POLITECNICO DI TORINO
Repository ISTITUZIONALE

Expectation propagation on the diluted Bayesian classifier

Original
Expectation propagation on the diluted Bayesian classifier / Braunstein, A., Gueudre, T., Pagnani, A., Pieropan, M.. - In: PHYSICAL REVIEW. E. - ISSN 2470-0045. - ELETTRONICO. - 103:4(2021). [10.1103/PhysRevE.103.043301]

Availability:
This version is available at: 11583/2885205 since: 2021-04-07T18:43:38Z

Publisher:
American Physical Society

Published
DOI:10.1103/PhysRevE.103.043301

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

11 November 2026

Expectation propagation on the diluted Bayesian classifier

Alfredo Braunstein,[1,2,3] Thomas Gueudré,[1,*] Andrea Pagnani,[1,2,3] and Mirko Pieropan[1,†]

[1]*Department of Applied Science and Technologies, Politecnico di Torino, Corso Duca Degli Abruzzi 24, Torino, Italy*

[2]*Italian Institute for Genomic Medicine, IRCCS Candiolo, SP-142, I-10060 Candiolo, Italy*

[3]*Istituto Nazionale di Fisica Nucleare Sezione di Torino, Via Pietro Giuria 1, I-10125 Torino, Italy*

(Received 29 September 2020; accepted 15 February 2021; published 2 April 2021)

Efficient feature selection from high-dimensional datasets is a very important challenge in many data-driven fields of science and engineering. We introduce a statistical mechanics inspired strategy that addresses the problem of sparse feature selection in the context of binary classification by leveraging a computational scheme known as expectation propagation (EP). The algorithm is used in order to train a continuous-weights perceptron learning a classification rule from a set of (possibly partly mislabeled) examples provided by a teacher perceptron with diluted continuous weights. We test the method in the Bayes optimal setting under a variety of conditions and compare it to other state-of-the-art algorithms based on message passing and on expectation maximization approximate inference schemes. Overall, our simulations show that EP is a robust and competitive algorithm in terms of variable selection properties, estimation accuracy, and computational complexity, especially when the student perceptron is trained from correlated patterns that prevent other iterative methods from converging. Furthermore, our numerical tests demonstrate that the algorithm is capable of learning online the unknown values of prior parameters, such as the dilution level of the weights of the teacher perceptron and the fraction of mislabeled examples, quite accurately. This is achieved by means of a simple maximum likelihood strategy that consists in minimizing the free energy associated with the EP algorithm.

DOI: 10.1103/PhysRevE.103.043301

I. INTRODUCTION

The problem of extracting sparse information from high dimensional data is among the most interesting challenges in theoretical computer science with many applications ranging from computational biology to combinatorial chemistry, neuroscience, and natural language processing [1,2]. As a specific example, next generation sequencing and, in general, the ongoing technological revolution related to high-throughput technologies in biology pose very stringent requirements to the algorithmic techniques that are supposed to analyze the data that are produced and made publicly available through easily accessible databases. Just to give some orders of magnitude, a typical genetic screening for cancer-related pathologies—freely available from The Cancer Genome Atlas web site [3]—involves measurement of activity or genetic sequence variation over ≈23 000 genes measured on patient cohorts that typically count around 1000 individuals divided into cases and controls (lung and colorectal cancer are an exception, with ≈10 000 individuals screened in each dataset). Here, a typical task is to determine the genotypic signature related to the disease that typically involves $O(10^2)$ genes from 23 000 measured probes. Such problem can be simply formulated in terms of the following classification problem: given the activity and/or the genetic alterations of an individual, find a simple rule involving a small—possibly the smallest—subset of genes to assess the probability for the individual to develop the disease. There are two main difficulties in this task: (i) typically genes act in a combinatorial and nonlinear manner and (ii) individual samples turn out to be statistically very correlated.

Historically, the problem of sparse feature selection in classification tasks has been divided into two complementary computational methods [2]: (i) *wrappers* that exploit the learning mechanism to produce a prediction value related score for the sought signature and (ii) *filters* where the signature extraction is a data preprocessing, typically unrelated to the classification task.

From the point of view of information theory, the problem of sparse feature selection in classification is strictly related to compressive sensing (CS), one of the most studied methods for data acquisition, with interesting applications in several other research fields [4,5]. CS was originally proposed as a new low-rate signal acquisition technique for compressible signals [4,6,7] and is formulated as follows: given $M < N$, a vector $\boldsymbol{z} \in \mathbb{R}^M$, and a linear operator of maximal rank $\mathbf{X} \in \mathbb{R}^{M\times N}$ often referred to as the *measurement* or *sensing* matrix, the CS problem consists in determining the unknown sparse vector $\boldsymbol{w} \in \mathbb{R}^N$ that is linked to its compressed projection $\boldsymbol{z}$ by means of the linear transformation $\boldsymbol{z} = \mathbf{X}\boldsymbol{w}$, where $\mathbf{X}$ and $\boldsymbol{z}$ are assumed to be known. Although research in CS has still many open challenges to face, very stringent results are known about the general conditions for the existence and uniqueness of the solution. Among the different algorithms that have been proposed in order to reconstruct efficiently the signal, many use techniques borrowed from the statistical mechanics of disordered systems [8–11].

More recently, the so-called 1-bit CS (1BCS) has been proposed as a strategy to deal with the problem of inferring a sparse signal knowing only the sign of the data of the linear

*Currently at Amazon Alexa, Torino.

†Corresponding author: mirko.pieropan@polito.it

 ©2021 American Physical Society

measurements: $\boldsymbol{\sigma} = \mathrm{sgn}(\mathbf{X}\boldsymbol{w})$, where $\mathrm{sgn}(z)$ is a vector with elements $z_i/|z_i|$ for $z_i \neq 0$. Besides being of interest for signal transmission related problems where discarding the amplitude of the signal can significantly reduce the amount of information to be stored or relayed [12,13], this problem can also be interpreted in terms of sparse boolean classification tasks. The most widely adopted inference scheme in CS is the so-called lasso regression or L_1-norm minimization [14], as originally proposed in the context of 1BCS in [12]. However, it is clear that the most efficient solution from the point of view of optimal dilution of the problem should be achieved by a L_0 pseudonorm, where nonzero parameters are indeed penalized independently of their nonzero value. Unfortunately, dealing with the nonconvex L_0 regularization is not so simple as it typically leads to phase transitions that make the problem computationally intractable. A practical solution to the problem is to restrict the space of parameters to a discrete set, where effectively the L_0 pseudonorm is equivalent to the more amenable L_1 case [15–19]. As far as continuous parameters are concerned, different strategies have been proposed. First, from the statistical physics community side, an approach pursuing this direction consists in a perceptron the continuous parameters of which are masked by boolean variables mimicking dilution [20–23]. Attempts to characterize theoretically the phase space diagram and the structure of the transition through the replica method have been reported in [24–26]. Variations of the generalized approximate message passing technique (GAMP) were employed in [27], as it provides a tractable and efficient way to perform minimum mean squared error (MMSE) estimation on the variables to be retrieved when the matrix of patterns is large and Gaussian independent and identically distributed (i.i.d.). However, for more general pattern matrices, GAMP convergence is not guaranteed, which has led to the extension of algorithms of the vector AMP (VAMP) type [28] to generalized linear models [29,30], including perceptron learning.

On the computer science side, many other algorithms for 1BCS combining the enforcement of sparsity and of sign consistency constraints were also proposed, building upon analogous algorithms developed for standard CS. Examples of methods for error-free sparse signal retrieval from one-bit quantized measurements include greedy approaches which iteratively determine the most appropriate sparse support given the sign measurements, such as matching sign pursuit [31], as well as binary iterative hard thresholding [32], where an L_1-based convex consistency-enforcing objective function minimization is alternated with a thresholding operation that selects the K largest elements. The problem of noisy 1BCS was addressed, for instance, in [33–35]. However, among these examples, only [35] proposes an algorithm which does not require the prior knowledge of the number of corrupted sign measurements. Here, the one-bit measurement errors are modeled by introducing a sparse vector $\boldsymbol{s}$ the nonzero components of which produce the sign mismatches as $\boldsymbol{\sigma} = \mathrm{sgn}(\mathbf{X}\boldsymbol{w} + \boldsymbol{s})$. The algorithm attempts to identify the sign errors and to retrieve the sparse signal $\boldsymbol{w}$ using a variational expectation-maximization (EM) based inference scheme.

In this paper we propose a *wrapper* strategy where both the variable selection and the classification tasks are simultaneously performed through expectation propagation (EP), an iterative scheme to approximate intractable distributions that was introduced first in the field of statistical physics [36,37] and shortly after in the field of theoretical computer science [38]. In analogy to what was presented in [39] in the context of sampling the space of high dimensional polytopes, we show that, by approximating the computationally intractable posterior distribution $P(\boldsymbol{w}|\boldsymbol{\sigma}, \mathbf{X})$ through a tractable multivariate probability density $Q(\boldsymbol{w}|\boldsymbol{\sigma}, \mathbf{X})$, we are able to solve both efficiently and accurately the 1BCS problem. We compare our results to those obtained from the AMP and VAMP based schemes proposed in [27] and [30], respectively, and to those given by the EM based approach of [35]. We provide the factor graphs associated with these algorithms in Fig. 1. We show through simulations that one of the main strengths of the EP-based approach is that it is effective on a wider family of measurement matrices with respect to other relatively similar algorithms such as VAMP and AMP.

The paper has the following structure: after this introduction, in Sec. II we define the problem, and introduce the EP algorithm. In Sec. III we present extensive numerical simulations both in the noiseless and noisy case. Here both i.i.d. and correlated measurement matrices are analyzed. Finally, in Sec. IV we summarize the results of the paper and draw the conclusions.

