ are represented as squares. An example of factor graph is shown in Fig. 2.3.

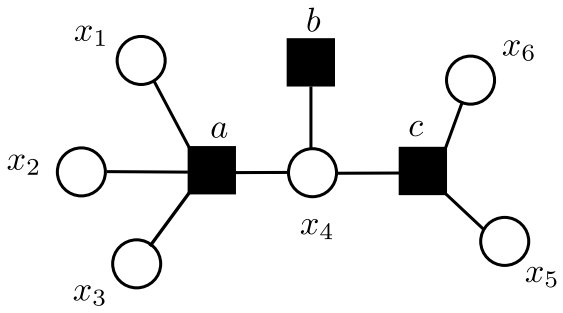


Figure 2.3: An example of factor graph.

2.3.4 Connections between statistical physics and statistical inference

Statistical inference tasks can be turned into statistical physics problems by interpreting the joint probability distribution $p(\boldsymbol{w}|\boldsymbol{y})$ as a Boltzmann distribution of a thermodynamical system with N degrees of freedom with a Hamiltonian that accounts for suitable interactions among the variables $\boldsymbol{w}$. Therefore, from a statistical physics standpoint, statistical inference can be viewed as computing the free energy and the marginals of

the following Boltzmann distribution:

$$P(\boldsymbol{w}|\boldsymbol{y}) = \frac{1}{Z}\,\mathrm{e}^{-\beta H(\boldsymbol{w},\boldsymbol{y})}, \tag{2.28}$$

where $\beta = 1$ and the Hamiltonian is expressed in terms of the prior and of the likelihood as [29]:

$$H(\boldsymbol{w}, \boldsymbol{y}) = -\ln(P(\boldsymbol{y}|\boldsymbol{w})) - \ln(P(\boldsymbol{w})). \tag{2.29}$$

If the prior is factorized over the variables:

$$P(\boldsymbol{w}) = \prod_{i=1}^{N} \psi_i(x_i), \tag{2.30}$$

then the Hamiltonian reads:

$$H(\boldsymbol{w}, \boldsymbol{y}) = -\ln(P(\boldsymbol{y}|\boldsymbol{w})) - \sum_{i=1}^{N} \ln(\psi_i(w_i)). \tag{2.31}$$

The likelihood term in Eq. (2.31) accounts for the interactions among the variables, whereas the terms associated with the factors play the role of local (magnetic) fields (or biases) and drive each variable towards values that are compatible with the additional constraints encoded in the prior distribution. From the statistical physics point of view presented in this section, as observed in [29], the MAP estimate (2.22) can be interpreted as the ground state of the Hamiltonian (2.31).

2.4 Modeling sparsity with regularizers and prior distributions

In the context of linear estimation and reconstruction problems, sparsity arises in many different systems and needs to be taken into account when building physical models. For example, natural images are often sparse in the frequency domain, in particular in the wavelet basis [30]. Other examples include systems with sparse connectivity and/or interactions, such as financial networks, gene regulatory networks in immunology and neural networks in the brain.

Sparsity is often imposed using a regularization approach that consists in formulating a minimization problem of the kind given in Eq. (2.8), where the additional penalty term appearing in the cost function favors sparse solutions. The additional regularizer can be interpreted as a way to incorporate prior knowledge in a probability distribution, as minimizing the cost function, which we shall here denote as $E(\boldsymbol{w})$ for brevity, is equivalent to maximize an associated posterior distribution proportional to $\exp(-E(\boldsymbol{w}))$. This is the same mapping as the one presented in the statistical physics interpretation

of section 2.3.4. In this thesis, a Bayesian point of view will be preferred, but it should be clear that the two approaches are related.

In a Bayesian setting, sparsity is enforced by means of sparsity encouraging prior distributions. Examples of these priors include Laplace, Student's *t*, spike-and-slab priors [31–33] as well as more complicated hierarchical priors, such as, for example, the one used in reference [34].

In particular, L_p norm minimization can be enforced using specific priors. For example, L_1 regularization [35] corresponds to introducing a Laplace prior, given by:

$$\psi(w;\tau) = \frac{\tau}{2}\exp\left(-\frac{\tau}{2}|w|\right), \tag{2.32}$$

whereas L_2 regularization (or ridge regression):

$$\min_{\boldsymbol{w}}\left(\|\mathbf{F}\boldsymbol{w}-\boldsymbol{y}\|_2^2+\tau\|\boldsymbol{w}\|_2^2\right). \tag{2.33}$$

is equivalent to assigning to each variable a Gaussian prior:

$$\psi(w;\tau) = \sqrt{\frac{\tau}{2\pi}}\exp\left(-\frac{\tau}{2}w^2\right). \tag{2.34}$$

Among these two priors, we notice that only the Laplace prior leads to sparse solutions, as only L_p norms with $p \leq 1$ encourage sparseness of the solution sought.

The *spike and slab* (or Gauss-Bernoulli) prior [31, 32] corresponds to L_0 regularization and is defined as:

$$\psi(w;\rho,\lambda) = (1-\rho)\delta(w) + \rho\sqrt{\frac{\lambda}{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\lambda w^2}, \tag{2.35}$$

where ρ is a parameter between 0 and 1 often referred to as the *density* parameter, $\delta(w)$ denotes the Dirac delta distribution (the "spike" part of the prior) and $\lambda > 0$ corresponds to the inverse variance of the Gaussian part of the distribution. As the latter part becomes a uniform distribution in the limit $\lambda \to 0$, it is referred to as the "slab" part of the prior. Thus, the spike-and-slab prior is a convex[3] combination of two contributions: the first one enforces the target variable to be zero (with relative frequency ρ), whereas the other one allows nonzero values (with relative frequency $1-\rho$) but penalizes very large realizations of w by assigning an exponentially small probability. The complementary parameter $1-\rho$ which weights the "spike" part is the *sparsity* level. Intuitively, one way to see the equivalence between the spike-and-slab prior in a probabilistic setting and L_0 regularization is the following: one can consider the product

$$\prod_{i=1}^{N}\left[(1-\rho)\delta(w_i)+\rho\phi(w_i)\right],$$

[3]The combination is convex with respect to ρ.

where $\phi(w) = \sqrt{\lambda/2\pi}\exp\left(-\lambda w^2/2\right)$, and expand it in powers of ρ, resulting in a mixture of measures weighted by factors of the form $(1-\rho)^k \rho^{N-k}$, for $k = 1, \ldots, N$. If the parameter ρ is small (respectively, large), then the mixture weights will be larger the larger the number of delta (respectively, ϕ) factors appearing in the corresponding mixture components, favoring sparse (respectively, dense) configurations of the entries of the vector $\boldsymbol{w}$ when intersecting with the space of solutions that are consistent with the linear constraints $\mathbf{F}\boldsymbol{w} = \boldsymbol{y}$. A rigourous proof of the equivalence between the use of the spike-and-slab prior and L_0 regularization can be found in reference [36].

In the following, we will focus on the spike-and-slab prior, which not only has good shrinkage properties, but also allows analytical computations when multiplied by Gaussian distributions and taking integrals. From this point of view, the advantage of using spike-and-slab priors will be clear in chapters 5 and 6, where the problems of compressed sensing and one-bit compressed sensing, respectively, will be studied using Gaussian expectation propagation, a method for approximate inference that will be presented in detail in chapter 4.

2.5 Examples of linear estimation problems in physics

- **Medical imaging.**

  Magnetic resonance imaging (MRI) is a noninvasive technique used to acquire medical images of soft tissues and based "on the interaction[4] of a strong magnetic field with the hydrogen nuclei contained in the body's water molecules" [4]. In MRI, a main magnet generates a strong static magnetic field, which polarizes the object by inducing an initial magnetization of the sample. As this magnetization vector is perturbed around equilibrium, it undergoes precession around the direction of the static magnetic field with a precession frequency known as *Larmor frequency*. Radiofrequency (RF) pulses of a desired duration and at the precession frequency are sent from the transmit part of a RF system to the target object in order to excite it. The precessing excited magnetization, in turn, causes fast variations in the magnetic field, which are detected by the receiver part of the RF system. The locations of the signal sources are encoded by changing the local magnetic field using spatial magnetic encoding fields, one for each of the three directions in space, which change the precession frequency at each location. Contrast is determined by the relative concentration of excited water protons and by the relaxation times of the tissues under examination. Each measurement taken by the RF detector is a time varying Fourier transform of the magnitude of the magnetization of the body to be imaged as computed along a known trajectory $\boldsymbol{k}(t)$ in Fourier space related to the applied spatial magnetic field gradient [38, 39]. By leveraging sparsity of the

[4] A detailed description of the physics of the nuclear magnetic resonance is beyond the scope of this dissertation. See, e.g., [37] for further details.

image (either in the image domain or in the Fourier domain), compressed sensing can be used in order to reconstruct the image from a relatively small number of measurements, as first shown in [40]. This allows one to reduce the time of MRI scans considerably as compared to what is possible using classical sampling, which is beneficial as it results in lower costs and lower sensitivity to factors, such as, for example, respiratory motion.

Computed X-ray tomography (CT) [41] is a diagnostic imaging technique which consists in producing pictures of cross sections ("slices") of a body by sending X-rays through it and by measuring the intensity attenuation of the radiation as it exits from the object. Let us assume that a monochromatic beam having intensity I_0 is sent through an homogeneous object. As the beam travels through the bulk of the object, its intensity is reduced, due to absorption, as:

$$I = I_0 \, \mathrm{e}^{-wF}, \tag{2.36}$$

where w is the attenuation coefficient of the medium and F denotes the distance traveled by the radiation inside the object. If the object is inhomogeneous, then one can discretize the cross section through which the radiation is transmitted, resulting in a set of a given number N of pixels. Thus, as the radiation illuminating the object leaves the body, the intensity collected by the detector is:

$$I = I_0 \, \mathrm{e}^{-w_1 F_1} \, \mathrm{e}^{-w_2 F_2} \ldots \mathrm{e}^{-w_N F_N}. \tag{2.37}$$

By taking M measurements at different angles and considering logarithms, one obtains the following linear estimation problem:

$$y_m = \sum_{n=1}^{N} F_{mn} w_n, \tag{2.38}$$

where the measurements are given by

$$y_m := \ln(I/I_0) \tag{2.39}$$

and the entry F_{mn} of the so-called *tomographic matrix* is the distance traveled by the mth ray across the nth pixel.

- **Photon limited imaging.** In many imaging applications, it is necessary to accurately extract information in conditions where it is only possible to detect an extremely low number of photons [22]. Examples of these applications include remote sensing [42], night vision, biological imaging [43, 44], fluorescence microscopy [45, 46], and so forth. The number of observed photons obeys the Poisson statistics:

$$\boldsymbol{y} \sim \mathrm{Poisson}(T\mathbf{F}\boldsymbol{w}) \tag{2.40}$$

where y_i denotes the number of photons detected at the i-th element of the detector, T is the time of acquisition, $\mathbf{F}$ is a linear distortion operator (such as blur, compressed sensing or tomographic projection matrix) and $\boldsymbol{w}$ is the image of interest. Photon-limited compressed sensing and image reconstruction pose additional challenges and depend on different thresholds [22, 47] than standard compressed sensing problems.

- **Single pixel camera.** Contrary to digital cameras, where each sensor measures a different pixel of an image (e.g. 10^7 sensors for a 10 Megapixel resolution camera), in single pixel cameras [48, 49] only one sensor, i.e. the *single pixel*, measures the entire image. *Digital micromirror device* (DMD): an array of microscopic mirrors individually tilted at a low angle, either at $+\theta$ (corresponding to *active* mirrors) or at $-\theta$ (corresponding to *inactive* mirrors), with $\theta \sim 15^\circ$. At any given time, for a fixed DMD arrangement, only some of the pixel of the image to be measured are focused to a light detector, whereas all other pixels are directed to a light absorber. By considering a very large number of arrangements of the mirrors and taking light measurements in series (one for each of such arrangements), it is possible to construct a compressed sensing matrix, the row of which being the DMD arrangements, and to retrieve the original image. As the measurements need to be taken in series, rather than simultaneously, the single pixel camera is not suited for video applications.

- **Phase retrieval.** One common problem in applications such as electron microscopy [50], speckle imaging [21, 51] and X-ray crystallography [51] is the so-called *phase retrieval problem*, which consists in retrieving a N-dimensional signal given a sensing matrix and M observations obtained as:

$$y_\mu = \left| \sum_{i=1}^{N} F_{\mu i} w_i \right| . \tag{2.41}$$

  In applications, the measurements are typically given by the magnitude of the Fourier transform of the signal of interest.

  For instance, in X-ray crystallography [52, 53], the signal to be estimated is the crystal electron density, which is sparse as it is nonzero only at the positions of the atoms, the measurements are the Fourier amplitudes of the electron density and the measurement matrix is given by a discrete Fourier transform (DFT) matrix. Indeed, as the magnitude of the Fourier transform of the electron density turns out to be non zero only very close to the points of the reciprocal lattice, it is possible to express the electron density in terms of a Fourier series, where each Fourier component is associated with a point in the reciprocal lattice. Experimentally, one measures the intensities of the diffraction peaks, which are related to the squared modulus of these Fourier components. Therefore, the phases are lost in the measurement procedure and need to be retrieved as well if one wishes to compute the structure of the crystal.

Another example is astronomical imaging, where the phase of the radiation to be detected is distorted by layers of air having different densities in the atmosphere, resulting in a decreased resolution at the level of the acquired image. There, the squared modulus of the Fourier tranform of the image of interest can be computed using speckle interferometry [54] by averaging a large number of short-exposure images, with exposure time smaller than the typical time of fluctuations of the atmosphere[5]. However, once again, all phase information cannot be accessed experimentally.

One more example of a phase retrieval problem is coherent diffractive imaging (CDI) [55], a lensless imaging technique that consists in sending a coherent wave to a non-periodic object – such as an inorganic [56] or a biological [57] specimen – in order to obtain a diffraction pattern, from which one wishes to characterize the structure of the object itself with a high resolution (typically ranging from nanometers to picometers). It can be realized using, e.g., X-rays, electrons, high harmonic generation or optical lasers. The measured diffraction intensity is proportional to the squared modulus of the Fourier transform of the wave taken at the object plane, with suitable spatial scale factors to be taken into account [58]. As in the previous examples, the phase problem arises from the fact that only the intensity of the diffracted wave can be experimentally accessed, whereas the phase information is lost. As a consequence, phase retrieval algorithms are needed to recover the scattering function from the diffraction pattern when reconstructing the image of the sample.

[5] In this way, the atmospheric refractive index is "frozen", resulting in each image being perturbed by a random speckle structure obeying the same statistic.

Chapter 3

Advanced mean field methods

This chapter discusses the mean field method and its improvements in statistical physics and in probabilistic modeling. In particular, we will describe the theoretical ideas and the approximations underlying the Thouless-Anderson-Palmer (TAP) approach, on which expectation propagation, which will be extensively discussed in chapter 4, is based. We first recall the variational free energy principle, which allows one to derive the mean field methods discussed in this dissertation, including expectation propagation, as will be clear in the next chapter. As this thesis addresses topics at the interface between statistical physics and machine learning, care was taken to attempt to bridge the gap between the formulation of the variational principle found in physics and the one often found in the variational inference literature. Subsequently, after introducing the naive mean field approximation, several improvements are discussed, starting from belief propagation and finishing with the adaptive TAP approach (adaTAP), approximate message passing (AMP) and vector approximate message passing (VAMP). The relationship between these methods and expectation propagation will be further explored in chapter 4.

3.1 Variational free energy principle

We recall the thermodynamical variational principle for the Helmholtz free energy. Consider a physical system described by a vector of N variables $\boldsymbol{x} = (x_1, \dots, x_N)$, either discrete or continuous. In the following, we will assume that variables are discrete, but the case of continuous variables is completely analogous. If the system is in thermodynamical equilibrium with a reservoir at temperature T, then the global minimum of the variational Helmholtz free energy functional $\mathscr{F}$:

$$\mathscr{F}[q] = \mathscr{U}[q] - T\mathscr{S}[q], \tag{3.1}$$

is attained when $q(\boldsymbol{x})$ is equal to the Boltzmann distribution of the system, where the minimum is taken over all (normalized) trial probability distributions $q(\boldsymbol{x})$. In Eq. (3.1),

$\mathcal{U}$ is a *variational internal energy*:

$$\mathcal{U}[q] = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) E(\boldsymbol{x}), \tag{3.2}$$

and S is the *variational entropy*[1]:

$$\mathcal{S}[q] = -k_B \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln q(\boldsymbol{x}). \tag{3.3}$$

The variational principle for the free energy can then be stated as:

$$F(\beta, \boldsymbol{\theta}) = \min_q \mathcal{F}[q], \quad p(\boldsymbol{x}) = \frac{1}{Z} \mathrm{e}^{-\beta H(\boldsymbol{\theta})} = \arg\min_q \mathcal{F}[q], \tag{3.4}$$

where β is the inverse temperature and $\boldsymbol{\theta}$ is the vector of all parameters that characterize the system (e.g. the couplings and the external fields for a system of spins). The variational free energy is also called *functional* free energy and, in reference to out-of-equilibrium statistical mechanics, *non-equilibrium free energy*[2].

It is interesting to recast the variational free energy in another form, as this clarifies the connection with the formulation commonly found in information theory and variational inference [59]. Starting from Eq. (3.1), we have:

$$\beta\mathcal{F}[q] = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln\left(\frac{q(\boldsymbol{x})}{\mathrm{e}^{-\beta E(\boldsymbol{x})}}\right) = -\ln Z + \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) = \beta F + D_{KL}\left(q(\boldsymbol{x}) \| p(\boldsymbol{x})\right), \tag{3.5}$$

from which one sees that the variational free energy is an upper bound to the equilibrium Helmholtz free energy due to the non-negativity of the Kullback-Leibler (KL) divergence, D_{KL}, and that, since βF is a constant, minimizing $\mathcal{F}[q]$ with respect to $q(\boldsymbol{x})$ implies that D_{KL} is minimized as well, thus forcing q to become as close as possible to p. In variational inference [20, 60], one is normally interested in computing a posterior distribution given by:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{y})}, \tag{3.6}$$

which cannot be computed because the normalization $p(\boldsymbol{y})$ (the evidence) is intractable. Therefore, the log evidence is decomposed as[3]:

$$\ln p(\boldsymbol{y}) = \mathcal{L}[q] + D_{KL}(q(\boldsymbol{x}) \| p(\boldsymbol{x}|\boldsymbol{y})), \tag{3.7}$$

[1] We recall that the *thermodynamical entropy* corresponds to the *statistical* or *information entropy* in the statistical physics interpretation.

[2] The variational free energy is often called "Gibbs free energy" as well. However, since this expression also refers to the thermodynamic potential that is obtained by performing the Legendre transform of the Helmholtz free energy, we will avoid it in this thesis.

[3] Notice that this derivation is the same of expectation maximization (EM), the only difference being that in EM the evidence $p(\boldsymbol{y})$ is replaced by the likelihood of the parameters $p(\boldsymbol{y}|\boldsymbol{\theta})$ and one is interested in estimating $\boldsymbol{\theta}$ via maximum likelihood estimation.

or lower bounded using Jensen's inequality:

$$\ln p(\boldsymbol{y}) \geq \mathscr{L}[q], \tag{3.8}$$

where $\mathscr{L}[q]$ is called evidence lower bound (ELBO) and is given by:

$$\mathscr{L}[q] = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{y})}{q(\boldsymbol{x})}. \tag{3.9}$$

Both Eq. (3.7) and Eq. (3.8) hold for any trial distribution q. Analogously to the case of Eq. (3.5), maximizing $\mathscr{L}[q]$ with respect to $q(\boldsymbol{x})$ automatically minimizes $D_{KL}(q(\boldsymbol{x}) \| p(\boldsymbol{x}|\boldsymbol{y}))$ and has the advantage of not requiring the computation of the posterior appearing in the Kullback-Leibler divergence. Recalling the mapping described in section 2.3.4, in which $H(\boldsymbol{x}, \boldsymbol{y}) = -\ln p(\boldsymbol{x}, \boldsymbol{y})$ and $\beta = 1$, one can readily identify:

$$\mathscr{U}[q] = -\sum_{\boldsymbol{x}} q(\boldsymbol{x}) \ln p(\boldsymbol{x}, \boldsymbol{y}), \tag{3.10}$$

and relate the ELBO to the variational free energy:

$$\mathscr{L}[q] = -\mathscr{F}[q], \tag{3.11}$$

which makes the equivalence between the formulation used in statistical physics and the one encountered in variational inference explicit.

The variational principle can be used to approximately estimate the partition function or the free energy of a system, which is desirable because if these are known then all thermodynamical properties can be easily accessed by computing their derivatives. In order to do so, the true probability distribution of the system is replaced by a trial distribution belonging to some tractable family $\mathbb{Q}$ and the minimization of the functional $\mathscr{F}$ is restricted over all the distributions q in $\mathbb{Q}$. Again, "tractable" means that it is possible to compute the marginals and the partition function exactly and in polynomial time as a function of the size of the system. Usually, the family $\mathbb{Q}$ is specified by a given functional form and parameterized by some set of parameters, so that the minimization over $q(\boldsymbol{x}; \boldsymbol{\theta})$ translates into a minimization over the values of the parameters $\boldsymbol{\theta}$. Outside statistical physics, the variational approximation method presented in this section is also called *variational Bayes* (VB) or *ensemble learning* (see, e.g., Refs. [20, 61]).

3.2 Naive mean field approximation

Choosing a family of probability distributions that are fully factorized in terms of single variable marginals:

$$q(\boldsymbol{x}) = \prod_{i=1}^{N} q_i(x_i) \tag{3.12}$$

results in the so called *naive* (or *classical*) *mean field* approximation. It is often referred to simply as *mean field* approximation and sometimes as *Bragg-Williams* approximation [62]. Minimizing $\mathscr{F}[q]$ with respect to the single variable marginals yields [63]:

$$q_i^*(x_i) \propto \exp\left(\langle -\beta E(\boldsymbol{x})\rangle_{q^*_{\setminus i}}\right), \tag{3.13}$$

when using the definition (3.2) for the energetic term or:

$$q_i^*(x_i) \propto \exp\left(\langle \log p(\boldsymbol{x}, \boldsymbol{y})\rangle_{q^*_{\setminus i}}\right), \tag{3.14}$$

when using the definition (3.10), where, in both cases, $\langle \dots \rangle_{q^*_{\setminus i}}$ denotes the expectation value with respect to $\prod_{j\neq i} q_j^*(x_j)$.

Historically, the mean field method was developed in order to explain the macroscopic properties of condensed matter systems in terms of the interactions of their microscopic constituents, where each component of the system is thought as a non interacting particle in an effective external field that summarizes all interactions from the other components (hence the expression "mean field"). The naive mean field approximation is widely used in statistical physics and has been especially useful in order to predict phase transitions, both in discrete and in continuous models. One of the best known examples is the Curie-Weiss model of ferromagnetism, which is a mean field theory for the Ising model: in this case,

$$\langle -\beta E(\boldsymbol{x})\rangle_{q^*_{\setminus i}} = \beta\left(B_i + \sum_{j\in\partial i} J_{ij}\langle x_j\rangle\right) x_i, \tag{3.15}$$

where ∂i denotes the set of the neighbors of the ith site, B_i is the external magnetic field acting on the ith site and J_{ij} is the interaction strength between sites i and j. Inserting Eq. (3.15) in Eq. (3.13) and using it to compute the magnetization yields the well known self-consistency equation:

$$\langle x_i\rangle = \tanh\left[\beta\left(B_i + \sum_{j\in\partial i} J_{ij}\langle x_j\rangle\right)\right]. \tag{3.16}$$

However, since the approximating distribution that describes the system is expressed as a product of single variable marginals, this approach assumes that the variables are statistically independent, thus failing to capture any correlations among them. Therefore, despite its mathematical and computational convenience, a naive mean field treatment is often not sufficiently accurate for large coupled systems and one needs to retain some information about the most relevant dependencies among the components.

3.3 Belief propagation

3.3.1 Belief propagation equations on a tree

Belief propagation (or *sum-product*) is a distributed "message-passing" algorithm used to approximately compute the marginals and the free energy of the Boltzmann distribution. Its equations were independently derived many times in several fields. In physics, the ideas underlying this method date back to an article by Bethe [64], where a simple form of the belief propagation equations was applied to the ferromagnetic Ising model on a lattice, and to a work by Guggenheim [65] about the theory of regular binary liquid mixtures. In coding theory, belief propagation was first used by Gallager for error correction in low density parity check codes [66] and then rediscovered in the 1990s. As an algorithm for inference in tree-like Bayesian networks, it was developed by Pearl [67].

In this section, the method will be briefly presented as applied to factor graphs using a modern formalism similar to the one found in [68]. We wish to compute the marginals and the partition function (or, equivalently, the free energy) of a joint probability distribution expressed as:

$$P(\boldsymbol{x}) = \frac{1}{Z}\prod_{a=1}^{M}\psi_a(\boldsymbol{x}_a), \tag{3.17}$$

where $\boldsymbol{x}_a$ denotes the set of variables associated with the set ∂a of variables nodes that are first neighbors of the factor node a in the factor graph, i.e. $\boldsymbol{x}_a = \{x_i | i \in \partial a\}$. The belief propagation equations are given by:

$$m_{i\to a}^{(t+1)}(x_i) = \frac{1}{Z_{i\to a}}\prod_{b\in\partial i\setminus a} m_{b\to i}^{(t)}(x_i), \tag{3.18}$$

$$m_{a\to i}^{(t+1)}(x_i) = \frac{1}{Z_{a\to i}}\sum_{\boldsymbol{x}_{\partial a\setminus i}}\psi_a(\boldsymbol{x}_a)\prod_{j\in\partial a\setminus i} m_{j\to a}^{(t)}(x_j). \tag{3.19}$$

and are pictorially depicted in Fig. 3.1. The functions $m_{i\to a}(x_i)$ and $m_{a\to i}(x_i)$ are called *messages*. Their fixed point values, which we denote as $m^*_{i\to a}(x_i)$ and $m^*_{a\to i}(x_i)$, correspond to single variable marginals in a modified graphical model. In particular, $m^*_{i\to a}(x_i)$ is the marginal distribution of the variable x_i in a factor graph where the factor a has been erased and $m^*_{a\to i}(x_i)$ is the marginal distribution of the same variable in a factor graph where, among the factor nodes that are neighbors of i, only a is kept and all other factors are removed [68].

After a certain number of iterations, the single variable marginals of the distribution (3.17) are estimated as the product of the messages from all neighboring factors:

$$b_i(x_i) = \frac{1}{Z_i}\prod_{a\in\partial i} m_{a\to i}(x_i). \tag{3.20}$$

The messages also allow one to estimate the marginal distributions of subsets of variables as:

$$b_a(\boldsymbol{x}_a) = \frac{1}{Z_a}\psi_a(\boldsymbol{x}_a)\prod_{i\in\partial a} m_{i\to a}(x_i)\,. \tag{3.21}$$

We shall refer to the estimates $b_i(x_i)$ and $b_a(\boldsymbol{x}_a)$ as *marginal beliefs* (or simply as *beliefs*).

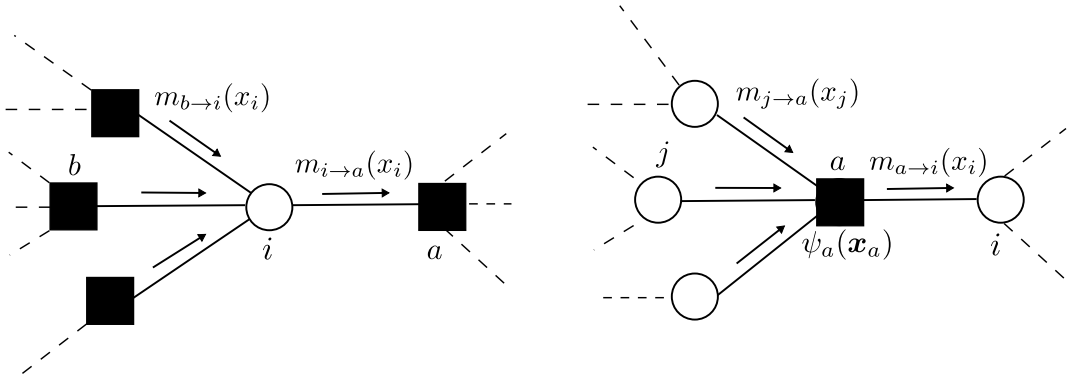


Figure 3.1: Graphical depiction of the belief propagation update rules.

If the factor graph is a tree, i.e. if it has no loops, then the belief propagation equations (3.18) and (3.19) are exact and the messages converge to fixed point values in a finite number of iterations. More precisely, the number of updates needed for convergence is given by the maximum distance between any two variable nodes in the tree, as can be easily verified by induction (see [68] for a proof). In this case, the marginals of the distribution (3.17) are exactly equal to the marginal beliefs (3.20) and (3.21).

For a tree, the junction tree theorem, which can also be easily proved by induction [26], states that any joint distribution of the form (3.17) can be factorized in terms of its marginals as:

$$P(\boldsymbol{x}) = \prod_a p(\boldsymbol{x}_a)\prod_i p(x_i)^{1-|\partial i|} = \tag{3.22}$$

$$= \prod_a \left(\frac{p(\boldsymbol{x}_a)}{\prod_{i\in\partial a} p(x_i)}\right)\prod_i p(x_i). \tag{3.23}$$

The free energy can be obtained directly from $F = -\ln Z$, where, comparing (3.17) and (3.22), one obtains:

$$Z = \frac{\prod_a \psi_a(\boldsymbol{x}_a)}{\prod_a p(\boldsymbol{x}_a)\prod_i p(x_i)^{1-|\partial i|}}, \tag{3.24}$$

which is valid for any $\boldsymbol{x}$. By multiplying by $P(\boldsymbol{x})$ and summing over $\boldsymbol{x}$, one has the following expression for the free energy of a tree:

$$F = -\sum_a \sum_{\boldsymbol{x}_a} p_a(\boldsymbol{x}_a)\ln\left(\frac{\psi_a(\boldsymbol{x}_a)}{p_a(\boldsymbol{x}_a)}\right) + \sum_j (1-|\partial j|)\sum_{x_j} p_j(x_j)\ln p_j(x_j). \tag{3.25}$$

This result is exact and can also be obtained from the decomposition in terms of internal energy and entropy, namely $F = U - S$, where it can be easily verified that the average energy reads:

$$U = -\sum_a \sum_{\boldsymbol{x}_a} p_a(\boldsymbol{x}_a) \ln \psi_a(\boldsymbol{x}_a), \tag{3.26}$$

and that, using Eq. (3.22), the entropy is given by:

$$S = \sum_a \sum_{\boldsymbol{x}_a} p_a(\boldsymbol{x}_a) \ln p_a(\boldsymbol{x}_a) - \sum_j (1 - |\partial j|) \sum_{x_j} p_j(x_j) \ln p_j(x_j). \tag{3.27}$$

Alternatively, the free energy can be expressed in terms of the normalizations of the messages [69]. More precisely, inserting the expressions for the beliefs given by Eq. (3.20) and by Eq. (3.21) in Eq. (3.24) yields the following expression for the partition function:

$$Z = \prod_a Z_a \prod_{(a,i)} \frac{Z_{i\to a}}{Z_i} \prod_i Z_i, \tag{3.28}$$

where Z_i, Z_a and $Z_{i\to a}$ are the normalization constants appearing in Eq. (3.20), (3.21) and (3.18), respectively. Defining:

$$Z_{ai} := \sum_{x_i} m_{a\to i}(x_i) m_{i\to a}(x_i), \tag{3.29}$$

and noticing that $Z_i / Z_{i\to a}$ is equal to the right hand side of the last definition[4], we finally obtain:

$$Z = \prod_a Z_a \prod_{(a,i)} Z_{ai}^{-1} \prod_i Z_i, \tag{3.30}$$

Accordingly, the associated expression for the free energy reads:

$$F = \sum_a F_a - \sum_{(a,i)} F_{ai} + \sum_i F_i, \tag{3.31}$$

[4]Indeed, one has:

$$b_i(x_i) = \frac{1}{Z_i} m_{a\to i}(x_i) \prod_{b\in\partial i\setminus a} m_{b\to i}(x_i) = \frac{Z_{i\to a}}{Z_i} m_{a\to i}(x_i) m_{i\to a}(x_i),$$

and the conclusion follows directly from the normalization condition on the marginal belief $b_i(x_i)$.

where:

$$F_a = \ln Z_a = \ln \left(\sum_{\boldsymbol{x}_a} \psi_a(\boldsymbol{x}_a) \prod_{i \in \partial a} m_{i \to a}(x_i) \right), \tag{3.32}$$

$$F_{ai} = \ln Z_{ai} = \ln \left(\sum_{x_i} m_{a \to i}(x_i) m_{i \to a}(x_i) \right), \tag{3.33}$$

$$F_i = \ln Z_i = \ln \left(\sum_{x_i} \prod_{a \in \partial i} m_{a \to i}(x_i) \right). \tag{3.34}$$

On a tree, the expression provided by Eq. (3.31) is exact and coincides with Eq. (3.25) when the messages appearing in Eq. (3.32), (3.33) and (3.34) are evaluated at their fixed point.

3.3.2 Bethe approximation and loopy belief propagation

The function defined by the right hand side of Eq. (3.31) is called *Bethe free energy*. Another way to write the same quantity consists in replacing the exact marginals with the marginal beliefs in Eq. (3.25):

$$F = -\sum_a \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln \left(\frac{\psi_a(\boldsymbol{x}_a)}{b_a(\boldsymbol{x}_a)} \right) + \sum_j (1 - |\partial j|) \sum_{x_j} b_j(x_j) \ln b_j(x_j). \tag{3.35}$$

As already mentioned, for an acyclic factor graph, the Bethe free energy is exact at the belief propagation fixed point, because the beliefs are exactly equal to the true marginals. Therefore, in this case, both Eq. (3.31) and Eq. (3.35) yield the exact free energy of the system.

However, if the graph is not tree-like, one can still approximate the free energy of the system by means of the Bethe free energy under the assumption that the graph can be considered locally acyclic. This is called the *Bethe approximation* [26, 70]. More precisely, given a set of marginals (beliefs) that satisfy the following conditions:

$$b_i(x_i) \geq 0, \quad b_a(\boldsymbol{x}_a) \geq 0, \tag{3.36}$$

$$\sum_{x_i} b_i(x_i) = 1, \tag{3.37}$$

$$\sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) = 1, \tag{3.38}$$

$$\sum_{\boldsymbol{x}_{a \setminus i}} b_a(\boldsymbol{x}_a) = b_i(x_i), \tag{3.39}$$

we associate a Bethe free energy (3.35) with it. The stationary points of the resulting

variational problem:

$$\mathscr{L}(\boldsymbol{b}, \boldsymbol{\lambda}) = F_{Bethe}[\boldsymbol{b}] - \sum_a \lambda_i \left[\sum_{x_i} b_i(x_i) - 1 \right] - \sum_{(i,a)} \sum_{x_i} \lambda_{ai}(x_i) \left[\sum_{\boldsymbol{x}_{a \setminus i}} b_a(\boldsymbol{x}_a) - b_i(x_i) \right], \tag{3.40}$$

where λ_i (λ_{ai}, respectively) are the Lagrange multipliers that impose the constraints specified by Eq. (3.37) (Eq. (3.39), respectively), are in one-to-one correspondence with the fixed points of the belief propagation algorithm. It can be proved that solving the variational problem and defining the messages in terms of the Lagrange multipliers as:

$$m_{i \to a}(x_i) \propto \mathrm{e}^{-\lambda_{ai}(x_i)}, \tag{3.41}$$

$$m_{a \to i}(x_i) \propto \sum_{\boldsymbol{x}_{a \setminus i}} \psi_a(\boldsymbol{x}_a) \prod_{j \in \partial a \setminus i} \mathrm{e}^{-\lambda_{aj}(x_j)}, \tag{3.42}$$

leads to the belief propagation equations (3.18) and (3.19). When applied to graphical models with cycles, the belief propagation recursion is referred to as *loopy belief propagation*. While the Bethe free energy is convex and there exists a unique fixed point on trees, in the general case of factor graphs with loops F_{Bethe} is, instead, non convex and there may be multiple belief propagation fixed points. Therefore, in the loopy case, convergence is not guaranteed, nor it is guaranteed that the iterative procedure will converge to the global minimum of F_{Bethe}.

3.4 The Thouless-Anderson-Palmer approach

The Thouless-Anderson-Palmer approach (TAP) [71–73] is an improvement of the naive mean field approximation that consists in including linear response corrections at leading order in the mean field equations. These corrections give rise to an additional term, called the *Onsager reaction term*, in the self-consistent equations of the model of interest.

The TAP equations were originally introduced for the Sherrington-Kirkpatrick (SK) model of a spin glass [71]:

$$\langle x_j \rangle = \tanh\left(B_j + \sum_{k \in \partial j} J_{jk} \langle x_k \rangle - \langle x_j \rangle \sum_{k \in \partial j} J_{jk}^2 \left(1 - \langle x_k \rangle^2\right) \right), \tag{3.43}$$

By comparing with Eq. (3.16), one sees that the two approximations differ as the Onsager correction is taken into account in Eq. (3.43). The TAP equations for the SK model can be derived from a Plefka expansion [74] by retaining all terms up to the second order.

The TAP equations can also be obtained by means of the *cavity method*, as we shall now outline, following [72, 75], for a general model with pairwise interactions between

the variables:

$$P(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{B}, \boldsymbol{J})} \exp\left[\sum_{i<j} x_i J_{ij} x_j + \sum_i x_i B_i\right] P_0(\boldsymbol{x}), \tag{3.44}$$

where $P_0(\boldsymbol{x})$ plays the role of a prior and will be assumed to be factorized over the variables $\boldsymbol{x}$:

$$P_0(\boldsymbol{x}) = \prod_i \psi_i(x_i). \tag{3.45}$$

Using the cavity method, we will now show how to compute the single variable marginal distribution:

$$P_i(x_i) = \frac{\int \prod_{j,j\neq i} \mathrm{d}x_j \psi_i(x_i) \exp\left[x_i\left(\sum_{j\in\partial i} J_{ij} x_j + B_i\right)\right] P(\boldsymbol{x}_{\setminus i})}{\int \prod_j \mathrm{d}x_j \psi_i(x_i) \exp\left[x_i\left(\sum_{j\in\partial i} J_{ij} x_j + B_i\right)\right] P(\boldsymbol{x}_{\setminus i})}, \tag{3.46}$$

where we have separated the part of $P(\boldsymbol{x})$ that depends on x_i, whereas $P(\boldsymbol{x}_{\setminus i})$ is the distribution of a system where the ith variable has been removed. Notice that variable x_i only interacts with all other variables by means of the effective field $h_i = \sum_{j\in\partial i} J_{ij} x_j$. Therefore, we introduce the *cavity distribution* of the field h_i at the location of the empty site x_i:

$$P(h_i \backslash x_i) = \int \prod_{j\neq i} \mathrm{d}x_j \delta\left(h_i - \sum_{j\in\partial i} J_{ij} x_j\right) P(\boldsymbol{x}_{\setminus i}). \tag{3.47}$$

In the large connectivity limit, under the assumption that correlations are weak, one can use the central limit theorem and approximate Eq. (3.47) by a Gaussian distribution:

$$P(h_i \setminus x_i) \approx \frac{1}{\sqrt{2\pi V_i}} \exp\left[-\frac{\left(h_i - \langle h_i\rangle_{\setminus i}\right)^2}{2V_i}\right], \tag{3.48}$$

where $V_i = \langle h_i^2\rangle_{\setminus i} - \langle h_i\rangle_{\setminus i}^2$. This is equivalent to neglecting all cumulants of the cavity distribution having order larger than 2.

By introducing the effective single variable Hamiltonian:

$$H_i(x) = -\ln\langle \mathrm{e}^{x(h_i+B_i)}\rangle_{\setminus i}, \tag{3.49}$$

where the average $\langle \dots \rangle_{\setminus i}$ denotes the expectation value computed with respect to the cavity field distribution (3.47), Eq. (3.46) can be rewritten as:

$$P_i(x_i) = \frac{\psi_i(x_i)}{Z_0^{(i)}} \mathrm{e}^{-H_i(x_i)} \approx \frac{1}{Z_0^{(i)}} \psi_i(x_i) \exp\left[x_i\left(\langle h_i\rangle_{\setminus i} + B_i\right) + \frac{V_i}{2} {x_i}^2\right], \tag{3.50}$$

where the partition function is given by:

$$Z_0^{(i)} = \int \mathrm{d}x \psi_i(x)\, \mathrm{e}^{-H_i(x)} \approx \int \mathrm{d}x \psi_i(x) \exp\left[x\left(\langle h_i\rangle_{\setminus i} + B_i\right) + \frac{V_i}{2}x^2\right], \tag{3.51}$$

and we have used the Gaussian approximation of the cavity field distributions (3.48) in the last step.