II. METHODS

A. The diluted perceptron as a linear estimation problem and its statistical mechanics setup

We consider a student perceptron with N input units and continuous weights $\boldsymbol{w} \in \mathbb{R}^N$. We assume that the connections are diluted and that only a fraction ρ of them are nonzero. We also assume that M real-valued patterns $\boldsymbol{x}_\tau \in \mathbb{R}^N$ are presented to the perceptron and that a binary label σ_τ, $\tau = 1, \ldots, M$ has already been assigned to each of them as a result of the classification performed by a teacher perceptron with sparse continuous weights $\boldsymbol{B}$. The task of the student perceptron is to learn the input-output association based on the examples $(\boldsymbol{x}_\tau, \sigma_\tau)$, $\tau = 1, \ldots, M$ provided by the teacher:

$$\sigma_\tau = \mathrm{sgn}(\boldsymbol{w}^T \boldsymbol{x}_\tau), \quad \tau = 1, \ldots, M, \tag{1}$$

where we use the convention that $\mathrm{sgn}(0) = 1$. For each example τ, the rule (1) is equivalent to the condition

$$\left(\sigma_\tau \boldsymbol{x}_\tau^T\right)\boldsymbol{w} \geqslant 0\,. \tag{2}$$

We now introduce the auxiliary variables, $y_\tau := (\sigma_\tau \boldsymbol{x}_\tau^T)\boldsymbol{w}$, and the data matrix

$$\mathbf{X}_\sigma = \begin{pmatrix} \sigma_1 \boldsymbol{x}_1^T \\ \sigma_2 \boldsymbol{x}_2^T \\ \vdots \\ \sigma_M \boldsymbol{x}_M^T \end{pmatrix}. \tag{3}$$

Through the previous definitions, we can define the following linear estimation problem:

$$\boldsymbol{y} = \mathbf{X}_\sigma \boldsymbol{w}, \tag{4}$$

where the variables to be inferred are both $\boldsymbol{y}$ and $\boldsymbol{w}$. As we will show below, the positivity constraints in Eq. (2) will be enforced in terms of a prior distribution on the $\boldsymbol{y}$ variables.

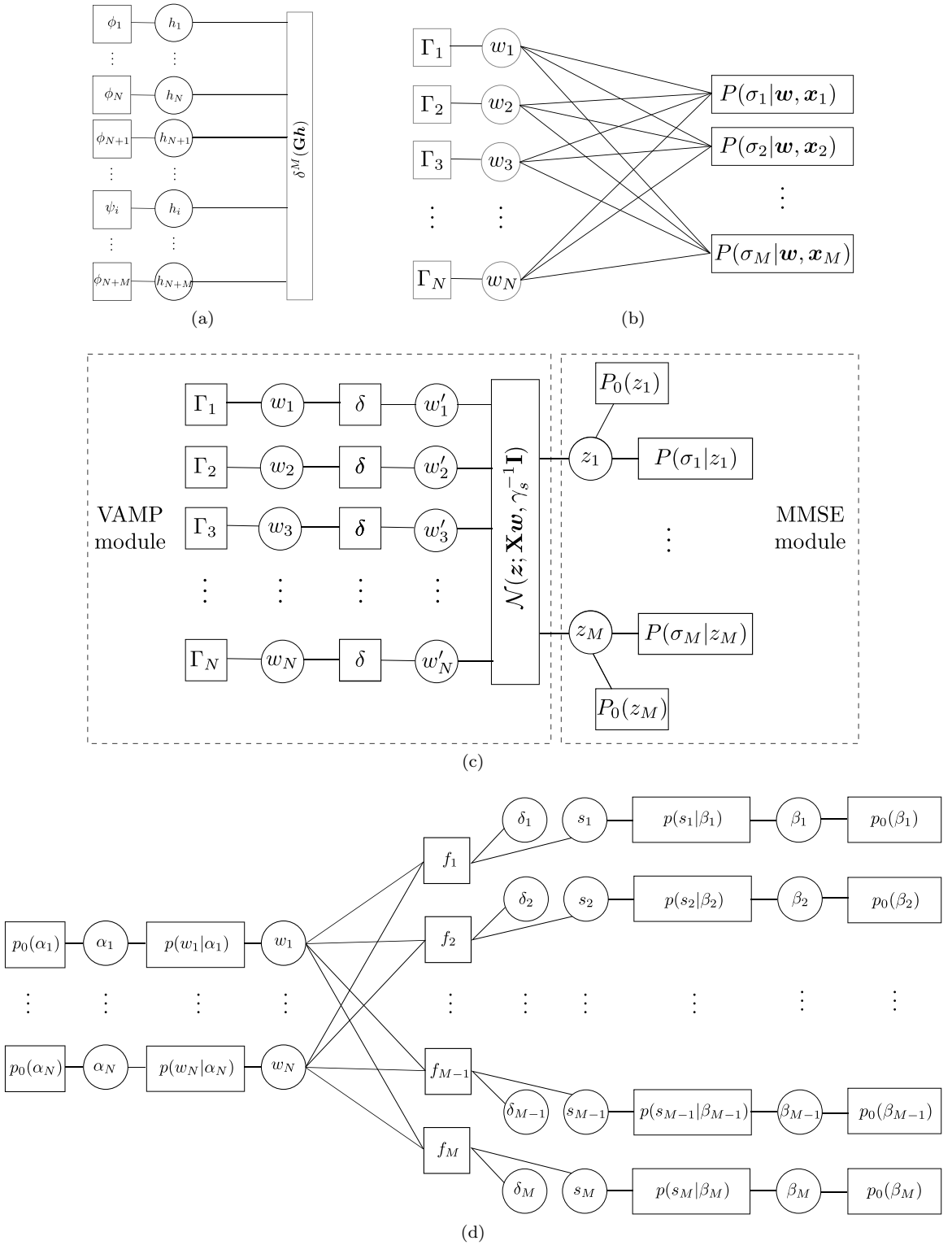

FIG. 1. Factor graphs associated with (a) EP, (b) AMP, (c) grVAMP, and (d) R1BCS. Variable nodes are represented as circles and function nodes are represented as squares and the notation has been made consistent with that employed in this paper. (a) Tilted distributions in EP (see Sec. II B for more details), where ψ denotes the exact prior and ϕ the approximated Gaussian prior factors in the EP approximation. (b) Factor graph related to AMP. In 1bitAMP, Γ are spike-and-slab priors and $P(\sigma_\mu|\boldsymbol{w},\boldsymbol{x}_\mu)=\Lambda(\sigma_\mu \boldsymbol{x}_\mu^T\boldsymbol{w})$. (c) Factor graph related to the grVAMP approximation, where we have emphasized the VAMP and MMSE modules composing the algorithm, Γ are the same as in (b), and $P(\sigma_\mu|z_\mu)=\Lambda(\sigma_\mu z_\mu)$. In (b) and (c), Λ denotes either the pseudoprior in Eq. (8) or the one in Eq. (10). (d) Factor graph of the function appearing in the lower bound maximized in R1BCS. The distributions $p(w_i|\alpha_i)p(\alpha_i)$ and $p(s_\mu|\beta_\mu)p(\beta_\mu)$ are hierarchical Gaussian-inverse-Gamma priors assigned to the weights and to the noise components s_μ, respectively, appearing in $\boldsymbol{\sigma}=\mathrm{sgn}(\mathbf{X}\boldsymbol{w}+\boldsymbol{s})$. The quantities α_i and β_μ are hyperparameters, whereas the quantities δ_μ are variational parameters optimized in the maximization step of the algorithm.

The linear estimation problem expressed in Eq. (4) can be addressed in a Bayesian setting: by introducing the variable vector $\boldsymbol{h} = (w_1, \ldots, w_N, y_1, \ldots, y_M)^T$ and the energy function,

$$E(\boldsymbol{w}, \boldsymbol{y}) = \|\boldsymbol{y} - \mathbf{X}_\sigma \boldsymbol{w}\|^2 = \boldsymbol{h}^T \mathbf{E}^{-1} \boldsymbol{h},$$
$$\mathbf{E}^{-1} = \begin{pmatrix} \mathbf{X}_\sigma^T \mathbf{X}_\sigma & -\mathbf{X}_\sigma^T \\ -\mathbf{X}_\sigma & \mathbf{I} \end{pmatrix}, \tag{5}$$

the likelihood of the set of N weights of the perceptron can be expressed as the Boltzmann distribution associated with $E(\boldsymbol{w}, \boldsymbol{y})$, which reads

$$\mathcal{L}(\boldsymbol{w}) = P(\sigma_1, \ldots, \sigma_M | \boldsymbol{w}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{w}, \boldsymbol{y})}, \tag{6}$$

where, from a statistical physics standpoint, β plays the role of an inverse temperature. In the absence of noise, it is convenient to consider the zero temperature limit of this likelihood $\mathcal{L}(\boldsymbol{w}) \xrightarrow{\beta \to \infty} \delta(\boldsymbol{y} - \mathbf{X}_\sigma \boldsymbol{w})$, where $\delta(x)$ denotes the Dirac delta distribution.

We also introduce prior distributions in order to encode the constraints to which the variables w_i, $i = 1, \ldots, N$, and y_τ, $\tau = 1, \ldots, M$, are subject. The sparsity assumption on the weights $\boldsymbol{w}$ is expressed in terms of a spike-and-slab prior [40]:

$$\Gamma(w_i) = (1 - \rho)\delta(w_i) + \rho \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda w_i^2}{2}}, \quad i = 1, \ldots, N. \tag{7}$$

If the labels of the teacher are not corrupted by noise, then the auxiliary variables $\boldsymbol{y}$ need to fulfill the positivity constraint (2), which can be expressed in terms of the pseudoprior:

$$\Lambda(y_\tau) = \Theta(y_\tau), \quad \tau = 1, \ldots, M. \tag{8}$$

On the other hand, if noise at the output of the teacher perceptron is present, one may assume that the labels provided by the teacher perceptron are assigned according to the following process [41],

$$\tilde{\sigma} = \begin{cases} \operatorname{sgn}(\boldsymbol{B}^T \boldsymbol{x}) & \text{with probability} \quad \eta \\ -\operatorname{sgn}(\boldsymbol{B}^T \boldsymbol{x}) & \text{with probability} \quad 1 - \eta, \end{cases} \tag{9}$$

and that the student receives the altered examples $(\boldsymbol{x}_\mu, \tilde{\sigma}_\mu)$, $\mu = 1, \ldots, M$. In this case, if the process that flips the labels is known, then it may be encoded in the pseudoprior Λ as follows:

$$\Lambda(y_\mu) = \eta \Theta(y_\mu) + (1 - \eta)\Theta(-y_\mu). \tag{10}$$

In general, the parameters ρ, λ, and η are not known and need to be learned by the student perceptron in the training phase. Finally, by Bayes's rule, the posterior distribution of both weights and auxiliary variables reads

$$P(\boldsymbol{w}, \boldsymbol{y}) = \frac{1}{Z_P} \delta(\boldsymbol{y} - \mathbf{X}_\sigma \boldsymbol{w}) \prod_{i=1}^{N} \Gamma_i(w_i) \prod_{\tau=1}^{M} \Lambda_\tau(y_\tau). \tag{11}$$

B. Learning the weights via expectation propagation

Zero temperature formulation

We wish to infer the values of the weights by estimating the expectation values of the marginals of the distribution (11), as this strategy minimizes the associated mean squared error. However, the latter marginalizations are intractable and we need to resort to approximation methods. Here we propose an expectation propagation scheme based on the zero temperature formulation presented in [42] in order to solve the problem.

Starting from the linear system $\mathbf{X}_\sigma \boldsymbol{w} = \boldsymbol{y}$, we notice that it can be written as the homogeneous system

$$\mathbf{G}\boldsymbol{h} = \mathbf{0}, \tag{12}$$

where $\mathbf{G} = (-\mathbf{X}_\sigma | \mathbf{I})$ and $\mathbf{I}$ is the $M \times M$ identity matrix.

The intractable posterior distribution reads

$$P(\boldsymbol{h}) = \frac{1}{Z_P} \delta^M(\mathbf{G}\boldsymbol{h}) \prod_{i \in W} \Gamma_i(h_i) \prod_{\tau \in Y} \Lambda_\tau(h_\tau), \tag{13}$$

where $W = \{1, \ldots, N\}$, $Y = \{N + 1, \ldots, N + M\}$ and $\delta^M(\boldsymbol{z})$ denotes the M-dimensional Dirac delta distribution. We introduce Gaussian approximating factors,

$$\phi_i(h_i) = \exp\left(-\frac{(h_i - a_i)^2}{2d_i}\right), \tag{14}$$

and a fully Gaussian approximation of the posterior distribution (13), in which all priors Γ and Λ are replaced by factors of the form (14):

$$Q(\boldsymbol{h}) = \frac{1}{Z_Q} \delta^M(\mathbf{G}\boldsymbol{h}) \prod_{i \in W} \phi(h_i; a_i, d_i) \prod_{\tau \in Y} \phi(h_\tau; a_\tau, d_\tau). \tag{15}$$

$Q(\boldsymbol{h})$ can be equivalently expressed as

$$Q(\boldsymbol{h}) = \frac{1}{Z_Q} \delta^M(\mathbf{G}\boldsymbol{h}) \exp\left(-\frac{1}{2}(\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{\Sigma}_W^{-1} (\boldsymbol{w} - \bar{\boldsymbol{w}})\right), \tag{16}$$

where the covariance matrix $\boldsymbol{\Sigma}_W$ and the mean $\bar{\boldsymbol{w}}$ in Eq. (16) are given, respectively, by

$$\boldsymbol{\Sigma}_W^{-1} = \sum_{i \in W} \frac{1}{d_i} \boldsymbol{e}_i \boldsymbol{e}_i^T + \mathbf{X}_\sigma^T \left(\sum_{i \in Y} \frac{1}{d_i} \boldsymbol{e}_i \boldsymbol{e}_i^T \right) \mathbf{X}_\sigma, \tag{17}$$

and by

$$\bar{\boldsymbol{w}} = \boldsymbol{\Sigma}_W \left(\sum_{i \in W} \frac{a_i}{d_i} \boldsymbol{e}_i + \sum_{i \in Y} \frac{a_i}{d_i} \mathbf{X}_\sigma^T \boldsymbol{e}_i \right). \tag{18}$$

Here, $\boldsymbol{e}_i$ denotes the ith basis vector of the standard basis of $\mathbb{R}^N$ (resp. $\mathbb{R}^M$) if $i \in W$ (resp. $i \in Y$). Notice that the marginal distributions of $Q(\boldsymbol{h})$ for each variable h_i are also Gaussian, with means $\bar{h}_i$ given by

$$\bar{h}_i = \begin{cases} \bar{w}_i, & i \in W \\ \boldsymbol{e}_i^T \mathbf{X}_\sigma \bar{\boldsymbol{w}}, & i \in Y, \end{cases} \tag{19}$$

and variances Σ_{ii} given by

$$\Sigma_{ii} = \begin{cases} \boldsymbol{e}_i^T \boldsymbol{\Sigma}_W \boldsymbol{e}_i, & i \in W, \\ \left(\boldsymbol{e}_i^T \mathbf{X}_\sigma\right) \boldsymbol{\Sigma}_W \left(\mathbf{X}_\sigma^T \boldsymbol{e}_i\right), & i \in Y, \end{cases} \tag{20}$$

where for $i \in Y$ we took advantage of the linear constraints $\boldsymbol{y} = \mathbf{X}_\sigma \boldsymbol{w}$. Notice that the full $(N + M) \times (N + M)$ covariance matrix $\boldsymbol{\Sigma}$ (the diagonal entries of which are defined in the previous equation) reads

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_W & \boldsymbol{\Sigma}_W \mathbf{X}_\sigma^T \\ \mathbf{X}_\sigma \boldsymbol{\Sigma}_W & \mathbf{X}_\sigma \boldsymbol{\Sigma}_W \mathbf{X}_\sigma^T \end{pmatrix}. \tag{21}$$

We now introduce $N+M$ tilted distributions $Q^{(i)}(\boldsymbol{h})$ for $i=1,\ldots,N+M$. In particular, if $i\in W$, we have

$$Q^{(i)}(\boldsymbol{h})=\frac{1}{Z_{Q^{(i)}}}\delta^M(\mathbf{G}\boldsymbol{h})\Gamma_i(h_i)\prod_{i\in W\setminus\{i\}}\phi(h_i;a_i,d_i)\prod_{\tau\in Y}\phi(h_\tau;a_\tau,d_\tau), \tag{22}$$

whereas, if $i\in Y$,

$$Q^{(i)}(\boldsymbol{h})=\frac{1}{Z_{Q^{(i)}}}\delta^M(\mathbf{G}\boldsymbol{h})\Lambda_i(h_i)\prod_{i\in W}\phi(h_i;a_i,d_i)\prod_{\tau\in Y\setminus\{i\}}\phi(h_\tau;a_\tau,d_\tau). \tag{23}$$

The tilted distributions can be expressed as the product of one of the priors and a Gaussian cavity distribution:

$$Q^{(i)}(\boldsymbol{h})=\psi_i(h_i)\tilde{Q}^{(i)}(\boldsymbol{h}), \tag{24}$$

where $\psi\in\{\Gamma,\Lambda\}$ and we have denoted the cavity distribution associated with the ith variable by $\tilde{Q}^{(i)}$:

$$\tilde{Q}^{(i)}(\boldsymbol{h})=\frac{1}{Z_{Q^{(i)}}}\delta^M(\mathbf{G}\boldsymbol{h})\exp\left(-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}}^{(i)})^T\left(\boldsymbol{\Sigma}_W^{(i)}\right)^{-1}(\boldsymbol{w}-\bar{\boldsymbol{w}}^{(i)})\right). \tag{25}$$

A factor graph representation of the tilted approximation to the posterior distribution is given in Fig. 1(a).

The cavity covariance matrices are given by the following expressions:

$$\left(\boldsymbol{\Sigma}_W^{(i)}\right)^{-1}=\begin{cases}\sum_{j\in W\setminus\{i\}}\frac{1}{d_j}\boldsymbol{e}_j\boldsymbol{e}_j^T+\mathbf{X}_\sigma^T\left(\sum_{j\in Y}\frac{1}{d_j}\boldsymbol{e}_j\boldsymbol{e}_j^T\right)\mathbf{X}_\sigma, & \text{if } i\in W,\\ \sum_{j\in W}\frac{1}{d_j}\boldsymbol{e}_j\boldsymbol{e}_j^T+\mathbf{X}_\sigma^T\left(\sum_{j\in Y\setminus\{i\}}\frac{1}{d_j}\boldsymbol{e}_j\boldsymbol{e}_j^T\right)\mathbf{X}_\sigma, & \text{if } i\in Y.\end{cases} \tag{26}$$

whereas the cavity means read

$$\bar{\boldsymbol{w}}^{(i)}=\begin{cases}\boldsymbol{\Sigma}_W^{(i)}\left(\sum_{j\in W\setminus\{i\}}\frac{a_j}{d_j}\boldsymbol{e}_j+\sum_{j\in Y}\frac{a_j}{d_j}\mathbf{X}_\sigma^T\boldsymbol{e}_j\right), & \text{if } i\in W,\\ \boldsymbol{\Sigma}_W^{(i)}\left(\sum_{j\in W}\frac{a_j}{d_j}\boldsymbol{e}_j+\sum_{j\in Y\setminus\{i\}}\frac{a_j}{d_j}\mathbf{X}_\sigma^T\boldsymbol{e}_j\right), & \text{if } i\in Y.\end{cases} \tag{27}$$