In order to self-consistently compute $\langle h_i\rangle_{\setminus i}$ for $i = 1, \dots, N$, we start from the identity:

$$\langle h_i\rangle = \frac{1}{Z_0^{(i)}} \int \mathrm{d}x \psi_i(x) \int \mathrm{d}h_i h_i p(h_i \setminus x)\, \mathrm{e}^{x(h_i+B_i)} = \frac{1}{Z_0^{(i)}} \int \mathrm{d}x \psi_i(x) \frac{\partial}{\partial x} \mathrm{e}^{-H_i(x)}\,. \tag{3.52}$$

On the one hand, one obtains:

$$\langle h_i\rangle = \sum_{j\in\partial i} J_{ij}\langle x_j\rangle, \tag{3.53}$$

and, on the other hand, using again the Gaussian approximation (3.48) one has:

$$\langle h_i\rangle \approx \langle h_i\rangle_{\setminus i} + V_i\langle x_i\rangle. \tag{3.54}$$

Thus, the $\langle h_i\rangle_{\setminus i}$ can be determined by:

$$\langle h_i\rangle_{\setminus i} = \sum_{j\in\partial i} J_{ij}\langle x_j\rangle - V_i\langle x_i\rangle, \tag{3.55}$$

where the correction $V_i\langle x_i\rangle$ to the naive mean-field term $\sum_{j\in\partial i} J_{ij}\langle x_j\rangle$ is the Onsager reaction term. By definition, the V_i are given by:

$$V_i = \langle h_i^2\rangle_{\setminus i} - \langle h_i^2\rangle_{\setminus i} = \sum_{j,k} J_{ij}J_{ik}(\langle x_j x_k\rangle_{\setminus i} - \langle x_j\rangle_{\setminus i}\langle x_k\rangle_{\setminus i}). \tag{3.56}$$

In the thermodynamic limit $N \to \infty$, for independent random couplings and assuming $\langle \dots\rangle_{\setminus i} \approx \langle \dots\rangle$, the non-diagonal correlations appearing in V_i can be neglected and one has:

$$V_i = \sum_{j\in\partial i} J_{ij}^2(1 - \langle x_j\rangle^2). \tag{3.57}$$

From Eq. (3.50), the magnetizations for Ising binary variables $x_i \in \{-1, +1\}$ read:

$$\langle x_i\rangle = \tanh((B_i + \langle h_i\rangle_{\setminus i})x_i), \tag{3.58}$$

so that, substituting Eq. (3.55) and Eq. (3.57), one finally recovers the TAP equations for the Sherrington-Kirkpatrick model (3.43).

3.5 The adaptive Thouless-Anderson-Palmer approach

The adaptive Thouless-Anderson-Palmer approach [75] introduced by Opper differs from standard TAP in the way the variances V_i are computed. In fact, noticing that:

$$\langle x_i \rangle = \frac{\partial \ln Z_0^{(i)}}{\partial B_i} \tag{3.59}$$

and using the linear response relation:

$$\chi_{ij} \equiv \frac{\partial \langle x_i \rangle}{\partial B_j} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle, \tag{3.60}$$

where χ_{ij} denotes the susceptibility, we obtain:

$$\chi_{ij} = \left(\delta_{ij} + \frac{\partial \langle h_i \rangle_{\setminus i}}{\partial B_j} \right) \frac{\partial \langle x_i \rangle}{\partial B_i}. \tag{3.61}$$

Moreover, inserting Eq. (3.55) and computing its derivative, the susceptibility finally reads:

$$\chi_{ij} = \frac{\partial \langle x_i \rangle}{\partial B_i} \left[\delta_{ij} + \sum_k (J_{ik} - V_k \delta_{ik}) \chi_{kj} \right], \tag{3.62}$$

or, in matrix form:

$$(\boldsymbol{\Lambda} - \mathbf{J}) \boldsymbol{\chi} = \mathbf{I}, \tag{3.63}$$

where $\boldsymbol{\Lambda}$ is defined as:

$$\Lambda = \mathrm{diag}(\Lambda_1, \dots, \Lambda_N), \Lambda_i \equiv V_i + \frac{1}{\chi_{ii}}. \tag{3.64}$$

Rewriting the last relation as:

$$\boldsymbol{\chi} = (\boldsymbol{\Lambda} - \mathbf{J})^{-1} \tag{3.65}$$

and taking the diagonal elements of the susceptibility matrix, one finally obtains an implicit equation for the variances V_i:

$$\frac{1}{\Lambda_i - V_i} = \chi_{ii} = [(\boldsymbol{\Lambda} - \mathbf{J})^{-1}]_{ii}, \tag{3.66}$$

which replaces Eq. (3.57) and that should be solved self-consistently. Once the updated values $\langle h_i \rangle^*_{\setminus i}$ and V_i^* are available, the means and the variances are given by:

$$\langle x_i \rangle = \frac{\partial}{\partial B_i} \ln Z_0^{(i)}(B_i, \langle h_i \rangle^*_{\setminus i}, V_i^*), \tag{3.67}$$

and by:

$$\langle x_i^2 \rangle - \langle x_i \rangle^2 = \frac{\partial^2}{\partial B_i^2} \ln Z_0^{(i)}(B_i, \langle h_i \rangle^*_{\setminus i}, V_i^*). \tag{3.68}$$

3.6 Approximate Message Passing (AMP)

Approximate message passing (AMP) is an efficient algorithm originally proposed by Donoho, Maleki and Montanari [76] and inspired by belief propagation that is able to perform compressed sensing reconstruction significantly faster than convex optimization techniques. It has the remarkable property that the dynamics of the algorithm can be rigorously tracked via a scalar state-evolution that holds in the case of large i.i.d. sub-Gaussian measurement matrices. In the following, a derivation based on loopy belief propagation is presented, as found e.g. in references [29] and [77]. With regard to the scalar state-evolution property, as it will not be used in this thesis, we refer the interested reader to references [78–80], where a proof is provided.

3.6.1 Loopy belief propagation for the GLM

Given a GLM and its related posterior distribution:

$$P(\boldsymbol{w}|\ \boldsymbol{y}, \mathbf{F}) = \frac{1}{Z}\prod_{\mu=1}^{M} P_\mu\left(y_\mu\middle|\ z_\mu\right)\prod_{i=1}^{N}\psi_i\left(w_i\right), \qquad \text{where } z_\mu = \sum_{i=1}^{N} F_{\mu i} w_i, \tag{3.69}$$

we can associate a factor graph with it, where, as usual, factor nodes correspond to interactions and variable nodes to the variables w_i.

For the factor graph in Fig. 3.2, it is straightforward to write the following set of $2NM$ BP update equations, one for each variable-factor pair (i, μ):

$$m_{i\to\mu}(w_i) = \frac{1}{Z_{i\to\mu}}\psi_i(w_i)\prod_{\gamma\neq\mu} m_{\gamma\to i}(w_i) \tag{3.70}$$

and one for each interaction-variable pair (μ, i):

$$m_{\mu\to i}(w_i) = \frac{1}{Z_{\mu\to i}}\int\prod_{j\neq i}\mathrm{d}w_j P_\mu(y_\mu|\boldsymbol{F}_\mu^\top\boldsymbol{w})\prod_{j\neq i} m_{j\to\mu}(w_j) \tag{3.71}$$

Unfortunately, without further approximations, this BP formulation is computationally intractable, even assuming that a suitable discretization of the continuous variables w_i induces a reliable approximation of the messages and of the resulting posterior distribution. The intractability arises from the fact that one needs to solve a set of coupled integral equations where the integrals are taken over variables interacting through the factors $P_\mu(y_\mu|\boldsymbol{F}_\mu^\top\boldsymbol{w})$.

3.6.2 Relaxed belief propagation

Under the assumption that the entries of the signal are independent (but not necessarily identically distributed) and that the scaling of the entries of the matrix $\mathbf{F}$ is $1/\sqrt{N}$, the

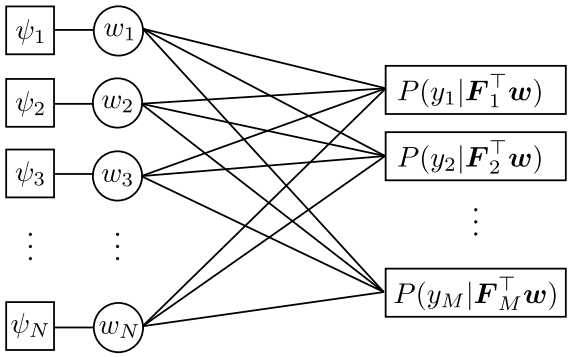


Figure 3.2: Factor graph associated with the posterior distribution of the GLM. Circles correspond to the unknowns, whereas squares correspond to interactions.

relaxed belief propagation (rBP) scheme allows one to obtain a tractable recursion by exploiting the central limit theorem and by projecting the messages to a pair of means and variances which parameterize the marginals for every given factor-variable pair of nodes of the factor graph.

In order to see this, we first note that Eq. (3.71) can be rewritten as:

$$m_{\mu\to i}(w_i) = \frac{1}{Z_{\mu\to i}} \int \mathrm{d}s P_\mu \left(y_\mu | s + F_{\mu i} w_i\right) \int \prod_{j\neq i} \mathrm{d}w_j \delta\left(s - \sum_{j\neq i} F_{\mu j} w_j\right) \prod_{j\neq i} m_{j\to\mu}(w_j), \quad (3.72)$$

where the integral with respect to the signal entries expresses the probability that the weighted sum $\sum_{j\neq i} F_{\mu j} w_j$ takes values s when each variable w_j, with $j \neq i$, is independently distributed according to its own associated message $m_{j\to\mu}(w_j)$. By the central limit theorem, the sum s is then Gaussian distributed, with mean $\omega_{\mu\to i}$ and variance $V_{\mu\to i}$ given by:

$$\begin{aligned} \omega_{\mu\to i} &= \textstyle\sum_{j\neq i} F_{\mu j} a_{j\to\mu} \\ V_{\mu\to i} &= \textstyle\sum_{j\neq i} F_{\mu j}^2 v_{j\to\mu}, \end{aligned} \quad (3.73)$$

where $a_{j\to\mu}$ and $v_{j\to\mu}$ are the mean and the variance of the message $m_{j\to\mu}$, respectively:

$$\begin{aligned} a_{j\to\mu} &= \textstyle\int \mathrm{d}w_j w_j m_{j\to\mu}\left(w_j\right) \\ v_{j\to\mu} &= \textstyle\int \mathrm{d}w_j w_j^2 m_{j\to\mu}\left(w_j\right) - a_{j\to\mu}^2 \end{aligned} \quad (3.74)$$

As a consequence, in terms of the intermediate reconstruction variable $z_\mu = s + F_{\mu i} w_i$, one can approximate Eq. (3.71) as:

$$m_{\mu\to i}(w_i) = \frac{1}{Z_{\mu\to i}} \int \mathrm{d}z_\mu P_\mu(y_\mu|z_\mu)\, \mathrm{e}^{-\frac{(z_\mu - F_{\mu i} w_i - \omega_{\mu\to i})^2}{2V_{\mu\to i}}}. \quad (3.75)$$

Using the fact that $F_{\mu i}$ is $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, one can expand the exponential in Eq. (3.75) around small values of $F_{\mu i} w_i$. Defining:

$$\begin{aligned} Z_{out}(y, \omega, V) &= \int \mathrm{d}z P(y|z)\, \mathrm{e}^{-\frac{(z-\omega)^2}{2V}} \\ g_{out}(y, \omega, V) &= \frac{1}{Z_{out}} \int \mathrm{d}z \left(\frac{z-\omega}{V}\right) P(y|z)\, \mathrm{e}^{-\frac{(z-\omega)^2}{2V}} \end{aligned} \tag{3.76}$$

and noticing that:

$$\begin{aligned} \frac{\partial_\omega Z_{out}(y, \omega, V)}{Z_{out}(y, \omega, V)} &= g_{out}(y, \omega, V) \\ \partial_\omega g_{out}(y, \omega, V) &= \frac{1}{Z_{out}} \int \mathrm{d}z \left(\frac{z-\omega}{V}\right)^2 P(y|z)\, \mathrm{e}^{-\frac{(z-\omega)^2}{2V}} - \frac{1}{V} - g_{out}^2(y, \omega, V) \end{aligned} \tag{3.77}$$

one can insert the definition (3.6.2), obtaining:

$$m_{\mu \to i}(w_i) \propto 1 + g_{out} F_{\mu i} w_i + \frac{1}{2} \left(\partial_\omega g_{out} + g_{out}^2\right) F_{\mu i}^2 w_i^2 + \mathcal{O}\left(N^{-\frac{3}{2}}\right) \tag{3.78}$$

and re-exponentiate the terms that depend on w_i. Doing so leads to a Gaussian form of the messages:

$$m_{\mu \to i}(w_i) \propto \mathrm{e}^{-\frac{1}{2} A_{\mu \to i} w_i^2 + B_{\mu \to i} w_i}, \tag{3.79}$$

where:

$$\begin{aligned} A_{\mu \to i} &= -\partial_\omega g_{out}(y_\mu, \omega_{\mu \to i}, V_{\mu \to i}) F_{\mu i}^2 \\ B_{\mu \to i} &= g_{out}(y_\mu, \omega_{\mu \to i}, V_{\mu \to i}) F_{\mu i} \end{aligned} \tag{3.80}$$

The Gaussian form of the messages (3.79) implies that the product of messages appearing in Eq. (3.70) is also Gaussian, with mean:

$$R_{i \to \mu} = \frac{\sum_{\gamma \neq \mu} B_{\gamma \to i}}{\sum_{\gamma \neq \mu} A_{\gamma \to i}} \tag{3.81}$$

and variance:

$$\Sigma_{i \to \mu} = \frac{1}{\sum_{\gamma \neq \mu} A_{\gamma \to i}}. \tag{3.82}$$

Thus, the messages in Eq. (3.70) can be expressed as:

$$m_{i \to \mu}(w_i) = \frac{1}{Z_{i \to \mu}} \psi_i(w_i)\, \mathrm{e}^{-\frac{(w_i - R_{i \to \mu})^2}{2\Sigma_{i \to \mu}}} \tag{3.83}$$

and the set of equations can be closed by inserting the latter expression in the means and variances (3.74)

$$\begin{aligned} a_{i\to\mu} &= f_a(R_{i\to\mu}, \Sigma_{i\to\mu}) \\ v_{i\to\mu} &= f_v(R_{i\to\mu}, \Sigma_{i\to\mu}) \end{aligned} \tag{3.84}$$

where the functions f_a and f_v are given by:

$$f_a(R, \Sigma) = \frac{1}{\mathcal{Z}} \int \mathrm{d}w w \psi_i(w)\, \mathrm{e}^{-\frac{(w-R)^2}{2\Sigma}} \tag{3.85}$$

$$f_v(R, \Sigma) = \frac{1}{\mathcal{Z}} \int \mathrm{d}w w^2 \psi_i(w)\, \mathrm{e}^{-\frac{(w-R)^2}{2\Sigma}} - f_a^2(R, \Sigma) \tag{3.86}$$

$$\mathcal{Z} = \int \mathrm{d}w \psi_i(w)\, \mathrm{e}^{-\frac{(w-R)^2}{2\Sigma}} \tag{3.87}$$

and are related by the identity:

$$\partial_R f_a(R, \Sigma) = \frac{1}{\Sigma} f_v(R, \Sigma). \tag{3.88}$$

In conclusion, the rBP equations can be cast to one set of self-consistent equations for each variable-interaction pair, leading to $\mathcal{O}(MN)$ message updates per iteration. This set of iterative equations is given by (3.73), (3.80), (3.81), (3.82) and (3.84) and involves $\mathcal{O}(NM^2 + MN^2)$ operations per iteration. The resulting algorithm gives as output the means and variances of the estimated BP marginals:

$$a_i = f_a(R_i, \Sigma_i), \quad v_i = f_v(R_i, \Sigma_i) \tag{3.89}$$

where:

$$R_i = \frac{\sum_\mu B_{\mu\to i}}{\sum_\mu A_{\mu\to i}}, \quad \Sigma_i = \frac{1}{\sum_\mu A_{\mu\to i}}. \tag{3.90}$$

3.6.3 Approximate message passing

The computational cost of solving the rBP equations presented in section 3.6.2 can be further reduced by using the fact that the contribution of the target factor node in Eq. (3.84) can be considered to some extent negligible in the large Nlimit. As a consequence, the parameters a_i and v_i of the estimated marginals can be used directly in the iterative procedure, allowing to reduce the computational cost of the whole recursive scheme. The algorithm that results from Taylor expanding Eq. (3.84) and conserving the leading order contributions to the means and variances of the marginals corresponds to the TAP form of the belief propagation equations and goes under the name of approximate message passing (AMP) [76] or generalized approximate message passing (GAMP)

[81] in the compressed sensing literature. Following [29], we start by introducing the quantities:

$$\omega_\mu = \sum_i F_{\mu i} a_{i\to\mu}, \quad V_\mu = \sum_i F_{\mu i}^2 v_{i\to\mu} \tag{3.91}$$

and by recalling the definitions of R_i and Σ_i given in Eq. (3.90).

Before proceeding with the Taylor expansion of $a_{i\to\mu}$ and $v_{i\to\mu}$, we notice that

$$\omega_{\mu\to i} = \omega_\mu - F_{\mu i} a_{i\to\mu}, \quad V_{\mu\to i} = V_\mu - F_{\mu i}^2 v_{i\to\mu} \tag{3.92}$$

and that the difference $\Sigma_{i\to\mu} - \Sigma_i$ can be neglected as it is $\mathcal{O}(N^{-2})$, whereas for the difference $R_{i\to\mu} - R_i$ we have:

$$R_{i\to\mu} - R_i = -B_{\mu\to i}\Sigma_i + \mathcal{O}(N^{-1}). \tag{3.93}$$

For V_μ, one simply has:

$$V_\mu = \sum_i F_{\mu i}^2 v_{i\to\mu} \approx \sum_i F_{\mu i}^2 v_i \tag{3.94}$$

Performing a Taylor expansion of g_{out} around (y, ω_μ, V_μ) yields:

$$\begin{aligned} g_{out}(y, \omega_{\mu\to i}, V_{\mu\to i}) &= g_{out}(y, \omega_\mu, V_\mu) - \partial_\omega g_{out}(y, \omega_\mu, V_\mu) F_{\mu i} a_{i\to\mu} + \mathcal{O}(N^{-1}) \\ &= g_{out}(y, \omega_\mu, V_\mu) - \partial_\omega g_{out}(y, \omega_\mu, V_\mu) F_{\mu i} a_i + \mathcal{O}(N^{-1}), \end{aligned} \tag{3.95}$$

where the term that is linear in the matrix element $F_{\mu i}$ plays the role of an Onsager reaction term. We also expand Σ_i and R_i around the same point:

$$\begin{aligned} \Sigma_i &= \left(-\partial_\omega g_{out}(y, \omega_{\mu\to i}, V_{\mu\to i}) F_{\mu i}^2\right)^{-1} \approx \left(-\partial_\omega g_{out}(y, \omega_\mu, V_\mu) F_{\mu i}^2\right)^{-1} \\ R_i &= \left(-\sum_\mu \partial_\omega g_{out}(y, \omega_{\mu\to i}, V_{\mu\to i}) F_{\mu i}^2\right)^{-1} \left(\sum_\mu g_{out}(y_\mu, \omega_{\mu\to i}, V_{\mu\to i}) F_{\mu i}\right) \\ &\approx \left(-\sum_\mu \partial_\omega g_{out}(y, \omega_\mu, V_\mu) F_{\mu i}^2\right)^{-1} \left(\sum_\mu g_{out}(y_\mu, \omega_\mu, V_\mu) F_{\mu i} - \sum_\mu \partial_\omega g_{out}(y, \omega_\mu, V_\mu) F_{\mu i}^2 a_i\right) \\ &= a_i + \Sigma_i \sum_\mu g_{out}(y_\mu, \omega_\mu, V_\mu) F_{\mu i}. \end{aligned} \tag{3.96}$$

Finally, we Taylor expand $a_{i\to\mu}$ and keep the Onsager correction:

$$\begin{aligned} a_{i\to\mu} &= f_a(R_{i\to\mu}, \Sigma_{i\to\mu}) \approx f_a(R_{i\to\mu}, \Sigma_i) \approx f_a(R_i, \Sigma_i) - \partial_R f_a(R_i, \Sigma_i) B_{\mu\to i}\Sigma_i \\ &= f_a(R_i, \Sigma_i) - B_{\mu\to i} f_v(R_i, \Sigma_i) = a_i - g_{out}(y, \omega_\mu, V_\mu) F_{\mu i} v_i, \end{aligned} \tag{3.97}$$

which, in turn, allows to rewrite ω_μ as:

$$\omega_\mu \approx \sum_i F_{\mu i}\left(a_i - g_{out}(y, \omega_\mu, V_\mu)F_{\mu i}v_i\right) = \sum_i F_{\mu i}a_i - V_\mu g_{out}(y, \omega_\mu, V_\mu). \tag{3.98}$$

Overall, the TAP form of the belief propagation scheme consists in a closed set of equations that are given by (3.94), (3.98) and, after having evaluated g_{out} and the derivative $\partial_\omega g_{out}$, by (3.96) and (3.89). The algorithm involves only $\mathcal{O}(N + M)$ message updates per iteration, thus lowering the computational cost to $\mathcal{O}(MN)$. In order for this recursion to converge properly, it is crucial that the Onsager correction in equation is evaluated one time step back. This aspect is a peculiar feature of the TAP equations that needs to be taken into account when constructing iterative TAP based iterative schemes.

3.7 Vector approximate message passing

The AMP algorithm for standard linear estimation models introduced above is extremely efficient, yet it can diverge when the entries of the measurement matrix are not i.i.d. [82, 83]. Vector approximate message passing algorithm (VAMP) [84] is an extension of the AMP scheme that is able to deal with a larger class of measurement matrices. It has a rigorous state-evolution that holds under large right-orthogonally invariant random matrices, namely, matrices $\mathbf{F}$ that have a singular value decomposition (SVD) of the form $\mathbf{F} = \mathbf{USV}^\top$, where $\mathbf{U}$ is a deterministic $M \times M$ orthogonal matrix, $\mathbf{V}$ is a $N \times N$ matrix uniformly distributed over the set of orthogonal matrices (a *Haar distributed* matrix), $\mathbf{S} = \text{Diag}(\mathbf{s})$ is a $M \times N$ diagonal matrix and $\mathbf{s}$ is the vector of singular values.

VAMP can be derived in terms of an approximation based on belief propagation, where messages are passed on the factor graph associated with a joint distribution $p(\boldsymbol{y}, \boldsymbol{x})$ and where the vector variable $\boldsymbol{x}$ is "duplicated" as $\boldsymbol{x}_1 = \boldsymbol{x}_2$. As a consequence, the joint distribution $p(\boldsymbol{y}, \boldsymbol{x})$ is rewritten as:

$$p(\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_1)\delta(\boldsymbol{x} - \boldsymbol{x}_2)\mathcal{N}(\boldsymbol{y}; \mathbf{F}\boldsymbol{x}, \gamma_w^{-1}\mathbf{I}). \tag{3.99}$$

The factor graph related to (3.99) is shown in Fig. 3.3, where we assume that the prior $p(\boldsymbol{x})$ is i.i.d.:

$$p(\boldsymbol{x}) = \prod_{l=1}^{N} \psi_l(x_l), \tag{3.100}$$

and messages are computed according to the following variation of the belief propagation scheme:

Beliefs. At variable node i, the variable marginal belief $b_{BP}(\boldsymbol{x}_i) \propto \prod_{a\in\partial i} m_{a\to i}(\boldsymbol{x}_i)$ is projected onto a Gaussian distribution with mean vector $\hat{\boldsymbol{x}} = \int \boldsymbol{x} b_{BP}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and

covariance matrix proportional to the identity matrix with common variance η^{-1} given by the average variance of $b_{BP}(\boldsymbol{x}_i)$, namely $\eta^{-1} = \frac{1}{N}\sum_{l=1}^{N} \mathrm{Var}_{b_{BP}}(x_{i,l})$. Thus, the approximated belief is: $b_{app}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{x}}, \eta^{-1}\mathbf{I})$.

Variable-to-factor. In belief propagation, the message from a variable node to a neighboring factor node would be given by Eq. (3.18): $m_{i\to a}(\boldsymbol{x}_i) = b_{BP}(\boldsymbol{x}_i)/m_{a\to i}(\boldsymbol{x}_i)$. In VAMP, it is computed from the approximate belief as: $m_{i\to a}(\boldsymbol{x}_i) \propto b_{app}(\boldsymbol{x}_i)/m_{a\to i}(\boldsymbol{x}_i)$.

Factor-to-variable. The message from a factor node to an adjacent variable node is computed using $m_{a\to \boldsymbol{x}_i}(\boldsymbol{x}_i) = \int \prod_{j\in\partial a\setminus i} \mathrm{d}\boldsymbol{x}_j f_a(\boldsymbol{x}_a) \prod_{j\in\partial a\setminus i} m_{a\to \boldsymbol{x}_j}(\boldsymbol{x}_j)$ as in belief propagation (cfr. Eq. (3.19)).

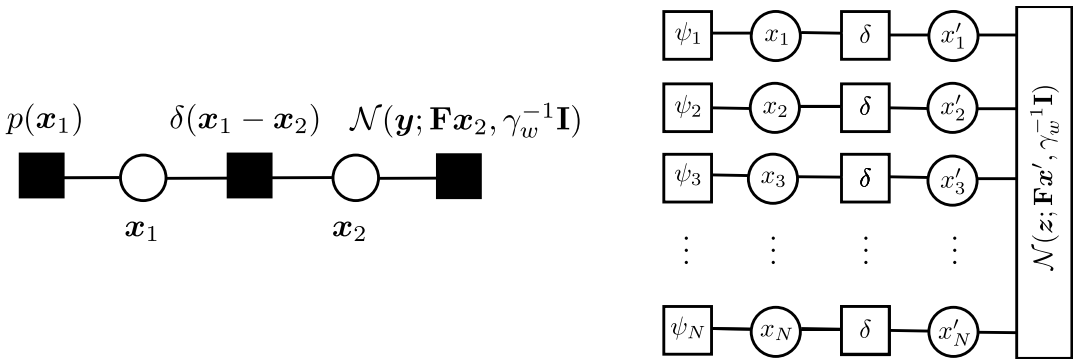


Figure 3.3: Representation of the VAMP factor graph with vector valued nodes (left) and with scalar nodes (right).

Passing the messages along the factor graph in Fig. 3.3 gives rise to the linear minimum mean squared error (LMMSE) formulation of the VAMP algorithm. In order to see this, we shall here highlight the correspondence between the steps of the algorithm and the messages passed along the factor graph, as detailed in Appendix A of reference [84], to which we refer for further details:

- **Initialization**. At iteration 0, the message from the delta factor to variable $\boldsymbol{x}_1$ is initialized as:

$$m_{\delta\to \boldsymbol{x}_1}(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{r}_{10}, \gamma_{10}^{-1}\mathbf{I}); \tag{3.101}$$

- **(→) Message from variable $\boldsymbol{x}_1$ to factor δ.**
Given the BP belief $b_{BP}(\boldsymbol{x}_1) \propto p(\boldsymbol{x}_1)\mathcal{N}(\boldsymbol{x}_{1k}; \boldsymbol{r}_{1k}, \gamma_{1k}^{-1}\mathbf{I})$, let its mean be denoted by $\hat{\boldsymbol{x}}_{1k}$ and its average variance by η_{1k}. Then, we have for the associated approximate belief:

$$b_{app}(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1; \hat{\boldsymbol{x}}_{1k}, \eta_{1k}^{-1}\mathbf{I}) \tag{3.102}$$

and the message from $\boldsymbol{x}_1$ to δ reads:

$$m_{\boldsymbol{x}_1\to\delta}(\boldsymbol{x}_1) \propto \frac{b_{app}(\boldsymbol{x}_1)}{m_{\delta\to\boldsymbol{x}_1}(\boldsymbol{x}_1)} \propto \mathcal{N}(\boldsymbol{x}_1;\boldsymbol{r}_{2k},\gamma_{2k}^{-1}\mathbf{I}), \tag{3.103}$$

where:

$$\boldsymbol{r}_{2k} = \frac{\hat{\boldsymbol{x}}_{1k}\eta_{1k} - \boldsymbol{r}_{1k}\gamma_{1k}}{\eta_{1k}-\gamma_{1k}}; \tag{3.104}$$

$$\gamma_{2k} = \eta_{1k} - \gamma_{1k}. \tag{3.105}$$

- **(→) Message from factor δ to variable $\boldsymbol{x}_2$.** Due to the Dirac delta distribution, computing the message $m_{\delta\to\boldsymbol{x}_2}$ results in a "copy" of message $m_{\boldsymbol{x}_1\to\delta}$, namely:

$$m_{\delta\to\boldsymbol{x}_2}(\boldsymbol{x}_2) \propto \mathcal{N}(\boldsymbol{x}_2;\boldsymbol{r}_{2k},\gamma_{2k}^{-1}\mathbf{I}), \tag{3.106}$$

with $\boldsymbol{r}_{2k}$ and γ_{2k} given by Eq. (3.104) and Eq. (3.105), respectively.

- **(→) Message from variable $\boldsymbol{x}_2$ to factor $\mathcal{N}(\boldsymbol{y};\mathbf{F}\boldsymbol{x},\gamma_w^{-1}\mathbf{I})$.** Here the BP belief is $b_{BP}(\boldsymbol{x}_2) \propto \mathcal{N}(\boldsymbol{x}_2;\boldsymbol{r}_{2k},\gamma_{2k}^{-1}\mathbf{I})\mathcal{N}(\boldsymbol{y};\mathbf{F}\boldsymbol{x}_{2k},\gamma_w^{-1}\mathbf{I})$, namely a Gaussian with mean:

$$\hat{\boldsymbol{x}}_{2k} = \left(\gamma_w\mathbf{F}^\top\mathbf{F} + \gamma_{2k}\mathbf{I}\right)^{-1}\left(\gamma_w\mathbf{F}^\top\boldsymbol{y} + \gamma_{2k}\boldsymbol{r}_{2k}\right), \tag{3.107}$$

and covariance matrix:

$$\boldsymbol{\Sigma}_{VAMP} = \left(\gamma_w\mathbf{F}^\top\mathbf{F} + \gamma_{2k}\mathbf{I}\right)^{-1}. \tag{3.108}$$

The approximate belief is then obtained by projecting the BP belief onto an isotropic Gaussian with mean $\hat{\boldsymbol{x}}_{2k}$ and shared precision parameter η_{2k} computed as the inverse of the average of the variances in (3.108):

$$b_{app}(\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{x}_2;\hat{\boldsymbol{x}}_{2k},\eta_{2k}^{-1}\mathbf{I}). \tag{3.109}$$

- **(←) Message from variable $\boldsymbol{x}_2$ to factor δ:**

$$m_{\boldsymbol{x}_2\to\delta}(\boldsymbol{x}_2) \propto \frac{\mathcal{N}(\boldsymbol{x}_2;\hat{\boldsymbol{x}}_{2k},\eta_{2k}^{-1}\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_2;\boldsymbol{r}_{2k},\gamma_{2k}^{-1}\mathbf{I})} \propto \mathcal{N}(\boldsymbol{x}_2;\boldsymbol{r}_{1,k+1},\gamma_{1,k+1}^{-1}\mathbf{I}), \tag{3.110}$$

where, by the quotient rule, the means and precision are expressed as:

$$\boldsymbol{r}_{1,k+1} = \frac{\hat{\boldsymbol{x}}_{2k}\eta_{2k} - \boldsymbol{r}_{2k}\gamma_{2k}}{\eta_{2k}-\gamma_{2k}}, \tag{3.111}$$

$$\gamma_{1,k+1} = \eta_{2k} - \gamma_{2k}. \tag{3.112}$$

- **(←) Message from factor δ to variable $\boldsymbol{x}_1$.** As before, the message "flowing" through the delta factor is left unchanged on the other side. Therefore, one has:

$$m_{\delta\to\boldsymbol{x}_1}(\boldsymbol{x}_1) = \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{r}_{1,k+1}, \gamma_{1,k+1}^{-1}\mathbf{I}). \tag{3.113}$$

In VAMP, two functions $\boldsymbol{g}_1(\boldsymbol{r}_{1k}, \gamma_{1k})$ and $\boldsymbol{g}_2(\boldsymbol{r}_{2k}, \gamma_{2k})$ are defined. The function $\boldsymbol{g}_1(\boldsymbol{r}_{1k}, \gamma_{1k})$ plays the role of a denoising function and is assumed to be separable, meaning that $(\boldsymbol{g}_1(\boldsymbol{r}_{1k}, \gamma_{1k}))_l = g_1(r_{1k,l}, \gamma_{1k})$ for all $l = 1, \dots, N$. It was denoted as g_{out} in the presentation of the AMP algorithm in Section 3.6.2. For the MMSE problem, the l-th component of $\boldsymbol{g}_1(\boldsymbol{r}_{1k}, \gamma_{1k})$ returns the expectation value:

$$g_1(r_{1k,l}, \gamma_{1k}) := \frac{\int x_l \psi(x_l)\mathcal{N}(x_l; r_{1k,l}, \gamma_{1k}^{-1})\mathrm{d}x_l}{\int \psi(x_l)\mathcal{N}(x_l; r_{1k,l}, \gamma_{1k}^{-1})\mathrm{d}x_l}, \tag{3.114}$$

which is used to compute the mean $\hat{\boldsymbol{x}}_{1k}$ of the approximate belief (3.102). The function $\boldsymbol{g}_2(\boldsymbol{r}_{2k}, \gamma_{2k})$ can be recognized as a MMSE estimate of $\boldsymbol{x}_2$ under the Gaussian likelihood $L(\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{y}; \mathbf{F}\boldsymbol{x}_2, \gamma_w^{-1}\mathbf{I})$ and under the Gaussian prior $p(\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{x}_2; \boldsymbol{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I})$. It is referred to as the *linear* MMSE (LMMSE) estimator [84], because the related estimate is linear in $\boldsymbol{r}_{2k}$:

$$\boldsymbol{g}_2(r_{2k}, \gamma_{2k}) := \left(\gamma_w \mathbf{F}^\top\mathbf{F} + \gamma_{2k}\mathbf{I}\right)^{-1}\left(\gamma_w \mathbf{F}^\top \boldsymbol{y} + \gamma_{2k}\boldsymbol{r}_{2k}\right) \tag{3.115}$$

The *divergence* of $\boldsymbol{g}(\boldsymbol{r}, \gamma)$, where $\boldsymbol{g} \in \{\boldsymbol{g}_1, \boldsymbol{g}_2\}$, is the diagonal of the Jacobian matrix $\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{r}) := \nabla_{\boldsymbol{r}}\boldsymbol{g}(\boldsymbol{r}, \gamma)$:

$$\boldsymbol{g}'(\boldsymbol{r}, \gamma) := \mathrm{diag}[\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{r})]. \tag{3.116}$$

For the MMSE problem, the componentwise derivatives $g_1'((r_{1k})_l, \gamma_{1k})$ are related to the variances $\mathrm{var}(x_l|(r_{1k})_l, \gamma_{1k})$ by means of the identity:

$$g_1'((r_{1k})_l, \gamma_{1k}) = \gamma_{1k}\mathrm{var}(x_l|(r_{1k})_l, \gamma_{1k}). \tag{3.117}$$

The precision parameter η_{1k} of the tilted distribution is extracted from Eq. (3.117) by taking the empirical average on both sides. On the other hand, one immediately realizes that the divergence of $\boldsymbol{g}_2(\boldsymbol{r}_{2k}, \gamma_{2k})$ is proportional to the diagonal of $\boldsymbol{\Sigma}_{VAMP}$ (3.108). As a consequence, the empirical average of $\boldsymbol{g}_2'(\boldsymbol{r}_{2k}, \gamma_{2k})$, which is proportional to the trace of $\boldsymbol{\Sigma}_{VAMP}$:

$$\frac{1}{N}\sum_{n=1}^{N}[g_2'(\boldsymbol{r}_{2k}, \gamma_{2k})]_n = \frac{\gamma_{2k}}{N}\,\mathrm{tr}\,\boldsymbol{\Sigma}_{VAMP}, \tag{3.118}$$

is used to compute the precision parameter η_{2k} in lines 11-12 of Algorithm 1. More details on the interpretation of the VAMP quantities will be given in Section 4.8, where

Algorithm 1 VAMP (LMMSE formulation).

procedure VAMP(a, b)
  for $k = 1, \dots,$ maxiter **do**
    # Denoising
    $\hat{\boldsymbol{x}}_{1k} = \boldsymbol{g}_1(\boldsymbol{r}_{1k}, \gamma_{1k})$ ▷ Denoising on $\boldsymbol{r}_1$
    $\alpha_{1k} = \mathtt{mean}(\boldsymbol{g}_1'(\boldsymbol{r}_{1k}, \gamma_{1k}))$ ▷ Divergence step
    $\eta_{1k} = \gamma_{1k}/\alpha_{1k}$
    $\gamma_{2k} = \eta_{1k} - \gamma_{1k}$
    $\boldsymbol{r}_{2k} = (\hat{\boldsymbol{x}}_{1k}\eta_{1k} - \boldsymbol{r}_{1k}\gamma_{1k})/\gamma_{2k}$ ▷ Onsager correction step
    # LMMSE estimation
    $\hat{\boldsymbol{x}}_{2k} = \boldsymbol{g}_2(\boldsymbol{r}_{2k}, \gamma_{2k})$ ▷ MMSE estimation of $\boldsymbol{x}_2$
    $\alpha_{2k} = \mathtt{mean}(\boldsymbol{g}_2'(\boldsymbol{r}_{2k}, \gamma_{2k}))$ ▷ Divergence step
    $\eta_{2k} = \gamma_{2k}/\alpha_{2k}$
    $\gamma_{1,k+1} = \eta_{2k} - \gamma_{2k}$
    $\boldsymbol{r}_{1,k+1} = (\hat{\boldsymbol{x}}_{2k}\eta_{2k} - \boldsymbol{r}_{2k}\gamma_{2k})/\gamma_{1,k+1}$ ▷ Onsager correction step
  return $\hat{\boldsymbol{x}}_{1,\text{maxiter}}$.

VAMP will be shown to be a particular instance of the expectation propagation algorithm.

The computational complexity of VAMP is $O(N^3)$, as it involves the inversion of the $N \times N$ matrix Σ_{VAMP} at each iteration. However, we notice that if a one-time singular value decomposition (SVD) of the matrix $\mathbf{F}$ is precomputed at the beginning, then the rest of the algorithm is dominated by matrix-vector multiplications and thus shares the same cost as AMP. Indeed, by taking advantage of the *standard* SVD $\mathbf{F} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, where $\mathbf{U}$ is a $M \times M$ orthogonal matrix, $\mathbf{S}$ is a rectangular diagonal matrix of size $M \times N$ and $\mathbf{V}$ is a $N \times N$ orthogonal matrix, starting from Eq. (3.115) and performing some algebraic manipulations leads to the expression:

$$\boldsymbol{x}_{2k} = \mathbf{V}\mathbf{D}^{-1}\left(\gamma_w \mathbf{S}^\top \mathbf{U}^\top \boldsymbol{y} + \gamma_{2k} \boldsymbol{V}^\top \boldsymbol{r}_{2k}\right), \tag{3.119}$$

which only involves inverting the $N \times N$ diagonal matrix:

$$\mathbf{D} = \gamma_w \mathbf{S}^\top \mathbf{S} + \gamma_{2k} \mathbf{I}. \tag{3.120}$$

Another option is to use the *compact* (or *"economy"*) SVD, leading to the SVD formulation of VAMP presented in Reference [84]. Either way, although the initial SVD still requires $O(N^3)$ elementary operations, it only needs to be computed once, while the cost of the remaining part of the VAMP scheme is $O(N^2)$.