Similarly to what we obtained for the marginals of Eq. (16), we have that the marginals of Eq. (25) are Gaussian distributions with means

$$\bar{h}_i^{(i)}=\begin{cases}\bar{w}_i^{(i)}, & \text{if } i\in W\\ \boldsymbol{e}_i^T\mathbf{X}_\sigma\bar{\boldsymbol{w}}^{(i)}, & \text{if } i\in Y,\end{cases} \tag{28}$$

and variances

$$\Sigma_{ii}^{(i)}=\begin{cases}\boldsymbol{e}_i^T\boldsymbol{\Sigma}_W^{(i)}\boldsymbol{e}_i, & \text{if } i\in W,\\ \left(\boldsymbol{e}_i^T\mathbf{X}_\sigma\right)\boldsymbol{\Sigma}_W^{(i)}\left(\mathbf{X}_\sigma^T\boldsymbol{e}_i\right), & \text{if } i\in Y.\end{cases} \tag{29}$$

The yet to be determined means $\boldsymbol{a}$ and variances $\boldsymbol{d}$ of the Gaussian approximating factors (14) are determined by minimizing the Kullback-Leibler divergence $D_{KL}(Q^{(i)}||Q)$ for all $i=1,\ldots,N+M$. It can be shown that each of these minimizations is equivalent to matching the first and second moments of the tilted and of the fully Gaussian approximated distributions:

$$\langle h_i\rangle_{Q^{(i)}}=\langle h_i\rangle_Q,\quad \langle h_i^2\rangle_{Q^{(i)}}=\langle h_i^2\rangle_Q. \tag{30}$$

The EP update equations follow from the moment matching conditions (30). In particular, recalling that the marginals of $Q(\boldsymbol{h})$ are Gaussian distributions, one can express a_i and d_i in terms of the means and variances of $Q^{(i)}$ and in terms of the means and variances of the associated tilted distributions. Indeed, using the fact that the product of Gaussians is a Gaussian and the moment matching conditions, we obtain the EP update rules for the variances $\boldsymbol{d}$ and the means $\boldsymbol{a}$:

$$d_i=\left(\frac{1}{\langle h_i^2\rangle_{Q^{(i)}}-\langle h_i\rangle_{Q^{(i)}}^2}-\frac{1}{\Sigma_{ii}^{(i)}}\right)^{-1}, \tag{31}$$

$$a_i=\langle h_i\rangle_{Q^{(i)}}+\frac{d_i}{\Sigma_{ii}^{(i)}}\left(\langle h_i\rangle_{Q^{(i)}}-\bar{h}_i^{(i)}\right), \tag{32}$$

for all $i=1,\ldots,N+M$. Following [39,42], the cavity variances $\Sigma_{ii}^{(i)}$ and means $\bar{h}_i^{(i)}$ appearing in Eqs. (31) and (32) can be computed in terms of the variances Σ_{ii} and means $\bar{h}_i$ using a low rank update rule:

$$\Sigma_{ii}^{(i)}=\frac{\Sigma_{ii}}{1-\frac{1}{d_i}\Sigma_{ii}}, \tag{33}$$

$$\bar{h}_i^{(i)}=\frac{\bar{h}_i-\Sigma_{ii}\frac{a_i}{d_i}}{1-\frac{\Sigma_{ii}}{d_i}}, \tag{34}$$

which allows us to perform only one matrix inversion per iteration.

EP repeatedly estimates the vectors $\boldsymbol{a}$ and $\boldsymbol{d}$ until a fixed point is eventually reached. From a practical point of view, the algorithm returns the means and the variances of the marginal tilted distributions as soon as the convergence criterion

$$\varepsilon_t:=\max_i\left\{\left|\langle h_i\rangle_{Q_t^{(i)}}-\langle h_i\rangle_{Q_{t-1}^{(i)}}\right|+\left|\langle h_i^2\rangle_{Q_t^{(i)}}-\langle h_i^2\rangle_{Q_{t-1}^{(i)}}\right|\right\}<\varepsilon_{\text{stop}} \tag{35}$$

is fulfilled, where t denotes the current iteration and $\varepsilon_{\text{stop}}$ is a convergence threshold. In particular, the posterior mean value of weights learned by the student perceptron is estimated as given by $\langle w_i\rangle_{Q^{(i)}}$, with a standard deviation equal to $\sqrt{\langle w_i^2\rangle_{Q^{(i)}}-\langle w_i\rangle_{Q^{(i)}}^2}$.

The zero temperature formulation of EP presented in this section is computationally advantageous compared to the finite temperature one presented in Appendix A, as its complexity is dominated by the computation of the N^2 scalar products between vectors of length M that appear in the

second term of the right-hand side of Eq. (17) and by the inversion of the $N \times N$ matrix given by the same equation, resulting in a cost $O(MN^2 + N^3)$, rather than $O[(N+M)^3]$. In general, in order to reduce the computational burden related to the inversion of the covariance matrix (17), we perform a Cholesky decomposition before inverting. For more details about the finite temperature formulation of EP, we refer to Appendix A.

C. Moments of the tilted distributions

1. Moments of the spike-and-slab prior

In this section, we shall compute the first and second moments of the leave-one-out distributions when the prior is of the spike-and-slab type. We recall the expression of the spike-and-slab prior already introduced in Eq. (7) for the sake of convenience:

$$\Gamma(h_k) = (1-\rho)\delta(h_k) + \rho\sqrt{\frac{\lambda}{2\pi}}e^{-\frac{1}{2}\lambda h_k^2}, \quad k = 1, \dots, N. \tag{36}$$

The marginal tilted distribution of each weight of the student perceptron reads

$$Q^{(k)}(h_k) = \frac{1}{Z_{Q^{(k)}}}\tilde{Q}_k(h_k)\Gamma(h_k), \tag{37}$$

where we have introduced the marginalized cavity Gaussian distribution $\tilde{Q}_k$:

$$\tilde{Q}_k(h_k; \mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi\,\Sigma_k}}e^{-\frac{(h_k-\mu_k)^2}{2\Sigma_k}}. \tag{38}$$

From Eq. (37), computing the partition function of the tilted distribution $Q^{(k)}$ yields

$$Z_{Q^{(k)}} = (1-\rho)\frac{1}{\sqrt{2\pi\,\Sigma_k}}e^{-\frac{\mu_k^2}{2\Sigma_k}} + \frac{\rho}{\sqrt{2\pi}}\sqrt{\frac{\lambda}{1+\lambda\Sigma_k}}e^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}. \tag{39}$$

Finally, the first moment and the second moment of the same distribution are given by

$$\langle h_k\rangle_{Q^{(k)}} = \frac{1}{Z_{Q^{(k)}}}\frac{\rho}{\sqrt{2\pi}}e^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}\sqrt{\frac{\lambda}{1+\lambda\Sigma_k}}\frac{\mu_k}{1+\lambda\Sigma_k}, \tag{40}$$

and by

$$\langle h_k^2\rangle_{Q^{(k)}} = \frac{1}{Z_{Q^{(k)}}}\frac{\rho}{\sqrt{2\pi}}e^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}\sqrt{\frac{\lambda}{1+\lambda\Sigma_k}}\left(\frac{\Sigma_k+\lambda\Sigma_k^2+\mu_k^2}{(1+\lambda\Sigma_k)^2}\right), \tag{41}$$

respectively.

2. Moments of the theta pseudoprior

We now repeat the same reasoning for the case of the theta pseudoprior, which was defined as

$$\Lambda(h_k) = \Theta(h_k), \quad k = N+1, \dots, N+M. \tag{42}$$

The associated tilted distribution of the kth variable is given by

$$Q^{(k)}(h_k) = \frac{1}{Z_{Q^{(k)}}}\tilde{Q}_k(h_k)\Lambda(h_k), \quad k = N+1, \dots, N+M, \tag{43}$$

where the expression for $\tilde{Q}_k$ is the same as in Eq. (38). The normalization of (43) is the partition function of the tilted distribution and reads

$$Z_{Q^{(k)}} = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right], \tag{44}$$

where erf denotes the error function, defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-z^2}dz. \tag{45}$$

Computing the first moment of the marginal tilted distribution leads to the expression

$$\langle h_k\rangle = \mu_k + \sqrt{\frac{\Sigma_k}{2\pi}}\frac{e^{-\frac{\mu_k^2}{2\Sigma_k}}}{\Phi\left(\frac{\mu_k}{\sqrt{\Sigma_k}}\right)} = \mu_k\left(1 + \frac{R(\alpha_k)}{\alpha_k}\right), \tag{46}$$

where $\Phi(x) = \frac{1}{2}[1 + \mathrm{erf}(\frac{x}{\sqrt{2}})]$ is the cumulative density function of the standard normal distribution, $R(x) = \frac{1}{\sqrt{2\pi}}\frac{e^{-x^2/2}}{\Phi(x)}$, and $\alpha_k = \frac{\mu_k}{\sqrt{\Sigma_k}}$. Finally, concerning the second moment of the marginal tilted distribution, one obtains

$$\langle h_k^2\rangle = \mu_k^2 + \Sigma_k + \mu_k\sqrt{\Sigma_k}R(\alpha_k), \tag{47}$$

implying that the variance of h_k with respect to the kth marginal tilted distribution can be expressed in a compact way by

$$\mathrm{Var}(h_k) = \Sigma_k[1 - \alpha_k R(\alpha_k) - R^2(\alpha_k)]. \tag{48}$$

3. Moments of the theta mixture pseudoprior

When the pseudoprior Λ is of the theta mixture type,

$$\Lambda(h_k) = \eta\Theta(h_k) + (1-\eta)\Theta(-h_k), \quad 0 \leqslant \eta \leqslant 1, \quad k = N+1, \dots, N+M, \tag{49}$$

we have for the partition function $Z_{Q^{(k)}}$ of the tilted distributions (43)

$$Z_{Q^{(k)}} = \eta\left[\frac{1}{2}\mathrm{erfc}\left(-\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right] + (1-\eta)\left[\frac{1}{2}\mathrm{erfc}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right] = \sqrt{\frac{\pi\,\Sigma_k}{2}}\left[\frac{1}{2} + \left(\eta - \frac{1}{2}\right)\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right]. \tag{50}$$