3.7.1 Generalized vector approximate message passing

While the formulation of VAMP presented in the previous Section applies to the standard linear model, the algorithm can be extended to generalized linear estimation models as well. The approach pursued in Ref. [85] consists in alternating two inference steps, which are implemented as two nested loops, until convergence is achieved: the outer loop corresponds to MMSE estimation of the intermediate reconstruction variable $\boldsymbol{z} = \mathbf{F}\boldsymbol{x}$ under a Gaussian prior and likelihood $p(\boldsymbol{y}|\boldsymbol{z})$, whereas the inner loop implements a standard linear model (SLM) inference step, in which VAMP is run for a predefined number of iterations given the current estimate of $\boldsymbol{z}$. The two modules and the associated factor graph are shown in Fig. 3.4.

More explicitly, MMSE module takes as input the mean $z_{a,A}^{ext}$ and the variance $v_{a,A}^{ext}$ of the auxiliary variable $\boldsymbol{z}$ as estimated by the VAMP module. Given the Gaussian prior:

$$P_0(\boldsymbol{z}) = \prod_{a=1}^{M} \mathcal{N}(z_a; z_{a,A}^{ext}, v_{a,A}^{ext}),$$

and a likelihood with componentwise factorization:

$$p(\boldsymbol{y}|\boldsymbol{z}) = \prod_{a=1}^{M} p(y_a|z_a),$$

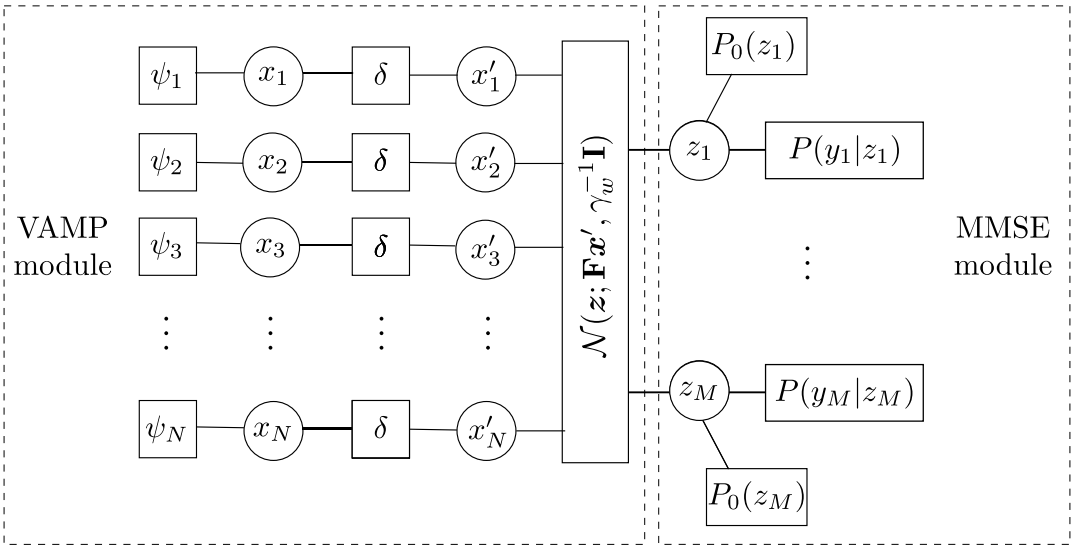


Figure 3.4: Representation of the grVAMP factor graph with scalar nodes.

the posterior of $\boldsymbol{z}$ reads:

$$q(z_a) \propto p(y_a|z_a)\mathcal{N}(z_a; z_{a,A}^{ext}, v_{a,A}^{ext}),$$

and we denote its mean and variances with $z_{a,A}^{post}$ and $v_{a,A}^{post}$, respectively. Then, at iteration t, the output of the MMSE estimation module is given by the so-called *extrinsic* mean and variance of $\boldsymbol{z}$, denoted as $\tilde{y}_a(t)$ and $\tilde{\sigma}_a^2(t)$ respectively, which are obtained by excluding the contribution of input messages $z_{a,A}^{ext}(t-1)$ and $v_{a,A}^{ext}(t-1)$ from $z_{a,A}^{post}(t)$ and from $v_{a,A}^{post}(t)$, as prescribed by the turbo principle [86]. Therefore, we have that the extrinsic variance of z_a is obtained as:

$$\frac{1}{\tilde{\sigma}_a^2(t)} = \frac{1}{v_{a,B}^{post}(t)} - \frac{1}{v_{a,A}^{ext}(t-1)},$$

and its extrinsic mean reads:

$$\tilde{y}_a(t) = \tilde{\sigma}_a^2(t)\left(\frac{z_{a,B}^{post}(t)}{v_{a,B}^{post}(t)} - \frac{z_{a,A}^{ext}(t-1)}{v_{a,A}^{ext}(t-1)}\right).$$

The extrinsic means and variances are then sent as input to the VAMP module, where $\tilde{y}_a(t)$ is interpreted as a pseudo-observation of z_a corrupted by Gaussian noise with variance $\tilde{\sigma}_a^2(t)$. Here, the SLM inference step in consists in solving the standard linear estimation problem for $\boldsymbol{x}$:

$$\tilde{\boldsymbol{y}} = \mathbf{F}\boldsymbol{x} + \tilde{\boldsymbol{w}}(t), \quad \tilde{\boldsymbol{w}}(t) \sim \mathcal{N}(\tilde{\boldsymbol{w}}(t); 0, \mathrm{diag}(\tilde{\boldsymbol{\sigma}}^2(t)), \tag{3.121}$$

resulting in the VAMP estimate of the posterior mean and of the variance of $\boldsymbol{x}$. Once these are determined, the posterior mean $z_{a,A}^{post}(t)$ and the posterior variance $v_{a,A}^{post}(t)$ for $\boldsymbol{z} = \mathbf{F}\boldsymbol{x}$ are computed and, by applying again the turbo principle, the extrinsic mean and variance of $\boldsymbol{z}$ are extracted from $z_{a,A}^{post}(t)$ and from $v_{a,A}^{post}(t)$ by excluding the contribution of the input messages $\tilde{\sigma}_a^2(t)$ and $\tilde{y}_a(t)$:

$$\frac{1}{v_{a,A}^{ext}(t)} = \frac{1}{v_{a,A}^{post}(t)} - \frac{1}{\tilde{\sigma}_a^2(t)}$$

$$z_{a,A}^{ext}(t) = v_{a,A}^{ext}(t)\left(\frac{z_{a,A}^{post}(t)}{v_{a,A}^{post}(t)} - \frac{\tilde{y}_a(t)}{\tilde{\sigma}_a^2(t)}\right).$$

Finally, the VAMP module outputs the estimates $z_{a,A}^{ext}(t)$ and $v_{a,A}^{ext}(t)$, which are sent back to the MMSE module, and the procedure is repeated iteratively until the estimate of the vector $\boldsymbol{x}$ to be inferred converges within a specified threshold.

Chapter 4

Expectation propagation

In this Chapter, we present the general framework on which the results presented in this PhD thesis build. The framework is called expectation propagation (EP) and was proposed by Thomas Minka in his seminal work [87], although, historically, the Gaussian case of EP that will be used in this dissertation was originally developed by Manfred Opper and Ole Winther under the name of *adaTAP* [75, 88, 89]. For the sake of generality, we will first briefly introduce EP using the formalism of exponential families, in terms of which the EP update rules can be cast in a very natural way as operations on so-called *natural parameters* and on their associated *moment parameters*, and then specialize to the Gaussian case of EP. We will highlight the connections between EP and some advanced mean field methods developed in the statistical physics of disordered systems, such as adaTAP and message passing algorithms including belief propagation and vector approximate message passing. Finally, we will review EP under the lens of variational problems, by introducing its variational free energy and by relating its stationary points to the fixed points of the algorithm. We will also show how this variational free energy can be used in order to construct an expectation-maximization like scheme allowing one to estimate unknown parameters by maximum likelihood.

4.1 Exponential families

In order to set the notation for the next sections, we introduce the exponential families of probability distributions [90] and highlight some parallels with statistical physics. As a thorough treatment of the topic is beyond the scope of this PhD thesis, we refer the reader to Refs. [26, 90] for additional details.

An exponential family is a family of probability distributions taking the form:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = h(\boldsymbol{x})\, \mathrm{e}^{\boldsymbol{\theta}^\top \boldsymbol{T}(\boldsymbol{x}) - \Phi(\boldsymbol{\theta})}, \tag{4.1}$$

where $\boldsymbol{T}(\boldsymbol{x})$ is a vector of sufficient statistics and $\boldsymbol{\theta}$ is a vector of parameters called *natural* (or canonical) parameters. The factor $h(\boldsymbol{x})$ is often absorbed in the exponential

function by adding a term $T_0(\boldsymbol{x}) = \ln(h(\boldsymbol{x}))$ to the set of sufficient statistics and by defining the corresponding natural parameter as $\theta_0 = 1$. The function $\Phi(\boldsymbol{\theta})$ is a normalization constant and is called the *log partition function*. In statistical mechanics, it corresponds to a negative free energy and is often called *free entropy* [68].

Exponential families play an important role in statistical physics. In fact, any Boltzmann distribution with Hamiltonian $\beta H = -\sum_{i=1}^{N} \theta_i T_i(\boldsymbol{x})$ belongs to an exponential family of distributions specified by the choice of the sufficient statistics $T_i(\boldsymbol{x})$, for $i = 1, \dots, N$. We here give a few examples:

- The probability distribution of the *Ising model* on a graph is:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_{(i,j)\in E} J_{ij} x_i x_j + \sum_{i \in V} h_i x_i \right),$$

  where $x_i \in \{-1, +1\}$ and we have absorbed the inverse temperature β in the couplings J_{ij} and in the external fields h_i. The vector of canonical parameters is

$$\boldsymbol{\theta} = \left(\{h_i\}_{i\in V}, \{J_{ij}\}_{(i,j)\in E} \right)^\top$$

  and the vector of sufficient statistics is

$$\boldsymbol{T}(\boldsymbol{x}) = \left(\{x_i\}_{i\in V}, \{x_i x_j\}_{(i,j)\in E} \right);$$

- In the *Potts model* on a graph (see, e.g., [91]), we have:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \frac{1}{Z} \exp\left(\sum_{(i,j)\in E} J_{ij} \delta_{x_i, x_j} + \sum_{i \in V} h_i \delta_{x_i, 1} \right) = \\ &= \frac{1}{Z} \exp\left(\sum_{u=1}^{q} \sum_{v=1}^{q} \sum_{(i,j)\in E} J_{ij} \mathbb{I}((x_i, x_j) = (u, v)) + \sum_{i \in V} h_i \mathbb{I}(x_i = 1) \right), \end{aligned}$$

  where $x_i \in \{1, \dots, q\}$ and we have absorbed the inverse temperature β in the couplings J_{ij} and in the external fields h_i. The vector of canonical parameters is

$$\boldsymbol{\theta} = \left(\{h_i\}_{i\in V}, \{J_{ij}\}_{(i,j)\in E} \right)^\top$$

  and the vector of sufficient statistics is

$$\boldsymbol{T}(\boldsymbol{x}) = \left(\{\mathbb{I}(x_i = 1)\}_{i\in V}, \{\mathbb{I}((x_i, x_j) = (u, v))\}_{(i,j)\in E, u\in\{1,\dots,q\}, v\in\{1,\dots,q\}} \right),$$

  where $\mathbb{I}$ denotes the indicator function;

- Similarly, in the *generalized Potts model* on a graph [92], we have:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{(i,j)\in E} J_{ij}(x_i, x_j) + \sum_{i\in V} h_i(x_i)\right) = \\ = \frac{1}{Z}\exp\left(\sum_{a=1}^{q}\sum_{b=1}^{q}\sum_{(i,j)\in E} J_{ij}^{ab}\mathbb{I}((x_i, x_j) = (a, b)) + \sum_{a=1}^{q}\sum_{i\in V} h_i^a \mathbb{I}(x_i = a)\right),$$

  where J_{ij}^{ab} is a $N \times N \times q \times q$ rank-4 tensor (N being the number of nodes) such that $J_{ij}^{ab} = 0$ if $(i, j) \notin E$ and, as before, $x_i \in \{1, \dots, q\}$. The vector of canonical parameters is

$$\boldsymbol{\theta} = \left(\left\{h_i^a\right\}_{i\in V, a\in\{1,\dots,q\}}, \left\{J_{ij}^{ab}\right\}_{(i,j)\in E, a\in\{1,\dots,q\}, b\in\{1,\dots,q\}}\right)^\top$$

  and the vector of sufficient statistics is

$$\boldsymbol{T}(\boldsymbol{x}) = \left(\left\{\mathbb{I}(x_i = a)\right\}_{i\in V, a\in\{1,\dots,q\}}, \left\{\mathbb{I}((x_i, x_j) = (a, b))\right\}_{(i,j)\in E, a\in\{1,\dots,q\}, b\in\{1,\dots,q\}}\right).$$

As observed in [93], a very useful property is the fact that the product of densities from a given exponential family $\mathscr{F}$ is proportional to a member of the same family:

$$\prod_{r=1}^{R} p(\boldsymbol{x}|\boldsymbol{\theta}_r) \propto p\left(\boldsymbol{x}\middle|\sum_{r=1}^{R}\boldsymbol{\theta}_r\right). \tag{4.2}$$

However, closure with respect to marginalization only holds for some exponential families, such as, for instance, multivariate Gaussian distributions.

Any exponential family admits an alternative parameterization in terms of *moment parameters* (also called *mean parameters*) [26], defined as:

$$\boldsymbol{\eta} = \langle \boldsymbol{T}(\boldsymbol{x})\rangle_{p_{\boldsymbol{\theta}}} = \int \boldsymbol{T}(\boldsymbol{x}) p_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{4.3}$$

These can be simply obtained by computing the derivatives of the log partition function, as customary in statistical physics (using the Helmholtz free energy). Using the notation of this section, we have:

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\nabla}_{\boldsymbol{\theta}}\Phi, \tag{4.4}$$

which allows one to obtain the moment parameters from the knowledge of the natural ones. The mapping from moment parameters to natural parameters is established by means of the Legendre transform of Φ with respect to the canonical parameters, which is equal to the negative Shannon entropy [26] and given by:

$$\Phi^* = \langle \ln p_{\boldsymbol{\eta}}(\boldsymbol{x})\rangle_{p_{\boldsymbol{\eta}}}, \tag{4.5}$$

where we used the notation $p_{\boldsymbol{\eta}}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\eta}(\boldsymbol{\theta}))$. Then, using Φ^* and assuming that the moments $\boldsymbol{\eta}$ are known, the mapping reads [26]

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \Phi^*. \tag{4.6}$$

In statistical physics, one example is the case of the inverse Ising model on a graph $\mathcal{G} = (V, E)$ [94]: there, the function Φ^* corresponds to the Legendre transform of $\ln Z$ with respect to the external fields and couplings, which yields the entropy S changed of sign:

$$\Phi^*(\boldsymbol{m}, \boldsymbol{\chi}) \equiv -S(\boldsymbol{m}, \boldsymbol{\chi}) = \min_{\mathbf{J}, \boldsymbol{h}} \left(-\sum_i h_i m_i - \sum_{(i,j)\in E} J_{ij}\chi_{ij} \right),$$

where $\boldsymbol{m}$ and $\boldsymbol{\chi}$ denote the magnetizations and the correlations, respectively. Finally, the canonical parameters are obtained from the knowledge of the magnetizations and of the correlations by evaluating the derivatives of S at the known values $\boldsymbol{m}$ and $\boldsymbol{\chi}$:

$$J_{ij} = -\left.\frac{\partial S}{\partial \chi_{ij}}\right|_{(\boldsymbol{\chi}, \boldsymbol{m})}, \quad h_i = -\left.\frac{\partial S}{\partial m_i}\right|_{(\boldsymbol{\chi}, \boldsymbol{m})}.$$

4.2 Expectation propagation

EP is an efficient approximation to compute posterior probabilities in Bayesian inference and was introduced in the machine learning community by Minka in its seminal paper [87]. However, as mentioned at the beginning of this chapter, it is rooted in ideas from the statistical physics of disordered systems and the Gaussian special case of EP was first developed as an improved mean-field method by Opper and Winther [75, 89]. Here, we will present the algorithm using the formalism of exponential families, as it is done in Matthias Seeger's PhD thesis [95] and in [93], before specializing to the Gaussian case. It is interesting to remark that the name "expectation propagation" is related to the fact that, from a message passing point of view, EP can be interpreted as an algorithm sending expected sufficient statistics from one region of a factor graph to another.

4.2.1 Assumed density filtering

Consider an intractable posterior distribution of the form:

$$P(\boldsymbol{x}) = \frac{1}{Z} G(\boldsymbol{x}) \prod_a \psi_a(\boldsymbol{x}), \tag{4.7}$$

where $G(\boldsymbol{x})$ could be a likelihood term and $\prod_a \psi_a(\boldsymbol{x})$ a prior distribution[1] and Z is a normalization, which will be interpreted as a partition function from a statistical mechanics standpoint. We shall assume that the term $G(\boldsymbol{x})$ is tractable, while $\prod_a \psi_a(\boldsymbol{x})$

[1] Or $G(\boldsymbol{x})$ could be a prior and $\prod_a \psi_a(\boldsymbol{x})$ a factorized likelihood, depending on the application.

is not. The aim is to approximate $P(\boldsymbol{x})$ by a member $Q(\boldsymbol{x})$ from an exponential family $\mathcal{F}$ and we shall assume that the tractable term $G(\boldsymbol{x})$ belongs in $\mathcal{F}$ as well. While one could think of fitting each approximating factor to the corresponding exact factor independently, in practice this approximation tends not to be a good one (see, e.g., [96]). A better way to proceed consists in approximating each exact factor one at a time, so that each approximating factor can be adjusted in order to correct for errors in earlier terms. This is the idea underlying the so-called *assumed density filtering* (ADF) algorithm [88, 97].

In ADF, one starts from $Q(\boldsymbol{x}) = G(\boldsymbol{x})$ and updates the current approximation $Q(\boldsymbol{x})$ by including a factor in it by multiplication, as $\hat{P}(\boldsymbol{x}) \propto Q(\boldsymbol{x})\psi_a(\boldsymbol{x})$, and then projecting $\hat{P}(\boldsymbol{x})$ back to the chosen exponential family[2]. The projection step is performed by choosing Q^{new} such that the Kullback-Leibler (KL) divergence between $\hat{P}$ and Q^{new} is minimal:

$$Q^{new}(\boldsymbol{x}) = \arg\min_{\hat{Q}\in\mathcal{F}} D_{KL}\left(\hat{P}(\boldsymbol{x})\|\hat{Q}(\boldsymbol{x})\right). \tag{4.8}$$

An important fact that will be used in the sequel is that, *for exponential families*, minimization of $D_{KL}\left(\hat{P}(\boldsymbol{x})\|\hat{Q}(\boldsymbol{x})\right)$ is equivalent to matching the expected sufficient statistics with respect to the distributions $\hat{P}$ and $\hat{Q}$ (*moment matching*), namely:

$$\langle \boldsymbol{T}(\boldsymbol{x})\rangle_{\hat{P}} = \langle \boldsymbol{T}(\boldsymbol{x})\rangle_{\hat{Q}}, \tag{4.9}$$

as it can be easily verified by requiring that the gradient of $D_{KL}\left(\hat{P}(\boldsymbol{x})\|\hat{Q}(\boldsymbol{x})\right)$ computed with respect to the natural parameters $\hat{\boldsymbol{\theta}}$ of the $\hat{Q}(\boldsymbol{x})$ distribution is equal to zero. To see this, let us start from the expressions of $\hat{P}(\boldsymbol{x})$ and $\hat{Q}(\boldsymbol{x})$:

$$\hat{P}(\boldsymbol{x}) = \frac{1}{Z_Q} h(\boldsymbol{x})\, \mathrm{e}^{\boldsymbol{\theta}^\top \boldsymbol{T}(\boldsymbol{x})}\, \psi_a(\boldsymbol{x})$$
$$\hat{Q}(\boldsymbol{x}) = \frac{1}{Z_{\hat{Q}}} h(\boldsymbol{x})\, \mathrm{e}^{\hat{\boldsymbol{\theta}}^\top \boldsymbol{T}(\boldsymbol{x})}$$

and write their KL divergence:

$$D_{KL}\left(\hat{P}(\boldsymbol{x})\|\hat{Q}(\boldsymbol{x})\right) = \int \mathrm{d}\boldsymbol{x}\hat{P}(\boldsymbol{x})\left[\ln Z_{\hat{Q}}(\hat{\boldsymbol{\theta}}) - \ln Z_Q(\boldsymbol{\theta}) + \ln\psi_a(\boldsymbol{x}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top \boldsymbol{T}(\boldsymbol{x})\right].$$

As only the first and last terms in the square bracket on the right hand side depend on $\hat{\boldsymbol{\theta}}$, we have:

$$\nabla_{\hat{\boldsymbol{\theta}}} D_{KL}\left(\hat{P}(\boldsymbol{x})\|\hat{Q}(\boldsymbol{x})\right) = \frac{1}{Z_{\hat{Q}}(\hat{\boldsymbol{\theta}})}\nabla_{\hat{\boldsymbol{\theta}}} Z_{\hat{Q}}(\hat{\boldsymbol{\theta}}) - \int \mathrm{d}\boldsymbol{x}\hat{P}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{x}),$$

[2] If there is no factor $G(\boldsymbol{x}) \in \mathcal{F}$ in $P(\boldsymbol{x})$, one sets $Q(\boldsymbol{x})$ to a uniform distribution.

where:

$$\frac{1}{Z_{\hat{Q}}(\hat{\boldsymbol{\theta}})}\nabla_{\hat{\boldsymbol{\theta}}}Z_{\hat{Q}}(\hat{\boldsymbol{\theta}}) = \int \mathrm{d}\boldsymbol{x}\hat{Q}(\boldsymbol{x})\boldsymbol{T}(\boldsymbol{x}).$$

Setting the gradient to zero, we obtain the moment matching conditions, as anticipated:

$$\langle \boldsymbol{T}(\boldsymbol{x})\rangle_{\hat{P}} = \langle \boldsymbol{T}(\boldsymbol{x})\rangle_{\hat{Q}}. \tag{4.10}$$

The inclusion and projection procedure outlined above is repeated sequentially, until all factors have been included in the approximated distribution $Q(\boldsymbol{x})$.

4.2.2 From assumed density filtering to expectation propagation

By construction, ADF is an *online* inference algorithm, meaning that input is processed in sequence, in a piece-by-piece fashion, as opposed to *offline* or *batch* algorithms, where a complete input data set is assumed to be available at once. As a consequence, each factor can be included only once and the result depends on the order in which these factors are presented. The expectation propagation algorithm is based on a revisited version of ADF, which was established by Minka in [87], where replacing $Q(\boldsymbol{x})$ by $Q^{new}(\boldsymbol{x})$ is interpreted as a *refinement* operation on approximating factors:

$$\phi_a(\boldsymbol{x}) \propto \frac{Q^{new}(\boldsymbol{x})}{Q(\boldsymbol{x})}. \tag{4.11}$$

Each factor $\psi_a(\boldsymbol{x})$ can be thought as being included in the approximated distribution by replacing $Q(\boldsymbol{x})$ with the product $Q^{new}(\boldsymbol{x}) = Q(\boldsymbol{x})\phi_a(\boldsymbol{x})$. In terms of natural parameters, one has:

$$\boldsymbol{\theta}^{new} = \boldsymbol{\theta} + \boldsymbol{\theta}_a,$$

where $\boldsymbol{\theta}$ (resp., $\boldsymbol{\theta}^{new}$) is the vector of natural parameters of $Q(\boldsymbol{x})$ (resp., $Q^{new}(\boldsymbol{x})$) and $\boldsymbol{\theta}_a$ is the one of $\phi_a(\boldsymbol{x})$. At the beginning, the approximating factors are initialized as $\phi_a(\boldsymbol{x}) = 1$ and the approximating distribution is given by $Q(\boldsymbol{x}) = G(\boldsymbol{x})$. Accordingly, the natural parameters are initialized as $\boldsymbol{\theta}_a = 0$ and as $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ denotes the vector of canonical parameters of $G(\boldsymbol{x})$. The relevance of this formulation of the ADF update lies in the fact that it allows multiple passes over the factors, which makes the resulting computational scheme iterative.

Keeping in mind the interpretation of ADF given above, the *EP update* is defined as deleting a factor $\phi_a(\boldsymbol{x})$ followed by an inclusion step in which the deleted factor is updated. More explicitly, we have:

Deletion: Given the current approximating distribution $Q(\boldsymbol{x})$, compute the *cavity distribution* [89] $Q^{\setminus a}(\boldsymbol{x})$:

$$Q^{\setminus a}(\boldsymbol{x}) \propto G(\boldsymbol{x})\prod_{b\neq a}\phi_b(\boldsymbol{x}). \tag{4.12}$$

In terms of natural parameters, one has $\boldsymbol{\theta}^{\setminus a} = \boldsymbol{\theta} - \boldsymbol{\theta}_a$, where $\boldsymbol{\theta}^{\setminus a}$ denotes the natural parameters of the cavity distribution.

Inclusion: Construct the *tilted distribution*, defined as:

$$Q^{(a)}(\boldsymbol{x}) \coloneqq \frac{1}{Z_{Q^{(a)}}} \psi_a(\boldsymbol{x}) Q^{\setminus a}(\boldsymbol{x}), \tag{4.13}$$

and compute its moments:

$$\boldsymbol{\eta}^{new} = \langle \boldsymbol{T}(\boldsymbol{x}) \rangle_{Q^{(a)}}. \tag{4.14}$$

Note that the tilted distribution differs from the approximated posterior $Q(\boldsymbol{x}) \in \mathscr{F}$ only in one factor, as it contains the exact factor $\psi_a(\boldsymbol{x})$ instead of the approximated one $\phi_a(\boldsymbol{x})$. Then, choose $Q^{new}(\boldsymbol{x})$ with these moments (*moment matching*) and replace $\phi_a(\boldsymbol{x})$ with $\phi_a^{new}(\boldsymbol{x})$ using Eq. (4.11). In terms of natural parameters, one has $\boldsymbol{\theta}_a = \boldsymbol{\theta}^{new} - \boldsymbol{\theta}^{\setminus a}$.

Analogously to the case of ADF, we remark that the matching of the moment parameters in the inclusion step is equivalent to minimizing a KL divergence, which, in the EP case, is the one between the tilted distribution $Q^{(a)}(\boldsymbol{x})$ and the EP approximation $Q(\boldsymbol{x})$ with respect to the canonical parameters $\boldsymbol{\theta}$ of the latter. By iteratively refining the approximating factors $\phi_a(\boldsymbol{x})$, EP attempts to mimick the way each factor $\psi_a(\boldsymbol{x})$ influences the tilted distribution $Q^{(a)}(\boldsymbol{x})$ as a whole, rather than to simply fit $\psi_a(\boldsymbol{x})$. Notice that, contrary to variational Bayes (see Section 3.1), both ADF and EP minimize the *forward* KL divergence $D_{KL}\left(p(\boldsymbol{x}) \| q(\boldsymbol{x})\right)$, rather than the *reverse* one, i.e. $D_{KL}\left(q(\boldsymbol{x}) \| p(\boldsymbol{x})\right)$, where $q(\boldsymbol{x})$ denotes the chosen approximating distribution. However, while in variational Bayes $p(\boldsymbol{x})$ is the true intractable distribution, we recall that in EP $p(\boldsymbol{x})$ corresponds to the tilted distribution, as computing the moments of the exact posterior distribution is an intractable problem. The fact that the two approximation schemes minimize two different KL divergences has relevant consequences on the properties of the support of the approximating distributions obtained as a result of the minimization [60]. In particular, notice that, on the one hand, the reverse KL divergence is *zero forcing* for $q(\boldsymbol{x})$: in fact, if $p(\boldsymbol{x})$ is zero for some $\boldsymbol{x}$, then $q(\boldsymbol{x})$ must be zero as well, otherwise $D_{KL}(q\|p)$ would become infinite. On the other hand, the forward KL is *zero avoiding* for $q(\boldsymbol{x})$, meaning that whenever $p(\boldsymbol{x})$ is strictly greater than zero, so must be $q(\boldsymbol{x})$ in order to prevent $D_{KL}(p\|q)$ from diverging. As a consequence, minimizing the reverse KL divergence results in $q(\boldsymbol{x})$ *underestimating* the support of $p(\boldsymbol{x})$, whereas, on the contrary, minimizing the forward KL divergence, as done in ADF and EP, leads to an approximating $q(\boldsymbol{x})$ which *overestimates* the support of the target distribution.

To conclude, notice that while EP is more robust and accurate than ADF as an approximation scheme, at the same time going from ADF to EP implies that the online nature of ADF is lost, making EP not suitable for online inference.

4.3 The Gaussian case of EP with univariate approximating factors

We will now introduce the Gaussian case of expectation propagation, in which the exponential family of distributions $\mathscr{F}$ is chosen to be the set of multivariate Gaussian distributions. Besides being closed under products, the Gaussian exponential family has the additional property of being closed under marginalization, as already mentioned in Sec. 4.1. Here, the EP moment matching condition reduces to matching the first and second moments of the tilted distribution and of the approximated (Gaussian) distribution. We will focus on the special case of Gaussian EP with univariate approximating factors, which is the computational framework used in the context of the linear estimation problems studied in this thesis. The formulation of Gaussian EP that will be presented in this section is published in Refs. [3, 96, 98].

Given a probability distribution expressed as in Eq. (4.7), namely:

$$P(\boldsymbol{x}) = \frac{1}{Z} G(\boldsymbol{x}) \prod_{i=1}^{N} \psi_i(x_i), \tag{4.15}$$

where $G(\boldsymbol{x})$ is a multivariate Gaussian distribution:

$$G(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{m}\right) \tag{4.16}$$

and $\psi_i(x_i)$ are factors the product of which makes the full distribution $P(\boldsymbol{x})$ intractable, we aim at computing its approximate marginals by means of the EP approximation. The fully approximated distribution $Q(\boldsymbol{x})$ is obtained by replacing each exact factor $\psi_i(x_i)$ with a univariate Gaussian distribution $\phi_i(x_i) = \mathcal{N}(w_i; a_i, d_i)$ having mean a_i and variance d_i. We have:

$$Q(\boldsymbol{x}) = \frac{1}{Z_Q} G(\boldsymbol{x}) \prod_{i=1}^{N} \phi_i(x_i) \tag{4.17}$$

$$:= \frac{1}{Z_Q} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})}, \tag{4.18}$$

where:

$$Z_Q = (2\pi)^{\frac{N}{2}} (\det \boldsymbol{\Sigma})^{\frac{1}{2}}, \tag{4.19}$$

$$\boldsymbol{\Sigma}^{-1} := \mathbf{A} + \boldsymbol{D}, \quad \bar{\boldsymbol{x}} := \boldsymbol{\Sigma}(\boldsymbol{m} + \boldsymbol{D}\boldsymbol{a}), \tag{4.20}$$

and $\boldsymbol{D}$ is a diagonal matrix having diagonal elements $d_1^{-1}, \dots, d_N^{-1}$.

In order to refine the mean a_n and the variance d_n of each factor $\phi_n(x_n)$, for $n = 1, \dots, N$, we perform an EP update consisting of a deletion of $\phi_n(x_n)$ followed by an

inclusion of the same term. To do so, we construct a *tilted* distribution $Q^{(n)}$, which, in the Gaussian case, is defined as:

$$Q^{(n)}(\boldsymbol{x}) \coloneqq \frac{1}{Z_{Q^{(n)}}} G(\boldsymbol{x}) \psi_n(x_n) \prod_{l \neq n} \phi_l(x_l; a_l, d_l) \tag{4.21}$$

$$= \frac{1}{Z_{Q^{(n)}}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}^{(n)})^\top \left(\boldsymbol{\Sigma}^{(n)}\right)^{-1} (\boldsymbol{x}-\bar{\boldsymbol{x}}^{(n)})} \psi_n(x_n), \tag{4.22}$$

where

$$\left(\boldsymbol{\Sigma}^{(n)}\right)^{-1} = \mathbf{A} + \boldsymbol{D}^{(n)}, \quad \bar{\boldsymbol{x}}^{(n)} = \boldsymbol{\Sigma}^{(n)} \left(\boldsymbol{m} + \boldsymbol{D}^{(n)} \boldsymbol{a}\right) \tag{4.23}$$

are Gaussian, and, analogously to Eq. (4.20), $\boldsymbol{D}^{(n)}$ is a diagonal matrix the entries of which are d_m^{-1} for all diagonal elements such that $m \neq n$ and zero for $m = n$. In order to complete the inclusion step, the tilted distribution is projected onto a multivariate Gaussian distribution by moment matching:

$$\langle x_n \rangle_{Q^{(n)}} = \langle x_n \rangle_Q \,, \quad \left\langle x_n^2 \right\rangle_{Q^{(n)}} = \left\langle x_n^2 \right\rangle_Q \,. \tag{4.24}$$

The computation of the moments of the *tilted* distribution on the left-hand side of Eq. (4.24) is dependent on the specific functional form of the distribution $\psi_i(x)$. In many cases, these can be computed analytically. When this is not possible, one can resort to numerical methods, such as Gaussian quadrature. On the other hand, the moments of Q can be easily computed by means of the product rule for univariate Gaussian distributions, which we recall here: given two univariate Gaussian distributions having means m_1 and m_2 and having variances S_1 and S_2, it is straightforward to verify that their product is given by:

$$\exp\left(\frac{(x-m_1)^2}{2S_1}\right) \exp\left(\frac{(x-m_2)^2}{2S_2}\right) \propto \exp\left(-\frac{(x-M)^2}{2S}\right), \tag{4.25}$$

where the variance S is expressed as:

$$S = \left(\frac{1}{S_1} + \frac{1}{S_2}\right)^{-1} \tag{4.26}$$

and the mean M reads:

$$M = S\left(\frac{m_1}{S_1} + \frac{m_2}{S_2}\right). \tag{4.27}$$

Indeed, considering the right hand side of Eq. (4.24), one sees that by integrating with respect to all variables x_m for $m \neq n$ and by using Eq. (4.25), the first and second moment

of Q are expressed as:

$$\langle x_n \rangle_Q = \left(\frac{1}{d_n} + \frac{1}{\Sigma_{nn}^{(n)}} \right)^{-1} \left(\frac{a_n}{d_n} + \frac{\bar{x}_n}{\Sigma_{nn}^{(n)}} \right), \tag{4.28}$$

$$\langle x_n^2 \rangle_Q = \frac{1}{\frac{1}{d_n} + \frac{1}{\Sigma_{nn}^{(n)}}} + \langle x_n \rangle_Q^2. \tag{4.29}$$

Since the two matrices Σ^{-1}, and $\left(\Sigma^{(n)}\right)^{-1}$ in Eqs. (4.20) and (4.23), respectively, only differ in one diagonal entry, we can leverage a low-rank update property to relate the two inverses, which follows from applying Sherman-Morrison formula [99] (see Sec. 4.5). As a consequence, the tilted parameters can be expressed in terms of the approximated ones:

$$\bar{x}_n^{(n)} = \left(\Sigma_{nn}^{(n)}\right) \left(\frac{\bar{x}_n}{\Sigma_{nn}} - \frac{a_n}{d_n} \right), \quad \left(\Sigma_{nn}^{(n)}\right) = \frac{\Sigma_{nn}}{1 - \frac{\Sigma_{nn}}{d_n}}, \tag{4.30}$$

which allows us to perform only one matrix inversion per iteration rather than N. After imposing the moment matching condition (4.24), we obtain an update rule for the parameters a_n, d_n of the approximating Gaussian factors $\phi_n(x_n)$, for $n = 1, \dots, N$:

$$d_n = \left(\frac{1}{\langle x_n^2 \rangle_{Q^{(n)}} - \langle x_n \rangle_{Q^{(n)}}^2} - \frac{1}{\Sigma_{nn}^{(n)}} \right)^{-1}, \tag{4.31}$$

$$a_n = \langle x_n \rangle_{Q^{(n)}} + \frac{d_n}{\Sigma_{nn}^{(n)}} \left(\langle x_n \rangle_{Q^{(n)}} - \left(\bar{x}^{(n)} \right)_n \right). \tag{4.32}$$

The iterations of the algorithm stop as soon as the EP parameters converge to a fixed point. When this happens, the tilted distributions provide the best approximation to the marginal densities of the posterior distribution in Eq. (4.15). In practice, from a numerical point of view, convergence at each iteration t (i.e. for each update of the $\boldsymbol{a}, \boldsymbol{d}$ vectors) is verified by checking the quantity:

$$\epsilon_t = \max_{n=1,\dots,N} \left\{ \left| \langle x_n \rangle_{Q_t^{(n)}} - \langle x_n \rangle_{Q_{t-1}^{(n)}} \right| + \left| \langle x_n^2 \rangle_{Q_t^{(n)}} - \langle x_n^2 \rangle_{Q_{t-1}^{(n)}} \right| \right\},$$

where $Q_t^{(n)}$ is the tilted distribution with parameters computed at iteration t, and by setting a convergence threshold ϵ_{stop}. Thus, as soon as $\epsilon_t < \epsilon_{\text{stop}}$, Gaussian EP returns the means and variances of the marginal tilted distributions. The means $\langle x_i \rangle_{Q^{(i)}}$ provide the EP estimate of the variables to be inferred, whereas the variances allow to extract the standard deviations $\sqrt{\langle x_i^2 \rangle_{Q^{(i)}} - \langle x_i \rangle_{Q^{(i)}}^2}$, which provide the estimate of the uncertainties associated with the EP inferred variables.

4.4 Gaussian EP with a Dirac delta factor

It is interesting to consider distributions of the kind:

$$P(\boldsymbol{x}) = \frac{1}{Z_P}\delta^M(\mathbf{G}\boldsymbol{x} - \tilde{\boldsymbol{y}})\prod_{i=1}^{N}\psi_i(x_i), \tag{4.33}$$

where $\delta^M(\boldsymbol{z})$ denotes the M-dimensional Dirac delta distribution, $\mathbf{G} \in \mathbb{R}^{M\times N}$ and $\tilde{\boldsymbol{y}} \in \mathbb{R}^M$, because they arise when taking a suitable infinite precision matrix limit of the multivariate Gaussian distribution $G(\boldsymbol{x})$ appearing in Eq. (4.7). This fact will be justified in the context of linear estimation problems in Sec. 5.1, where $G(\boldsymbol{x})$ and, consequently, $\delta^M(\mathbf{G}\boldsymbol{x})$ are interpreted as likelihood functions. There, the matrix $\mathbf{G}$ is $M\times N$ and has the form $\mathbf{G} = (-\mathbf{F}|\mathbf{I})$ and the vector $\boldsymbol{x}$ is composed of N variables, such that $M < N$ variables depend on all other variables. Indeed, without loss of generality, we may assume that $\boldsymbol{v} \coloneqq (x_{N-M+1}, \dots, x_N)^\top$ depends on the set of variables $\boldsymbol{u} \coloneqq (x_1, \dots, x_{N-M})^\top$ as:

$$\boldsymbol{v} = \mathbf{F}\boldsymbol{u} + \tilde{\boldsymbol{y}}. \tag{4.34}$$

The aim of this Section will be to outline the derivation of a Gaussian EP algorithm able to deal with this scenario. This formulation of EP will be used to obtain the results in Chapters 5 and 6 and is published in Ref. [100] in the homogeneous case $\tilde{\boldsymbol{y}} = 0$.

To set the notation, let us define the sets $U \coloneqq \{1, \dots, N-M\}$ and $V \coloneqq \{N-M+1, \dots, N\}$. Furthermore, let $\boldsymbol{e}_i$ denote the i-th basis vector of the standard basis of $\mathbb{R}^{N-M}$ if $i \in U$ or that of $\mathbb{R}^M$ in the case where $i \in V$. In order to adapt the EP scheme introduced in Sec. 4, we start by defining the Gaussian approximating factors:

$$\phi_i(x_i) = \exp\left(-\frac{(x_i - a_i)^2}{2d_i}\right), \tag{4.35}$$

and a fully Gaussian approximation of the posterior distribution (4.33), in which all priors ψ_i are replaced by factors of the form (4.35):

$$Q(\boldsymbol{x}) = \frac{1}{Z_Q}\delta^M(\mathbf{G}\boldsymbol{x} - \tilde{\boldsymbol{y}})\prod_{i=1}^{N}\phi_i(x_i). \tag{4.36}$$

From the linear dependence (4.34), it follows that $Q(\boldsymbol{x})$ can be rewritten as:

$$Q(\boldsymbol{x}) = \frac{1}{Z_Q}\delta^M(\mathbf{G}\boldsymbol{x} - \tilde{\boldsymbol{y}})\exp\left(-\frac{1}{2}(\boldsymbol{u} - \bar{\boldsymbol{u}})^\top\boldsymbol{\Sigma}_U^{-1}(\boldsymbol{u} - \bar{\boldsymbol{u}})\right), \tag{4.37}$$

where the covariance matrix $\boldsymbol{\Sigma}_U$ and the mean $\bar{\boldsymbol{u}}$ in Eq. (4.37) are given by:

$$\boldsymbol{\Sigma}_U^{-1} = \sum_{i\in U}\frac{1}{d_i}\boldsymbol{e}_i\boldsymbol{e}_i^\top + \mathbf{F}^\top\left(\sum_{i\in V}\frac{1}{d_i}\boldsymbol{e}_i\boldsymbol{e}_i^T\right)\mathbf{F}, \tag{4.38}$$

and by:

$$\bar{\boldsymbol{u}} = \Sigma_U \left(\sum_{i \in U} \frac{a_i}{d_i} \boldsymbol{e}_i + \sum_{i \in V} \frac{a_i + \tilde{y}_i}{d_i} \mathbf{F}^\top \boldsymbol{e}_i \right), \tag{4.39}$$

respectively. Notice that, since the marginal distribution of $Q(\boldsymbol{u})$ is Gaussian with mean $\bar{\boldsymbol{u}}$ and covariance matrix Σ_U and since the random vector $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{v})^\top$ is an affine transformation of $\boldsymbol{u}$, given by:

$$\boldsymbol{x} = \begin{pmatrix} \mathbf{I} \\ \mathbf{F} \end{pmatrix} \boldsymbol{u} + \begin{pmatrix} 0 \\ \tilde{\boldsymbol{y}} \end{pmatrix},$$

the joint distribution $Q(\boldsymbol{x})$ is Gaussian, with mean $\bar{\boldsymbol{x}} = (\bar{\boldsymbol{u}}, \mathbf{F}\bar{\boldsymbol{u}} + \tilde{\boldsymbol{y}})^\top$ and $N \times N$ covariance matrix Σ expressed as:

$$\Sigma = \begin{pmatrix} \mathbf{I} \\ \mathbf{F} \end{pmatrix} \Sigma_U \begin{pmatrix} \mathbf{I} & \mathbf{F} \end{pmatrix} = \begin{pmatrix} \Sigma_U & \Sigma_U \mathbf{F}^\top \\ \mathbf{F}\Sigma_U & \mathbf{F}\Sigma_U \mathbf{F}^\top \end{pmatrix}.$$

Moreover, as $Q(\boldsymbol{x})$ is Gaussian, the marginal distributions $Q(x_i)$ for each x_i ($i = 1, \dots, N$) are also Gaussian, with means:

$$\bar{x}_i = \begin{cases} \bar{u}_i, & i \in U \\ \tilde{y}_i + \boldsymbol{e}_i^\top \mathbf{F} \bar{\boldsymbol{u}}, & i \in V, \end{cases} \tag{4.40}$$

and variances:

$$\Sigma_{ii} = \begin{cases} \boldsymbol{e}_i^\top \Sigma_U \boldsymbol{e}_i, & i \in U, \\ \left(\boldsymbol{e}_i^\top \mathbf{F} \right) \Sigma_U \left(\mathbf{F}^\top \boldsymbol{e}_i \right), & i \in V. \end{cases} \tag{4.41}$$

We now need to introduce one tilted distribution $Q^{(i)}(\boldsymbol{x})$ for each $i = 1, \dots, N$:

$$Q^{(i)}(\boldsymbol{x}) \coloneqq \frac{1}{Z_{Q^{(i)}}} \delta^M(\mathbf{G}\boldsymbol{x} - \tilde{\boldsymbol{y}}) \psi_i(x_i) \prod_{j \neq i} \phi(x_j) = \psi_i(x_i) Q^{\setminus i}(\boldsymbol{x}), \tag{4.42}$$

where, in the last equality, we have isolated the Gaussian cavity distribution $Q^{\setminus i}(\boldsymbol{x})$:

$$Q^{\setminus i}(\boldsymbol{x}) = \frac{1}{Z_{Q^{(i)}}} \delta^M(\mathbf{G}\boldsymbol{x} - \tilde{\boldsymbol{y}}) \exp\left(-\frac{1}{2} \left(\boldsymbol{u} - \bar{\boldsymbol{u}}^{(i)} \right)^\top \left(\Sigma_U^{(i)} \right)^{-1} \left(\boldsymbol{u} - \bar{\boldsymbol{u}}^{(i)} \right) \right) \tag{4.43}$$

having cavity covariance matrix:

$$\left(\Sigma_U^{(i)} \right)^{-1} = \begin{cases} \sum_{j \in U \setminus \{i\}} \frac{1}{d_j} \boldsymbol{e}_j \boldsymbol{e}_j^\top + \mathbf{F}^\top \left(\sum_{j \in V} \frac{1}{d_j} \boldsymbol{e}_j \boldsymbol{e}_j^\top \right) \mathbf{F}, & \text{if } i \in U, \\ \sum_{j \in U} \frac{1}{d_j} \boldsymbol{e}_j \boldsymbol{e}_j^\top + \mathbf{F}^\top \left(\sum_{j \in V \setminus \{i\}} \frac{1}{d_j} \boldsymbol{e}_j \boldsymbol{e}_j^\top \right) \mathbf{F}, & \text{if } i \in V, \end{cases} \tag{4.44}$$

and cavity mean:

$$\bar{\boldsymbol{x}}^{(i)} = \begin{cases} \Sigma_U^{(i)} \left(\sum_{j \in U \setminus \{i\}} \frac{a_j}{d_j} \boldsymbol{e}_j + \sum_{j \in V} \frac{\tilde{y}_j + a_j}{d_j} \mathbf{F}^\top \boldsymbol{e}_j \right), & \text{if } i \in U, \\ \Sigma_U^{(i)} \left(\sum_{j \in U} \frac{a_j}{d_j} \boldsymbol{e}_j + \sum_{j \in V \setminus \{i\}} \frac{\tilde{y}_j + a_j}{d_j} \mathbf{F}^\top \boldsymbol{e}_j \right), & \text{if } i \in V. \end{cases} \tag{4.45}$$

The marginals of the cavity distribution in Eq. (4.43) are Gaussian distributions having means:

$$\bar{x}_i^{(i)} = \begin{cases} \bar{u}_i^{(i)}, & \text{if } i \in U \\ \tilde{y}_i + \boldsymbol{e}_i^\top \mathbf{F} \bar{\boldsymbol{u}}^{(i)}, & \text{if } i \in V, \end{cases} \tag{4.46}$$

and variances:

$$\Sigma_{ii}^{(i)} = \begin{cases} \boldsymbol{e}_i^\top \Sigma_U^{(i)} \boldsymbol{e}_i, & \text{if } i \in U, \\ \left(\boldsymbol{e}_i^\top \mathbf{F} \right) \Sigma_U^{(i)} \left(\mathbf{F}^\top \boldsymbol{e}_i \right), & \text{if } i \in V. \end{cases} \tag{4.47}$$

We are now able to determine the means $\boldsymbol{a}$ and variances $\boldsymbol{d}$ of the Gaussian approximating factors (4.35) by moment matching of the Gaussian approximation of the posterior distribution and of each tilted distribution for all $i = 1, \dots, N$:

$$\langle x_i \rangle_{Q^{(i)}} = \langle x_i \rangle_Q, \quad \langle x_i^2 \rangle_{Q^{(i)}} = \langle x_i^2 \rangle_Q, \tag{4.48}$$

which, in turn, allows us to obtain the EP update equations. Indeed, proceeding exactly as we did in Sec. 4.3, namely, using the product rule for Gaussian distributions and imposing the moment matching conditions, the EP update rules for the variances $\boldsymbol{d}$ and the means $\boldsymbol{a}$ are given by Eq. (4.32), which we rewrite here for the sake of completeness:

$$\begin{aligned} d_i &= \left(\frac{1}{\langle x_i^2 \rangle_{Q^{(i)}} - \langle x_i \rangle_{Q^{(i)}}^2} - \frac{1}{\Sigma_{ii}^{(i)}} \right)^{-1}, \\ a_i &= \langle x_i \rangle_{Q^{(i)}} + \frac{d_i}{\Sigma_{ii}^{(i)}} \left(\langle x_i \rangle_{Q^{(i)}} - \bar{x}_i^{(i)} \right), \end{aligned}$$

for all $i = 1, \dots, N$. Furthermore, the cavity variances $\Sigma_{ii}^{(i)}$ and means $\bar{x}_i^{(i)}$ appearing in the EP update equations can be computed in terms of the variances Σ_{ii} and means $\bar{x}_i$ using the low rank update rule given in Eq. (4.30), which allows to perform only one matrix inversion per iteration.

The main advantage of the formulation of EP presented in this section, which was obtained exploiting the linear relationship between the variables to be inferred, as compared to that of the previous one is the fact that one only needs to invert the $(N - M) \times (N - M)$ matrix (4.38), rather than a larger $N \times N$ matrix.

4.5 Sequential and parallel update schemes for Gaussian EP

In this section, we show two ways of implementing Gaussian EP: the first one involves a *sequential update scheme* which requires a matrix inversion for each variable $k = 1, \ldots, N$, whereas the second one uses the low rank update mentioned in Secs. 4.3 and 4.4, resulting in a *parallel update scheme*.

In Alg. 2, we show a pseudocode describing the sequential update scheme as applied to the formulation of Gaussian EP of Sec. 4.3 and notice that this version of Gaussian EP requires that N matrices of size $N \times N$ are inverted at each iteration. The number of operations needed to invert each matrix $(\mathbf{A} + \mathbf{D}^{(k)})$ is $O(N^3)$, so the computational cost of each iteration is $O(N^4)$.

Algorithm 2 Expectation Propagation: sequential update

procedure EP($\mathbf{A}$, $\boldsymbol{m}$, $\{\psi_1, ..., \psi_N\}$)
  Initialize $\boldsymbol{a}^{\text{old}}$ and $\boldsymbol{d}^{\text{old}}$
  for *iter* $<$ *maxiter* **do**
    $\boldsymbol{av} = \mathbf{0}$, $\Delta av = 0$
    for $k = 1, ..., N$ **do**
      $\mathbf{D}^{(k)} = \text{Diag}\left(\frac{1}{d_1} ... \frac{1}{d_N}\right) - \frac{1}{d_k}\boldsymbol{e}_k\boldsymbol{e}_k^T$
      $\boldsymbol{\Sigma}^{(k)} = \left(\mathbf{A} + \mathbf{D}^{(k)}\right)^{-1}$
      $\bar{\boldsymbol{x}}^{(k)} = \boldsymbol{\Sigma}^{(k)}(\boldsymbol{m} + \mathbf{D}^{(k)}\mathbf{A})$
      $\langle x_k \rangle_{Q^{(k)}}, \langle x_k^2 \rangle_{Q^{(k)}} = \textbf{moments}\left(\mu_k^{(k)}, \Sigma_{kk}^{(k)}, \psi_k\right)$
      $\Delta av \leftarrow \max(\Delta av, |\langle x_k \rangle_{Q^{(k)}} - av_k|)$
      $av_k \leftarrow \langle x_k \rangle_{Q^{(k)}}$
      $var_k \leftarrow \langle x_k^2 \rangle_{Q^{(k)}} - \langle x_k \rangle_{Q^{(k)}}^2$
      $d_k^{\text{new}} = \left(\frac{1}{var_k} - \frac{1}{\Sigma_{kk}^{(k)}}\right)^{-1}$
      $a_k^{\text{new}} = \langle x_k \rangle_{Q^{(k)}} + d_k\left(\frac{\langle x_k \rangle_{Q^{(k)}}}{\Sigma_{kk}^{(k)}} - \frac{\mu_k^{(k)}}{\Sigma_{kk}^{(k)}}\right)$
      $d_k^{\text{old}} \leftarrow \gamma d_k^{\text{old}} + (1-\gamma) d_k^{\text{new}}$
      $a_k^{\text{old}} \leftarrow \gamma a_k^{\text{old}} + (1-\gamma) a_k^{\text{new}}$
    if $\Delta av < \epsilon$ **then**
      return $\boldsymbol{av}$, $\boldsymbol{var}$

In order to reduce the computational complexity of the algorithm, a parallel update scheme, shown in Alg. 3, was proposed in Ref. [96], in which only one inversion per iteration is required. In both Algs. 2 and 3, *maxiter* denotes the maximum number of iterations allowed, the function **moments** computes the first and second moments of

the kth tilted distribution, γ is a damping factor used in the refinement step of the EP update, Δav denotes the maximal absolute difference between the tilted means at the current iteration *iter* and at the previous iteration and ϵ is the convergence threshold to which Δav is compared in order to decide whether convergence has been achieved.

A simple way to derive the relation between the variances Σ_{kk} of the Gaussian approximation $Q(\boldsymbol{x})$ and the associated cavity variances $\Sigma_{kk}^{(k)}$, for $k = 1, \ldots, N$, consists in using the Sherman-Morrison formula (see, e.g., [101]):

$$\left(\mathbf{C} + \boldsymbol{u}\boldsymbol{v}^\top\right)^{-1} = \mathbf{C}^{-1} - \frac{\mathbf{C}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\mathbf{C}^{-1}}{1 + \boldsymbol{v}^\top\mathbf{C}^{-1}\boldsymbol{u}}. \tag{4.49}$$

In particular, considering Eqs. (4.20) and (4.23) and recalling the definitions of $\mathbf{D}$ and $\mathbf{D}^{(k)}$, we have:

$$\boldsymbol{\Sigma}^{(k)} = \left(\boldsymbol{\Sigma}^{-1} - \frac{1}{d_k}\boldsymbol{e}_k\boldsymbol{e}_k^\top\right)^{-1}. \tag{4.50}$$

Therefore, identifying $\mathbf{C}$ with $\boldsymbol{\Sigma}^{-1}$ as well as, e.g., $\boldsymbol{u}$ with $d_k^{-1}\boldsymbol{e}_k$ and $\boldsymbol{v}$ with $\boldsymbol{e}_k$ in Eq. (4.49), we obtain:

$$\Sigma_{kk}^{(k)} = \frac{\Sigma_{kk}}{1 - \frac{1}{d_k}\Sigma_{kk}}$$

which is the second relation in Eq. (4.30), whereas the first one is obtained using the product rule for univariate Gaussian distributions given in Eq. (4.27), which in our case reads:

$$\frac{\mu_k}{\Sigma_{kk}} = \frac{\mu_k^{(k)}}{\Sigma_{kk}^{(k)}} + \frac{a_k}{d_k} \tag{4.51}$$

and substituting the above relation for $\Sigma_{kk}^{(k)}$.

4.6 Relationship to loopy belief propagation

It is interesting to notice that the loopy belief propagation algorithm discussed in Sec. 3.3.2 can be interpreted as a particular instance of expectation propagation, as reported in [26, 60, 63, 87, 102].

In order to see this, consider a factor graph $\mathcal{G} = (V, F, E)$ and the following exponential family of distributions defined on $\mathcal{G}$:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\,\mathrm{e}^{\boldsymbol{\theta}^\top\boldsymbol{T}(\boldsymbol{x})} = \frac{1}{Z(\boldsymbol{\theta})}\exp\left(\sum_{i\in V}\theta_i(x_i) + \sum_{a\in F}\theta_a(\boldsymbol{x}_a)\right), \tag{4.52}$$

Algorithm 3 Expectation Propagation: parallel update

procedure EP($\mathbf{A}$, $\boldsymbol{m}$, $\{\psi_1, ..., \psi_N\}$)
  Initialize $\boldsymbol{a}^{\text{old}}$ and $\boldsymbol{d}^{\text{old}}$
  for $iter = 1, ..., maxiter$ **do**
    $\boldsymbol{av} = \mathbf{0}$, $\Delta av = 0$
    $\boldsymbol{\Sigma} = (\mathbf{A} + \mathbf{D})^{-1}$
    $\bar{\boldsymbol{x}} = \boldsymbol{\Sigma}(\boldsymbol{m} + \mathbf{DA})$
    for $k = 1, ..., N$ **do**
      $\bar{x}_k^{(k)} = \frac{\bar{x}_k - \Sigma_{kk}\frac{a_k}{d_k}}{1 - \frac{\Sigma_{kk}}{d_k}}$
      $\Sigma_{kk}^{(k)} = \frac{\Sigma_{kk}}{1 - \frac{1}{d_k}\Sigma_{kk}}$
      $\langle x_k \rangle_{Q^{(k)}}, \langle x_k^2 \rangle_{Q^{(k)}} = \textbf{moments}\left(\mu_k^{(k)}, \Sigma_{kk}^{(k)}, \psi_k\right)$
      $\Delta av \leftarrow \max(\Delta av, |\langle x_k \rangle_{Q^{(k)}} - av_k|)$
      $av_k \leftarrow \langle x_k \rangle_{Q^{(k)}}$
      $var_k \leftarrow \langle x_k^2 \rangle_{Q^{(k)}} - \langle x_k \rangle_{Q^{(k)}}^2$
      $d_k^{\text{new}} = \left(\frac{1}{var_k} - \frac{1}{\Sigma_{kk}^{(k)}}\right)^{-1}$
      $a_k^{\text{new}} = \langle x_k \rangle_{Q^{(k)}} + d_k\left(\frac{\langle x_k \rangle_{Q^{(k)}}}{\Sigma_{kk}^{(k)}} - \frac{\mu_k^{(k)}}{\Sigma_{kk}^{(k)}}\right)$
      $d_k^{\text{old}} \leftarrow \gamma d_k^{\text{old}} + (1 - \gamma) d_k^{\text{new}}$
      $a_k^{\text{old}} \leftarrow \gamma a_k^{\text{old}} + (1 - \gamma) a_k^{\text{new}}$
    if $\Delta av < \epsilon$ **then**
      return $\boldsymbol{av}, \boldsymbol{var}$

where we have used the shorthand notations:

$$\theta_i(x_i) = \sum_{j \in \mathcal{X}} \theta_{i;j} \mathbb{I}(x_i = j), \tag{4.53}$$

$$\theta_a(\boldsymbol{x}_a) = \sum_{\boldsymbol{k} \in \mathcal{X}^{|\partial a|}} \theta_{a;\boldsymbol{k}} \mathbb{I}(\boldsymbol{x}_a = \boldsymbol{k}). \tag{4.54}$$

Here, the vector of sufficient statistics is given by the set of indicator functions:

$$\boldsymbol{T}(\boldsymbol{x}) = \left(\{\mathbb{I}(x_i = j)\}_{i \in V, j \in \mathcal{X}}, \{\mathbb{I}(\boldsymbol{x}_a = \boldsymbol{k})\}_{a \in F, \boldsymbol{k} \in \mathcal{X}^{|\partial a|}} \right), \tag{4.55}$$

and $\boldsymbol{\theta} = \left(\{\theta_{i;j}\}_{i \in V, j \in \mathcal{X}}, \{\theta_{a;\boldsymbol{k}}\}_{a \in F, \boldsymbol{k} \in \mathcal{X}^{|\partial a|}} \right)$ is the vector of natural parameters. For this choice of sufficient statistics, the mean parameters of the model are given by the variable marginals and by the factor marginals:

$$\langle \boldsymbol{T}(\boldsymbol{x}) \rangle = \left(\{p(x_i = j)\}_{i \in V}, \{p(\boldsymbol{x}_a = \boldsymbol{k})\}_{a \in F} \right). \tag{4.56}$$

The tractable part of the distribution is fully factorized over all the variables and reads:

$$G(\boldsymbol{x}) = \exp\left(\sum_{i \in V} \theta_i(x_i) \right), \tag{4.57}$$

whereas, concerning the intractable part, we shall approximate:

$$\psi_a(\boldsymbol{x}_a) = \exp\left(\theta_a(\boldsymbol{x}_a)\right) \tag{4.58}$$

by means of fully factorized approximating terms given by:

$$\phi_a(\boldsymbol{x}_a) \propto \prod_{i \in \partial a} h_{ai}(x_i). \tag{4.59}$$

Within the EP framework, this corresponds to identifying:

$$h_{ai}(x_i) = \exp(\lambda_{ai}(x_i)), \tag{4.60}$$

thereby approximating Eq. (4.52) by means of *another* exponential family of distributions:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \mathrm{e}^{\boldsymbol{\theta}^\top \boldsymbol{T}(\boldsymbol{x})} = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{i \in V} \theta_i(x_i) + \sum_{a \in F} \sum_{i \in \partial a} \lambda_{ai}(x_i) \right), \tag{4.61}$$

where all sufficient statistics are now given by *single variable* indicator functions and where we have used the shorthand notation:

$$\lambda_{ai}(x_i) = \sum_{j \in \mathcal{X}} \lambda^a_{i;j} \mathbb{I}(x_i = j). \tag{4.62}$$

By defining:

$$q_i(x_i) \propto \prod_{a \in \partial i} h_{ai}(x_i), \tag{4.63}$$

we thus obtain a fully factorized approximated distribution:

$$Q(\boldsymbol{x}) \propto G(\boldsymbol{x}) \prod_{a \in F} \prod_{i \in \partial a} h_{ai}(x_i) = \prod_{i \in V} \left(\exp\left(\theta_i(x_i)\right) \prod_{a \in \partial i} h_{ai}(x_i) \right) \propto \prod_{i \in V} q_i(x_i), \tag{4.64}$$

After performing the removal step, the cavity distribution reads:

$$Q^{\setminus a}(\boldsymbol{x}) = \frac{Q(\boldsymbol{x})}{\phi_a(\boldsymbol{x}_a)} \propto G(\boldsymbol{x}) \prod_{b \neq a} \prod_{i \in \partial b} h_{bi}(x_i) \propto \prod_{i \in \partial a} q_i^{\setminus a}(x_i), \tag{4.65}$$

where, given a fixed $a \in F$ and $i \in \partial a$, we have defined $q_i^{\setminus a}(x_i)$ as:

$$q_i^{\setminus a}(x_i) \propto \frac{q_i(x_i)}{h_{ai}(x_i)} \propto \prod_{b \in \partial i \setminus a} h_{bi}(x_i). \tag{4.66}$$

As a consequence, for the tilted distribution, we have that:

$$Q^{(a)}(\boldsymbol{x}) \propto \psi_a(\boldsymbol{x}_a) Q^{\setminus a}(\boldsymbol{x}) \propto \psi_a(\boldsymbol{x}_a) \prod_{i \in \partial a} q_i^{\setminus a}(x_i). \tag{4.67}$$

While in Gaussian EP the projection step consists in the matching of the first and second moments of $Q(\boldsymbol{x})$ and of $Q^{(a)}(\boldsymbol{x})$, for the sufficient statistics given in Eq. (4.55) we have a condition on the matching of their marginals:

$$Q_i(x_i) = Q_i^{(a)}(x_i). \tag{4.68}$$

On the one hand, if the variable node i is not connected to the factor node a, then the matching condition (4.68) leaves the approximating factors $h_{bi}(x_i)$ (with $b \in \partial i$) unchanged. On the other hand, if $i \in \partial a$, then the marginal of the tilted distribution can be expressed as:

$$Q_i^{(a)}(x_i) = \sum_{\boldsymbol{x}_{\setminus i}} Q^{(a)}(\boldsymbol{x}) \propto \sum_{\boldsymbol{x}_{\partial a \setminus i}} \psi_a(\boldsymbol{x}_a) \prod_{j \in \partial a} q_j^{\setminus a}(x_j) =: q_i^{(a)}(x_i) \tag{4.69}$$

and the approximating factors $\phi_a(\boldsymbol{x}_a)$ are updated at the refine step by computing the new terms $h_{ai}(x_i)$ as follows:

$$h_{ai}(x_i) \propto \frac{q_i(x_i)}{q_i^{\setminus a}(x_i)} = \frac{q_i^{(a)}(x_i)}{q_i^{\setminus a}(x_i)} = \sum_{\boldsymbol{x}_{\partial a \setminus i}} \psi_a(\boldsymbol{x}_a) \prod_{j \in \partial a \setminus i} q_j^{\setminus a}(x_j), \tag{4.70}$$

where we took advantage of the projection step.

Equations (4.66) and (4.70) coincide with the belief propagation equations. In particular, from Eq. (4.66), one sees that the $\{q_i^{\setminus a}(x_i)\}_{a\in F}$ play the role of variable to factor messages, whereas, from Eq. (4.70) the factors $h_{ai}(x_i)$ can be interpreted as factor to variable messages. Finally, we see that each $q_i(x_i)$ corresponds to the marginal belief of the variable x_i:

$$q_i(x_i) \propto q_i^{\setminus a}(x_i) h_{ai}(x_i) = \prod_{b\in\partial i} h_{bi}(x_i), \tag{4.71}$$

namely, the product of all incoming messages $h_{bi}(x_i)$.

4.7 Relationship to adaTAP

Gaussian EP is equivalent to the adaTAP approach described in Sec. 3.5 [103], as will now be made precise. Consider the marginal distribution $P_i(x_i)$ given in Eq. (3.50) and let us replace the factor $\psi_i(x_i)$ with a Gaussian $\phi_i(x_i) \propto \exp\left(-\frac{1}{2}\Lambda_i x_i^2 + \gamma_i x_i\right)$:

$$Q_i(x_i) = \frac{1}{Z_Q}\phi_i(x_i) \exp\left[x_i\left(\langle h_i\rangle_{\setminus i} + B_i\right) + \frac{V_i}{2}{x_i}^2\right], \tag{4.72}$$

the parameters Λ_i and γ_i of which are chosen so that the first two moments of $P_i(x_i)$ and of $Q_i(x_i)$ are matched. In other words, we require that the following equalities hold:

$$\langle x_i\rangle_{P_i} = \frac{\gamma_i + \langle h_i\rangle_{\setminus i}}{\Lambda_i - V_i}, \tag{4.73}$$

$$\langle x_i^2\rangle_{P_i} - \langle x_i\rangle_{P_i}^2 = \frac{1}{\Lambda_i - V_i}, \tag{4.74}$$

where the quantities on the right hand side are the mean and the variance of $Q_i(x_i)$ expressed in terms of the cavity parameters and of the parameters of $\phi_i(x_i)$. However, the moments of $Q_i(x_i)$ can also be directly obtained starting from the multivariate Gaussian approximation $Q(\boldsymbol{x})$:

$$Q(\boldsymbol{x}) \propto \exp\left(\frac{1}{2}\boldsymbol{x}^\top \mathbf{J}\boldsymbol{x}\right)\prod_i \phi_i(x_i), \tag{4.75}$$

resulting in:

$$\langle x_i\rangle_{Q_i} = \left((\boldsymbol{\Lambda} - \mathbf{J})^{-1}\boldsymbol{\gamma}\right)_i, \tag{4.76}$$

$$\langle x_i^2\rangle_{Q_i} - \langle x_i\rangle_{Q_i}^2 = \left((\boldsymbol{\Lambda} - \mathbf{J})^{-1}\right)_{ii}. \tag{4.77}$$

In particular, from the relations for the variance in Eqs. (4.74) and (4.77) one recovers the adaTAP relation (3.66).

4.8 Vector approximate message passing as a special case of Gaussian EP

The vector approximate message passing (VAMP) algorithm presented in Sec. 3.7 can be seen as a special case of Gaussian EP with univariate Gaussian approximating factors too. In this case, the main difference is that in VAMP the approximating Gaussian factors have the same variance. Thus, the exponential family of distributions considered in VAMP is that of isotropic Gaussian distributions. In order to highlight the relationship to Gaussian EP, we shall consider the LMMSE formulation of VAMP as presented in Algorithm 1, where all but one factor nodes are connected to only one variable. More precisely, we have $P(\boldsymbol{x}) \propto \psi_0(\boldsymbol{x}) \prod_{a=1}^N \psi_a(x_a)$ or, equivalently, $P(\boldsymbol{x}, \boldsymbol{x}') = \psi_0(\boldsymbol{x}')\delta(\boldsymbol{x} - \boldsymbol{x}') \prod_{a=1}^N \psi_a(x_a)$ for the intractable distribution.

In VAMP, each marginalized tilted distribution, for $l = 1, \dots, N$, has mean given by the output of the MMSE scalar denoiser $g_1((r_{1k})_l, \gamma_{1k})$, namely:

$$(\hat{x}_{1k})_l = \int x\psi(x)\mathcal{N}\big(x; (r_{1k})_l, \gamma_{1k}^{-1}\big)\,\mathrm{d}x, \tag{4.78}$$

where $\boldsymbol{r}_{1k}$ and γ_{1k} play the role of cavity means and cavity precisions and are associated with the projection of the multivariate factor $\psi_0(\boldsymbol{x})$ onto an isotropic Gaussian distribution $\mathcal{N}(\boldsymbol{x}; \boldsymbol{r}_{1k}, \gamma_{1k}^{-1}\mathbf{I})$. At iteration k, the resulting tilted distributions are also projected onto a single isotropic Gaussian distribution $\mathcal{N}(\mathbf{x}; \hat{\boldsymbol{x}}_1, \eta_{1k}^{-1}\mathbf{I})$. The Gaussian approximation of the product $\prod_{a=1}^N \psi_a(x_a)$ appearing in the intractable distribution $P(\boldsymbol{x})$ is computed from the projected tilted distribution $\mathcal{N}(\mathbf{x}; \hat{\boldsymbol{x}}_1, \eta_{1k}^{-1}\mathbf{I})$ and from $\mathcal{N}(\boldsymbol{x}; \boldsymbol{r}_{1k}, \gamma_{1k}^{-1}\mathbf{I})$ by means of the quotient rule, yielding the isotropic Gaussian distribution $\mathcal{N}(\mathbf{x}; \mathbf{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I})$. Therefore, in VAMP, the EP-like approximating Gaussian factors $\phi(x_i; a_i, b_i)$ have mean $a_i = r_{2k,i}$ and common precision $b_i^{-1} = \gamma_{2k}$. The fully approximate Gaussian approximation of the posterior distribution, which in EP was denoted as $Q(\boldsymbol{x})$, is obtained by first combining the multivariate Gaussian factor $\psi_0(\boldsymbol{x})$ and the approximation $\mathcal{N}(\mathbf{x}; \mathbf{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I})$ of the product $\prod_{a=1}^N \psi_a(x_a)$ using the product rule for Gaussian distributions and then by projecting back onto the family of isotropic Gaussian distributions. The result of the projection is used to recompute the approximation of $\psi_0(\boldsymbol{x})$ at iteration $k+1$ by dividing by $\mathcal{N}(\mathbf{x}; \mathbf{r}_{2k}, \gamma_{2k}^{-1}\mathbf{I})$ by means of the quotient rule. Notice that, contrary to VAMP, the multivariate factor $\psi_0(\boldsymbol{x})$ is treated without further approximations in Gaussian EP. The steps are then repeated until a predefined maximum number of iterations is reached. The VAMP quantities are compared to the associated Gaussian EP quantities in Table 4.1.

The EP moment matching conditions are not explicitly imposed but, at the fixed point of the algorithm, they are implied by self-consistency, as the two Gaussian distributions $\mathcal{N}(\boldsymbol{x}_1; \hat{\boldsymbol{x}}_{1k}, \eta_{1k}^{-1}\mathbf{I})$ and $\mathcal{N}(\boldsymbol{x}_2; \hat{\boldsymbol{x}}_{2k}, \eta_{2k}^{-1}\mathbf{I})$ that are updated at each iteration both aim at approximating the intractable distribution $P(\boldsymbol{x})$. As a consequence, at the fixed point, one must have that $\boldsymbol{r}_{1k} = \boldsymbol{r}_{2k}$ and that $\eta_{1k} = \eta_{2k}$.

Quantity	Expression in VAMP	Expression in Gaussian EP
Tilted distribution	$\mathcal{N}(\boldsymbol{x}_1; \hat{\boldsymbol{x}}_{1k}, \eta_{1k}^{-1}\mathbf{I})$	$\mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}^{(l)}, \boldsymbol{\Sigma}^{(l)})$
Base distribution $\psi_0(\boldsymbol{x})$	$\mathcal{N}(\boldsymbol{x}_1; \boldsymbol{r}_{1k}, \gamma_{1k}^{-1})$	$G(\boldsymbol{x})$
Approximating factors	$\mathcal{N}(\boldsymbol{x}_2; \hat{\boldsymbol{x}}_{2k}, \gamma_{2k}^{-1}\mathbf{I})$	$\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi d_n}} \exp\left(-\frac{(x-a_n)^2}{2d_n}\right)$
Full Gaussian approximation	$\mathcal{N}(\boldsymbol{x}_1; \hat{\boldsymbol{r}}_{2k}, \eta_{2k}^{-1}\mathbf{I})$	$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

Table 4.1: Relationship of the VAMP quantities to the Gaussian EP related ones.

4.9 Expectation propagation as a variational problem

Expectation propagation can be viewed as a variational problem where the approximating factors are the minimizers of a variational free energy associated with the algorithm. In this section, we will introduce the EP free energy having in mind Gaussian EP and show that the fixed points of the algorithm are stationary points of this variational free energy.

4.9.1 EP free energy

In order to obtain the variational free energy of Gaussian EP, we can proceed by analogy with the case of loopy belief propagation (A. Braunstein, personal communication, 2018). Indeed, let us consider the distribution:

$$P(\boldsymbol{x}) = \frac{1}{Z} \prod_a \psi_a(\boldsymbol{x}_a), \tag{4.79}$$

and recall the definitions of its Gaussian EP approximation:

$$Q(\boldsymbol{x}) = \frac{1}{Z_Q} \prod_a \phi_a(\boldsymbol{x}_a) \tag{4.80}$$

and of the tilted distributions:

$$Q^{(a)}(\boldsymbol{x}) = \frac{1}{Z_Q} \psi_a(\boldsymbol{x}_a) \prod_{b \neq a} \phi_b(\boldsymbol{x}_b), \tag{4.81}$$

for $a = 1, \ldots, N$.

As anticipated above, let us first consider the case of loopy belief propagation. It can be proved that the following identity for the intractable distribution $P(\boldsymbol{x})$ holds:

$$P(\boldsymbol{x}) = \alpha_{BP} \hat{P}(\boldsymbol{x}), \tag{4.82}$$

where the proportionality constant reads $\alpha_{BP} := Z_{\text{Bethe}}/Z$, with Z_{Bethe} given by Eq. (3.30), and the distribution $\hat{P}(\boldsymbol{x})$ is a Bethe factorization in terms of the marginal beliefs

(3.20) and (3.21):

$$\hat{P}(\boldsymbol{x}) \coloneqq \prod_a \frac{b_a(\boldsymbol{x}_a)}{\prod_{i\in\partial a} b_i(x_i)} \prod_i b_i(x_i). \tag{4.83}$$

Note, that in the case of a tree, the identity (4.82) holds with $\alpha_{BP} = 1$ by the junction tree theorem [26].

Considering now EP, we define the factorization:

$$\hat{P}(\boldsymbol{x}) \coloneqq Q(\boldsymbol{x}) \prod_a \frac{Q^{(a)}(\boldsymbol{x})}{Q(\boldsymbol{x})}, \tag{4.84}$$

and we may assume that:

$$P(\boldsymbol{x}) = \alpha_{EP}\hat{P}(\boldsymbol{x}), \tag{4.85}$$

where we identify $\alpha_{EP} \coloneqq \frac{Z_{EP}}{Z}$. Indeed, in the Gaussian case, all $\psi_a(\boldsymbol{x}_a)$ are Gaussian factors:

$$\psi_a(\boldsymbol{x}_a) = \phi_a(\boldsymbol{x}_a), \tag{4.86}$$

so the equalities $P(\boldsymbol{x}) = \hat{P}(\boldsymbol{x})$ and $\alpha_{EP} = 1$ are identically satisfied. On the other hand, if the priors are not Gaussian, we can express the constant α_{EP} in terms of the partition functions (or, equivalently, just take the expression of $\hat{P}(\boldsymbol{x})$, rearrange it and isolate Z_{EP} by comparison with the assumption (4.85)):

$$\begin{aligned} 1 &= \int \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) = \alpha_{EP} \int \mathrm{d}\boldsymbol{x} \hat{P}(\boldsymbol{x}) = \alpha_{EP} \int \mathrm{d}\boldsymbol{x} Q(\boldsymbol{x}) \prod_a \frac{Q^{(a)}(\boldsymbol{x})}{Q(\boldsymbol{x})} \\ &= \alpha_{EP} \int \mathrm{d}\boldsymbol{x} Q(\boldsymbol{x}) \prod_a \frac{\frac{\psi_a(\boldsymbol{x}_a)}{Z_{Q^{(a)}}}}{\frac{\phi_a(\boldsymbol{x}_a)}{Z_Q}} = \alpha_{EP} \frac{1}{Z_Q} \left(\prod_a \frac{Z_Q}{Z_{Q^{(a)}}} \right) \left(\int \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) \right) Z \\ &= \alpha_{EP} Z_Q^{N-1} \left(\prod_a Z_{Q^{(a)}}^{-1} \right) Z \end{aligned} \tag{4.87}$$

Therefore, we have that:

$$\alpha_{EP} = \frac{1}{Z} \left(Z_Q^{1-N} \prod_a Z_{Q^{(a)}} \right), \tag{4.88}$$

which, by the definition of α_{EP}, implies:

$$Z_{EP} = Z_Q^{1-N} \prod_a Z_{Q^{(a)}}. \tag{4.89}$$

Finally, we obtain that the associated EP free energy is expressed as:

$$F_{EP} = -\ln Z_{EP} = (N-1) \ln Z_Q - \sum_a \ln \left(Z_{Q^{(a)}} \right). \tag{4.90}$$

4.9.2 EP free energy and EP fixed points

The fixed points $(\boldsymbol{a}^*, \boldsymbol{d}^*)$ of Gaussian EP, which fulfill the moment matching conditions (4.24) for all $n = 1, \dots, N$, are stationary points of the EP free energy (4.90), i.e., they satisfy the conditions:

$$0 = \left.\frac{\partial F_{EP}}{\partial a_n}\right|_{\boldsymbol{a}^*,\boldsymbol{d}^*} = \left[(N-1)\langle w_n\rangle_Q - \sum_{l\neq n}\langle w_n\rangle_{Q^{(l)}}\right]_{\boldsymbol{a}^*,\boldsymbol{d}^*}, \tag{4.91}$$

$$0 = \left.\frac{\partial F_{EP}}{\partial d_n}\right|_{\boldsymbol{a}^*,\boldsymbol{d}^*} = \left[(N-1)\langle w_n^2\rangle_Q - \sum_{l\neq n}\langle w_n^2\rangle_{Q^{(l)}}\right]_{\boldsymbol{a}^*,\boldsymbol{d}^*}, \tag{4.92}$$

for $n = 1, \dots, N$.

In order to show this, we shall prove that the moment matching conditions imply that (4.91) and (4.92) are satisfied. This proof is published in Ref. [98].

We start from the expressions of the tilted moments $\langle w_n\rangle_{Q^{(l)}}$ and $\langle w_n^2\rangle_{Q^{(l)}}$ and write them as:

$$\begin{aligned}\langle w_n^\alpha\rangle_{Q^{(l)}} &= \int \frac{Z_Q}{Z_{Q^{(l)}}} Q(\boldsymbol{w})\frac{\psi_l(w_l)}{\phi_l(w_l)} w_n^\alpha \mathrm{d}\boldsymbol{w} = \int Q(\boldsymbol{w})\frac{Q^{(l)}(w_l)}{Q(w_l)} w_n^\alpha \mathrm{d}\boldsymbol{w} \\ &= \left\langle \int_{-\infty}^{+\infty} \frac{Q(w_n, w_l)}{Q(w_l)} w_n^\alpha \mathrm{d}w_n \right\rangle_{Q^{(l)}(w_l)}, \quad \alpha = 1,2,\end{aligned}$$

where, in the last equality, for $\alpha = 1$ (resp., $\alpha = 2$), the integral that appears in the average with respect to $Q^{(l)}(w_l)$ is the first (resp., second) moment of w_n, conditioned on w_l and computed with respect to Q. These moments depend on w_l through the mean (resp., squared mean) of $Q(w_n|w_l)$. In turn, the dependence of such mean on w_l is linear, implying that $\langle w_n\rangle_{Q(w_n|w_l)}$ and $\langle w_n^2\rangle_{Q(w_n|w_l)}$ depend on w_l linearly and quadratically, respectively. As a consequence, by the moment matching conditions, we have that:

$$\left\langle \int_{-\infty}^{+\infty} \frac{Q(w_n, w_l)}{Q(w_l)} w_n^\alpha \mathrm{d}w_n \right\rangle_{Q^{(l)}(w_l)} = \left\langle \int_{-\infty}^{+\infty} \frac{Q(w_n, w_l)}{Q(w_l)} w_n^\alpha \mathrm{d}w_n \right\rangle_{Q(w_l)}, \tag{4.93}$$

for $\alpha = 1,2$, implying that $\langle w_n^\alpha\rangle_{Q^{(l)}} = \langle w_n^\alpha\rangle_Q$ and, therefore, that the conditions (4.91) and (4.92) are identically true.