For the first moment, one obtains

$$\begin{aligned}\langle h_k\rangle_{Q^{(k)}} &= \frac{1}{Z_{Q^{(k)}}}\left\{\frac{\eta}{\sqrt{2\pi\,\Sigma_k}}\left[\Sigma_k e^{-\frac{\mu_k^2}{2\Sigma_k}} + \mu_k\sqrt{\frac{\pi\,\Sigma_k}{2}}\mathrm{erfc}\left(-\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right] + \frac{1-\eta}{\sqrt{2\pi\,\Sigma_k}}\left[-e^{-\frac{\mu_k^2}{2\Sigma_k}} + \mu_k\sqrt{\frac{\pi\,\Sigma_k}{2}}\mathrm{erfc}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right]\right\} \\ &= \mu_k + \sqrt{\frac{2\Sigma_k}{\pi}}\frac{(2\eta-1)e^{-\frac{\mu_k^2}{2\Sigma_k}}}{\eta\mathrm{erfc}\left(-\frac{\mu_k}{\sqrt{2\Sigma_k}}\right) + (1-\eta)\mathrm{erfc}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)},\end{aligned} \tag{51}$$

and the second moment with respect to the marginal tilted distribution (43) reads

$$\langle h_k^2 \rangle_{Q^{(k)}} = \frac{1}{Z_{Q^{(k)}}} \left\{ \frac{\eta}{\sqrt{2\pi\Sigma_k}} \left[\mu_k \Sigma_k e^{-\frac{\mu_k^2}{2\Sigma_k}} + \sqrt{\frac{\pi\Sigma_k}{2}} (\mu_k^2 + \Sigma_k) \mathrm{erfc}\left(-\frac{\mu_k}{\sqrt{2\Sigma_k}} \right) \right] \right.$$
$$\left. + \frac{1-\eta}{\sqrt{2\pi\Sigma_k}} \left[-\mu_k \Sigma_k e^{-\frac{\mu_k^2}{2\Sigma_k}} + \sqrt{\frac{\pi\Sigma_k}{2}} (\mu_k^2 + \Sigma_k) \mathrm{erfc}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}} \right) \right] \right\}$$
$$= \mu_k^2 + \Sigma_k + \mu_k \sqrt{\frac{2\Sigma_k}{\pi}} \frac{(2\eta - 1) e^{-\frac{\mu_k^2}{2\Sigma_k}}}{\eta \mathrm{erfc}\left(-\frac{\mu_k}{\sqrt{2\Sigma_k}} \right) + (1-\eta) \mathrm{erfc}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}} \right)}. \tag{52}$$

III. RESULTS

A. Sparse perceptron learning from noiseless examples

In this section, we will present some results obtained from numerical simulations in the presence of noiseless examples, both in the case where patterns are i.i.d. and in a simple case of correlated patterns. For the sake of simplicity, in all the situations described in the following, we have chosen a Bayes-optimal setting, where the prior information provided by the spike-and-slab prior mirrors the actual distribution of the weights to be retrieved.

First, we performed numerical experiments with i.i.d. patterns drawn from a Gaussian distribution having zero mean and unit variance. As a performance measure, we consider the mean squared error between the normalized weights of the student perceptron at the end of the learning process and those of the teacher perceptron:

$$\mathrm{MSE}(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{B}}) = \frac{1}{N} \sum_{k=1}^{N} (\tilde{w}_k - \tilde{B}_k)^2, \tag{53}$$

where $\tilde{\boldsymbol{w}} = \boldsymbol{w}/\|\boldsymbol{w}\|$ are the rescaled weights of the student and $\tilde{\boldsymbol{B}} = \boldsymbol{B}/\|\boldsymbol{B}\|$ denote those of the teacher. In our results, this metric is expressed in decibel (dB) and used to compare expectation propagation to the 1-bit approximate message passing (1bitAMP) algorithm introduced in [27] and to the generalized vector approximate message passing (grVAMP) algorithm proposed within the unified Bayesian framework of general linear models published in [30], which the authors show to yield equivalent results to the VAMP algorithm for the generalized linear model described in [29]. We refer to Figs. 1(b) and 1(c) for the factor graphs related to the 1bitAMP and to the grVAMP approximations, respectively. The computational cost of 1bitAMP is $O(N^2)$, while the cost of grVAMP is $O(N^3)$, as it involves a one-time initial singular value decomposition of $\mathbf{X}$. However, since this computation is not needed again in the following part of the algorithm, its cost can be neglected for small enough values of N. The remaining part of the grVAMP scheme shares the same per iteration computational cost of 1bitAMP, as both are dominated by a matrix-vector product.

We considered the average of the MSE (53) over $N_{\mathrm{samples}} = 100$ simulations. The simulations correspond to sparse perceptron learning of different instances of the weights of the teacher perceptron, each from a different set of i.i.d. Gaussian patterns fed to the student perceptron. We considered the case in which the total number of weights is $N = 128$ and their density level is fixed to $\rho = 0.25$. The Gaussian part of the spike-and-slab prior was set to a standard Gaussian distribution in 1bitAMP, EP and grVAMP. The EP convergence threshold was set to $\epsilon_{\mathrm{stop}} = 10^{-4}$ and the value of the damping parameter of the EP algorithm was set equal to 0.9995 (although good results can be obtained using a lower damping too, e.g., 0.99). The results of the simulations for different values of α are reported in Fig.2 and show that EP, 1bitAMP, and grVAMP based learning from i.i.d. Gaussian patterns have roughly the same performance regardless of the specific value of α. The convergence criterion was 10^{-4} in the 1bitAMP simulations and 10^{-8} in the grVAMP simulations. All the simulations performed using EP, 1bitAMP, and grVAMP converged within the thresholds we considered. The error bars in Fig. 2 were estimated as $\sigma/\sqrt{N_{\mathrm{samples}}}$, where σ denotes the sample standard deviation of the MSE.

We also considered the problem of sparse perceptron learning from correlated patterns drawn from a multivariate normal distribution, in the simple case where the mean is $\boldsymbol{m} = \mathbf{0}$ and

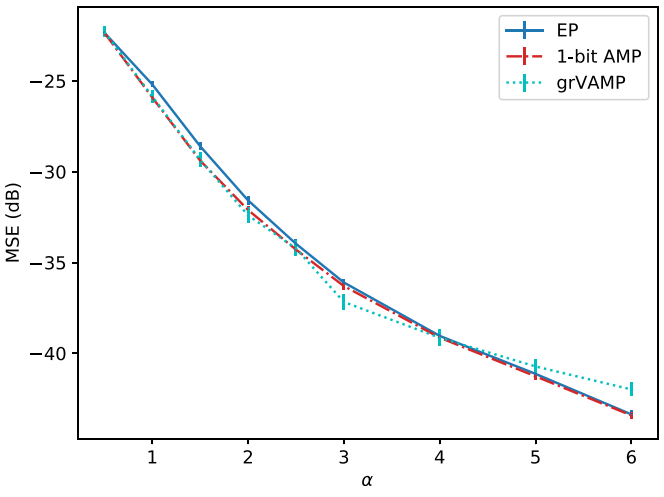

FIG. 2. MSE resulting from sparse weight learning from i.i.d. patterns using EP, 1bitAMP, and grVAMP based estimation as a function of α. The parameters considered for the perceptron are $N = 128$ and $\rho = 0.25$ and the number of instances is $N_{\mathrm{samples}} = 100$. All simulations converged and the MSE shown is averaged over all the considered instances. The error bars are estimated as $\sigma/\sqrt{N_{\mathrm{samples}}}$, where σ is the sample standard deviation of the MSE computed over all the instances.

TABLE I. Fraction of converged trials over a set of 100 different instances of the weights of the teacher perceptron and of the training set of examples. The patterns were sampled from the multivariate Gaussian distribution with covariance matrix (54). The number of variables is $N = 128$ and the density of the weights of the teacher is $\rho = 0.25$.

(a) $N = 128$					(b) $N = 256$				
α	f_{EP}	$\sigma_{f_{\mathrm{EP}}}$	f_{grVAMP}	σ_{grVAMP}	α	f_{EP}	$\sigma_{f_{\mathrm{EP}}}$	f_{grVAMP}	σ_{grVAMP}
0.5	1	0	1	0	0.5	0.99	0.01	1	0
1.0	1	0	0.96	0.02	1.0	0.73	0.04	0.99	0.01
1.5	1	0	0.89	0.03	1.5	0.92	0.03	0.94	0.02
2.0	1	0	0.83	0.04	2.0	0.96	0.02	0.88	0.03
2.5	1	0	0.87	0.03	2.5	0.88	0.03	0.83	0.04
3.0	1	0	0.87	0.03	3.0	0.88	0.03	0.73	0.04
4.0	1	0	0.85	0.04	4.0	0.92	0.03	0.80	0.04
5.0	1	0	0.82	0.04	5.0	0.94	0.02	0.76	0.04
6.0	1	0	0.80	0.04	6.0	0.96	0.02	0.70	0.05

the covariance matrix is constructed according to

$$\mathbf{S} = \mathbf{Y}^T\mathbf{Y} + \mathbf{\Delta}, \tag{54}$$

where $\mathbf{Y} \in \mathbb{R}^{u\times N}$ is an i.i.d. matrix with entries drawn from a standard univariate Gaussian distribution and $\mathbf{\Delta}$ is a diagonal matrix the eigenvalues of which are given by the absolute value of i.i.d. random entries drawn from the same distribution. By construction, this matrix is symmetric and positive definite and, therefore, is a proper covariance matrix. The diagonal matrix $\mathbf{\Delta}$ is added in order to ensure that $\mathbf{S}$ has full rank. As an extreme case, we choose $u = 1$ for the matrix $\mathbf{Y}$.

We find that even in this case the student perceptron is able to estimate the weights of the teacher, although, under the same values of the parameters N, ρ, and α of the model and under the same values of the EP parameters (i.e., damping, ϵ_{stop}, and maximum number of iterations), the accuracy of the estimation is lower than the one achieved by learning from i.i.d. Gaussian patterns, as one might expect. Still, in the presence of the correlated patterns considered here, expectation propagation based learning proves to be advantageous as compared with other algorithms for 1-bit compressed sensing such as 1bitAMP, the estimates of which of the means and variances of the weights to be retrieved diverge. In addition, in the same situation, EP outperforms grVAMP based learning, as shown in Fig. 6 for the set of parameters $N = 128$ and $\rho = 0.25$. The convergence thresholds of the grVAMP and of the EP algorithms were set to the same values as in the case of learning from i.i.d. Gaussian patterns and further lowering the value of the threshold parameter of grVAMP did not result in a noticeable improvement of the grVAMP results. In the case of EP, the damping factor was set to 0.999 and the maximum number of iterations for convergence was 50 000. The fraction of converged trials is shown in Table I for both algorithms in the case where $N = 128$ and $\rho = 0.25$. The EP led student perceptron is more accurate at determining the nonzero weights than the grVAMP led counterpart, as shown by the receiver operating characteristic curves (ROC curves) in Fig. 3(a) and by the sensitivity plots of Fig. 3(b). In order to construct these curves, each weight of the teacher was assigned a score given by its probability of being nonzero as estimated by EP and grVAMP. The weights of the teacher were sorted in decreasing order according to these probabilities, which are given by

$$P_k^{\neq 0} = \left[1 + \left(\frac{1}{\rho} - 1\right)\sqrt{\frac{1+\lambda\Sigma_k}{\lambda\Sigma_k}}\, e^{-\frac{\mu_k^2}{2\Sigma_k(1+\lambda\Sigma_k)}}\right]^{-1}. \tag{55}$$

In the case of EP, μ_k and Σ_k, for $k = 1, \ldots, N$, are the EP cavity means and variances, whereas, in the case of grVAMP, $\boldsymbol{\mu}$ corresponds to the VAMP quantity $\mathbf{r}_{1k}$ and $\Sigma_k = \gamma_{1k}^{-1}$, where γ_{1k} is the quantity that parametrizes the denoiser in VAMP [28] and we used the standard VAMP notation for $\mathbf{r}_{1k}$ and γ_{1k}, for which the index k refers to the number of the current iteration. In both cases, ρ denotes the density parameter of the spike-and-slab prior. Interestingly, the discrepancy between the accuracy of the two algorithms becomes larger as the number of patterns increases and, as a consequence, the difference between the mean squared errors of the two algorithms increases, as shown in Fig. 3(c), implying that the EP and the grVAMP approximations are very different in this Gaussian correlated pattern regime. This fact is confirmed by the heterogeneity of the variances $d_k, \quad k = 1, \ldots, N$ of the approximating univariate Gaussian factors ϕ_k when one considers the EP solution for instances where both EP and grVAMP converge. More precisely, it can be seen that the parameters d_k span several orders of magnitude, contrary to VAMP where these variances are constrained to be equal. In each plot in Fig. 3, we have shown the average of the quantities considered over the set of the N_{conv} instances for which each algorithm achieved convergence. Accordingly, the error bars were estimated as $\sigma/\sqrt{N_{\mathrm{conv}}}$, where σ denotes the standard deviation over the same set of instances.

A useful additional feature of the EP-based learning approach is the possibility to learn iteratively the value of ρ during the estimation of the weights of the teacher, as, unlike EP, many algorithms for 1-bit compressed sensing assume the density of the signal to be given *a priori*. The estimation of the density parameter is achieved by minimizing the EP free energy with respect to ρ and yields good results as long as the number of the patterns presented to the student is large enough. We refer to the Appendices for details concerning the EP free energy, its expression for the sparse perceptron learning problem, and free energy optimization based learning of the parameters of the prior. We mention here that a similar expectation maximization based strategy can be implemented

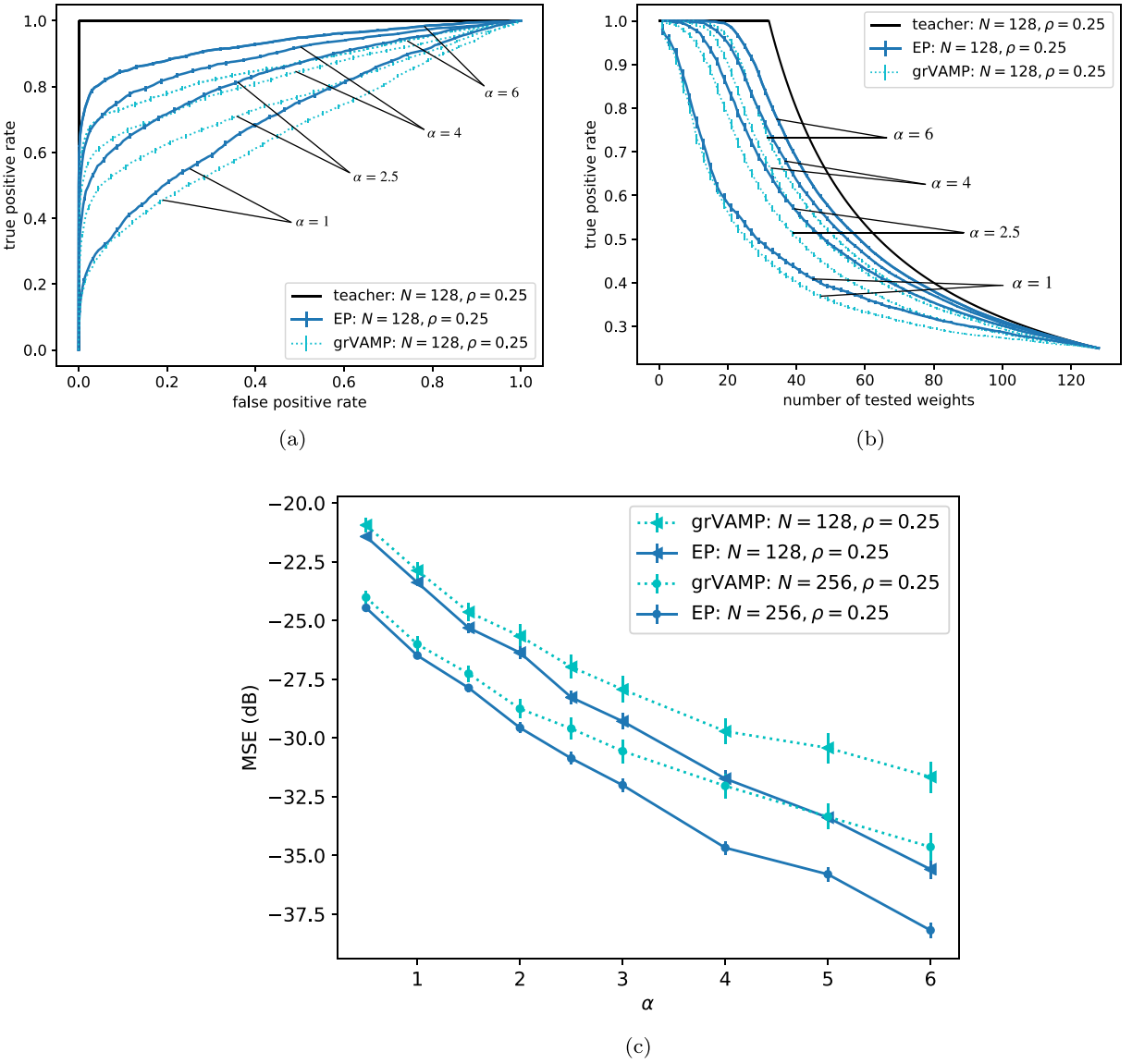

FIG. 3. Sparse perceptron learning from correlated patterns sampled from a multivariate Gaussian distribution. The values of the parameters are specified in each panel and we set $u = 1$. Comparison between grVAMP and EP based learning. (a) ROC curves. (b) Sensitivity plots. For reference, in (a) and (b) the case of ideal variable selection by the teacher perceptron that provided the examples is also shown. (c) Mean squared error in dB. 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 $\sigma/\sqrt{N_{\text{conv}}}$, where σ is the sample standard deviation over the same set of instances.

also in the case of 1bitAMP and grVAMP in order to estimate the density parameter ρ.

In order to show that our approach allows one to estimate the dilution level of the teacher perceptron, we performed a set of $N_{\text{samples}} = 100$ EP simulations on a system with $N = 128$ and $\rho = 0.25$, where the density parameter ρ_0 of the spike-and-slab prior assigned to each weight variable was randomly initialized by sampling its value from a uniform distribution over the interval $0.05 \leqslant \rho_0 \leqslant 0.95$ and where the learning rate was chosen to be $\delta\rho = 10^{-5}$. We show our results in Table II. For each value of α, we show the average value ρ_L of the density estimate over all samples and its associated statistical uncertainty, which was computed as $\delta\rho_L = \sigma_{\rho_L}/\sqrt{N_{\text{samples}}}$, as for these values of the parameters all simulations converged. We also show the relative difference $\Delta\rho/\rho$ between the true value of the density and the estimated one. Since $\Delta\rho \gg \delta\rho_L$, we omit the statistical uncertainty associated with $\Delta\rho$. Finally, we notice that, even when learning from correlated patterns constructed as described above, the student is able to estimate the density level of the weights of the teacher perceptron quite accurately, provided that a sufficient number of patterns is provided to the student perceptron. In Table II, we give an example of this fact when the teacher perceptron has $N = 128$ weights and density $\rho = 0.25$.

B. Sparse perceptron learning from a noisy teacher

We analyzed the performance of EP based sparse perceptron learning when a small fraction of the examples is mislabeled. The student perceptron is given the *a priori* information that a certain fraction of the labels is wrongly assigned. As in the noiseless case, we consider a Bayes-optimal

TABLE II. Learning of the density ρ of the weights of the teacher for a perceptron with parameters $N = 128$ and $\rho = 0.25$. The average and the standard deviation of the learned value of ρ at convergence over all the trials for which convergence was achieved during the training process are denoted by ρ_L and $\delta\rho_L$, respectively. In each trial, the initial condition ρ_0 was drawn uniformly from the interval $0.05 \leqslant \rho_0 \leqslant 0.95$.