4.10 Learning the parameters of the priors within the EP framework

Let us consider a Bayesian inference problem and let $\boldsymbol{\zeta}$ be the set of parameters of a prior distribution $P(\boldsymbol{x}|\boldsymbol{\zeta}) = \prod_a \psi_a(x_a|\boldsymbol{\zeta}_a)$. For example, the parameters $\boldsymbol{\zeta}_a$ could be the density ρ and the precision λ of the spike-and-slab prior introduced in Eq. (2.35), as it will be

the case in Chapters 5 and 6. Moreover, let us assume that $\boldsymbol{y}$ is a set of observations, from the knowledge of which we would like to infer a set of hidden variables $\boldsymbol{x}$. We can estimate the parameters $\boldsymbol{\zeta}$ of the prior by maximimum likelihood, where the likelihood function reads:

$$P(\boldsymbol{y}|\boldsymbol{\zeta}) = \int \mathrm{d}\boldsymbol{x} P(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\zeta}) = \int \mathrm{d}\boldsymbol{x} P(\boldsymbol{y}|\boldsymbol{x}) P(\boldsymbol{x}|\boldsymbol{\zeta}) = Z(\boldsymbol{\zeta}), \tag{4.94}$$

which is nothing but the intractable normalization of the posterior distribution in Eq. (4.7). Equivalently, one can associate a free energy to the partition function (4.94) by means of the definition $F = -\ln Z(\boldsymbol{\zeta})$. However, since this is intractable too, it is desirable to consider an approximate free energy to be optimized.

One way to approximate $F(\boldsymbol{\zeta})$ is by means of the EP free energy. Indeed, at the fixed point of the EP algorithm, $F(\boldsymbol{\zeta})$ can be approximated by Eq. (4.90), to be minimized F_{EP} via gradient descent:

$$\zeta_j^{(t+1)} = \zeta_j^{(t)} - \delta\zeta_j \frac{\partial F_{EP}}{\partial \zeta_j}, \tag{4.95}$$

where t denotes the current iteration, ζ_j denotes the j-th component of the parameter vector $\boldsymbol{\zeta}$ and $\delta\zeta_j$ is its corresponding learning rate.

Notice that, although the only contributions to F_{EP} depending explicitly on the parameters $\boldsymbol{\zeta}$ of the prior are the terms $F_{Q^{(a)}}$, the components of the gradient should, in principle, include other terms as well. However, since these contributions depend on the derivatives of the free energy with respect to the cavity parameters, which vanish at the EP fixed point, we can neglect those terms.

The EP inference algorithm can be combined with the estimation of the parameters of the prior by iteratively alternating an EP update step at fixed $\boldsymbol{\zeta}$ and an update of $\boldsymbol{\zeta}$ performed via gradient descent at fixed EP parameters. This is completely analogous to an expectation maximization (EM) scheme, where the optimization over the EP parameters corresponds to the expectation step (*E-step*) and the minimization of F_{EP} with respect to $\boldsymbol{\zeta}$ corresponds to the maximization step (*M-step*). Notice that a "standard" EM procedure could also be performed. In this case, one can alternate a complete EP estimation of the approximating posterior distributions at fixed prior parameters until convergence is reached (E-step) and a maximum likelihood update of the prior parameters (M-step). The fact that we employ an alternating minimization procedure of this kind justifies the fact that we only consider the explicit dependence of F_{EP} on the prior parameters.

Chapter 5

Expectation propagation on compressed sensing

As discussed in Sec. 2.1, the compressed sensing (CS) problem involves finding the K-dimensional support set of an N-dimensional sparse signal, with $K \ll N$, which is, in principle, an NP complete problem. As a consequence, a number of approximate techniques have been developed in order to find the non-zero entries of the signal and estimate their values. These methods are based on convex relaxation (e.g., LASSO), greedy approaches (e.g., matching pursuit) or Bayesian inference (e.g., message passing algorithms). In this Chapter, we present a Gaussian EP based scheme for the CS problem. After formulating the problem within a Bayesian framework and describing how Gaussian EP can be applied to it, we will present empirical results about the performance of the method as compared to other state-of-the-art techniques from statistical physics and signal processing: in particular, we will first consider CS reconstruction with random i.i.d. sensing matrices and then focus on a simple case of correlated random matrices. The work presented in this Chapter is published in Ref. [98] and partly develops some preliminary results appearing in Anna Paola Muntoni's PhD thesis [104]. The implementation of Gaussian EP used to obtain the results shown in this chapter can be found in Ref. [105].

5.1 Bayesian framework for the CS problem and EP approximation of the posterior distribution

We will now introduce the Bayesian setup that will be used to study the CS problem. Let us denote the unknown vector (or *signal*) to be retrieved as $\boldsymbol{w} \in \mathbb{R}^N$ and let $\mathbf{F} \in \mathbb{R}^{M \times N}$ be a matrix with maximum rank, called *sensing* or *measurement* matrix in the context of CS. We consider the standard linear estimation problem:

$$\boldsymbol{y} = \mathbf{F}\boldsymbol{w} + \boldsymbol{n}, \tag{5.1}$$

where $\boldsymbol{y} \in \mathbb{R}^M$ is a set of noisy measurements, $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}; 0, \beta^{-1}\mathbf{I})$ is additive white noise with variance $1/\beta$ and we are interested in the *undersampling*[1] regime $M < N$. Notice that Eq. (5.1) can be recast as the problem of minimizing an energy function:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w}), \tag{5.2}$$

where $E(\boldsymbol{w})$ is the quadratic form given by:

$$E(\boldsymbol{w}) \coloneqq \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{F}\boldsymbol{w}\right\|_2^2 = \frac{1}{2}(\boldsymbol{y} - \mathbf{F}\boldsymbol{w})^T(\boldsymbol{y} - \mathbf{F}\boldsymbol{w}). \tag{5.3}$$

In this sense, the problem is akin to that of finding the ground state of a system of interacting particles, where the interactions are determined by the matrix $\mathbf{F}$.

5.1.1 Case of nonrigid linear constraints

In a Bayesian framework, we can enforce the linear constraints in Eq. (5.1) by defining the *likelihood*:

$$L(\boldsymbol{w}) \coloneqq P(\boldsymbol{y}|\boldsymbol{w}) = \left(\frac{\beta}{2\pi}\right)^{\frac{M}{2}} \mathrm{e}^{-\beta E(\boldsymbol{w})}, \tag{5.4}$$

which, for fixed $\boldsymbol{w}$ and given $\mathbf{F}$, can be interpreted as the probability of observing an additive white noise vector $\boldsymbol{n} = \boldsymbol{y} - \mathbf{F}\boldsymbol{w}$ distributed according to $\mathcal{N}(\boldsymbol{n}; 0, \beta^{-1}\mathbf{I})$. The value of β determines the extent to which the linear constraints can be violated and is set to a very large value (e.g., $\beta \sim 10^9$) when dealing with the problem of CS reconstruction from noiseless measurements. The extra constraints on the entries of $\boldsymbol{w}$ will be imposed in terms of a *prior* distribution, which we assume to be factorized:

$$P(\boldsymbol{w}) = \prod_{i=1}^{N} \psi_i(w_i).$$

In particular, we choose factors $\psi_i(w_i)$ of the spike-and-slab kind [31] in order to promote sparsity of the solution sought:

$$\psi(w_i) = (1-\rho)\delta(w_i) + \rho\sqrt{\frac{2\pi}{\lambda}}\,\mathrm{e}^{-\frac{1}{2}\lambda w_i^2}, \tag{5.5}$$

Therefore, we have for the posterior distribution of the signal $\boldsymbol{w}$:

$$\begin{aligned} P(\boldsymbol{w}|\mathbf{F}, \boldsymbol{y}) &= \frac{P(\boldsymbol{y}|\mathbf{F}, \boldsymbol{w})\,P(\boldsymbol{w})}{P(\boldsymbol{y})} \\ &= \frac{1}{Z_P}\,\mathrm{e}^{-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})^T(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})} \prod_{i=1}^{N}\left[(1-\rho)\delta(w_i) + \rho\sqrt{\frac{2\pi}{\lambda}}\,\mathrm{e}^{-\frac{1}{2}\lambda w_i^2}\right]. \end{aligned} \tag{5.6}$$

[1] Notice that the *overdetermined* regime $M \geq N$ can also be addressed using the same computational framework and, in particular, using Gaussian EP. In the noiseless case, the system can be solved exactly using Gaussian elimination.

We are interested in determining the MMSE estimate $\hat{\boldsymbol{w}}$ of the signal $\boldsymbol{w}$. Since the posterior probability in equation (5.6) is intractable, we can approximate it using Gaussian EP.

In order to compute the approximated distribution $Q(\boldsymbol{w})$, we replace the intractable factors $\psi_i(w_i)$ with univariate Gaussian distributions $\phi_i(w_i) = \mathcal{N}(w_i; a_i, d_i)$, for $i = 1, \dots, N$:

$$Q(\boldsymbol{w}|\mathbf{F}, \boldsymbol{y}) = \frac{1}{Z_Q} \mathrm{e}^{-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})^\top(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})} \prod_{i=1}^{N} \phi_i(w_i),$$

and rewrite the exponent:

$$\begin{aligned} &\beta(\boldsymbol{y} - \mathbf{F}\boldsymbol{w})^\top(\boldsymbol{y} - \mathbf{F}\boldsymbol{w}) + (\boldsymbol{w} - \boldsymbol{a})^\top\mathbf{D}(\boldsymbol{w} - \boldsymbol{a}) = \\ &= \boldsymbol{w}^\top(\beta\mathbf{F}^\top\mathbf{F} + \mathbf{D})\boldsymbol{w} - 2\boldsymbol{w}^\top(\beta\mathbf{F}^\top\boldsymbol{y} + \mathbf{D}\boldsymbol{a}) + \text{cst}, \end{aligned} \tag{5.7}$$

where $\mathbf{D}$ is a diagonal matrix having diagonal elements $d_1^{-1}, \dots, d_N^{-1}$. Then, defining the covariance matrix

$$\boldsymbol{\Sigma}^{-1} := \beta\mathbf{F}^T\mathbf{F} + \boldsymbol{D}, \tag{5.8}$$

and the mean vector:

$$\bar{\boldsymbol{w}} := \boldsymbol{\Sigma}(\beta\mathbf{F}^T\boldsymbol{y} + \boldsymbol{D}\boldsymbol{a}), \tag{5.9}$$

we obtain a multivariate Gaussian distribution with this mean and covariance matrix after completing the square in Eq. (5.7):

$$Q(\boldsymbol{w}) := \frac{1}{Z_Q} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\bar{\boldsymbol{w}})}, \tag{5.10}$$

where the normalization constant is given by $\boldsymbol{\Sigma}$:

$$Z_Q = (2\pi)^{\frac{N}{2}} (\det \boldsymbol{\Sigma})^{\frac{1}{2}}. \tag{5.11}$$

Analogously, the tilted distributions $Q^{(i)}(\boldsymbol{w})$ are expressed as:

$$Q^{(i)}(\boldsymbol{w}) := \frac{1}{Z_{Q^{(i)}}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}})^T\left(\boldsymbol{\Sigma}^{(i)}\right)^{-1}(\boldsymbol{w}-\bar{\boldsymbol{w}})} \left[(1-\rho)\delta(w_i) + \rho\sqrt{\frac{2\pi}{\lambda}} \mathrm{e}^{-\frac{1}{2}\lambda w_i^2}\right], \tag{5.12}$$

where

$$\left(\boldsymbol{\Sigma}^{(i)}\right)^{-1} := \beta\mathbf{F}^T\mathbf{F} + \boldsymbol{D}^{(i)}, \quad \bar{\boldsymbol{w}}^{(i)} := \boldsymbol{\Sigma}^{(i)}(\beta\mathbf{F}^T\boldsymbol{y} + \boldsymbol{D}\boldsymbol{a}). \tag{5.13}$$

Given these definitions, one can apply the Gaussian EP scheme to the CS problem by iterating the EP moment matching conditions as described in Section 4.3.

5.1.2 Limit of rigid linear constraints

We can take the limit $\beta \to \infty$ of Eq. (5.6) in order to take advantage of the formulation of Gaussian EP described in Sec. 4.4. In this limit, the likelihood (5.4) becomes a Dirac delta factor, $\delta(\mathbf{F}\boldsymbol{w} - \boldsymbol{y})$, which enforces the linear constraints exactly. As we have N variables and M constraints, we can separate the signal $\boldsymbol{w}$ into two subvectors, one of which is made of $N - M$ variables, whereas the other one is made of the remaining M variables. We shall call the former $\boldsymbol{u}$ and the latter $\boldsymbol{v}$, where $\boldsymbol{v}$ depends on all components of the subvector $\boldsymbol{u}$. Without loss of generality, we can take, for instance, $\boldsymbol{u} = (w_1, \dots, w_{N-M})^\top$ and $\boldsymbol{v} = (w_{N-M+1}, \dots, w_N)^\top$. Then, by performing elementary row operations, the linear system of equations:

$$\boldsymbol{y} = \mathbf{F}\boldsymbol{w}$$

can be rewritten as:

$$\tilde{\boldsymbol{y}} = (-\mathbf{X}|\ \mathbf{I})\,\boldsymbol{w},$$

where $\mathbf{I}$ is the $M \times M$ identity matrix, or, equivalently, as:

$$\boldsymbol{v} = \mathbf{X}\boldsymbol{u} + \tilde{\boldsymbol{y}},$$

where we have isolated the dependent variables and expressed them as a function of the independent ones. As a consequence, the posterior distribution of $\boldsymbol{w}$ can be written as:

$$\begin{aligned} P(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{Z}\delta(\boldsymbol{v} - \mathbf{X}\boldsymbol{u} - \tilde{\boldsymbol{y}}) \prod_{i=1}^{N-M} \left[(1-\rho)\delta(u_i) + \rho\sqrt{\frac{2\pi}{\lambda}}\,\mathrm{e}^{-\frac{1}{2}\lambda u_i^2}\right] \times \\ \times \prod_{i=1}^{M} \left[(1-\rho)\delta(v_i) + \rho\sqrt{\frac{2\pi}{\lambda}}\,\mathrm{e}^{-\frac{1}{2}\lambda v_i^2}\right], \end{aligned} \tag{5.14}$$

which can be approximated using Gaussian EP as detailed in Sec. 4.4.

5.2 Moments of the tilted distributions in the CS problem

In this section, we give the expressions of the moments of the tilted distributions given by Eq. (5.12), which are needed when imposing the moment matching conditions in Gaussian EP.

5.2.1 Tilted moments with a spike-and-slab prior

Starting from the i-th marginal of the tilted distribution (4.22):

$$Q^{(i)}(w_i) = \frac{1}{Z_{Q^{(i)}}} Q^{\setminus i}(w_i)\psi_i(w_i) \tag{5.15}$$

where, given the cavity distribution:

$$Q^{\setminus i}(w_i) = \frac{1}{\sqrt{2\pi\Sigma_i'}} \mathrm{e}^{-\frac{(w_i - \bar{w}_i')^2}{2\Sigma_i'}}, \tag{5.16}$$

we have simplified the notation by replacing $\left(\bar{w}_{(i)}\right)_n$ with $\bar{w}_i'$ and $\left(\Sigma_{(i)}\right)_{i,i}$ with Σ_i'. After inserting the definition of the spike-and-slab prior prior, Eq. (5.5), in Eq. (5.15), the tilted partition function appearing in Eq. (5.15) reads:

$$Z_{Q^{(i)}} = (1-\rho)\frac{1}{\sqrt{2\pi\Sigma_i'}} \mathrm{e}^{-\frac{\bar{w}_i'^2}{2\Sigma_i'}} + \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma_i')}} \mathrm{e}^{-\frac{1}{2}\frac{\bar{w}_i'^2}{\lambda+\Sigma_i'}}, \tag{5.17}$$

and the first and second tilted moments of w_n yield the expressions:

$$\langle w_i \rangle_{Q^{(i)}} = \frac{1}{Z_{Q^{(i)}}} \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma_{(i)}')}} \frac{\lambda\bar{w}_i'}{\lambda+\Sigma_i'} \mathrm{e}^{-\frac{1}{2}\frac{\bar{w}_i'^2}{\lambda+\Sigma_i'}}, \tag{5.18}$$

$$\langle w_n^2 \rangle_{Q^{(n)}} = \frac{1}{Z_{Q^{(n)}}} \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma_n')}} \left(\frac{\lambda\Sigma_n'(\lambda+\Sigma_n') + \lambda^2\bar{w}_n'^2}{(\lambda+\Sigma_n')^2} \right) \mathrm{e}^{-\frac{1}{2}\frac{\bar{w}_n'^2}{\lambda+\Sigma_n'}}. \tag{5.19}$$

5.3 Learning of the density parameter

The density parameter of the sparsity prior can be estimated iteratively via maximum likelihood by following the procedure outlined in Sec. 4.10. In this section, we give the expression of the EP free energy and of its partial derivative with respect to the density parameter ρ, as this is needed for the gradient descent step given the current values of the EP parameters:

$$\rho^{(t+1)} \leftarrow \rho^{(t)} - \eta \frac{\partial F_{EP}}{\partial \rho}, \tag{5.20}$$

where the index t denotes the current iteration and η is a learning rate, which we set as $\eta = 5 \times 10^{-4}$ in the numerical experiments of this chapter.

5.3.1 EP free energy with spike-and-slab priors

We recall that the EP free energy is given by:

$$F_{EP} = (1-N)F_Q + \sum_{i=1}^{N} F_{Q^{(i)}}.$$

The contribution associated with the Gaussian approximate posterior $Q(\boldsymbol{w})$ is independent of the prior and reads:

$$F_Q = \ln Z_Q = \frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln(\det \Sigma), \tag{5.21}$$

whereas the expression of $F_{Q^{(i)}}$ depends on the type of prior considered:

$$F_{Q^{(i)}} = -\ln Z_{Q^{(i)}} = -\ln\left(\int \tilde{Q}^{(i)}(\boldsymbol{w}|\boldsymbol{y})\psi_i(w_i)\mathrm{d}\boldsymbol{w}\right) \tag{5.22}$$

We are interested in the case of spike-and-slab prior factors, for which we have:

$$\ln Z_{Q^{(i)}} = \ln\left((1-\rho)\frac{1}{\sqrt{2\pi\Sigma_i}}\,\mathrm{e}^{-\frac{\mu_i^2}{2\Sigma_i}} + \frac{\rho}{\sqrt{2\pi}}\sqrt{\frac{\lambda}{1+\lambda\Sigma_i}}\,\mathrm{e}^{-\frac{1}{2}\frac{\lambda\mu_i^2}{1+\lambda\Sigma_i}}\right). \tag{5.23}$$

Since the update of the parameters of the prior is assumed to be done at fixed EP parameters, as we already argued in Sec. 4.10, we only need to focus on the tilted contributions to the EP free energy. In fact, these are the only terms that depend on ρ explicitly. As a consequence, we simply have:

$$\frac{\partial F_{EP}}{\partial \rho} = \sum_{i=1}^{N}\frac{\partial F_{Q^{(i)}}}{\partial \rho}, \tag{5.24}$$

where the partial derivatives on the right hand side are computed as:

$$\frac{\partial F_{Q^{(i)}}}{\partial \rho} = -\frac{1}{Z_{Q^{(i)}}}\int Q^{\setminus i}(w_i)\frac{\partial}{\partial \rho}\psi_i(w_i)\mathrm{d}w_i. \tag{5.25}$$

Taking into account the fact that:

$$\frac{\partial}{\partial \rho}\psi_i(w_i) = -\delta(w_i) + \frac{1}{\sqrt{2\pi\lambda}}\,\mathrm{e}^{-\frac{w_i^2}{2\lambda}}, \tag{5.26}$$

we finally obtain:

$$\frac{\partial F_{EP}}{\partial \rho} = \sum_{i=1}^{N}\frac{\frac{1}{\sqrt{2\pi\Sigma_{i,i}}}\,\mathrm{e}^{-\frac{\tilde{w}_i^2}{2\Sigma_{i,i}}} - \frac{1}{\sqrt{2\pi(\lambda+\Sigma_{i,i})}}\,\mathrm{e}^{-\frac{1}{2}\frac{\tilde{w}_i^2}{\lambda+\Sigma_{i,i}}}}{(1-\rho)\frac{1}{\sqrt{2\pi\Sigma_{i,i}}}\,\mathrm{e}^{-\frac{\tilde{w}_i^2}{2\Sigma_{i,i}}} + \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma_{i,i})}}\,\mathrm{e}^{-\frac{1}{2}\frac{\tilde{w}_i^2}{\lambda+\Sigma_{i,i}}}}. \tag{5.27}$$

By computing the second derivative of F_{EP} with respect to ρ, we easily realize that F_{EP} is strictly convex in ρ for positive λ. In fact, by taking the derivative of the last equation

with respect to ρ one more time, we immediately have:

$$\frac{\partial^2 F_{EP}}{\partial \rho^2} = \sum_{n=1}^{N} \left[\frac{\frac{1}{\sqrt{2\pi\Sigma_{n,n}}} e^{-\frac{\bar{w}_n^2}{2\Sigma_{n,n}}} - \frac{1}{\sqrt{2\pi(\lambda+\Sigma_{n,n})}} e^{-\frac{1}{2}\frac{\bar{w}_n^2}{\lambda+\Sigma_{n,n}}}}{(1-\rho)\frac{1}{\sqrt{2\pi\Sigma_{n,n}}} e^{-\frac{\bar{w}_n^2}{2\Sigma_{n,n}}} + \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma_{n,n})}} e^{-\frac{1}{2}\frac{\bar{w}_n^2}{\lambda+\Sigma_{n,n}}}} \right]^2, \tag{5.28}$$

which ensures that the estimate of ρ is unique for fixed values of $\bar{w}_n$ and $\Sigma_{n,n}$.

5.4 Results on uncorrelated measurements

First of all, we consider the case of noiseless uncorrelated measurements, obtained by multiplying the signal to be reconstructed by an $M \times N$ sensing matrix with i.i.d. entries drawn from a standard Gaussian distribution $\mathcal{N}(w; 0,1)$. We assume the signal to be K-sparse, namely, with only K nonzero components, each of which is drawn from a standard Gaussian distribution. We consider a Bayes optimal scenario, where the generative model of the signal coincides with its prior distribution. We also define the *density of the signal* $\rho = K/N$ and the *measurement rate* $\alpha = M/N$. The parameter ρ is the true value of the density parameter of the spike-and-slab prior, which, in general, is not assumed to be known: it can be learned using the free energy minimization procedure outlined in Sec. 4.10.

We quantify the goodness of the reconstructed signal $\hat{\boldsymbol{w}}$ using the sample Pearson correlation coefficient r of the true and inferred vectors, defined as:

$$r = \frac{\sum_{k=1}^{N}(w_k - w_{sm})(\hat{w}_k - \hat{w}_{sm})}{\sqrt{\sum_{k=1}^{N}(w_k - w_{sm})^2}\sqrt{\sum_{k=1}^{N}(\hat{w}_k - \hat{w}_{sm})^2}}, \tag{5.29}$$

where w_{sm} is the sample mean of the signal and $\hat{w}_{sm}$ is that of the reconstructed vector. We also consider the within-sample mean squared error (MSE) as a measure of the reconstruction error:

$$MSE = \frac{1}{N}\sum_{k=1}^{N}(w_k - \hat{w}_k)^2. \tag{5.30}$$

In the limit $N \to \infty$, the Pearson correlation coefficient r and the MSE of the signal estimate reconstructed using Gaussian EP reveal a phase transition occurring as one of the two parameters ρ and α is varied while the other one is kept fixed. We demonstrate this fact at fixed α and increasing ρ in Fig. 5.1, for $N = 400$ and $N = 1600$, where we show the average Pearson coefficient over $N_t = 100$ instances of a signal reconstructed with Gaussian EP. Here and in the following sections of this chapter, we set the precision parameter of the spike-and-slab to $\lambda = 1$ and $\beta = 10^9$ when using the formulation of

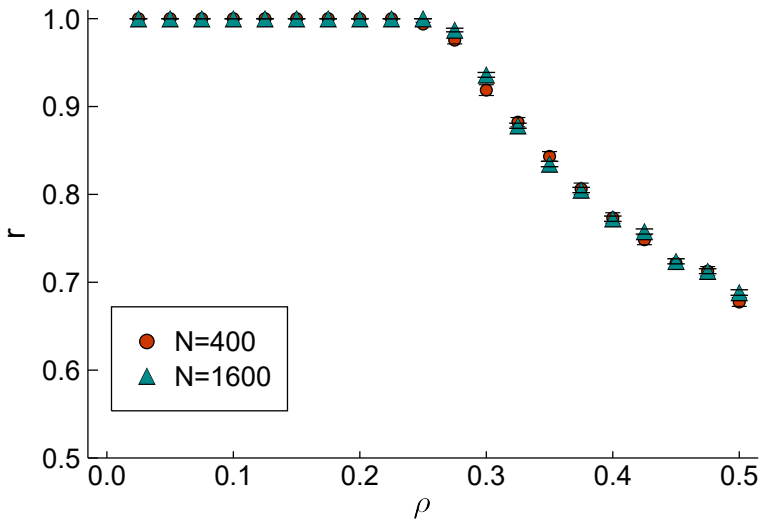


Figure 5.1: Pearson correlation as a function of ρ at $\alpha = 0.5$, with $N = 400$ and $N = 1600$. The error bars are estimated as $\sigma(r)/N_t$, where $N_t = 100$ is the number of instances considered.

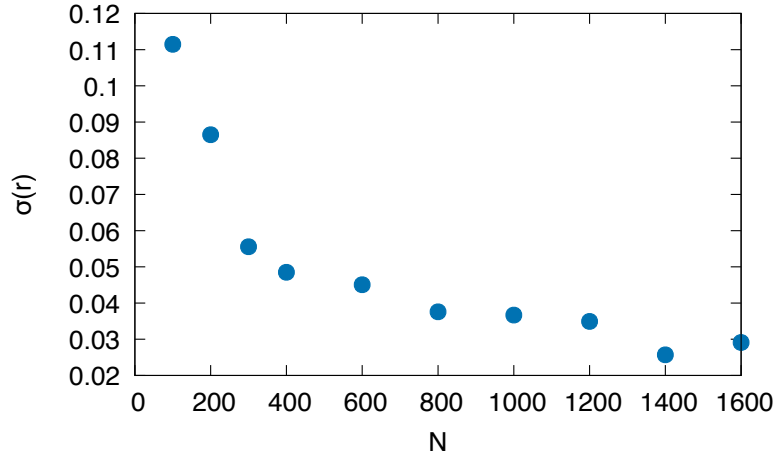


Figure 5.2: Sample standard deviation of r as a function of N. The value of the density and of the measurement rate are fixed and given by $\rho = 0.4$ and $\alpha = 0.55$.

section 5.1.1. The average value of r appears to be independent of the size N of the system, whereas its sample standard deviation decreases as a function of N, as shown in Fig. 5.2.

The fact that fluctuations are low for large N when considering different instances of the CS problem at fixed values of ρ and α allow us to obtain an EP dependent phase diagram numerically by considering only one instance for each pair of parameters (ρ, α)

and the associated Pearson coefficient. To this aim, we plot the value of r for various values of ρ and α and $N = 200$, $N = 400$, $N = 800$ and $N = 1600$ in Fig. 5.3. For each value of ρ and α, a signal with $K = \rho N$ nonzero entries and a i.i.d. matrix $\mathbf{F}$ with $M = \alpha N$ rows and N columns were generated and EP reconstruction of the signal was attempted. The resulting Pearson coefficient shows that the probability of achieving perfect reconstruction above a well defined transition line increases as the size of the system increases, as it can be appreciated from the fact that the separation between the two regions becomes sharper at larger values of N in Fig. 5.3. The L_0 line $\alpha(\rho) = \rho$, plotted in blue, is the information theoretic limit under which reconstructing the signal is not possible, the infeasibility of the region being due to the fact that its points correspond to the situation where one has less equations than nonzero unknowns to be retrieved. The L_1 line, plotted in black in the figure, is the one obtained by Kabashima in [106, 107] in the thermodynamic limit $N \to \infty$ and $M \to \infty$, with α finite, by means of the replica method. Finally, the white region is the one where EP simulations failed to converge.

Fig. 5.3 confirms the fact that there is a Gaussian EP dependent phase transition line located under the L_1 line, which separates the region where reconstruction is perfect – that is, where r is close to 1 (the yellow region in the figure) – from the one where, on the contrary, reconstruction is unsuccessful. We obtain the coordinates $(\rho, \alpha(\rho))$ of the points of the EP transition line using a *bisection*-like algorithm, which we summarize in Alg. 4. We first discretize the interval $0 \leq \rho \leq 1$ and then, for each discretized value ρ_0, we select two values α_0 and α_1 for the bisection algorithm, which we choose on the L_0 transition line and on the L_1 transition line, respectively. Thus, we have for α_0:

$$\alpha_0(\rho_0) = \rho_0, \tag{5.31}$$

and for α_1 [106]:

$$\alpha_1(\rho_0) = 2(1-\rho_0)H(\hat{\chi}^{-1/2}) + \rho_0, \tag{5.32}$$

where $\hat{\chi}$ is given by the solution of the equation [107]:

$$\hat{\chi} = \alpha^{-1}\left[2(1-\rho_0)\left((\hat{\chi}+1)H(\hat{\chi}^{-1/2}) - \hat{\chi}^{1/2}\frac{\mathrm{e}^{-1/(2\hat{\chi})}}{\sqrt{2\pi}}\right) + \rho_0(\hat{\chi}+1)\right], \tag{5.33}$$

and $H(x) = \int_x^{+\infty} \frac{1}{2\pi}\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t$. We perform EP simulations at points (ρ_0, α_0), (ρ_0, α_1) and (ρ_0, α^*), where $\alpha^* = (\alpha_0 + \alpha_1)/2$ and we compute the mean squared error of the related EP solutions, which we shall denote as $MSE(\alpha_0)$, $MSE(\alpha_1)$ and $MSE(\alpha^*)$, respectively. In order to iteratively restrict the interval $\alpha_0 < \alpha < \alpha_1$ via bisection, we define a threshold δ and set $\alpha_1 = \alpha^*$ if the absolute difference $|MSE(\alpha_1) - MSE(\alpha^*)| < \delta$ and $\alpha_0 = \alpha^*$ otherwise. Finally, we recompute $\alpha^* = (\alpha_0 + \alpha_1)/2$. By repeating these steps until a predefined accuracy for the estimate α^* is achieved, we obtain an EP dependent phase transition line, which we show in Fig. 5.4 for the case $N = 1600$.

Algorithm 4 Bisection algorithm

procedure BISECTION(N, ρ_0,α_0,α_1;δ,$\Delta\alpha_{min}$)
  Set $K = \rho_0 N$.
  Set $\alpha^* = (\alpha_0 + \alpha_1)/2$.
  repeat
    Set $M_1 = \alpha_1 N$.
    Generate signal $\boldsymbol{w}_1$ and sensing matrix $\mathbf{F}_1$.
    Infer $\hat{\boldsymbol{w}}_1$ using Parallel EP scheme with inputs $\boldsymbol{y}_1 = \mathbf{F}_1\boldsymbol{w}_1$, $\mathbf{F}_1$ and $\rho = \rho_0$.
    Compute $MSE(\alpha_1)$ between $\hat{\boldsymbol{w}}_1$ and $\boldsymbol{w}_1$.
    Set $M^* = \alpha^* N$.
    Generate signal $\boldsymbol{w}^*$ and sensing matrix $\mathbf{F}^*$.
    Infer $\hat{\boldsymbol{w}}^*$ using Parallel EP scheme with inputs $\boldsymbol{y}^* = \mathbf{F}^*\boldsymbol{w}^*$, $\mathbf{F}^*$ and $\rho = \rho_0$.
    Compute $MSE(\alpha^*)$ between $\hat{\boldsymbol{w}}^*$ and $\boldsymbol{w}^*$.
    if $|MSE(\alpha_1) - MSE(\alpha^*)| > \delta$ **then**
      $\alpha_0 = \alpha^*$.
    else
      $\alpha_1 = \alpha^*$.
    Reassign $\alpha^* = (\alpha_0 + \alpha_1)/2$.
  until $|\alpha_1 - \alpha_0|/2 < \Delta\alpha_{min}$
  return α^*

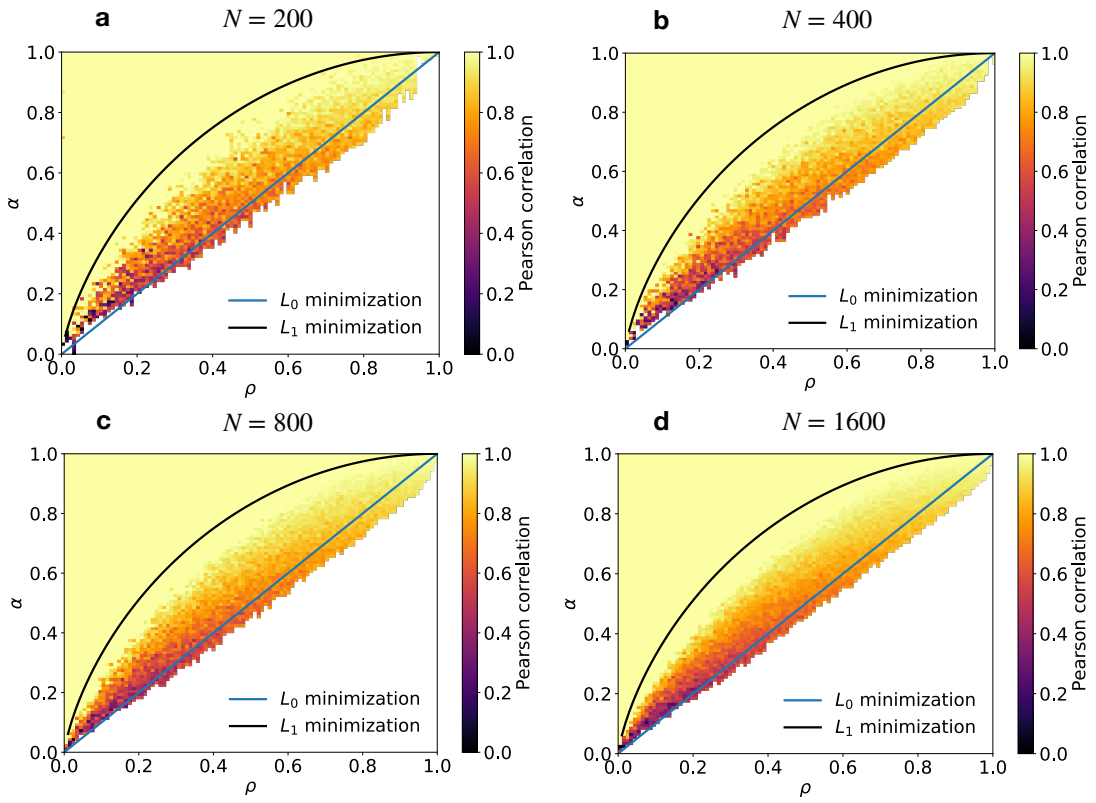


Figure 5.3: Compressed sensing phase diagram for $N = 200,400,800,1600$. Each point (ρ, α) corresponds to a single simulation over a different instance with signal density ρ and measurement rate α. The color refers to the Pearson correlation coefficient between the true signal and the EP reconstructed signal. The black line corresponds to the L_1 reconstruction, while the blue line corresponds to the L_0 condition. Adapted from [98]

Overall, the variable selection properties of EP tend to be quite good, despite the fact that the reconstruction of the signal is no more accurate below the EP phase transition line. In fact, by separating the MSE in two contributions:

$$MSE = \rho MSE_1 + (1-\rho)MSE_2, \tag{5.34}$$

the first of which (MSE_1) is related to the vector of the K nonzero components of the signal and the second of which (MSE_2) is associated with the vector of the remaining $N-K$ components, we see that the dominant contribution to the reconstruction error in the region where $r < 1$ comes from the *estimates* of the K nonzero components of the signal, as it can be deduced from Figs. 5.5a and 5.5b, where the two contributions appearing in Eq. (5.34) are compared to the total mean squared error. This fact implies that the sparse support of the signal is approximately retrieved by Gaussian EP and

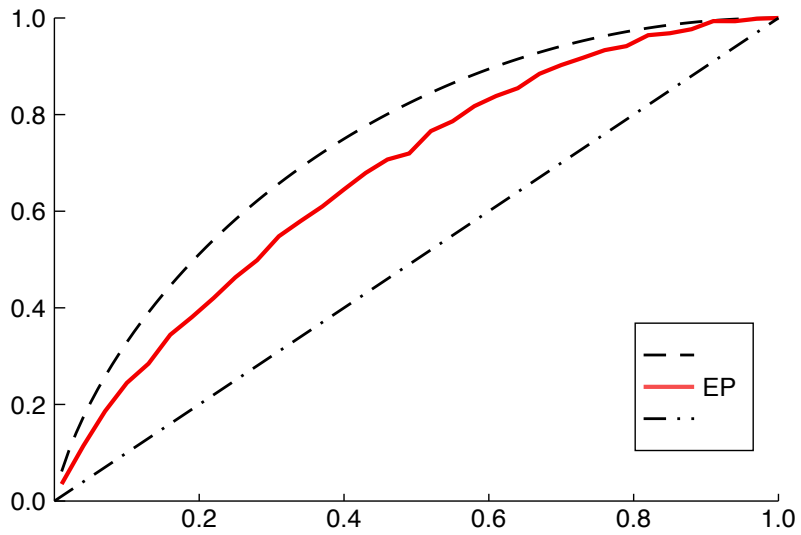


Figure 5.4: EP phase transition line resulting from the bisection-like algorithm given in Alg. 4. The length of the signals to be retrieved is $N = 1600$ and the threshold δ for the MSE difference between the evaluated points is 10^{-5}.

that the reconstruction error is mainly due to the nonzero values not being accurately estimated below the phase transition threshold.

5.5 Results on correlated measurements

We now consider a simple case of correlated sensing matrices $\mathbf{F}$, where the rows are sampled from a multivariate Gaussian distribution:

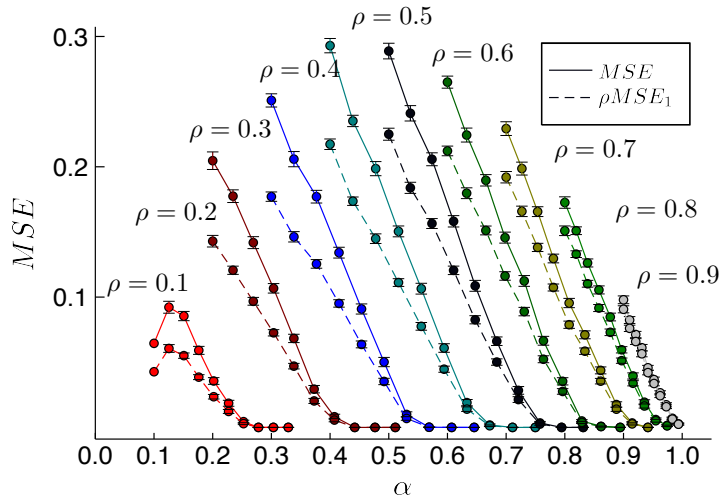
$$\mathbf{F} = (\boldsymbol{f}_1, ..., \boldsymbol{f}_M)^T \tag{5.35}$$

$$\boldsymbol{f}_i \sim \mathcal{N}(0, \mathbf{S}), \qquad i = 1, ..., M, \tag{5.36}$$

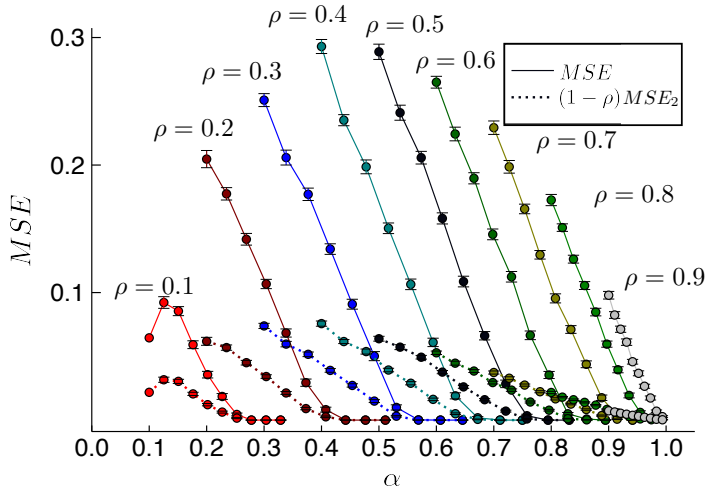
so that the entries of each given row $\boldsymbol{f}_i \in \mathbb{R}^N$ are correlated. We assume the covariance matrix $\mathbf{S}$ appearing in Eq. (5.36) to be constructed as:

$$\mathbf{S} = \mathbf{Y}^\top \mathbf{Y} + \boldsymbol{\Delta}, \tag{5.37}$$

where $\mathbf{Y}$ is a $k \times N$ matrix, the entries of which are i.i.d. and drawn from a standard Gaussian distribution $\mathcal{N}(y; 0, 1)$, and $\boldsymbol{\Delta}$ is a diagonal matrix with positive diagonal entries obtained by sampling again from a standard normal distribution and taking the absolute value. The Gram matrix $\mathbf{Y}^\top \mathbf{Y}$ is symmetric and positive semi-definite by definition and its rank k is a parameter that can be varied in order to change the degree of



(a)



(b)

Figure 5.5: (a) Contribution of the first K components to the MSE (dashed lines), compared to the MSE itself (solid lines). (b) Contribution of the last $N-K$ components (the "tail" of the vector) to the MSE (dotted lines), compared to the MSE itself (solid lines). In both panels a) and b), $N = 400$, the number of simulations is 100 and each curve corresponds to a different value of ρ, for $\rho = 0.1, 0.2, \ldots, 0.9$. The points are averages computed over the N_c converged simulations and the uncertanties are estimated from the sample standard deviations σ using $\sigma/\sqrt{N_c}$.

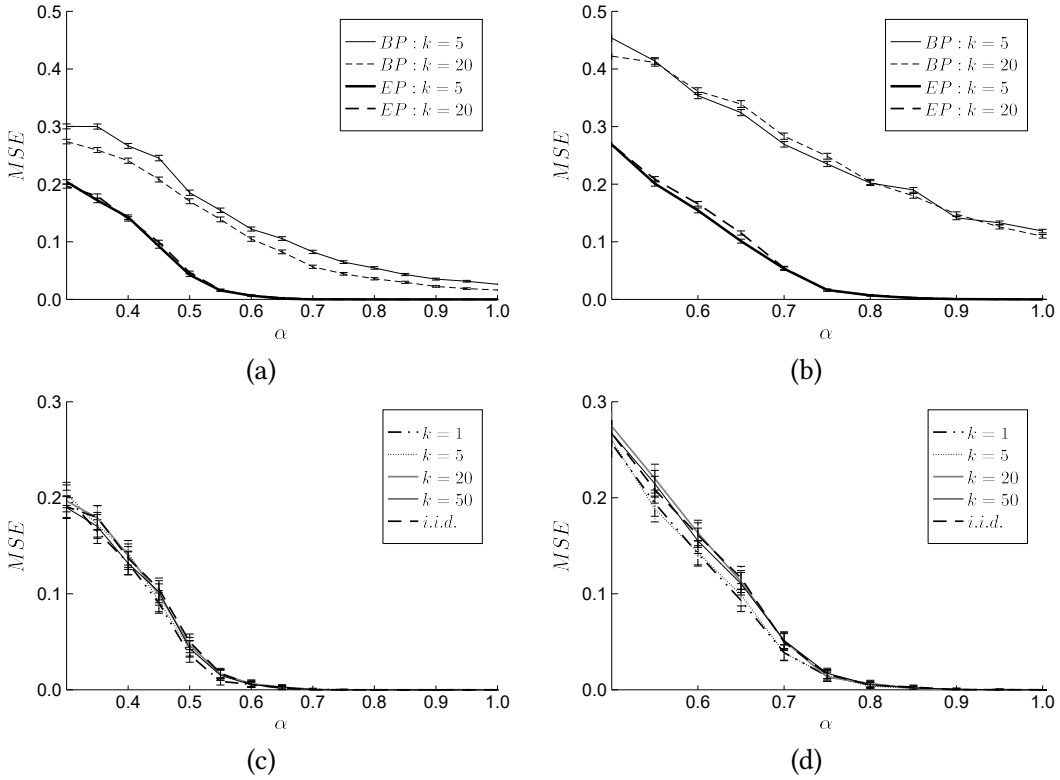


Figure 5.6: (a),(b) Comparison between the MSE obtained from Gaussian EP reconstruction and from EMBP reconstruction for correlated Gaussian sensing matrices. The MSE was evaluated over $N_t = 1000$ converged trials and the uncertainty was estimated as $\sigma/\sqrt{N_t}$, where σ is the sample standard deviation. For all values of α, the EMBP algorithm fails to reconstruct correctly the signal, whereas EP achieves zero MSE beyond a critical value $\alpha_c(\rho)$. (c),(d) MSE resulting from CS reconstruction with i.i.d. and correlated Gaussian sensing matrices. Each point was evaluated over $N_t = 1000$ trials. Lower values of k correspond to more correlated measurements. The length of the signal is $N = 50$ and its density is (a),(c) $\rho = 0.3$ and (b),(d) $\rho = 0.5$.

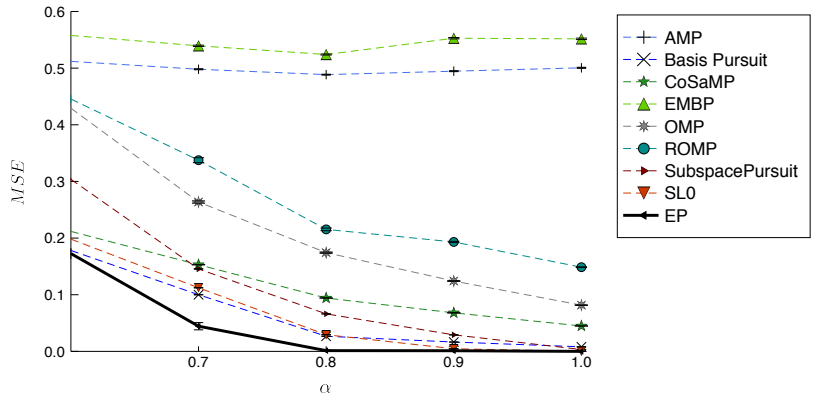
correlation among the entries of the rows of $\mathbf{F}$. Notice that we have added the matrix $\boldsymbol{\Delta}$ in order to guarantee that $\mathbf{S}$ has maximum rank, equal to N.

We compare the reconstruction performance of EP to that of Expectation Maximization Belief Propagation (EMBP) [108, 109]. In order to do so, we use the MATLAB implementation of Ref. [110]. The motivation for choosing EMBP in our comparisons lies on the fact that, in general, the entries of the measurement matrix are not independent for matrices constructed as in Eq. (5.36), especially for small values of the rank k of $\mathbf{Y}$, as, in this case, the covariances and the variances of $\mathbf{S}$ are comparable in terms of magnitude. Conversely, the case of i.i.d. entries is recovered in the large k limit: in

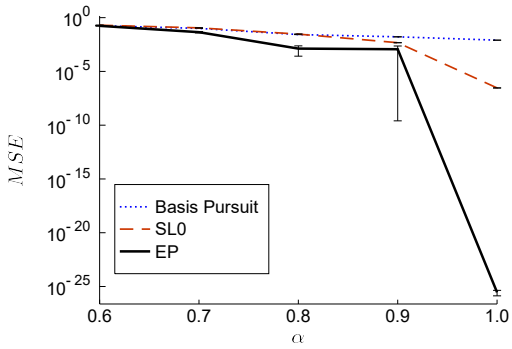
fact, due to cancellation effects, the off-diagonal entries of $\mathbf{S}$ become negligible as compared to its diagonal ones and the associated multivariate Gaussian measure resembles an isotropic Gaussian distribution, according to which the variables are approximately independent of each other. We show the MSE achieved by EP and EMBP in the presence of this kind of correlated matrices as a function of α for densities of the signal given by $\rho = 0.3$ and $\rho = 0.5$, respectively, in Figs. 5.6a and 5.6b, where we only include converged trials. We estimate the signal density using expectation maximization in the case of EMBP and using gradient descent on the EP free energy at each iteration in the case of EP (see Sec. 5.3.1). Remarkably, the results in Figs. 5.6a and 5.6b show that BP is not able to reconstruct the signal nor its density from the knowledge of $\mathbf{F}$ and of the observation vector $\boldsymbol{y}$, regardless of the specific value of α considered. In particular, notice that the BP reconstruction fails even in the case where the number of equations equals the number of variables, which corresponds to $\alpha = 1$, as it is deduced from the fact that the MSE is nonzero. However, as correlations become weaker, which corresponds to larger values of k, we see that the reconstruction accuracy of EMBP tends to improve. On the contrary, the MSE associated with EP does not appear to depend on the degree of correlation of the entries of the sensing matrix. This is shown in Figs. 5.6a and 5.6b and is further confirmed in Figs. 5.6c and 5.6d, where we consider the i.i.d. sensing matrix case $k \to \infty$ and the correlated case with $k = 50$, $k = 20$, $k = 10$, $k = 5$ and $k = 1$. In particular, notice that the reconstruction threshold separating the successful and unsuccessful reconstruction regions does not change.

Finally, we performed further numerical tests in order to assess how EP compares to state of the art algorithms for CS reconstruction of sparse signals in the setup considered in this section. The algorithms that we analyzed include L_1 based convex optimization, in particular Basis Pursuit [5], message passing algorithms from statistical physics such as EMBP [108] and approximate message passing (AMP) [76], signal processing algorithms of the matching pursuit type, namely, Orthogonal Matching Pursuit (OMP) [111], Regularized Orthogonal Matching Pursuit [112], Compressive Sampling Matching Pursuit (CoSaMP) [113] and Subspace Pursuit [114], and the smoothed L_0 norm regularization algorithm (SL0) from Ref. [115]. We used the implementation available in the C++ library KL1p [116], which is based on the linear algebra library *Armadillo* [117, 118].

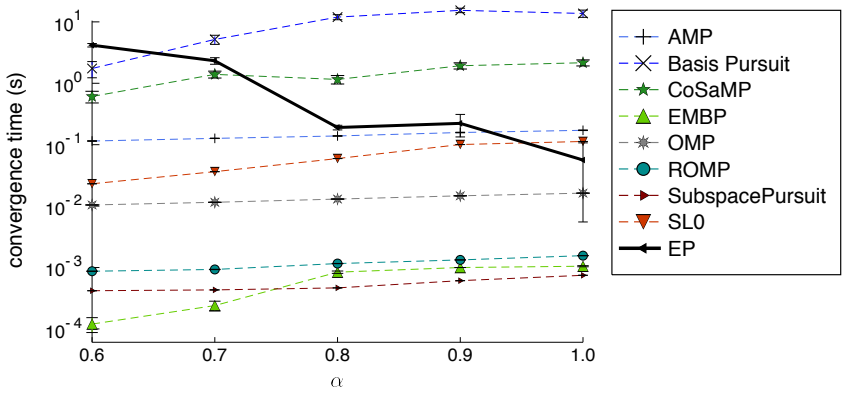
We generated $N_t = 100$ different Gaussian i.i.d. signals of length $N = 100$ and as many random correlated sensing matrices, with $k = 5$, for various values of the measurement rate α. For each pair of signal $\boldsymbol{w}$ and measurement matrix $\mathbf{F}$, we ran EP and all algorithms implemented in KL1p in order to solve the related CS reconstruction problem. The outcome of our numerical experiments are shown in Fig. 5.7, where we obtain that EP is the only algorithm displaying a reconstruction phase transition, as it can be seen in Fig. 5.7a and in the semi-logarithmic plot in Fig. 5.7b, contrary to all the other algorithms, the MSE of which is larger than zero at all values of α. Empirically, we find that the running time of EP is mostly comparable to several of the reconstruction methods examined in our tests, as Fig. 5.7c shows.



(a)



(b)



(c)

Figure 5.7: (a) Comparison of the MSE resulting from CS reconstruction with various state-of-the-art algorithms with correlated Gaussian sensing matrices. (b) Semi-logarithmic plot of the MSE associated with Basis Pursuit, SL0 and EP reconstruction. (c) Comparison of the elapsed running time of the algorithms considered in panel (a) with correlated Gaussian sensing matrices. In all panels, the instances considered are the same, their total number being $N_t = 100$, and the parameters of the generated signals are $N = 100$, $\rho = 0.5$ and $k = 5$. Both unconverged and converged simulations were considered.

Chapter 6

Expectation propagation on the sparse perceptron learning problem

In this chapter, we apply Gaussian EP to the problem of learning a binary classification rule in the teacher-student scenario and compare its performance to other variational algorithms. We will be mostly interested in assessing the variable selection properties of EP in various settings, where the examples provided to the student perceptron can be noiseless, noisy, uncorrelated or correlated. The work presented in this chapter is published in Ref. [100], from which figures and tables are either taken or adapted. Therefore, copyright (2021) by the American Physical Society applies to all figures and tables in this chapter, except for Figs. 6.3, 6.4 and 6.7, which were not published before.

6.1 The sparse perceptron learning problem

We consider two perceptrons having N units and continuous valued weights in the *teacher-student scenario* [119, 120]. Let us denote the weights of the teacher perceptron as $\boldsymbol{B} \in \mathbb{R}^N$ and those of the student perceptron as $\boldsymbol{w} \in \mathbb{R}^N$. We assume that the teacher has sparse weights, a fraction ρ of which are nonzero. In the teacher-student framework, a set of M binary labels σ_τ, $\tau = 1, \dots, M$ is assigned to as many pattern vectors $x_\tau \in \mathbb{R}^N$ according to the teacher's classification rule:

$$\sigma_\tau = \operatorname{sign}(\boldsymbol{B}^\top \boldsymbol{x}_\tau), \quad \tau = 1, \dots, M, \tag{6.1}$$

where we use the convention that $\operatorname{sign}(0) \coloneqq 1$, and the task of the student perceptron is to learn the rule (6.1) by adjusting the weights $\boldsymbol{w}$ based on the set of examples $\{(\boldsymbol{x}_1, \sigma_1), \dots, (\boldsymbol{x}_M, \sigma_M)\}$, so that the relation:

$$\sigma_\tau = \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x}_\tau), \quad \tau = 1, \dots, M. \tag{6.2}$$

is fulfilled. Using the terminology adopted in the neural network literature, we will refer to the process of learning the relationship between inputs and outputs by updating the weights of the perceptron based on a set of examples as *training*. Accordingly, we will use the expression *training set* to refer to the set of examples provided to the student.

In order to enforce the classification rule of the student, we notice that Eq. (6.2) can be equivalently written as a set of positivity constraints:

$$\left(\sigma_\tau \boldsymbol{x}_\tau^\top\right) \boldsymbol{w} \geq 0 \quad \tau = 1, \dots, M. \tag{6.3}$$

Moreover, we define the auxiliary variables $y_\tau := \left(\sigma_\tau \boldsymbol{x}_\tau^\top\right) \boldsymbol{w}$ and the data matrix:

$$\mathbf{X}_{\boldsymbol{\sigma}} := \begin{pmatrix} \sigma_1 \boldsymbol{x}_1^\top \\ \sigma_2 \boldsymbol{x}_2^\top \\ \vdots \\ \sigma_M \boldsymbol{x}_M^\top \end{pmatrix}. \tag{6.4}$$

Using these definitions, we immediately obtain the linear estimation problem:

$$\boldsymbol{y} = \mathbf{X}_\sigma \boldsymbol{w}, \tag{6.5}$$

where the dependent sets of variables $\boldsymbol{y}$ and $\boldsymbol{w}$ need to be jointly estimated.

6.1.1 Bayesian framework for the sparse perceptron learning problem and EP approximation of the posterior distribution

We introduce a Bayesian setup for the linear estimation problem given in Eq. (6.5), similarly to what was done in Sec. 5.1. In order to do so, we define the variable vector $\boldsymbol{h} = (w_1, \dots, w_N, y_1, \dots, y_M)^\top$ and the energy function:

$$E(\boldsymbol{w}, \boldsymbol{y}) = \|\boldsymbol{y} - \mathbf{X}_\sigma \boldsymbol{w}\|^2 = \boldsymbol{h}^\top \mathbf{E}^{-1} \boldsymbol{h}, \tag{6.6}$$

where the matrix $\mathbf{E}^{-1}$ is expressed as:

$$\mathbf{E}^{-1} = \begin{pmatrix} \mathbf{X}_\sigma^\top \mathbf{X}_\sigma & -\mathbf{X}_\sigma^\top \\ -\mathbf{X}_\sigma & \mathbf{I} \end{pmatrix}. \tag{6.7}$$

The likelihood function of the N weights of the student perceptron is given by:

$$L(\boldsymbol{w}) = P(\sigma_1, \dots, \sigma_M | \boldsymbol{w}) = \frac{1}{Z} \mathrm{e}^{-\beta E(\boldsymbol{w}, \boldsymbol{y})}. \tag{6.8}$$

If we wish to enforce the linear constraints that define the variables $\boldsymbol{y}$ exactly, we can consider the limit $\beta \to +\infty$ of $L(\boldsymbol{w})$:

$$\lim_{\beta \to +\infty} L(\boldsymbol{w}) = \delta^M(\boldsymbol{y} - \mathbf{X}_\sigma \boldsymbol{w}), \tag{6.9}$$

where $\delta^M(\boldsymbol{z})$ denotes the M-dimensional Dirac delta distribution.

Furthermore, we introduce prior distributions in order to enforce sparsity of the weights of the student and consistency (or lack thereof) between the patterns and the labels given the student classification rule. Therefore, we associate a spike-and-slab prior $\Gamma(w_i)$ with each weight w_i:

$$\Gamma(w_i) = (1-\rho)\delta(w_i) + \rho\sqrt{\frac{\lambda}{2\pi}}e^{-\frac{\lambda w_i^2}{2}}, \quad i = 1, \dots, N.$$

and a pseudoprior $\Lambda(y_\tau)$, which we express as:

$$\Lambda(y_\tau) = \Theta(y_\tau), \quad \tau = 1, \dots, M. \tag{6.10}$$

if the consistency relations (6.3) are fulfilled (*theta* pseudoprior) and as:

$$\Lambda(y_\mu) = \eta\Theta(y_\mu) + (1-\eta)\Theta(-y_\mu). \tag{6.11}$$

if some consistency relations are violated (*theta mixture* pseudoprior), in which case we assume that some examples are mislabeled according to:

$$\tilde{\sigma} = \begin{cases} \text{sign}(\boldsymbol{B}^\top \boldsymbol{x}) & \text{with probability} \quad \eta, \\ -\text{sign}(\boldsymbol{B}^\top \boldsymbol{x}) & \text{with probability} \quad 1-\eta, \end{cases} \tag{6.12}$$

where $0 \leq \eta \leq 1$, and that the student perceptron receives the set of corrupted examples $(x_\mu, \tilde{\sigma}_\mu)$, $\mu = 1, \dots, M$ [121]. In Eqs. (6.10) and (6.11), we used $\Theta(z)$ to denote the Heaviside step function:

$$\Theta(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

We will consider a Bayes-optimal scenario, where the priors to be incorporated in the student posterior distribution match both the model that generates the weights of the teacher and the process that assigns the labels to the patterns. Despite the fact that sparsity and the presence of inconsistencies between patterns and labels are modeled as prior knowledge, we stress that, in general, the values of the sparsity level ρ_0 of the teacher weights and of the consistency level η_0 of the examples are not assumed to be known a priori and, therefore, need to be estimated during the learning task performed by the student.

Considering the limit (6.9) and using Bayes' rule, we have for the joint posterior distribution of the variables $\boldsymbol{w}$ and $\boldsymbol{y}$:

$$P(\boldsymbol{w}, \boldsymbol{y}) = \frac{1}{Z_P}\delta(\boldsymbol{y} - \mathbf{X}_\sigma \boldsymbol{w}) \prod_{i=1}^{N} \Gamma_i(w_i) \prod_{\tau=1}^{M} \Lambda_\tau(y_\tau), \tag{6.13}$$

which we wish to approximate using Gaussian EP. We can straightforwardly apply the formulation of Gaussian EP with rigid linear constraints, which we derived in Sec. 4.4,

to the homogeneous linear system (6.5) by identifying $\boldsymbol{v} \coloneqq \boldsymbol{y}$, $\boldsymbol{u} \coloneqq \boldsymbol{w}$, $\mathbf{F} \coloneqq \mathbf{X}_\sigma$ and $\tilde{\boldsymbol{y}} \coloneqq 0$[1]. As we did in the CS problem, the estimate of the variables of interests will be provided by the MMSE estimator.

6.2 Moments of the tilted distributions in the sparse perceptron learning problem

In this section, we express the moments of the tilted distributions relevant to the sparse perceptron learning problems and entering in the EP moment matching updates. As the tilted moments related to the spike-and-slab prior were already given in Chapter 5, we do not repeat them here and refer to Sec. 5.2.1. Instead, we only give the expressions of the tilted moments related to the consistency variables y_τ, $\tau = 1, \dots, M$ and consider both the case of theta pseudoprior factors, Eq. (6.10), and the case of theta mixture pseudoprior factors, Eq. (6.11). For ease of notation, we will denote the means of the cavity distribution by μ_k, its variances by Σ_k and introduce the ratios $o_k \coloneqq \mu_k/\sqrt{\Sigma_k}$ for $k = N+1, \dots, N+M$.

6.2.1 Tilted moments with a theta pseudoprior factor

In the case of a theta pseudoprior the expression of the tilted distributions associated with the variables $\boldsymbol{y} = (h_{N+1}, \dots, h_{N+M})^\top$ is given by:

$$Q^{(k)}(h_k) = \frac{1}{Z_{Q^{(k)}}} Q^{\setminus k}(h_k)\Theta(h_k), \quad k = N+1, \dots, N+M, \tag{6.14}$$

where, as usual, $Q^{\setminus k}$ is the cavity Gaussian distribution. The normalization of the tilted distribution (6.14) reads:

$$Z_{Q^{(k)}} = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right], \tag{6.15}$$

where erf is the error function:

$$\operatorname{erf}(x) \coloneqq \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-z^2}\, \mathrm{d}z. \tag{6.16}$$

Computation of the first tilted moment yields:

$$\langle h_k \rangle = \mu_k + \sqrt{\frac{\Sigma_k}{2\pi}} \frac{\mathrm{e}^{-\frac{\mu_k^2}{2\Sigma_k}}}{\Phi\left(\frac{\mu_k}{\sqrt{\Sigma_k}}\right)} = \mu_k \left(1 + \frac{R(o_k)}{o_k}\right), \tag{6.17}$$

[1]Notice that in this chapter the total number of variables is $N+M$, whereas it is N in Sec. 4.4.

where $\Phi(x)$ denotes the cumulative density function (CDF) of the standard normal distribution:

$$\Phi(x) \coloneqq \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right],$$

and the function $R(x)$ is defined as:

$$R(x) = \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-x^2/2}}{\Phi(x)}.$$

With regard to the second tilted moment, we obtain:

$$\langle h_k^2 \rangle = \mu_k^2 + \Sigma_k + \mu_k \sqrt{\Sigma_k} R(o_k)\,, \tag{6.18}$$

so that the tilted variance can be compactly written as:

$$\operatorname{Var}(h_k) = \Sigma_k (1 - o_k R(o_k) - R^2(o_k)). \tag{6.19}$$

6.2.2 Tilted moments with a theta mixture pseudoprior factor

In the case of a theta mixture pseudoprior, the tilted distribution reads:

$$Q^{(k)}(h_k) = \frac{1}{Z_{Q^{(k)}}} Q^{\setminus k}(h_k) \left[\eta \Theta(h_k) + (1-\eta)\Theta(-h_k)\right], \quad k = N+1, \dots, N+M, \tag{6.20}$$

its normalization $Z_{Q^{(k)}}$ being:

$$Z_{Q^{(k)}} = \eta \left[\frac{1}{2}\operatorname{erfc}\left(-\frac{o_k}{\sqrt{2}}\right)\right] + (1-\eta)\left[\frac{1}{2}\operatorname{erfc}\left(\frac{o_k}{\sqrt{2}}\right)\right] = \sqrt{\frac{\pi\Sigma_k}{2}}\left[\frac{1}{2} + \left(\eta - \frac{1}{2}\right)\operatorname{erf}\left(\frac{o_k}{\sqrt{2}}\right)\right]. \tag{6.21}$$

For the first moment, one obtains:

$$\begin{aligned} \langle h_k \rangle_{Q^{(k)}} &= \frac{1}{Z_{Q^{(k)}}} \left\{ \frac{\eta}{\sqrt{2\pi\Sigma_k}} \left[\Sigma_k \, \mathrm{e}^{-\frac{o_k^2}{2}} + \mu_k \sqrt{\frac{\pi\Sigma_k}{2}} \operatorname{erfc}\left(-\frac{o_k}{\sqrt{2}}\right)\right] \right. \\ &\quad \left. + \frac{1-\eta}{\sqrt{2\pi\Sigma_k}} \left[-\Sigma_k \, \mathrm{e}^{-\frac{o_k^2}{2}} + \mu_k \sqrt{\frac{\pi\Sigma_k}{2}} \operatorname{erfc}\left(\frac{o_k}{\sqrt{2}}\right)\right] \right\} \\ &= \mu_k + \sqrt{\frac{2\Sigma_k}{\pi}} \frac{(2\eta - 1)\, \mathrm{e}^{-\frac{o_k^2}{2}}}{\eta \operatorname{erfc}\left(-\frac{o_k}{\sqrt{2}}\right) + (1-\eta)\operatorname{erfc}\left(\frac{o_k}{\sqrt{2}}\right)}, \end{aligned} \tag{6.22}$$

whereas for the second tilted moment we have:

$$\begin{aligned}\langle h_k^2\rangle_{Q^{(k)}} &= \frac{1}{Z_{Q^{(k)}}}\left\{\frac{\eta}{\sqrt{2\pi\Sigma_k}}\left[\mu_k\Sigma_k\,\mathrm{e}^{-\frac{o_k^2}{2}}+\sqrt{\frac{\pi\Sigma_k}{2}}\left(\mu_k^2+\Sigma_k\right)\mathrm{erfc}\left(-\frac{o_k}{\sqrt{2}}\right)\right]\right.\\ &\left.+\frac{1-\eta}{\sqrt{2\pi\Sigma_k}}\left[-\mu_k\Sigma_k\,\mathrm{e}^{-\frac{o_k^2}{2}}+\sqrt{\frac{\pi\Sigma_k}{2}}\left(\mu_k^2+\Sigma_k\right)\mathrm{erfc}\left(\frac{o_k}{\sqrt{2}}\right)\right]\right\}\\ &= \mu_k^2+\Sigma_k+\mu_k\sqrt{\frac{2\Sigma_k}{\pi}}\frac{(2\eta-1)\,\mathrm{e}^{-\frac{o_k^2}{2}}}{\eta\mathrm{erfc}\left(-\frac{o_k}{\sqrt{2}}\right)+(1-\eta)\mathrm{erfc}\left(\frac{o_k}{\sqrt{2}}\right)}.\end{aligned} \tag{6.23}$$

6.3 Learning of the parameters of the prior

As discussed in the previous chapters, Gaussian EP allows one to iteratively estimate the parameters of the priors by minimizing the EP free energy. In the sparse perceptron learning problem, the parameters of interest are the density level of the weights of the teacher and, if there is noise flipping some of the labels, the fraction of mislabeled examples. After writing the expression of the EP free energy for the problem that we are considering, we will give the expression of the gradient descent update of the parameter η of the theta mixture prior. Concerning the spike-and-slab prior, we already wrote the update in Sec. 5.3.1, to which the reader is referred for further details.

6.3.1 EP free energy for the diluted perceptron problem

In the case of the diluted classifier, the total number of variable is $N+M$ and the EP free energy is given by:

$$F_{EP} = (N+M-1)\ln Z_Q - \sum_{k=1}^{N+M}\ln Z_{Q^{(k)}}, \tag{6.24}$$

where:

$$\ln Z_Q = \frac{N+M}{2}\ln(2\pi)+\frac{1}{2}\ln(\det\Sigma), \tag{6.25}$$

and the expression of $\ln Z_{Q^{(k)}}$ reads:

$$\ln Z_{Q^{(k)}} = \ln\left((1-\rho)\frac{1}{\sqrt{2\pi\Sigma_k}}\,\mathrm{e}^{-\frac{\mu_k^2}{2\Sigma_k}}+\frac{\rho}{\sqrt{2\pi}}\sqrt{\frac{\lambda}{1+\lambda\Sigma_k}}\,\mathrm{e}^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}\right),\quad k=1,\dots,N \tag{6.26}$$

for the weights variables, as the associated priors are spike-and-slab priors, while for the consistency enforcing variables $h_{N+1},\dots,h_{N+M}$ we have two possibilities:

1. if, on the one hand, their associated pseudoprior is of the theta kind, then $Z_{Q^{(k)}}$ is expressed as:

$$Z_{Q^{(k)}} = \sqrt{\frac{\pi\Sigma_k}{2}}\left(1 + \mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right), \quad k = N+1, ..., N+M \tag{6.27}$$

2. if, on the other hand, their associated pseudoprior is of the theta mixture kind, then we have:

$$\begin{aligned} Z_{Q^{(k)}} &= \sqrt{\frac{\pi\Sigma_k}{2}}\left[\frac{\eta}{2}\mathrm{erfc}\left(-\frac{\mu_k}{\sqrt{2\Sigma_k}}\right) + \frac{1-\eta}{2}\mathrm{erfc}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right] \\ &= \sqrt{\frac{\pi\Sigma_k}{2}}\left[\frac{1}{2} + \left(\eta - \frac{1}{2}\right)\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right], \quad k = N+1, ..., N+M, \end{aligned} \tag{6.28}$$

Inserting these terms in the EP free energy, we finally obtain:

1. for the case $\Lambda(h_k) = \Theta(h_k)$:

$$\begin{aligned} F_{EP} = (N+M-1)&\left(\frac{N+M}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\boldsymbol{\Sigma})\right) \\ &- \sum_{k=1}^{N}\ln\left((1-\rho)\frac{1}{\sqrt{2\pi\Sigma_k}}\,\mathrm{e}^{-\frac{\mu_k^2}{2\Sigma_k}} + \frac{\rho}{\sqrt{2\pi}}\sqrt{\frac{\lambda}{1+\lambda\Sigma_k}}\,\mathrm{e}^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}\right) \\ &- \frac{M}{2}\ln(\pi/2) - \frac{1}{2}\sum_{k=N+1}^{N+M}\ln\Sigma_k - \sum_{k=N+1}^{N+M}\ln\left(1 + \mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right), \end{aligned} \tag{6.29}$$

2. for the case $\Lambda(h_k) = \eta\Theta(h_k) + (1-\eta)\Theta(-h_k)$:

$$\begin{aligned} F_{EP} = (N+M-1)&\left(\frac{N+M}{2}\ln(2\pi) + \frac{1}{2}\ln(\det\boldsymbol{\Sigma})\right) \\ &- \sum_{k=1}^{N}\ln\left((1-\rho)\frac{1}{\sqrt{2\pi\Sigma_k}}\,\mathrm{e}^{-\frac{\mu_k^2}{2\Sigma_k}} + \frac{\rho}{\sqrt{2\pi}}\sqrt{\frac{\lambda}{1+\lambda\Sigma_k}}e^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}\right) \\ &- \frac{M}{2}\ln(\pi/2) - \frac{1}{2}\sum_{k=N+1}^{N+M}\ln\Sigma_k - \sum_{k=N+1}^{N+M}\ln\left(\frac{1}{2} + \left(\eta - \frac{1}{2}\right)\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right). \end{aligned} \tag{6.30}$$

As a technical side comment, we notice that the $\ln(\det\boldsymbol{\Sigma})$ contribution in the EP free energy can be efficiently computed by means of the Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}$. In fact, the covariance matrix can be written as $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix with real positive diagonal elements L_{kk} for $k = 1, \dots, N+M$. Therefore, recalling that the determinant of a lower triangular matrix is nothing but the

product of its diagonal entries, the term $\frac{1}{2}\ln(\det \Sigma)$ can be numerically computed by first obtaining $\mathbf{L}$ and then taking its trace:

$$\frac{1}{2}\ln(\det \Sigma) = \sum_{k=1}^{N+M} \ln(L_{kk}).$$

6.3.2 Learning of the η parameter of the theta mixture pseudoprior

By following the reasoning of Sec. 4.10, and in analogy with what was done in the case of the spike-and-slab prior, we can easily write a gradient descent update for the parameter η of the theta mixture pseudoprior. All we need to do is to compute the partial derivative of the terms $F_{Q^{(k)}}$, $k = N+1, ..., N+M$ given in Eq. (6.28) with respect to η. In this way, we readily obtain:

$$\frac{\partial F_{EP}}{\partial \eta} = \sum_{k=N+1}^{N+M} \frac{\partial F_{Q^{(k)}}}{\partial \eta},$$

where:

$$\frac{\partial F_{Q^{(k)}}}{\partial \eta} = \frac{-2\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)}{1 + (2\eta - 1)\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)}. \tag{6.31}$$

Finally, the gradient descent update reads:

$$\eta^{new} \leftarrow \eta^{old} - \left.\frac{\partial F_{EP}}{\partial \eta}\right|_{\eta^{old}} \delta\eta, \tag{6.32}$$

where $\delta\eta$ is the learning rate.

6.4 Sparse perceptron learning from noiseless examples

In this section, we show the results obtained from applying Gaussian EP to sparse perceptron learning from noiseless examples. We will first analyze the case of i.i.d. Gaussian patterns and then move on to that of correlated Gaussian patterns. As anticipated above, we will consider a Bayes-optimal setup in order to perform the analysis under controlled conditions. Therefore, we assume that a fraction ρ_0 of the weights of the teacher are drawn from a standard normal distribution and the remaining fraction

$1 - \rho_0$ are zero and choose the prior associated with the weights of the students to be a spike-and-slab prior with density parameter ρ and precision of the slab $\lambda = 1$.

Since the amplitude information of the weights to be retrieved is lost due to the sign nonlinearity, given a solution of Eq. (6.2), all vectors obtained by multiplying the norm of this solution by a scale factor also fulfill the perceptron classification rule. Therefore, we will normalize the weights of the student and of the teacher to one when assessing the goodness of the student's estimate. This will be done by considering the MSE between the normalized weights of the student and the normalized weights of the teacher:

$$\text{MSE}\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \frac{\boldsymbol{B}}{\|\boldsymbol{B}\|}\right) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{w_k}{\|\boldsymbol{w}\|} - \frac{B_k}{\|\boldsymbol{B}\|}\right)^2, \tag{6.33}$$

which we will express in decibel (dB). The performance of EP based sparse perceptron learning will be compared to those related to the 1-bit approximate message passing technique (1bitAMP) from Ref. [122] and to the generalized vector approximate message passing algorithm (grVAMP) from Ref. [85], which was discussed in Sec. 3.7.1.

We carried out numerical simulations on i.i.d. patterns drawn from a standard normal distribution and evaluated the MSE (6.33) over $N_{samples} = 100$ instances, each of which consisting of a i.i.d. pattern matrix, its associated labels and a set of weights of the teacher perceptron. We set the number of weights as $N = 128$ and the density level $\rho_0 = 0.25$ Moreover, we set $\epsilon_{\text{stop}} = 10^{-4}$ for the EP convergence parameter and use 0.9995 for the EP damping parameter, although good results are obtained using lower values too, e.g. 0.99. Moreover, we set the maximum number of iteration to 50000. The resulting values of the MSE as a function of α are shown in Fig. 6.1 and demonstrate that using EP, 1-bit AMP and grVAMP to train the student perceptron from i.i.d. Gaussian patterns do not lead to appreciable differences in terms of training error. We adopted a convergence threshold equal to 10^{-4} in the case of 1-bit AMP and equal to 10^{-8} in the case of grVAMP. All simulations converged within the thresholds specified, regardless of the algorithm. The error bars in Fig. 6.1 were computed by dividing the sample standard deviation of the MSE by $\sqrt{N_{samples}}$.

Analogously to what was done when analyzing the performance of Gaussian EP on the CS problem (cfr. Sec. 5.5), we now consider the problem of training the student on correlated patterns drawn from a multivariate normal distribution, in the simple case where the Gaussian distribution has zero vector mean and covariance matrix given by:

$$\mathbf{S} = \mathbf{Y}^\top\mathbf{Y} + \boldsymbol{\Delta}, \tag{6.34}$$

where $\mathbf{Y} \in \mathbb{R}^{u\times N}$ is an i.i.d. matrix with entries drawn from a standard normal distribution and $\boldsymbol{\Delta}$ is a diagonal matrix, the diagonal entries of which are given by the absolute value of i.i.d. random numbers drawn from a standard Gaussian distribution too. We recall that the diagonal matrix $\boldsymbol{\Delta}$ is added so that $\mathbf{S}$ has full rank. As we did in the CS problem, we choose $u = 1$ for the matrix $\mathbf{Y}$, because this choice leads to the off diagonal

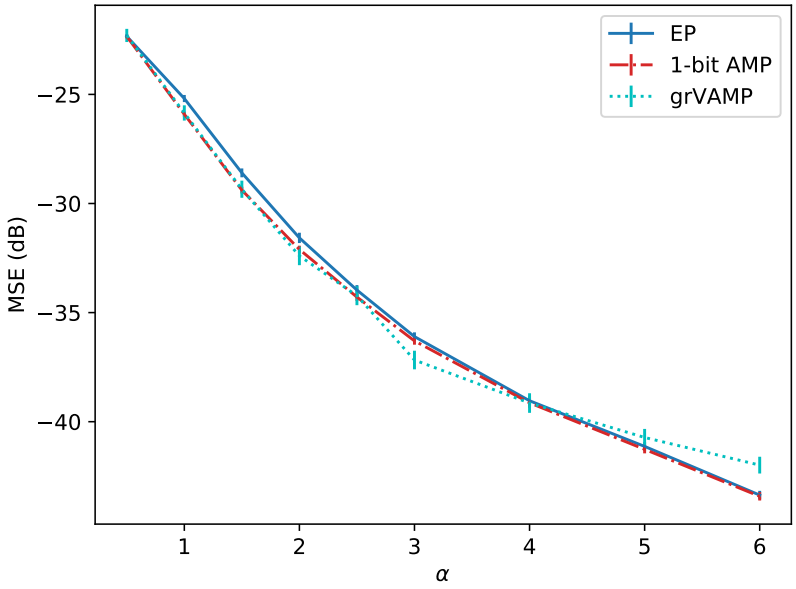


Figure 6.1: *MSE* resulting from sparse perceptron training from i.i.d. patterns using EP, 1-bit AMP and grVAMP based learning as a function of α. The parameters of the teacher are $N = 128$ and $\rho_0 = 0.25$ and the number of instances considered is $N_{samples} = 100$. All simulations converged and the MSE shown is averaged over all instances. The error bars are estimated as $\sigma/\sqrt{N_{samples}}$, where σ is the sample standard deviation of the *MSE* computed over all instances. Copyright (2021) by the American Physical Society. Reproduced with permission.

and diagonal elements of the covariance matrix having the same order of magnitude, so that correlations are not negligible.