α	i.i.d. patterns: $\rho_L \pm \delta\rho_L$	i.i.d. patterns: $\Delta\rho/\rho$	Patterns from MVN: $\rho_L \pm \delta\rho_L$	Patterns from MVN: $\Delta\rho/\rho$
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

setting and therefore we provide the student with such *a priori* knowledge using the theta mixture pseudoprior introduced in Eq. (49).

In order to test and evaluate the performance of EP-guided learning in this situation, we compared EP with grVAMP, in which we introduced a theta mixture measure as done in EP, and with the R1BCS algorithm for 1-bit compressed sensing with sign-flip errors proposed by Li *et al.* [35], which is based on an expectation maximization scheme involving both the signal to be retrieved and the noise. In grVAMP, the slab part of the spike-and-slab prior was set to a standard Gaussian distribution. When using R1BCS, we rescaled the pattern matrix so that each column had unit norm and we included a convergence threshold $\varepsilon_{\text{R1BCS}} = 10^{-4}$. Thus, the R1BCS iterations stop when the estimate $\boldsymbol{w}_{\text{R1BCS}}$ of the weights of the teacher is such that $\|\boldsymbol{w}_{\text{R1BCS}} - \boldsymbol{w}^{\text{old}}_{\text{R1BCS}}\| < \varepsilon_{\text{R1BCS}}$. In each experiment, a given number $K_{\text{label}} = (1-\eta)M$ of labels were flipped, where η is the fraction of unchanged labels. In the EP simulations a damping factor equal to 0.99 and a convergence threshold $\epsilon_{\text{stop}} = 10^{-4}$ were used. The parameter λ of the spike-and-slab prior was set equal to 10^4.

We first considered the case of i.i.d. Gaussian patterns drawn from a Gaussian distribution with zero mean and unit variance and, subsequently, the case of correlated patterns drawn from the zero mean multivariate Gaussian distribution with the covariance matrix expressed in (54), which we already introduced in the noiseless case. For both kinds of patterns, we conducted numerical experiments on a set of 100 different instances. In the case of i.i.d. patterns, convergence over all instances was achieved for all three algorithms within the convergence thresholds specified. The same was true in the case of correlated patterns for R1BCS and for EP, whereas grVAMP exhibited a failure rate up to 15% in terms of convergence in the correlated case.

In order to assess the variable selection capabilities of EP when attempting to learn the weights of the teacher from these kinds of patterns in the presence of mislabeled examples, we computed the ROC curves and the sensitivity plots related to both the former [Figs. 4(a) and 4(b), respectively] and to the latter kind of patterns [Figs. 4(c) and 4(d), respectively]. Such ROC curves and sensitivity plots are associated with the weights of the student after the training phase was completed and were obtained in the case where the number of weights was $N = 128$, the density of the weights $\boldsymbol{B}$ of the teacher was $\rho = 0.25$, and the fraction of uncorrupted labels was $\eta = 0.95$. The ordering criterion for the weights $\boldsymbol{B}$ adopted in the ROC curves in Figs. 4(a) and 4(c) and in the sensitivity plots in Figs. 4(b) and 4(d) was based on the absolute value of the weights $\boldsymbol{w}$ of the student. In the case of EP and grVAMP, we also plotted the ROC and sensitivity curves according to the sorting criterion based on the score expressed in Eq. (55), but the results did not exhibit noticeable differences with respect to those obtained using the previous sorting criterion, as the curves in Fig. 4 mostly overlap. The ROC curves and sensitivity plots associated with EP and grVAMP mostly exhibit similar values for the true positive ratio—except for large α, where the EP values are slightly smaller—and outperform significatively the ones related to R1BCS in the case of i.i.d. patterns, especially at low values of α. This is confirmed by the values of the areas under the curves, as shown in Table III, where the maximum relative discrepancy between the areas under the ROC curve (AUC) is 0.008 when considering EP and grVAMP and 0.03 when considering EP and R1BCS. However, in the case of correlated patterns, the ROC curves and sensitivity plots related to EP are mostly comparable to those associated with R1BCS and yield values of the true positive rate that are systematically larger than the ones of grVAMP, implying that the variable selection properties of EP tend to be affected to a far lesser extent than those of grVAMP in this regime, as confirmed by the AUC reported in Table III and analogously to what was observed in the noiseless case. In this case, the maximum relative discrepancy between the AUC of EP and of R1BCS is 0.05 and is attained at the lowest value of α, namely, $\alpha = 0.5$, whereas, when considering EP and grVAMP, the relative discrepancy is largest at large values of α, its maximum being 0.06 at $\alpha = 6$. The mean squared error values associated with the normalized student and teacher are plotted in Fig. 5 for both $\eta = 0.95$ and 0.9 and their values are shown both in dB (main plots) and using a linear scale (insets). In particular, the MSE values reported in Fig. 5(a) correspond to the ROC curves in Fig. 4(a). While the MSE of EP is definitely larger than that of grVAMP at large α in the presence of i.i.d. patterns [Figs. 5(a) and 5(b)], it should be noted that the differences of the values involved are very small and are not noticeable on a linear scale. Since the EP approximation is richer than the VAMP one, which, in turn, is due to the fact that EP reduces to VAMP when the approximate univariate prior factors are constrained to have the same variance, and because of the very small errors and differences observed, we believe that such observed discrepancy between grVAMP and EP in the i.i.d. pattern regime at large α is due to numerical effects rather than to intrinsic limitations of the EP scheme. Instead, this is not the case for the differences displayed by the MSE related to EP and grVAMP in Fig. 5(c) for $\eta = 0.95$, which are clearly noticeable on a linear scale as well. The discrepancy involved

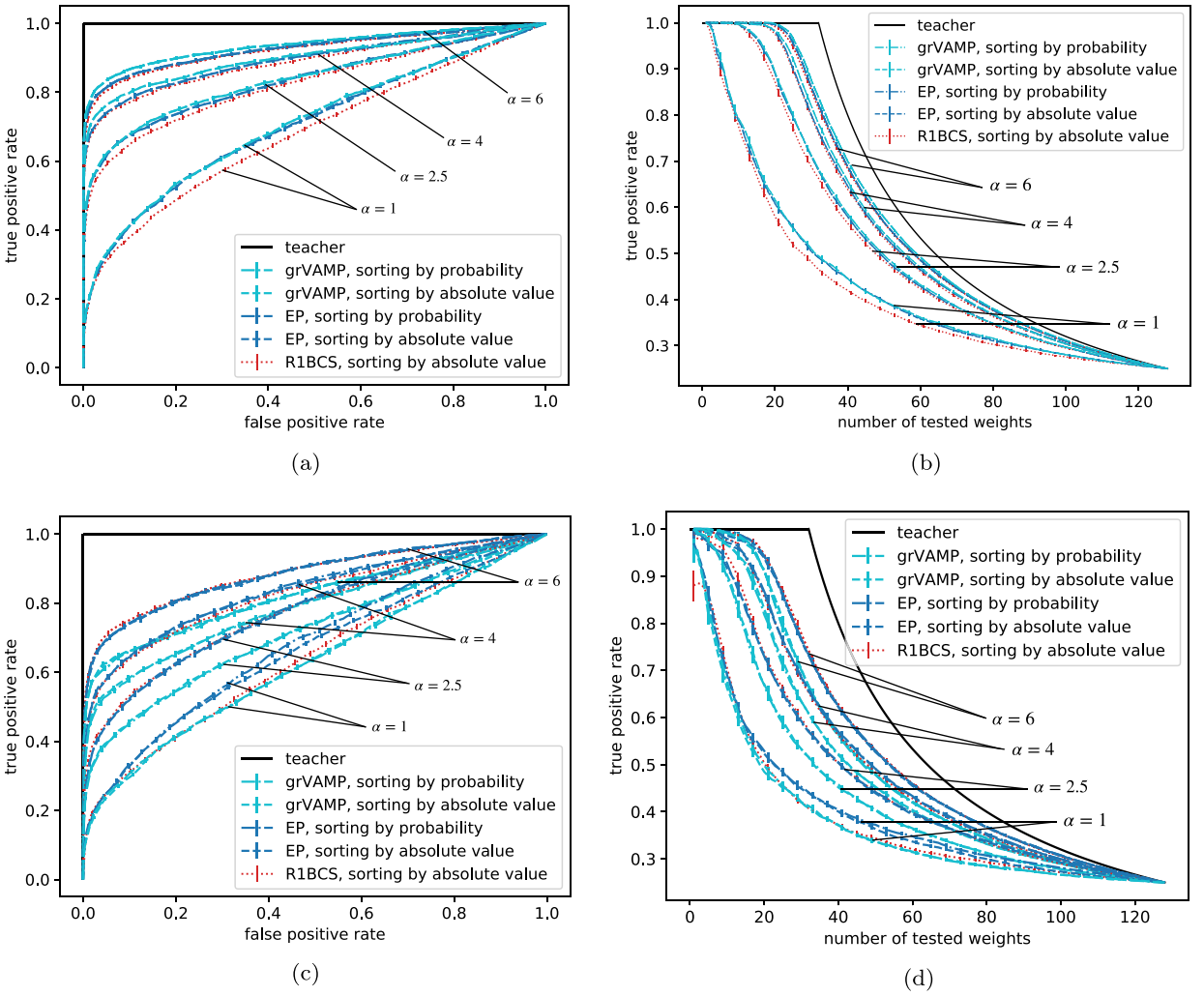

FIG. 4. Sparse perceptron learning from i.i.d. patterns sampled from a standard Gaussian distribution and from correlated patterns sampled from a multivariate Gaussian distribution, with $N = 128$, $\rho = 0.25$, $u = 1$, $\eta = 0.95$. A fraction $(1 - \eta)$ of the labels are mislabeled. Comparison between R1BCS, grVAMP, and EP in terms of their ROC curves (a, c) and of their sensitivity plots (b, d). For reference, the case of ideal variable selection by the teacher perceptron that provided the examples is shown. The plotted quantities are the mean values computed over the set of all N_{samples} instances and the error bars are estimated as $\sigma/\sqrt{N_{\text{samples}}}$ where σ denotes the sample standard deviation over all the instances considered.