Although, assuming the parameters N, ρ_0 and α to be equal and given the same values for ϵ_{stop} and for the maximum number of iterations in Gaussian EP, the estimation accuracy appears to be lower when considering this kind of patterns as compared to the previous case of i.i.d. Gaussian patterns, our results show that Gaussian EP still allows the student perceptron to learn the weights of the teacher fairly well. The increase in the MSE as compared to the case of CS reconstruction is to be expected, since the information reduction in the linear projections, due to the patterns being correlated, is worsened by the presence of the sign non-linearity. Both effects are combined in the data matrix $\mathbf{X}_\sigma$, as both patterns and labels enter in its definition, cfr. Eq. (6.4). However, it is worth noticing that Gaussian EP based training outperforms other message passing algorithms applied to the same problem. For example, the estimates of the means and of the variances of the weights of the teacher diverge if one uses 1bitAMP,

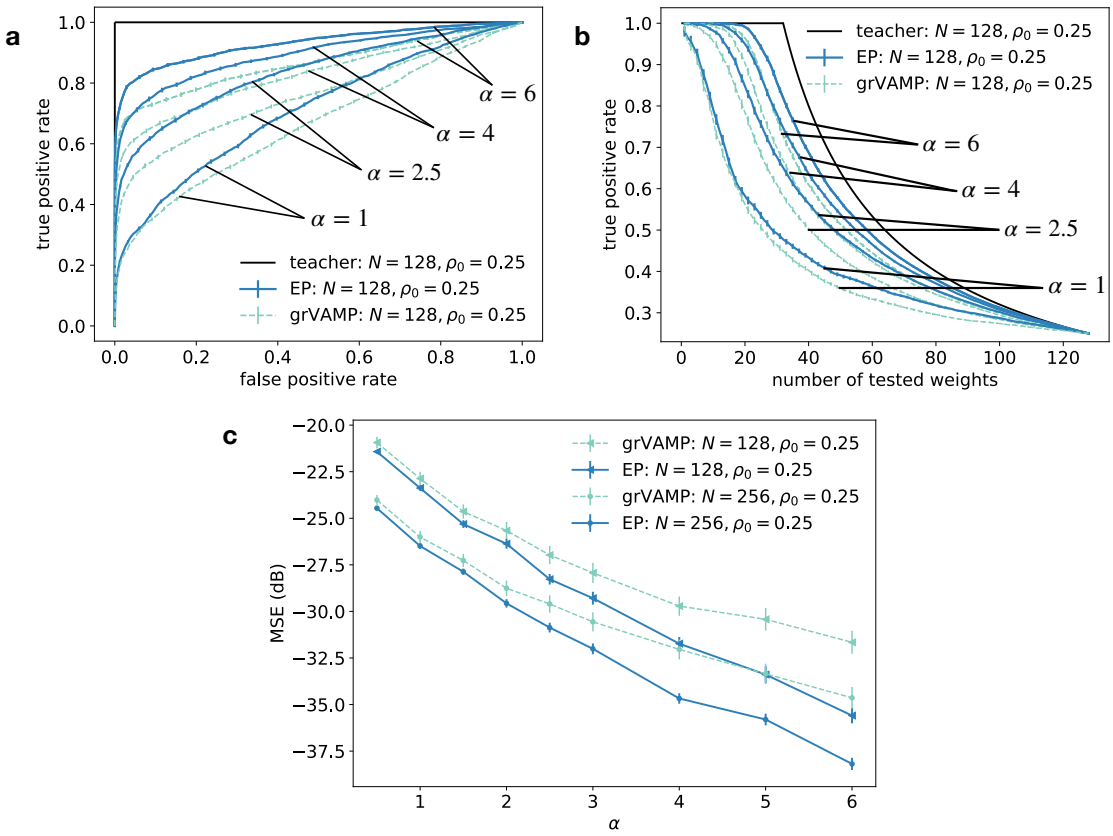


Figure 6.2: Sparse perceptron learning from correlated Gaussian patterns, with parameter $u = 1$, using EP and grVAMP. (a) ROC curves for various values of α. (b) Sensitivity plots for the same values of α considered in panel (a). For reference, in (a) and (b) the case of ideal variable selection by the teacher perceptron that provided the examples is also shown (black line). (c) MSE in dB for $N = 128$ and $N = 256$. In each plot, the mean values and the standard deviations are computed over the set of all N_{conv} instances for which convergence was achieved. The error bars are estimated as $\sigma/\sqrt{N_{conv}}$, where σ is the sample standard deviation over the same set of instances. Copyright (2021) by the American Physical Society. Reproduced with permission.

which prevents us from including it in our comparisons. Moreover, we find that Gaussian EP also outperforms grVAMP in this scenario and show this fact in Fig. 6.2, where we considered $N = 128$ weights, with density level of the teacher given by $\rho_0 = 0.25$. As in the previous case involving i.i.d. Gaussian patterns, the convergence parameters of grVAMP and EP were set to 10^{-8} and to 10^{-4}, respectively. Attempting to train the student perceptron setting a lower convergence threshold for grVAMP did not lead to a better estimate of the target weights. For EP, a damping equal to 0.999 was used. Under these values of the parameters, we show the fraction of converged trials achieved by EP and grVAMP in Table 6.1.

α	f_{EP}	$\sigma_{f_{EP}}$	f_{grVAMP}	σ_{grVAMP}
0.5	1	0	1	0
1.0	1	0	0.96	0.02
1.5	1	0	0.89	0.03
2.0	1	0	0.83	0.04
2.5	1	0	0.87	0.03
3.0	1	0	0.87	0.03
4.0	1	0	0.85	0.04
5.0	1	0	0.82	0.04
6.0	1	0	0.80	0.04

(a) $N = 128$

α	f_{EP}	$\sigma_{f_{EP}}$	f_{grVAMP}	σ_{grVAMP}
0.5	0.99	0.01	1	0
1.0	0.73	0.04	0.99	0.01
1.5	0.92	0.03	0.94	0.02
2.0	0.96	0.02	0.88	0.03
2.5	0.88	0.03	0.83	0.04
3.0	0.88	0.03	0.73	0.04
4.0	0.92	0.03	0.80	0.04
5.0	0.94	0.02	0.76	0.04
6.0	0.96	0.02	0.70	0.05

(b) $N = 256$

Table 6.1: Fraction of converged trials over a set of $N_{samples} = 100$ different instances of the teacher's weights and of the training set, with $N = 128$ and $\rho_0 = 0.25$. The patterns in the training set were drawn from a multivariate Gaussian distribution having covariance matrix given by (6.34). The quantity f_{EP} denotes the fraction of converged EP simulations and σ_{EP} is associated uncertainty, which was estimated as $\sqrt{f_{EP}(1 - f_{EP})/N_{samples}}$, whereas f_{grVAMP} and σ_{EP} denote the analogous quantities associated with grVAMP.

In terms of variable selection, we can see that EP based training of the student classifier leads to a more accurate estimate of the subset of nonzero weights as compared to grVAMP based training, as demonstrated by the *receiver operating characteristic* (ROC) curves shown in Fig. 6.2a and by the *sensitivity plots* shown in Fig. 6.2b, where the average of the quantities considered was computed over the set of the N_{conv} instances for which both algorithms achieved convergence and the error bars were estimated as the standard deviation over the same set of instances divided by $\sqrt{N_{conv}}$. We here give some details on how these curves are constructed for a single instance:

1. Given a set of weights of the teacher and given the parameters of the student at the

end of the training phase, we assign a score to each weight of the teacher, based either on their absolute value or on their estimated probability of being nonzero as provided by the parameters of EP and of grVAMP. In case one is interested in using the probability that a given weight is nonzero, the relevant expression is given by:

$$P_i^{\neq 0} = \left(1 + \left(\frac{1}{\rho} - 1\right)\sqrt{\frac{1+\lambda\Sigma_i}{\lambda\Sigma_i}}\, \mathrm{e}^{-\frac{\mu_i^2}{2\Sigma_i(1+\lambda\Sigma_i)}}\right)^{-1} . \tag{6.35}$$

If, on the one hand, the algorithm of interest is EP, then:

- μ_i, for $i = 1, \dots, N$, denotes the mean of the ith marginalized cavity distribution
- Σ_i denotes the variance of the ith marginalized cavity distribution

while if, on the other hand, the algorithm considered is grVAMP, then:

- μ_i is the ith component of the VAMP vector $\mathbf{r}_{1k}$,
- Σ_i is given by the VAMP variance γ_{1k}^{-1},

where the index i refers to vector components and the index k to the current iteration in VAMP (cfr. Sec. 3.7) and ρ and λ denote the values of the spike-and-slab prior.

2. We sort the weights of the teacher in decreasing order based on the scores specified at the previous point.

One interesting observation emerging from Fig. 6.2 is that the discrepancy between the ROC curves associated to EP and grVAMP increases as a function of α. Accordingly, the same behavior appears in the sensitivity plot, signaling a qualitative difference in the variable selection capabilities of the two algorithms in the Gaussian correlated pattern setup that we are considering here. This fact, in turn, implies that the fixed points of Gaussian EP describe the weights of the teacher more faithfully than those reached by grVAMP, as confirmed by the discrepancy between the MSEs associated with the two algorithms, which are shown in Fig. 6.2c, and allows us to conclude that, indeed, the EP and grVAMP approximations are significantly different. Notice that this fact is not (only) due to convergence issues, since the plots in Fig. were obtained by considering only instances where both algorithms converged. For instance, considering $\alpha = 2$, we obtain $MSE = (-25.7 \pm 0.2)$ dB in the case of grVAMP and $MSE = (-26.4 \pm 0.3)$ dB in the case of EP. In order to grasp this difference, let us consider the variances d_k of the EP approximating Gaussian factors ϕ_k, for $k = 1, \dots, N$, bearing in mind the relationship between EP and VAMP described in Sec. 4.8. In the EP approximation, these variances span several orders of magnitude, as it can be appreciated from Fig. 6.3, where we plot the histogram of the values taken by $d_1, \dots, d_N$ at the end of the training phase for a particular instance of the weights of the teacher and of the patterns given as input to

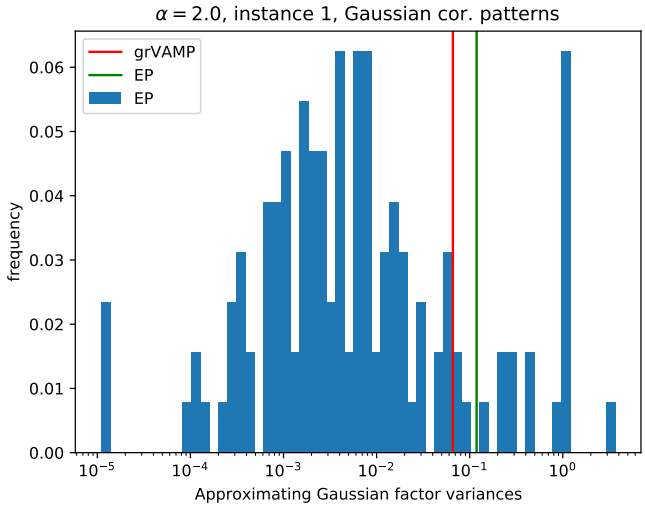
the student, for $\alpha = 2$. On the contrary, the VAMP ansatz in grVAMP constrains all variances to be equal, thus limiting the extent to which the posterior distribution can be well approximated. We show both the case of Gaussian correlated patterns and the case of i.i.d. Gaussian patterns in order to highlight the difference in their distributions in the two cases: in the Gaussian correlated case, the distribution of the variances is mostly spread over the range $10^{-4} \leq d \leq 1$, whereas in the i.i.d. Gaussian case the same distribution mostly concentrates around one peak. The discrepancy between the MSEs obtained from the two approximations becomes even larger as α is increased. The variances of the approximating factors still span several orders of magnitude, but, as we show in Fig. 6.4, tend to concentrate at lower values, in accordance with the better performance displayed by EP at large α. In both Figs. 6.3 and 6.4, the range of the parameters d_k is represented in logarithmic scale and the grVAMP estimate of the variances d_k is given by the vertical red line. For the sake of comparison, we also include the average EP estimate and its associated sample standard deviation, which correspond to the green vertical line and to the light green region, respectively.

We explained in Sec. 4.10 how the parameters of the prior can be approximately inferred during the EP iterations by minimizing the EP free energy using gradient descent and we applied the procedure to the online estimation of the density parameter of the spike-and-slab prior in the CS problem, as discussed in Sec. 5.3. The same strategy can be applied to the sparse perceptron learning problem in order to infer the value ρ_0 of the density level of the teacher weights, which one assumes not to be known a priori. Although we will only state results concerning the estimation of ρ_0 as obtained in the Gaussian EP framework, we notice that a similar expectation maximization strategy can also be implemented for 1bitAMP and grVAMP (see e.g. Ref. [29]).

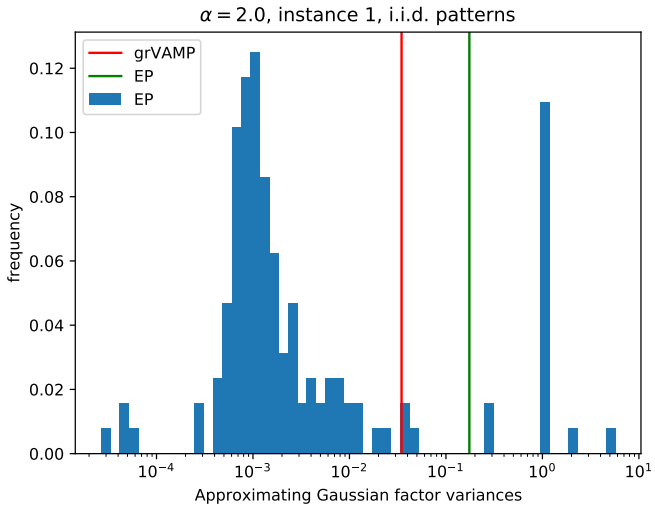
In Table 6.2, we show the EP estimate of the teacher density as obtained from a set of $N_{samples} = 100$ numerical experiments on a system with $N = 128$ and target value of the density level given by $\rho_0 = 0.25$. More precisely, we give its mean ρ_L and its standard deviation, obtained as its sample standard deviation divided by $\sqrt{N_{samples}}$. We also provide the relative discrepancy between ρ_0 and ρ_L, where we have not included the associated statistical uncertanties due to the fact that $\Delta\rho \gg \delta\rho_L$. In all simulations, the initial value of the parameter ρ of the spike-and-slab prior associated with the weights of the student was randomly drawn from a uniform distribution over the interval $0.05 \leq \rho \leq 0.95$ and we chose $\delta\rho = 10^{-5}$ for the learning rate of the gradient descent. For these values of the parameters, convergence was achieved in all simulations for all values of α. Overall, Table 6.2 shows that the EP estimate of ρ_0 tends to be quite accurate and that it improves as the number of patterns presented to the student perceptron increases.

6.5 Sparse perceptron learning from a noisy teacher

We now analyze the performance of Gaussian EP on the sparse perceptron learning problem when a small fraction of the binary labels is corrupted by noise. In this case, some patterns are mislabeled and the task of the student perceptron is to learn the



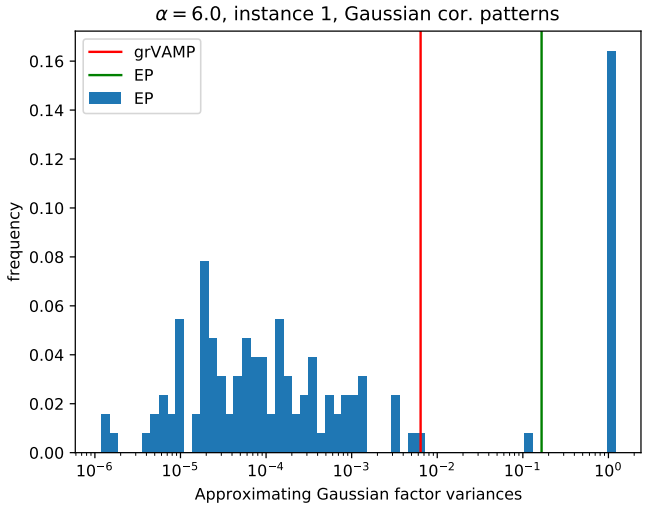
(a)



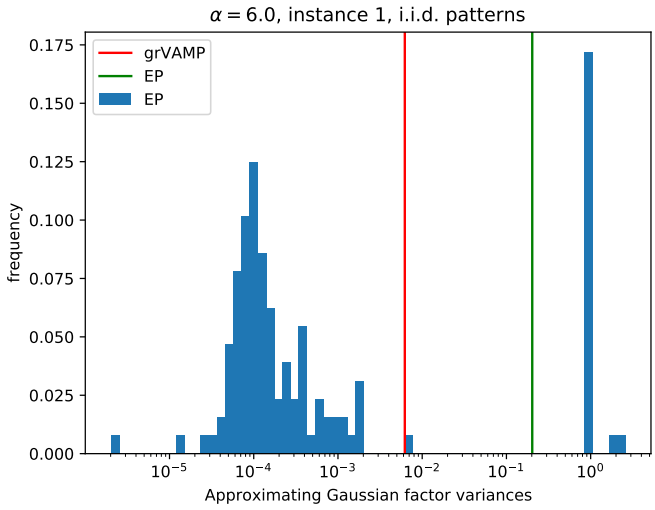
(b)

Figure 6.3: Examples of a set of variances in Gaussian EP and in VAMP at the end of the training on one instance for $N = 128$, $\rho_0 = 0.25$ and $\alpha = 2$.

correct classification rule from the knowledge of the corrupted examples and from the information that a (possibly unknown) fraction of the labels was wrongly assigned. As in the noiseless case, we consider a Bayes-optimal setting in order to avoid complicating the analysis. Therefore, we replace the theta pseudoprior with the theta mixture pseudoprior given in Eq. (6.11).



(a)



(b)

Figure 6.4: Examples of a set of variances in Gaussian EP and in VAMP at the end of the training on one instance for $N = 128$, $\rho_0 = 0.25$ and $\alpha = 6$.

In order to compare EP to grVAMP in this scenario, we implemented the theta mixture pseudoprior in both algorithms. Moreover, we include in our comparisons the variational algorithm proposed in Ref. [34], called R1BCS. This algorithm attempts to jointly recover the signal and a sparse noise vector producing the incorrect labels by means of an expectation maximization scheme. As in the previous section, we set the

α	$\rho_L \pm \delta\rho_L$ (i.i.d.)	$\Delta\rho/\rho_0$ (i.i.d.)	$\rho_L \pm \delta\rho_L$ (MVN)	$\Delta\rho/\rho_0$ (MVN)
2	0.191 ± 0.003	0.236	0.161 ± 0.004	0.341
3	0.220 ± 0.002	0.121	0.196 ± 0.004	0.206
4	0.234 ± 0.002	0.066	0.207 ± 0.003	0.182
5	0.240 ± 0.002	0.042	0.214 ± 0.003	0.144
6	0.242 ± 0.001	0.031	0.223 ± 0.003	0.115

Table 6.2: EP based estimation of the density ρ_0 of the teacher's weights for a perceptron with parameters $N = 128$ and $\rho_0 = 0.25$. The average and the standard deviation of the learned value of ρ_0 at convergence over all converged training simulations are denoted by ρ_L and $\delta\rho_L$, respectively, whereas $\Delta\rho$ is the absolute difference between ρ_0 and the estimate ρ_L. In each trial, the initial condition ρ_0 was drawn uniformly from the interval $0.05 \leq \rho \leq 0.95$.

precision parameter of the spike-and-slab to $\lambda = 1$ in grVAMP. In R1BCS, we set a convergence criterion such that $\|\boldsymbol{w}_{R1BCS} - \boldsymbol{w}^{old}_{R1BCS}\| < \epsilon_{stop}$, where $\boldsymbol{w}_{R1BCS}$ denotes the R1BCS estimate of the student perceptron after training is completed. In Gaussian EP, we set the parameter λ of the spike-and-slab prior to 10^{-4} and the damping factor to 0.99. In all algorithms, the convergence threshold ϵ_{stop} was set to 10^{-4}. In our numerical experiments, we mislabelled a number $K_{label} = (1 - \eta_0)M$ of examples, where the parameter η_0 denotes the true fraction of correctly assigned labels, which the students possibly needs to retrieve during the training process.

We consider both the case of perceptron learning from i.i.d. Gaussian patterns drawn from a standard Gaussian distribution and that from correlated Gaussian patterns constructed as in the previous section about noiseless examples, with covariance matrix given by (6.34). In both cases, we considered a set of 100 instances, each of which consisting of a different weight vector for the teacher and a different set of patterns. The rate of convergence within the threshold ϵ_{stop} was 1 for all algorithms in the case of i.i.d patterns and for R1BCS and EP in the case of multivariate Gaussian patterns. However, in the latter case, we observed a rate of failure in the convergence of grVAMP up to 15%.

We analyzed the variable selection capabilities of EP in the mislabeled examples scenario by computing the ROC curves and the sensitivity plots related to the same instances considered above, both those consisting of i.i.d. Gaussian patterns and those consisting of correlated Gaussian patterns. The ROC curves and sensitivity plots are associated with the weights of the student perceptron, where the number of weights was $N = 128$, the density level of the teacher was $\rho_0 = 0.25$ and the fraction of unflipped labels was taken as $\eta_0 = 0.95$. We show the resulting ROC curves for the i.i.d. pattern case in Fig. 6.5a and their associated sensitivity plots in Fig. 6.5b. Likewise, for the case of Gaussian correlated patterns, the relevant ROC curves are shown in Fig. 6.5a and the related sensitivity plots are shown in Fig. 6.5b.

In order to obtain the curves presented in Fig. 6.5, the weights of the teacher, $\boldsymbol{B}$, were sorted based on the absolute value of the weights of the student, $\boldsymbol{w}$. Concerning EP and

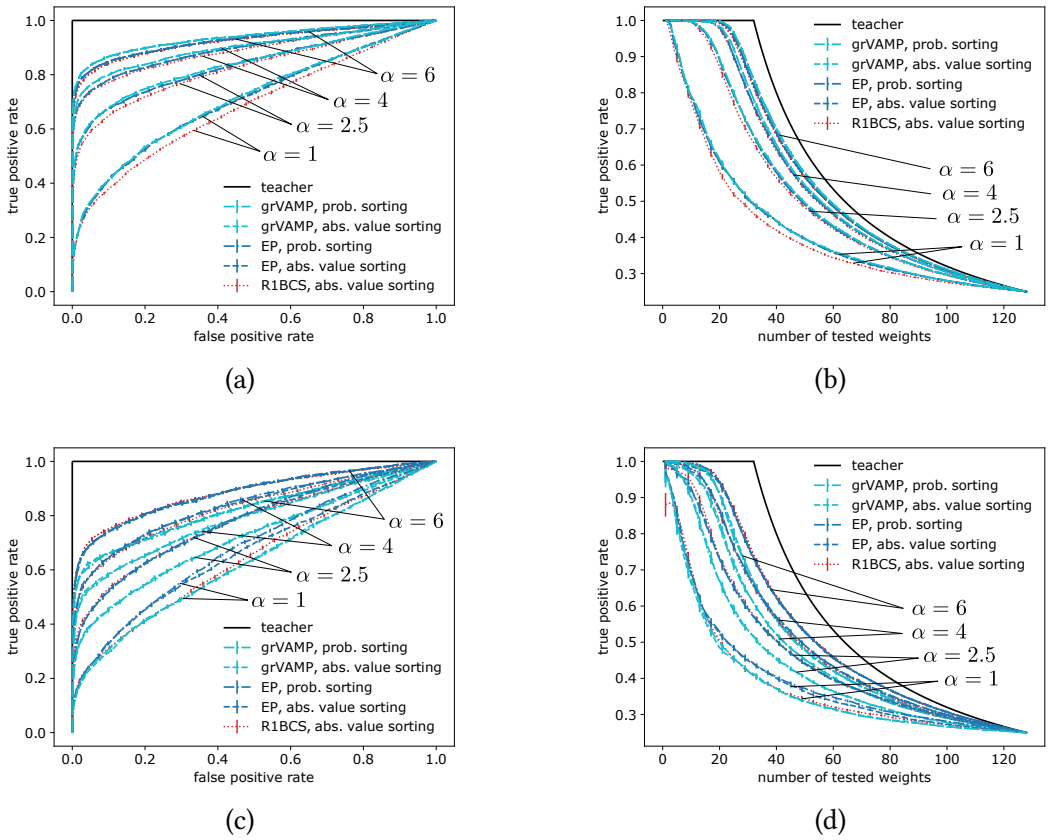


Figure 6.5: Sparse weight estimation from i.i.d. Gaussian patterns and from correlated Gaussian patterns, with parameters $N = 128$, $\rho_0 = 0.25$, $u = 1$, $\eta_0 = 0.95$. A fraction $(1 - \eta_0)$ of the labels are mislabeled. Comparison between the ROC curves associated with the student's estimate as obtained from R1BCS, grVAMP and EP (a,c) and between the related sensitivity plots (b,d). For reference, the case of ideal variable selection by the teacher perceptron that provided the examples is shown in black in all panels. The plotted quantities are the mean values computed over the set of all $N_{samples}$ instances and the error bars are estimated as $\sigma/\sqrt{N_{samples}}$, where $N_{samples} = 100$ and σ denotes the sample standard deviation over all instances. Copyright (2021) by the American Physical Society. Reproduced with permission.

grVAMP, we also considered the score expressed in Eq. (6.35) as a sorting criterion. However, this did not lead to noticeably different results, as the curves obtained with the two criteria approximately overlap.

From the ROC curves and sensitivity plots in the panels of Fig. 6.5, we see that EP and grVAMP are mostly comparable in terms of their true positive ratio. However, for large α, the values associated with EP are smaller in the i.i.d. case. This fact is most likely to be attributed to the presence of numerical effects, as we will argue in

α	AUC_{EP}	AUC_{grVAMP}	AUC_{R1BCS}
0.5	0.621 ± 0.005	0.627 ± 0.005	0.595 ± 0.005
1.0	0.706 ± 0.005	0.710 ± 0.005	0.682 ± 0.004
1.5	0.770 ± 0.005	0.777 ± 0.005	0.746 ± 0.005
2.0	0.806 ± 0.005	0.809 ± 0.005	0.792 ± 0.004
2.5	0.835 ± 0.004	0.840 ± 0.004	0.824 ± 0.005
3.0	0.860 ± 0.004	0.865 ± 0.005	0.854 ± 0.004
4.0	0.893 ± 0.004	0.899 ± 0.004	0.887 ± 0.004
5.0	0.913 ± 0.004	0.920 ± 0.004	0.91 ± 0.004
6.0	0.927 ± 0.003	0.936 ± 0.003	0.923 ± 0.003

(a) AUC (i.i.d. patterns)

α	AUC_{EP}	AUC_{grVAMP}	AUC_{R1BCS}
0.5	0.588 ± 0.006	0.579 ± 0.006	0.559 ± 0.005
1.0	0.661 ± 0.005	0.628 ± 0.006	0.641 ± 0.006
1.5	0.727 ± 0.006	0.685 ± 0.006	0.694 ± 0.006
2.0	0.732 ± 0.007	0.696 ± 0.006	0.734 ± 0.006
2.5	0.775 ± 0.007	0.719 ± 0.006	0.775 ± 0.006
3.0	0.788 ± 0.007	0.742 ± 0.007	0.793 ± 0.006
4.0	0.834 ± 0.007	0.788 ± 0.006	0.828 ± 0.005
5.0	0.856 ± 0.006	0.807 ± 0.006	0.851 ± 0.006
6.0	0.882 ± 0.005	0.825 ± 0.007	0.885 ± 0.005

(b) AUC (patterns from MVN)

Table 6.3: (a) AUC scores associated with the ROC curves shown in Fig. 6.5a, which correspond to EP, grVAMP and R1BCS based training from i.i.d. Gaussian patterns with a small fraction of mislabeled examples. (b) AUC scores associated with the ROC curves shown in Fig. 6.5c, corresponding to EP, grVAMP and R1BCS based training from correlated Gaussian patterns with mislabeled examples. In both (a) and (b), $N = 128$, $\rho_0 = 0.25$, $u = 1$ and the consistency level of the labels assigned to the patterns is $\eta_0 = 0.95$.

more detail below. In terms of performance as compared to R1BCS, we find that both EP and grVAMP significantly outperform R1BCS when training is conducted on i.i.d. patterns, especially at low α. In order to corroborate this fact, we give the values of the areas under the the ROC curves (AUC), which are reported in Table 6.3a. From the table, we can see that the relative discrepancy between the AUCs does not exceed 0.008 when considering EP and grVAMP, but can be as large as 0.03 when considering EP and R1BCS.

The situation appears to be qualitatively different in the case of correlated Gaussian patterns, where the ROC curves and sensitivity plots related to EP and to R1BCS are mostly comparable and exceed those related to grVAMP in terms of true positive rate.

Overall, these results show that the variable selection properties of EP tend to be more robust than those of grVAMP in the correlated Gaussian pattern regime. This observation is confirmed by the AUC reported in Tab. 6.3b and agrees with what we observed in the ideal case without mislabeling. Here, the discrepancy between EP and grVAMP is large for large α and is maximum at $\alpha = 6$.

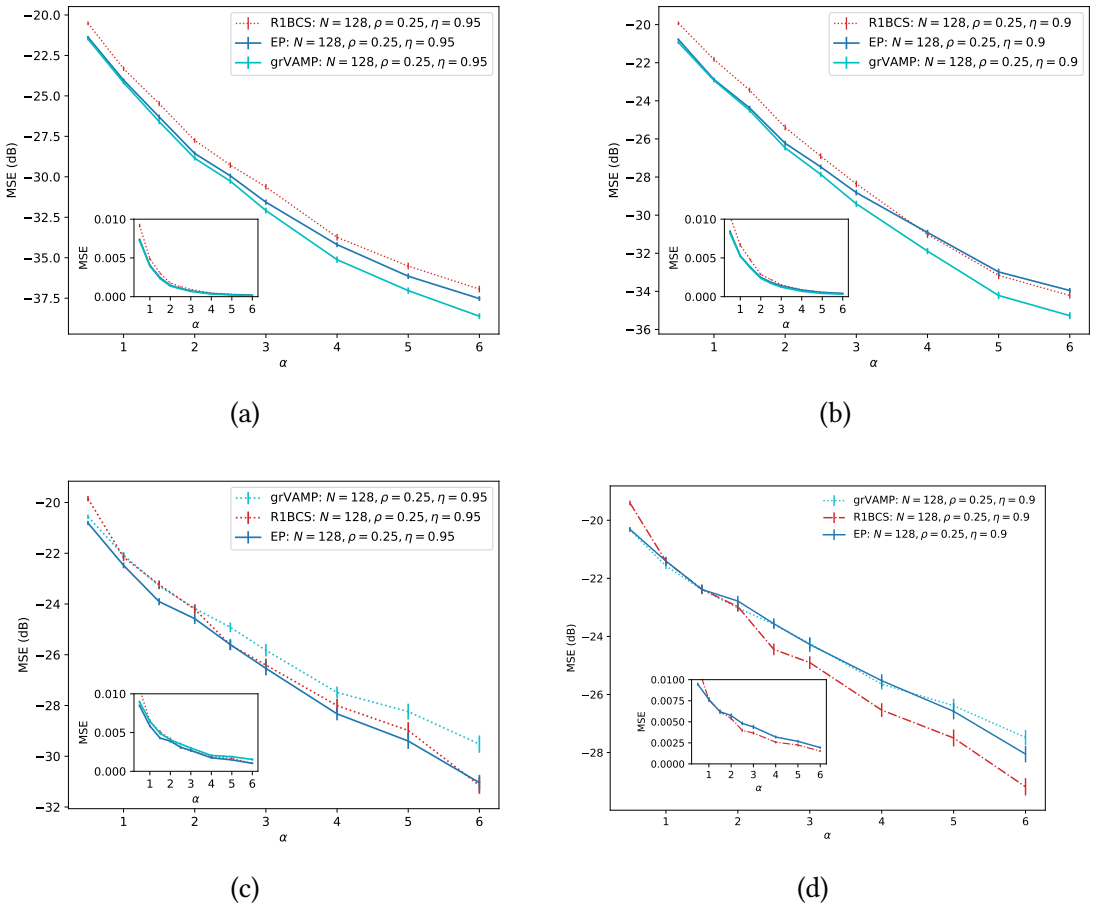


Figure 6.6: Sparse perceptron learning from a training set with a fraction $(1 - \eta_0)$ of mislabeled examples: comparison between EP, grVAMP and R1BCS in terms of their MSE for (a) i.i.d. Gaussian patterns, $\eta_0 = 0.95$, (b) i.i.d. Gaussian patterns, $\eta_0 = 0.9$, (c) correlated Gaussian patterns, $\eta_0 = 0.95$ and (d) correlated Gaussian patterns, $\eta_0 = 0.9$. In all panels, $N = 128$ and $\rho_0 = 0.25$. The mean squared errors plotted are averaged over the set of all $N_{samples}$ instances considered, which are the same as in Fig. 6.5, and the error bars are estimated as $\sigma/\sqrt{N_{samples}}$, where σ denotes the sample standard deviation over all instances. Copyright (2021) by the American Physical Society. Reproduced with permission.

In order to see how these variable selection properties affect the student's retrieval of the teacher's weights, we show the MSE in Fig. 6.6 for both $\eta_0 = 0.95$ and $\eta_0 = 0.9$.

Notice that, as the instances considered are the same, the case of i.i.d. Gaussian patterns in Fig. 6.6a corresponds to the ROC curves shown in Fig. 6.5a and to the sensitivity plots in Fig. 6.5b. Likewise, Fig. 6.6c, which refers to correlated Gaussian patterns, is related to Figs. 6.5c and 6.5d. We show the MSE both using a dB scale (see the main plots) and using a linear scale (insets). By inspecting both scales, we find that, although the MSE of EP is larger than that of grVAMP at large α in the i.i.d. pattern setup when looking at the dB scale (see Figs. 6.6a and 6.6b), the difference is very small and therefore is not noticeable on a linear scale. The slightly lower performance of EP is in agreement with Figs. 6.5a and 6.5b and should be ascribed to numerical effects. This belief is not only based to the smallness of the MSE and of their discrepancies, but also on the relationship of the VAMP approximation to Gaussian EP, which was discussed in Sec. 4.8. In particular, we recall that VAMP can be seen as a special case of Gaussian EP where all approximating factors are univariate Gaussian factors having the same variance. As such, the Gaussian EP approximation is expected to be more general than the VAMP ansatz, as we discussed above when interpreting our results concerning perceptron training from correlated Gaussian patterns in the noiseless regime. In turn, this observation explains the differences between the MSE of EP and grVAMP in Fig. 6.6c for $\eta_0 = 0.95$, which not only appear on a dB scale but also on a linear one. Once again, this fact agrees with the results found for the ROC curves in Fig. 6.5c. However, as soon as the fraction of mislabeled examples is large enough, EP and grVAMP become comparable in terms of MSE regardless of the specific value of α, as it can be deduced from Fig. 6.6d.

As already mentioned above, the fraction η_0 of correctly assigned labels is not known, in general, and the same applies to the parameter ρ_0. We here demonstrate that EP is able to find an accurate estimate of η_0 by iteratively updating the parameter η of the prior using the gradient descent update given in Eq. (6.32). In order to see this, we considered 100 instances and sampled a different initial condition for the parameter η by drawing it uniformly from the interval $0.5 < \eta < 1$. We checked that ρ_0 and η_0 can be learned simultaneously, where the initial condition for ρ was chosen as described at the end of Sec. 6.4. We show the EP estimate of the parameters ρ_0 and η_0 in the case of i.i.d. patterns in Tab. 6.4a and in the case of correlated Gaussian patterns in Tab. 6.4b.

Among the algorithms presented in our comparisons, EP proves to be quite efficient in terms of computational complexity. For comparison, R1BCS has a computational cost $O((1+\alpha^3)N^3)$, as both a $N\times N$ and a $M\times M$ matrix are inverted, whereas EP is dominated by $O((1+\alpha)N^3)$ elementary operations if the formulation with rigid linear constraints of Sec. 4.4 is used. In the case of EP, the cost is due both to the inversion of the covariance matrix of the approximated distribution Q and to the computation of N^2 scalar products between M-dimensional vectors in Eq. (4.38). This makes EP especially advantageous as compared to R1BCS in the large α regime. However, in general, EP tends to be slower than grVAMP for large sets of weights of the teacher and student perceptrons. In fact, in spite of the fact that grVAMP has cubic computational complexity as a function of N, the initial SVD associated with this cost needs only be performed once (cfr. Secs.

α	ρ_L	$\Delta\rho/\rho_0$	η_L	$\Delta\eta/\eta_0$
2.5	0.229 ± 0.004	0.08	0.964 ± 0.001	0.02
3.0	0.234 ± 0.003	0.07	0.957 ± 0.003	0.007
4.0	0.247 ± 0.004	0.01	0.9584 ± 0.0005	0.009
5.0	0.249 ± 0.003	0.003	0.9561 ± 0.0007	0.006
6.0	0.252 ± 0.003	0.007	0.9544 ± 0.0004	0.005

(a) i.i.d. patterns

α	ρ_L	$\Delta\rho/\rho_0$	η_L	$\Delta\eta/\eta_0$
2.5	0.206 ± 0.006	0.2	0.951 ± 0.002	0.0006
3.0	0.208 ± 0.006	0.2	0.953 ± 0.001	0.003
4.0	0.228 ± 0.005	0.09	0.953 ± 0.001	0.003
5.0	0.23 ± 0.005	0.08	0.9526 ± 0.0005	0.003
6.0	0.236 ± 0.004	0.05	0.9529 ± 0.0004	0.003

(b) patterns from MVN

Table 6.4: Values of the η_0 parameter of the theta mixture pseudoprior estimated by the student perceptron during the training phase when using EP to learn the weights of the teacher (a) from i.i.d. Gaussian patterns and (b) from correlated Gaussian patterns for $N = 128$. The estimated value of η_0 is denoted as η_L, the target value being $\eta_0 = 0.95$, whereas the target value of ρ is $\rho_0 = 0.25$. $\Delta\rho$ and $\Delta\eta$ denote the absolute difference between the target value and the estimate of the parameters ρ and η, respectively.

3.7) and can thus be neglected if N is small enough. Then, the remaining part of the algorithm is dominated by a matrix-vector product and, therefore, its computational cost is quadratic in N. In order to give an idea of the running times involved for the implementations considered in our comparisons, which can be found at Refs. [105, 123–125], we show the simulation times of R1BCS (t_{R1BCS}), EP (t_{EP}) and grVAMP (t_{grVAMP}) in Tab. 6.5 for $N = 128$. In grVAMP, we set the number of iterations of the inner VAMP module to 2000 and the maximum number of iterations of the outer MMSE module to 1000. The fact that t_{grVAMP} appears to be larger than t_{EP} for the simulated size that we considered is related to the VAMP module of grVAMP being run one time per iteration.

6.6 Sparse perceptron learning from temporally correlated patterns

We now consider another example of correlated patterns produced as follows. N diluted perceptrons connected to form a recurrent network with no self-loops receive binary inputs $\boldsymbol{x}$. generated according to a Glauber dynamics at zero temperature. Let $\mathbf{B} \in \mathbb{R}^{N\times(N-1)}$ be the weight matrix of the network and $\boldsymbol{B}_i$ its ith row, which corresponds to the set of weights that the ith perceptron of the network receives from all

α	t_{EP} (s)	t_{grVAMP} (s)	t_{R1BCS} (s)
0.5	2.7 ± 0.2	126.7 ± 0.8	6.5 ± 0.2
1.0	2.6 ± 0.03	134.6 ± 0.9	12.0 ± 0.3
1.5	3.74 ± 0.04	147.0 ± 1.0	22.2 ± 0.5
2.0	5.65 ± 0.08	159.0 ± 1.0	47.6 ± 1.0
2.5	6.97 ± 0.09	175.7 ± 1.0	94.9 ± 2.0
3.0	8.3 ± 0.1	212.9 ± 20.0	150.9 ± 3.0
4.0	10.2 ± 0.1	258.3 ± 10.0	373.4 ± 7.0
5.0	12.2 ± 0.1	311.6 ± 10.0	628.0 ± 10.0
6.0	15.6 ± 0.2	353.9 ± 10.0	1052.9 ± 20.0

(a) i.i.d. patterns

α	t_{EP} (s)	t_{grVAMP} (s)	t_{R1BCS} (s)
0.5	2.7 ± 0.2	139.2 ± 2.0	6.7 ± 0.2
1.0	3.7 ± 0.1	155.7 ± 2.0	14.4 ± 0.4
1.5	4.8 ± 0.1	206.8 ± 20.0	27.0 ± 0.7
2.0	6.2 ± 0.2	289.0 ± 40.0	53.6 ± 1.0
2.5	7.7 ± 0.2	236.3 ± 20.0	105.8 ± 3.0
3.0	9.5 ± 0.3	294.3 ± 30.0	158.5 ± 4.0
4.0	11.4 ± 0.3	366.9 ± 40.0	379.4 ± 8.0
5.0	14.1 ± 0.4	423.5 ± 40.0	592.7 ± 10.0
6.0	15.9 ± 0.4	498.5 ± 50.0	1057.2 ± 20.0

(b) patterns from MVN

Table 6.5: Simulation time for the EP and R1BCS based sparse weight learning from (a) i.i.d. Gaussian patterns and from (b) correlated Gaussian patterns. In both (a) and (b), a small fraction of examples were mislabeled. In simulations, parameters were set as $u = 1$ and $\eta_0 = 0.95$, while the EP damping factor was equal to 0.99. The uncertainties were estimated as $\sigma/\sqrt{N_{conv}}$, where σ is the sample standard deviation over the set of converged trials and N_{conv} denotes the number of converged simulations.

other perceptrons. A schematic representation of such a network is given in Fig. 6.7 Starting at $t = 0$ from a random vector $x_0 = \text{sign}(\xi_0)$, where $\xi_0 \sim \mathcal{N}(\xi; 0, \mathbf{I})$, the patterns at discrete times $t = 1, 2, \ldots, T$ can be dynamically generated either synchronously or asynchronously:

- **Synchronous update.** Given a pattern $\boldsymbol{x}^t$ at time t, all perceptrons compute their own outputs at time $t + 1$ according to the Glauber dynamics:

$$z_i^t = \boldsymbol{B}_i^\top \boldsymbol{x}_{\setminus i}^t, \tag{6.36}$$

$$x_i^{t+1} = \text{sign}\left(z_i^t\right), \tag{6.37}$$

  where $\boldsymbol{x}_{\setminus i}^t$ is the binary classification labels produced by all perceptrons except the

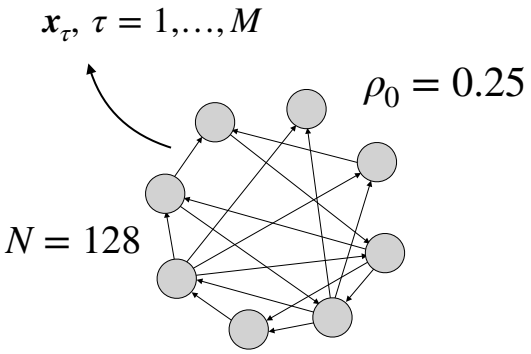


Figure 6.7: Schematic representation of a recurrent network of teacher perceptrons with diluted weights generating M pattern vectors dynamically.

i-th one at the current time t.

- **Asynchronous update.** At each time $t = 1,2,\ldots,T$, one perceptron is selected at random with uniform probability and generates a binary label according to Eq. (6.37). Thus, given a pattern vector at the current time t, the pattern vector at time $t+1$ has all components equal to those at the current time step, except for the ith one, which is replaced with the label generated by the perceptron selected. In order to adjust the extent to which consecutive pattern vectors are temporally correlated, we fix the Hamming distance between subsequent patterns to a desired level and, given the current pattern, run the dynamics (6.37) and only store the candidate pattern vector when the target Hamming distance is achieved.

In practice, it is important to select the nonzero weights of the recurrent network at random in order to avoid that the synchronous dynamics converges to a periodic attractor. Indeed, by selecting a number $N = 128$ of perceptrons with connection density level $\rho_0 = 0.25$, this problem did not occur.

Considering again the teacher-student paradigm and a noiseless setup, we compared EP and grVAMP in terms of training accuracy of the perceptrons in a student network from the examples generated by a teacher recurrent network. Both networks consist of $N = 128$ perceptrons and we set the teacher density level to $\rho_0 = 0.25$. The damping parameter of EP was set to 0.999 and the convergence threshold was set to 10^{-4} for both EP and grVAMP. In addition, the precision parameter of the spike-and-slab was set to $\lambda = 1$ both in EP and in grVAMP.

We quantified the accuracy of learning the target classification rule from synchronously generated patterns and from asynchronously generated patterns. In the presence of synchronously updated patterns, all simulations run on the perceptrons of the student network converged within the threshold specified both for EP and for grVAMP. We analyzed the variable selection properties and learning accuracy of the two algorithms in the synchronous and asynchronous pattern regimes by computing the average ROC

curves, the average MSEs and the statistical uncertanties associated with the perceptrons of the student network. The sorting criterion of the weights of the teacher perceptrons was based on their absolute values when computing the ROC curves.

In the synchronous update setup, both EP and grVAMP were able to properly identify the same number of teacher's weights. This can be seen from the overlapping vertical portion of the ROC curves shown in Fig. 6.8a. Therefore, the teacher-student MSE as a function of α exhibits comparable values for the two algorithms. As can be seen in Fig. 6.8b, a similar picture was observed when the patterns were obtained from an asynchronous generative process, where patterns vectors were included among the examples only after a "full sweep" of as many updates as the number of perceptrons in the teacher network. In this situation, patterns were weakly correlated and convergence of grVAMP was still very good. In fact, the observed convergence rate was larger than 95% for all values of α that we considered.

We now show our results on simulations performed with asynchronously generated patterns, where we set a Hamming distance $d_H(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t) = 10$ between subsequent pattern vectors. This choice corresponds to a Pearson correlation coefficient $r_{Pearson} = 0.84$ between pattern vectors $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$. In this case, the observed convergence rate of the student perceptrons dropped significantly for grVAMP, whereas the convergence rate for EP was mostly unaffected, as shown in the inset of Fig. 6.8d. Taking into account *all* perceptrons of the student network (both those for which grVAMP converged and those for which it did not), we find that grVAMP is not able to accurately estimate the weights of the teacher perceptrons with strongly correlated patterns. We show the ROC curves obtained in this case in Figs. 6.8c and the related MSE in Fig. 6.8d. The poor performance of the grVAMP estimate is related to the low convergence rate of the algorithm, as can be deduced from the fact that when both EP and grVAMP converged their estimates of the teacher weights were similar.

Analogously to what we saw in the previous sections, the student perceptrons were able to infer the density of the weights of their teachers quite accurately using EP on the temporally correlated patterns considered here. The parameter η_0 of the theta mixture pseudoprior could also be learned if noise was injected on the labels, provided that the noise level $1 - \eta_0$ was not too large, similarly to what we observed in Sec. 6.5 for i.i.d. Gaussian patterns and for the correlated Gaussian patterns. We provide the EP estimate of the noise level when $\eta_0 = 0.9$ and some examples of the estimated values of the density level ρ_0 at large α in Tab. 6.6. The parameter η_0 is inferred quite well for weakly correlated patterns (synchronous case), but is overestimated when patterns are strongly correlated (asynchronous case with $d_H = 10$).

α	η_L (synchr.)	$\Delta\eta/\eta_0$ (synchr.)	η_L (asynchr.)	$\Delta\eta/\eta_0$, (asynchr.)
0.5	0.86 ± 0.02	0.05	0.979 ± 0.004	0.09
1.0	0.96 ± 0.01	0.06	0.9945 ± 0.0009	0.1
1.5	0.927 ± 0.007	0.03	0.991 ± 0.001	0.1
2.0	0.899 ± 0.008	0.001	0.983 ± 0.001	0.09
2.5	0.893 ± 0.007	0.008	0.975 ± 0.001	0.08
3.0	0.901 ± 0.001	0.001	0.967 ± 0.001	0.07
4.0	0.9036 ± 0.0008	0.004	0.950 ± 0.001	0.06
5.0	0.9087 ± 0.0008	0.01	0.946 ± 0.001	0.05
6.0	0.9108 ± 0.0006	0.01	0.9389 ± 0.0008	0.04

(a)

Value of α	Estimated ρ^{synchr}	Estimated ρ^{asynchr}, $d_H = 10$
4.0	0.224 ± 0.002	0.208 ± 0.002
5.0	0.233 ± 0.002	0.223 ± 0.002
6.0	0.237 ± 0.001	0.231 ± 0.001

(b)

Table 6.6: (a) Estimate of the parameter η_0 of the theta mixture pseudoprior resulting from sparse weight learning from the same patterns as in Fig. 6.8. The true unknown value of η_0 is 0.9. (b) Estimated value of the density of the weights of single perceptrons for large α and with no mislabeling of the examples, both with synchronously updated patterns and with asynchronously updated patternes, where, in the latter case, the Hamming distance between pattern vectors at consecutive time steps was fixed as $d_H = 10$. In both (a) and (b), $N = 128$ and $\rho_0 = 0.25$.

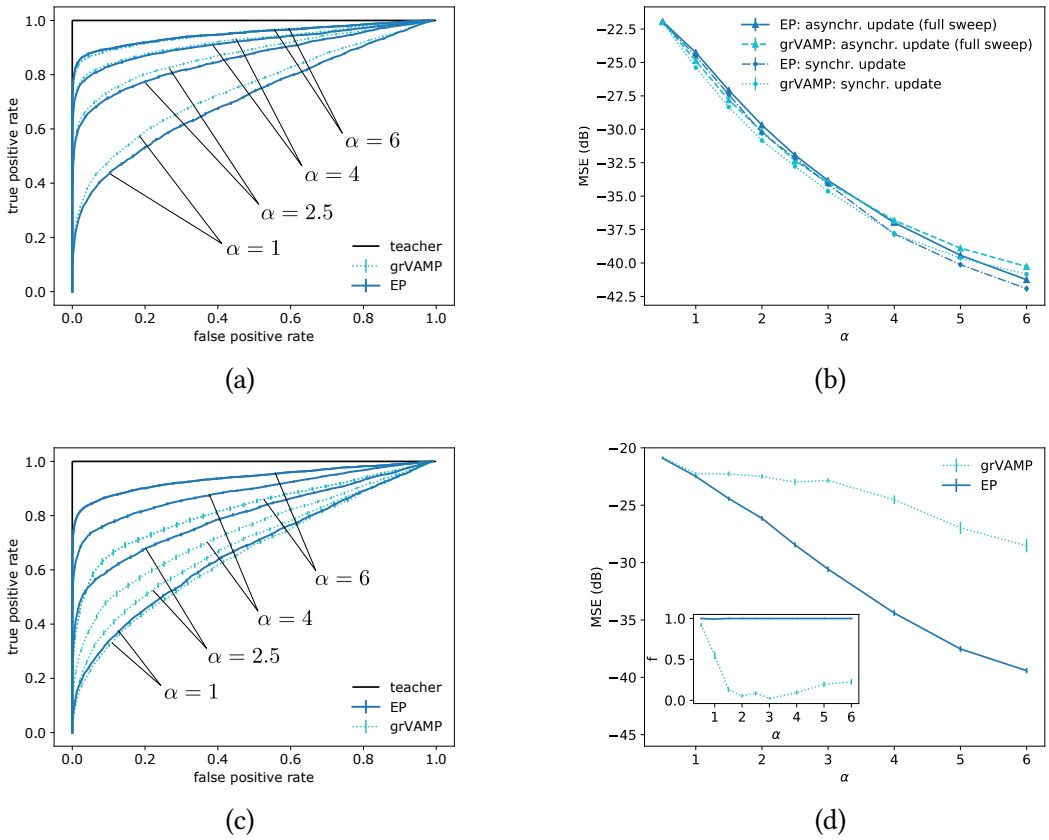


Figure 6.8: Performance of grVAMP and of EP led sparse weight retrieval from correlated binary patterns generated by means of a Glauber dynamics at zero temperature. Panels (a) and (b) refer to *weakly correlated* patterns, while panels (c) and (d) refer to *strongly correlated* patterns. Both in the teacher network and in the student network, $N = 128$ and $\rho_0 = 0.25$. In each panel, the mean values and the uncertainties were evaluated over the whole set of N perceptrons. The error bars were estimated as $\sigma/\sqrt{N}$, where σ denotes the sample standard deviation computed over the set of all N student perceptrons completing the learning task. (a) ROC curves associated with the learned weights of each student perceptron for several values of α for synchronously updated patterns. In this case, convergence was achieved by all perceptrons during the training task. (b) MSE (in dB) related to the synchronously updated patterns of (a) and to the case where patterns are updated asynchronously and included in the training set only after each perceptron is selected to yield the corresponding update ("full sweep' update'). (c) ROC curves related to the student's weights as learned from a training set consisting of asynchronously updated patterns, for the same values of α as in panel (a). The Hamming distance between pattern vectors at consecutive times was set to 10. (d) MSE (in dB) related to the simulations over the same instances considered in panel (c). The fraction of perceptrons for which the training task converged is shown in the inset of panel (d). Copyright (2021) by the American Physical Society. Reproduced with permission.

Chapter 7

Conclusions and open questions

In this PhD dissertation, we analyzed Gaussian EP as an efficient computational scheme to solve sparse linear estimation problems and tested it on two important applications: compressed sensing (CS) reconstruction and the problem of learning a binary classification rule using a diluted Bayesian classifier.

In the first problem, we studied the CS reconstruction threshold of Gaussian EP when incorporating a L_0 regularization in the approximation. Moreover, we showed in a simple case of correlated sensing matrices that the Gaussian EP threshold is robust against the presence of structure in the measurement matrix and that EP outperforms several state-of-the art algorithms, which either fail to converge or, despite converging, are unable to successfully retrieve the target signal.

In the sparse perceptron learning problem, we used Gaussian EP in order to train a sparse Bayesian perceptron from a set of examples under different conditions. We took advantage of a novel Gaussian EP implementation which allows to significantly reduce the computational cost of the EP algorithm by leveraging the linear dependence between variables. For the sparse perceptron, this improvement led to the cost being $O((1+\alpha)N^3)$ rather than $O((1+\alpha)^3N^3)$, which is particularly useful when training the perceptron on large training sets. We compared the performance of EP to algorithms of the AMP type and of the VAMP type when correlated inputs are provided to the perceptron to be trained and found a significantly better performance of EP in terms of convergence, variable selection properties and accuracy of the fixed points reached at the end of the training phase. Moreover, the algorithm appeared to be robust against noise as long as the noise level was small enough. In the opinion of the author, these results provide a firm step towards the development of effective approximate inference algorithms that can be applied to data sets with statistical structure.

The work presented in this dissertation complements other studies concerning EP [126] and opens the way to several research directions, both in terms of applications

and on the theoretical and computational side. A few relevant applications outside of optical and medical imaging are, for instance, supervised and unsupervised learning in neural networks.

Indeed, with regards to supervised learning, the application of EP to binary classification could be extended to the case of shallow neural networks with one hidden layer, for instance *soft committee machines* [127, 128], as well as to deeper architectures, where the hidden layer could have ReLU activation functions, as commonly found in deep learning. Here, one may be interested in estimating the inputs to the network and the states of the hidden units given the knowledge of the weights and of the output, similarly to the multilayer extension of the AMP algorithm [129] and of the VAMP algorithm [130, 131]. Moreover, assessing the generalization error associated with EP based training and understanding the possible effects of neural network overparameterization on the performance of EP in the teacher-student scenario are questions that were not addressed in this PhD thesis and that could be worth pursuing.

On the unsupervised learning side, an interesting application is provided by the restricted Boltzmann machine (RBMs) [132], which is a generative model defined on a bipartite graph with two sets of random variables called visible units and hidden units. As the graph is bipartite, there are no connections between units of the same kind. The visible units are assumed to reproduce the observed data, the statistical dependencies of which are set by means of the hidden layer. RBMs are trained by maximizing the likelihood of the parameters of the model (i.e., the connections between the visible layer and the hidden layer and the biases of the units) given the observations of a training set [133]. Computing the gradient of the likelihood requires matching the empirical moments of the data distribution to those of the RBM distribution. As the latter distribution is typically intractable, its moments are usually estimated using Monte Carlo sampling [134, 135]. Recently, some TAP based mean fields methods have been proposed [136], which suggests that the EP ansatz would be a promising candidate for RBM training too. In fact, as we argued in this PhD thesis, EP is a flexible and expressive approximation scheme and appears to be particularly advantageous due to its ability to take into account non trivial correlations present in the data to be modeled.

Finally, two very important theoretical issues to be addressed in future research consist in understanding the convergence of EP based schemes and in finding ways to modify the EP update rules in order to reduce the computational burden of the algorithm, while ensuring that the algorithm still converges to the same fixed points. In this sense, the derivation of alternative iterative schemes exploiting the fact that the fixed points of EP are stationary points of the EP free energy seems a promising direction worth pursuing. A tool which has proven to be effective in order to track the convergence of message passing algorithms is the framework known as *state evolution* (or *density evolution*). To date, this framework has mostly been used to predict the performance of AMP and VAMP based algorithms [82, 137–139] and has only recently begun to be applied to EP (see, e.g., Ref. [140]). Therefore, an interesting direction for future research concerns using state evolution to characterize the convergence and the performance

of the EP formulations presented in this thesis as applied to several linear estimation problems, including those studied in this dissertation.

Bibliography

[1] J. Tan, B. Mailhe, Q. Wang, and M. S. Nadar. "Generalized approximate message passing algorithms for sparse magnetic resonance imaging reconstruction." U.S. Patent 9542761. 2017.

[2] E. Gouillart, F. Krzakala, M. Mézard, and L. Zdeborová. "Belief-propagation reconstruction for discrete tomography." In: *Inverse Problems* 29.3 (2013), p. 035003.

[3] A. P. Muntoni, R. D. H. Rojas, A. Braunstein, A. Pagnani, and I. Pérez Castillo. "Non-convex image reconstruction via expectation propagation." In: *Phys. Rev. E* 100 (3 2019), p. 032134.

[4] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. 2013. ISBN: 978-0-8176-4947-0.

[5] E. Candes and J. Romberg. "L1-magic: recovery of sparse signals via convex programming." User's guide of the l1-MAGIC code (last accessed: July 8, 2021). 2005.

[6] G. K. Yonina C. Eldar. *Compressed Sensing: Theory and Applications*. 2012.

[7] C. Shannon. "Communication in the presence of noise." In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders. "Atomic decomposition by basis pursuit." In: *SIAM J. Sci. Comput.* 20.1 (1998), pp. 33–61. ISSN: 1064-8275.

[9] R. Tibshirani. "Regression shrinkage and selection via the lasso." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246.

[10] E. J. Candes and J. Romberg. "Quantitative robust uncertainty principles and optimally sparse decompositions." In: *Foundations of Computational Mathematics* 6.2 (2006), pp. 227–254.

[11] E. J. Candes, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information." In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509.

[12] E. J. Candès, J. K. Romberg, and T. Tao. "Stable signal recovery from incomplete and inaccurate measurements." In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223.

[13] E. J. Candès and T. Tao. "Near-optimal signal recovery from random projections: universal encoding strategies?" In: *IEEE Transactions on Information Theory* 52.12 (2006), pp. 5406–5425.

[14] E. Candes and T. Tao. "The Dantzig selector: statistical estimation when p is much larger than n." In: *Ann. Statist.* 35.6 (2007), pp. 2313–2351.

[15] E. J. Candès. "The restricted isometry property and its implications for compressed sensing." In: *Comptes Rendus Mathematique* 346.9 (2008), pp. 589–592. ISSN: 1631-073X.

[16] E. J. Candes and T. Tao. "Decoding by linear programming." In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215.

[17] E. J. Candes and P. A. Randall. "Highly robust error correction by convex programming." In: *IEEE Transactions on Information Theory* 54.7 (2008), pp. 2829–2840.

[18] J. A. Nelder and R. W. M. Wedderburn. "Generalized linear models." In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), pp. 370–384. ISSN: 00359238.

[19] S. Dirksen. "Quantized compressed sensing: a survey." In: *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*. Ed. by H. Boche, G. Caire, R. Calderbank, G. Kutyniok, R. Mathar, and P. Petersen. Cham: Springer International Publishing, 2019, pp. 67–95. ISBN: 978-3-319-73074-5.

[20] C. M. Bishop. *Pattern Recognition and Machine Learning Information Science and Statistics*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[21] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. "Phase retrieval with application to optical imaging: a contemporary overview." In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 87–109.

[22] R. M. Willett, R. F. Marcia, and J. M. Nichols. "Compressed sensing for practical optical imaging systems: a tutorial." In: *Optical Engineering* 50.7 (2011), pp. 1–14.

[23] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010. ISBN: 1441923225.

[24] A. Montanari. "Statistical estimation: from denoising to sparse regression and hidden cliques." In: *Statistical Physics, Optimization, Inference, and Message-Passing Algorithms*. Oxford: Oxford University Press, 2015. ISBN: 9780198743736.

[25] K. P. Burnham and D. R. Anderson, eds. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer New York, 2002, pp. 437–454. ISBN: 978-0-387-22456-5.

[26] M. J. Wainwright and M. I. Jordan. "Graphical models, exponential families, and variational inference." In: *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pp. 1–305. ISSN: 1935-8237.

[27] D. J. C. MacKay. *Information Theory Inference and Learning Algorithms*. USA: Cambridge University Press, 2003. ISBN: 9780521642989.

[28] F. Kschischang, B. Frey, and H.-A. Loeliger. "Factor graphs and the sum-product algorithm." In: *IEEE Trans. Inf. Theory* 47 (2001), pp. 498–519.

[29] L. Zdeborová and F. Krzakala. "Statistical physics of inference: thresholds and algorithms." In: *Advances in Physics* 65.5 (2016), pp. 453–552.

[30] S. Mallat and G. Peyré. *A Wavelet Tour of Signal Processing: The Sparse Way*. 3rd ed. Elsevier Science & Technology, 2008. ISBN: 9780123743701.

[31] T. J. Mitchell and J. J. Beauchamp. "Bayesian variable selection in linear regression." In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

[32] E. I. George and R. E. McCulloch. "Variable selection via Gibbs sampling." In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889.

[33] J. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez. "Expectation propagation in linear regression models with spike-and-slab priors." In: *Machine Learning* 99.3 (2015), pp. 437–487.

[34] F. Li, J. Fang, H. Li, and L. Huang. "Robust one-bit Bayesian compressed sensing with sign-flip errors." In: *IEEE Signal Processing Letters* 22.7 (2015), pp. 857–861.

[35] T. Park and G. Casella. "The Bayesian Lasso." In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686.

[36] N. G. Polson and L. Sun. "Bayesian l0-regularized least squares." In: *Applied Stochastic Models in Business and Industry* 35.3 (2019), pp. 717–731.

[37] G. Schultz. "Physical and technical background." In: *Magnetic Resonance Imaging with Nonlinear Gradient Fields: Signal Encoding and Image Reconstruction*. Wiesbaden: Springer Fachmedien Wiesbaden, 2013, pp. 11–38. ISBN: 978-3-658-01134-5.

[38] D. B. Twieg. "The k-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods." In: *Medical Physics* 10.5 (1983), pp. 610–621.

[39] S. Ljunggren. "A simple graphical representation of Fourier-based imaging methods." In: *Journal of Magnetic Resonance (1969)* 54.2 (1983), pp. 338–343. ISSN: 0022-2364.

[40] M. Lustig, D. Donoho, and J. M. Pauly. "Sparse MRI: the application of compressed sensing for rapid MR imaging." In: *Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195.

[41] A. C. Kak and M. Slaney. *Principles of computerized tomographic imaging*. Philadelphia: SIAM, 2001. ISBN: 978-0-898714-94-4.

[42] A. McCarthy et al. "Kilometer-range, high resolution depth imaging via 1560 nm wavelength single-photon detection." In: *Opt. Express* 21.7 (2013), pp. 8904–8915.

[43] D. M. McClatchy et al. "Wide-field quantitative imaging of tissue microstructure using sub-diffuse spatial frequency domain imaging." In: *Optica* 3.6 (2016), pp. 613–621.

[44] L. Nie and X. Chen. "Structural and functional photoacoustic molecular tomography aided by emerging contrast agents." In: *Chem. Soc. Rev.* 43 (20 2014), pp. 7132–7170.

[45] O. DeGuchy, L. Adhikari, A. Kim, and R. F. Marcia. "Photon-limited fluorescence lifetime imaging microscopy signal recovery with known bounds." In: *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. 2017, pp. 1–5.

[46] V. Studer, J. Bobin, M. Chahid, H. S. Mousavi, E. Candes, and M. Dahan. "Compressive fluorescence microscopy for biological and hyperspectral imaging." In: *Proceedings of the National Academy of Sciences* 109.26 (2012), E1679–E1687. ISSN: 0027-8424.

[47] X. Jiang, G. Raskutti, and R. Willett. "Minimax optimal rates for poisson inverse problems with physical constraints." 2015.

[48] M. F. Duarte et al. "Single-pixel imaging via compressive sampling." In: *IEEE Signal Processing Magazine* 25.2 (2008), pp. 83–91.

[49] G. M. Gibson, S. D. Johnson, and M. J. Padgett. "Single-pixel imaging 12 years on: a review." In: *Opt. Express* 28.19 (2020), pp. 28190–28208.

[50] A. Drenth, A. Huiser, and H. Ferwerda. "The problem of phase retrieval in light and electron microscopy of strong objects." In: *Optica Acta: International Journal of Optics* 22.7 (1975), pp. 615–628.

[51] R. P. Millane. "Phase retrieval in crystallography and optics." In: *J. Opt. Soc. Am. A* 7.3 (1990), pp. 394–411.

[52] V. Elser, T.-Y. Lan, and T. Bendory. "Benchmark problems for phase retrieval." 2018. arXiv: 1706.00399 [cs.IT].

[53] T. Bendory and D. Edidin. "Toward a mathematical theory of the crystallographic phase retrieval problem." In: *SIAM Journal on Mathematics of Data Science* 2.3 (2020), pp. 809–839.

[54] R. Bates. "Astronomical speckle imaging." In: *Physics Reports* 90.4 (1982), pp. 203–297. ISSN: 0370-1573.

[55] L. De Caro, E. Carlino, D. Siliqi, and C. Giannini. "Coherent diffractive imaging: from nanometric down to picometric resolution." In: *Handbook of Coherent-Domain Optical Methods: Biomedical Diagnostics, Environmental Monitoring, and Materials Science*. Ed. by V. V. Tuchin. New York, NY: Springer New York, 2013, pp. 291–314. ISBN: 978-1-4614-5176-1.

[56] I. Barke et al. "The 3d-architecture of individual free silver nanoparticles captured by x-ray scattering." In: *Nature Communications* 6.1 (2015), p. 6187.

[57] M. M. Seibert et al. "Single mimivirus particles intercepted and imaged with an x-ray laser." In: *Nature* 470.7332 (2011), pp. 78–81.

[58] M. C. T. Bahaa E. A. Saleh. *Fundamentals of photonics*. 2nd ed. Wiley series in pure applied optics. Hoboken: Wiley, 2007. ISBN: 9780471358329.

[59] D. Blei, A. Kucukelbir, and J. McAuliffe. "Variational inference: a review for statisticians." In: *Journal of the American Statistical Association* 112 (2017), pp. 859–877.

[60] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020.

[61] D. Barber and C. Bishop. "Ensemble learning in Bayesian neural networks." In: *Generalization in Neural Networks and Machine Learning*. Generalization in neural networks and machine learning. Springer Verlag, 1998, pp. 215–237.

[62] W. L. Bragg and E. J. Williams. "The effect of thermal agitation on atomic arrangement in alloys." In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 145.855 (1934), pp. 699–730.

[63] A. Coja-Oghlan et al. "Statistical physics, optimization, inference and message-passing algorithms." In: École de Physique des Houches, special issue, 30 September-11 October 2013. Oxford: Oxford University Press, 2016.

[64] H. A. Bethe and W. L. Bragg. "Statistical theory of superlattices." In: *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 150.871 (1935), pp. 552–575.

[65] E. A. Guggenheim and R. H. Fowler. "The statistical mechanics of regular solutions." In: *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 148.864 (1935), pp. 304–312.

[66] R. Gallager. "Low-density parity-check codes." In: *IRE Transactions on Information Theory* 8.1 (1962), pp. 21–28.

[67] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988. ISBN: 0934613737.

[68] M. Mezard and A. Montanari. *Information, Physics, and Computation*. USA: Oxford University Press, Inc., 2009. ISBN: 019857083X.

[69] A. Braunstein. "Algorithms for optimization, inference and learning." Politecnico di Torino. Unpublished lecture notes. 2020.

[70] J. S. Yedidia, W. T. Freeman, and Y. Weiss. "Understanding belief propagation and its generalizations." In: *Exploring Artificial Intelligence in the New Millennium*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 239–269. ISBN: 1558608117.

[71] D. J. Thouless, P. W. Anderson, and R. G. Palmer. "Solution of solvable model of a spin glass." In: *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics* 35.3 (1977), pp. 593–601.

[72] M. Mezard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, 1986.

[73] M. Shamir and H. Sompolinsky. "Thouless-Anderson-Palmer equations for neural networks." In: *Phys. Rev. E* 61 (2 2000), pp. 1839–1844.

[74] T. Plefka. "Convergence condition of the TAP equation for the infinite-ranged ising spin glass model." In: *Journal of Physics A: Mathematical and General* 15.6 (1982), pp. 1971–1978.

[75] M. Opper and O. Winther. "Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling." In: *Physical Review E* 64.5 (2001), p. 056131.

[76] D. L. Donoho, A. Maleki, and A. Montanari. "Message-passing algorithms for compressed sensing." In: *Proceedings of the National Academy of Sciences* 106.45 (2009), pp. 18914–18919. ISSN: 0027-8424.

[77] M. Gabrié. "Mean-field inference methods for neural networks." In: *Journal of Physics A: Mathematical and Theoretical* 53.22 (2020), p. 223002. ISSN: 1751-8121.

[78] M. Bayati and A. Montanari. "The dynamics of message passing on dense graphs with applications to compressed sensing." In: *IEEE Transactions on Information Theory* 57.2 (2011), pp. 764–785.

[79] A. Javanmard and A. Montanari. "State evolution for general approximate message passing algorithms with applications to spatial coupling." In: *Information and Inference: A Journal of the IMA* 2.2 (2013), pp. 115–144.

[80] P. Schniter. "A simple derivation of AMP and its state evolution via first-order cancellation." In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 4283–4292.

[81] S. Rangan. "Generalized approximate message passing for estimation with random linear mixing." In: *2011 IEEE International Symposium on Information Theory Proceedings*. 2011, pp. 2168–2172.

[82] F. Caltagirone, L. Zdeborová, and F. Krzakala. "On convergence of approximate message passing." In: *2014 IEEE International Symposium on Information Theory*. 2014, pp. 1812–1816.

[83] S. Rangan, P. Schniter, A. K. Fletcher, and S. Sarkar. "On the convergence of approximate message passing with arbitrary matrices." In: *IEEE Transactions on Information Theory* 65.9 (2019), pp. 5339–5351.

[84] S. Rangan, P. Schniter, and A. K. Fletcher. "Vector approximate message passing." In: *IEEE Transactions on Information Theory* 65.10 (2019), pp. 6664–6684.

[85] X. Meng, S. Wu, and J. Zhu. "A unified Bayesian inference framework for generalized linear models." In: *IEEE Signal Processing Letters* 25.3 (2018), pp. 398–402. ISSN: 1558-2361.

[86] C. Berrou and A. Glavieux. "Near optimum error correcting coding and decoding: turbo-codes." In: *IEEE Transactions on Communications* 44.10 (1996), pp. 1261–1271.

[87] T. P. Minka. "Expectation propagation for approximate Bayesian inference." In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[88] M. Opper. "A Bayesian approach to on-line learning." In: *On-Line Learning in Neural Networks*. Ed. by D. Saad. Publications of the Newton Institute. Cambridge University Press, 1999, pp. 363–378.

[89] M. Opper and O. Winther. "Gaussian processes for classification: mean-field algorithms." In: *Neural Computation* 12.11 (2000), pp. 2655–2684.

[90] L. D. Brown. "Fundamentals of statistical exponential families with applications in statistical decision theory." In: *Lecture Notes-Monograph Series* 9 (1986), pp. i–279. ISSN: 07492170.

[91] F. Y. Wu. "The Potts model." In: *Rev. Mod. Phys.* 54 (1 1982), pp. 235–268.

[92] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell. "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models." In: *Phys. Rev. E* 87 (1 2013), p. 012707.

[93] M. Seeger. *Expectation propagation for exponential families*. Tech. rep. Department of EECS, Berkeley, 2005.

[94] H. C. Nguyen, R. Zecchina, and J. Berg. "Inverse statistical problems: from the inverse ising problem to data science." In: *Advances in Physics* 66.3 (2017), pp. 197–261.

[95] M. W. Seeger. "Bayesian gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations." PhD thesis, University of Edinburgh. 2003.

[96] A. Braunstein, A. P. Muntoni, and A. Pagnani. "An analytic approximation of the feasible space of metabolic networks." In: *Nature communications* 8 (2017), p. 14915.

[97] "Chapter 12: nonlinear estimation." In: *Stochastic Models, Estimation, and Control: Volume 2*. Ed. by P. S. Maybeck. Vol. 141. Mathematics in Science and Engineering. Elsevier, 1982, pp. 212–271.

[98] A. Braunstein, A. P. Muntoni, A. Pagnani, and M. Pieropan. "Compressed sensing reconstruction using expectation propagation." In: *Journal of Physics A: Mathematical and Theoretical* 53.18 (2020), p. 184001.

[99] J. Sherman and W. J. Morrison. "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix." In: *The Annals of Mathematical Statistics* 21.1 (1950), pp. 124–127.

[100] A. Braunstein, T. Gueudré, A. Pagnani, and M. Pieropan. "Expectation propagation on the diluted Bayesian classifier." In: *Phys. Rev. E* 103.4 (2021), p. 043301.

[101] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013. ISBN: 9781421407944.

[102] C. Sutton. "Expectation propagation in factor graphs: a tutorial." Unpublished manuscript (last accessed: July 8, 2021). 2005.

[103] L. Csató, M. Opper, and O. Winther. "TAP Gibbs free energy belief propagation and sparsity." In: *Advances in Neural Information Processing Systems*. Ed. by T. Dietterich, S. Becker, and Z. Ghahramani. Vol. 14. MIT Press, 2002.

[104] A. P. Muntoni. "Statistical mechanics approaches to optimization and inference." 2017.

[105] A. Braunstein, A. Pagnani, and M. Pieropan. "GaussianEP." GitHub repository (last accessed: July 7, 2021).

[106] Y. Kabashima, T. Wadayama, and T. Tanaka. "A typical reconstruction limit for compressed sensing based on l_p-norm minimization." In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.09 (2009), p. L09003.

[107] Y. Kabashima, T. Wadayama, and T. Tanaka. "Erratum: a typical reconstruction limit of compressed sensing based on l_p -norm minimization." In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.07 (2012), E07001.

[108] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová. "Statistical-physics-based reconstruction in compressed sensing." In: *Phys. Rev. X* 2 (2 2012), p. 021005.

[109] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. "Probabilistic reconstruction in compressed sensing: algorithms phase diagrams and threshold achieving matrices." In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.08 (2012), P08009.

[110] F. Krzakala. "Aspics: applying statistical physics to inference in compressed sensing." Last accessed: July 7, 2021.

[111] J. A. Tropp and A. C. Gilbert. "Signal recovery from random measurements via orthogonal matching pursuit." In: *IEEE Transactions on Information Theory* 53.12 (2007), pp. 4655–4666.

[112] D. Needell and R. Vershynin. "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit." In: *IEEE Journal of Selected Topics in Signal Processing* 4.2 (2010), pp. 310–316.

[113] D. Needell and J. A. Tropp. "CoSaMP: iterative signal recovery from incomplete and inaccurate samples." In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 301–321.

[114] W. Dai and O. Milenkovic. "Subspace pursuit for compressive sensing signal reconstruction." In: *IEEE Transactions on Information Theory* 55.5 (2009), pp. 2230–2249.

[115] H. Mohimani, M. Babaie-Zadeh, and C. Jutten. "A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm." In: *IEEE Transactions on Signal Processing* 57.1 (2009), pp. 289–301.

[116] R. Gebel. "Kl1p – a portable C++ library for compressed sensing." C++ library (last accessed: July 9, 2021). 2012.

[117] C. Sanderson and R. Curtin. "A user-friendly hybrid sparse matrix class in C++." In: (2018). Ed. by J. H. Davenport, M. Kauers, G. Labahn, and J. Urban, pp. 422–430.

[118] C. Sanderson and R. Curtin. "Armadillo: a template-based C++ library for linear algebra." In: *Journal of Open Source Software* 1.2 (2016), p. 26.

[119] H. Sompolinsky, N. Tishby, and H. S. Seung. "Learning from examples in large neural networks." In: *Phys. Rev. Lett.* 65 (13 1990), pp. 1683–1686.

[120] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

[121] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. International series of monographs on physics. Oxford University Press, 2001. ISBN: 9780198509400.

[122] Y. Xu, Y. Kabashima, and L. Zdeborová. "Bayesian signal reconstruction for 1-bit compressed sensing." In: *Journal of Statistical Mechanics: Theory and Experiment* 2014.11 (2014), P11015.

[123] M. Pieropan. "EPDilutedClassifier." GitHub repository (last accessed: July 7, 2021).

[124] X. Meng and J. Zhu. "Glmcode." GitHub repository (last accessed: July 7, 2021).

[125] F. Li. "R1BCS: robust one-bit Bayesian compressed sensing with sign-flip errors." GitHub repository.

[126] T. Minka. "A roadmap to research on ep." Webpage (last accessed: July 7, 2021).

[127] M. Biehl and H. Schwarze. "Learning by on-line gradient descent." In: *Journal of Physics A: Mathematical and General* 28.3 (1995), pp. 643–656.

[128] M. Ahr, M. Biehl, and R. Urbanczik. "Statistical physics and practical training of soft-committee machines." In: *The European Physical Journal B - Condensed Matter and Complex Systems* 10.3 (1999), pp. 583–588.

[129] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová. "Multi-layer generalized linear estimation." In: *2017 IEEE International Symposium on Information Theory (ISIT)*. 2017, pp. 2098–2102.

[130] A. K. Fletcher and S. Rangan. "Inference in deep networks in high dimensions." 2017. arXiv: 1706.06549 [cs.LG].

[131] P. Pandit, M. Sahraee-Ardakan, S. Rangan, P. Schniter, and A. K. Fletcher. "Inference in multi-layer networks with matrix-valued unknowns." 2020. arXiv: 2001.09396 [cs.LG].

[132] R. Salakhutdinov, A. Mnih, and G. Hinton. "Restricted Boltzmann machines for collaborative filtering." In: *ICML '07: Proceedings of the 24th international conference on Machine learning*. Corvalis, Oregon: ACM, 2007, pp. 791–798. ISBN: 978-1-59593-793-3.

[133] P. Smolensky. "Information processing in dynamical systems: foundations of harmony theory." In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986, pp. 194–281. ISBN: 026268053X.

[134] T. Tieleman. "Training restricted Boltzmann machines using approximations to the likelihood gradient." In: ICML '08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 1064–1071. ISBN: 9781605582054.

[135] G. E. Hinton. "Training products of experts by minimizing contrastive divergence." In: *Neural Computation* 14.8 (2002), pp. 1771–1800. ISSN: 0899-7667.

[136] E. W. Tramel, M. Gabrié, A. Manoel, F. Caltagirone, and F. Krzakala. "Deterministic and generalized framework for unsupervised learning with restricted Boltzmann machines." In: *Phys. Rev. X* 8 (4 2018), p. 041006.

[137] M. Bayati, M. Lelarge, and A. Montanari. "Universality in polytope phase transitions and message passing algorithms." In: *The Annals of Applied Probability* 25.2 (2015), pp. 753–822. ISSN: 10505164.

[138] A. Javanmard and A. Montanari. "State evolution for general approximate message passing algorithms with applications to spatial coupling." In: *Information and Inference: A Journal of the IMA* 2.2 (2013), pp. 115–144. ISSN: 2049-8764.

[139] M. Bayati and A. Montanari. "The dynamics of message passing on dense graphs with applications to compressed sensing." In: *IEEE Transactions on Information Theory* 57.2 (2011), pp. 764–785.

[140] K. Takeuchi. "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements." In: *IEEE Transactions on Information Theory* 66.1 (2020), pp. 368–386.

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was LuaLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.