TABLE III. (a) AUC scores associated with the ROC curves shown in Fig. 4(a), which correspond to EP, grVAMP, and R1BCS based classification from i.i.d. Gaussian patterns in the presence of label noise, with $\eta = 0.95$. (b) AUC scores associated with the ROC curves shown in Fig. 4(c), which correspond to EP, grVAMP, and R1BCS based classification from multivariate Gaussian patterns in the presence of label noise, with $\eta = 0.95$.

(a) AUC (i.i.d. patterns)

α	AUC_{EP}	$\text{AUC}_{\text{grVAMP}}$	$\text{AUC}_{\text{R1BCS}}$
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

(b) AUC (patterns from MVN)

α	AUC_{EP}	$\text{AUC}_{\text{grVAMP}}$	$\text{AUC}_{\text{R1BCS}}$
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

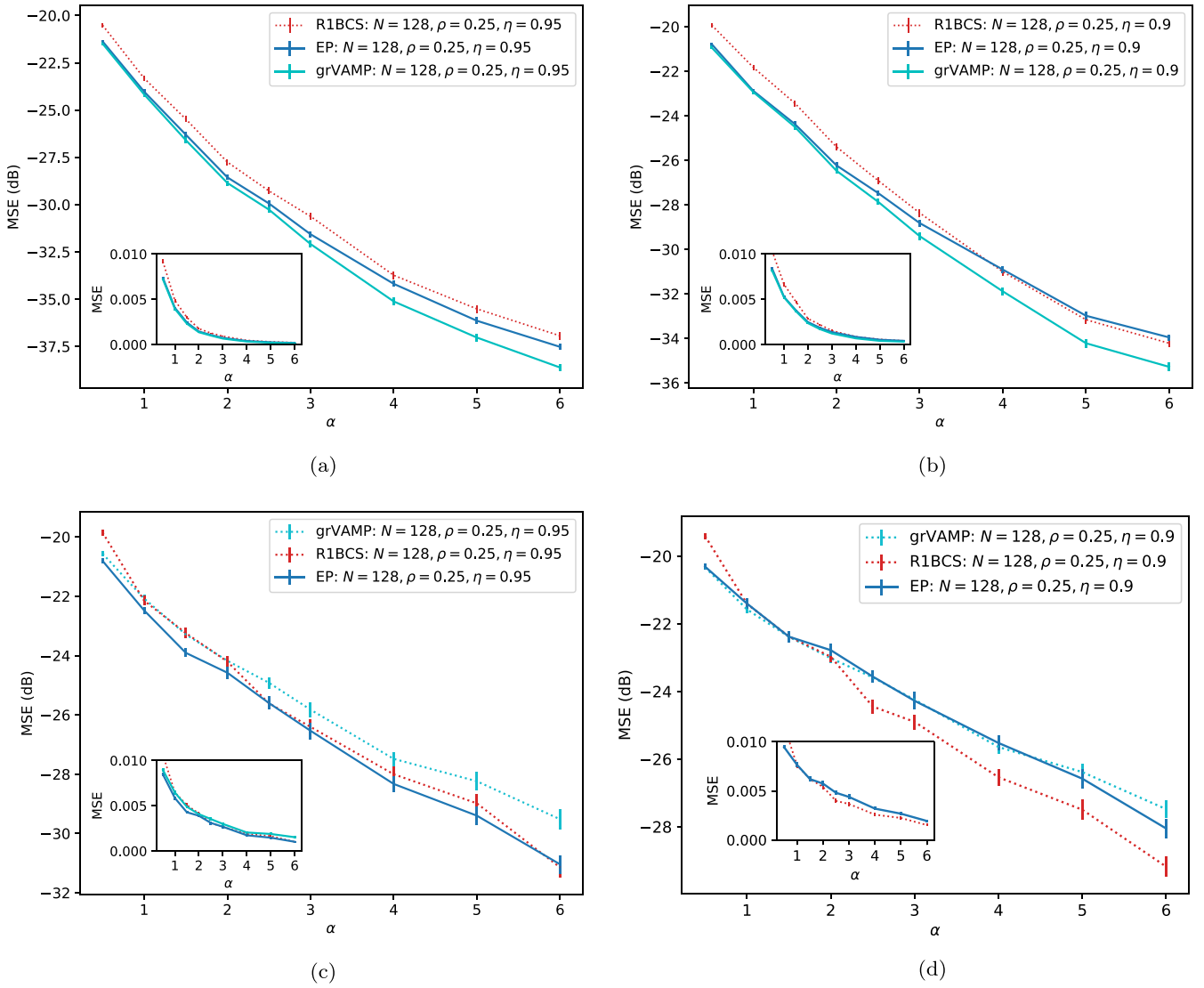

FIG. 5. Sparse perceptron learning from $(1-\eta)M$ mislabeled examples: comparison between EP, grVAMP, and R1BCS in terms of their mean squared errors in the case of (a) i.i.d. Gaussian patterns, $\eta = 0.95$; (b) i.i.d. Gaussian patterns, $\eta = 0.9$; (c) Gaussian correlated patterns, $\eta = 0.95$; and (d) Gaussian correlated patterns, $\eta = 0.9$. In all figures, $N = 128$ and $\rho = 0.25$. The mean squared errors plotted are averaged over the set of all N_{samples} instances and the error bars are estimated as $\sigma/\sqrt{N_{\text{samples}}}$ where σ denotes the sample standard deviation over all the instances considered.

reflects that observed at the level of the ROC curves shown in Fig. 4(c). Finally, in the Gaussian correlated scenario, as soon as the noise level affecting the labels becomes large enough, we see that EP and grVAMP yield similar results at all values of α, as shown in Fig. 5(d).

Similarly to R1BCS, EP was able to correctly estimate the noise level affecting the labels. This was achieved by sampling the initial condition η_0 for the parameter η of the theta mixture pseudoprior uniformly from the interval $0.5 < \eta < 1$ and performing one step of gradient descent on the EP free energy at each EP iteration as described in Appendix C. In addition, analogously to the noiseless case, it is possible to learn the ρ parameter and we verified that the two parameters can be learned simultaneously. We show the estimated values of the density of the weights and of the consistency level η of the labels in the case of i.i.d. patterns in Table IV and in the case of Gaussian correlated patterns in Table IV. We notice that, analogously to EP, an expectation maximization scheme can be implemented in the case of grVAMP in order to iteratively learn the density parameter ρ of the spike-and-slab prior and the η parameter of the theta mixture measure.

One important limitation of the R1BCS algorithm as compared with EP (and grVAMP) is that it involves the inversion both of a $N \times N$ matrix and of a $M \times M$ matrix at each iteration. As a consequence, the computational complexity of R1BCS is dominated by $O[(1+\alpha^3)N^3]$ operations. Therefore, from a computational point of view, EP is especially advantageous as compared to R1BCS when the number of patterns in the training set is large, as, in the EP formulation proposed in this paper, the computational cost is of order $O[(1+\alpha)N^3]$. However, in the large N regime, grVAMP will be faster than EP (and R1BCS) as the only operation that

TABLE IV. Values of the η 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 multivariate Gaussian patterns in the case $N = 128$. The estimated value of η is denoted as η_L, the true value being $\eta = 0.95$, whereas the true value of the density parameter of the spike-and-slab prior was given by $\rho = 0.25$.

(a) i.i.d. patterns					(b) Patterns from MVN				
α	ρ_L	$\Delta\rho/\rho$	η_L	$\Delta\eta/\eta$	α	ρ_L	$\Delta\rho/\rho$	η_L	$\Delta\eta/\eta$
2.5	0.229 ± 0.004	0.08	0.964 ± 0.001	0.02	2.5	0.206 ± 0.006	0.2	0.951 ± 0.002	0.0006
3.0	0.234 ± 0.003	0.07	0.957 ± 0.003	0.007	3.0	0.208 ± 0.006	0.2	0.953 ± 0.001	0.003
4.0	0.247 ± 0.004	0.01	0.9584 ± 0.0005	0.009	4.0	0.228 ± 0.005	0.09	0.953 ± 0.001	0.003
5.0	0.249 ± 0.003	0.003	0.9561 ± 0.0007	0.006	5.0	0.23 ± 0.005	0.08	0.9526 ± 0.0005	0.003
6.0	0.252 ± 0.003	0.007	0.9544 ± 0.0004	0.005	6.0	0.236 ± 0.004	0.05	0.9529 ± 0.0004	0.003

is $O(N^3)$ needs to be performed only once, whereas in EP the operations with cubic cost must be performed at each iteration. In fact, while the VAMP part in grVAMP is faster than EP for the sizes simulated in this paper, we recall that the grVAMP algorithm is composed of two modules [see also Fig. 1(c)] implemented as two nested loops: the external loop corresponds to a minimum mean squared error estimation of the linear projection vector $\boldsymbol{z} = \mathbf{X}\boldsymbol{w}$, where $\mathbf{X}$ denotes the matrix of the patterns, under a Gaussian prior and a likelihood having the same functional form of the Λ pseudoprior used in EP, whereas the inner loop consists of a VAMP module running on a standard linear model, the measurement vector of which is given by the current estimate of $\boldsymbol{z}$, for a predefined number of iterations. We show the running times of R1BCS, EP, and grVAMP in Table V for the simulated size $N = 128$. All the simulations were performed in parallel on a HP Proliant server with 64 cores clocked at 2.1 GHz. In grVAMP, the maximum number of iterations of the outer MMSE module loop was set to 1000, while the number of iterations of the inner VAMP module was set to 2000. The comparisons shown in Table V are only meant to give an idea of the running times observed for the implementations that we used, which can be found in [43]. In particular, grVAMP was adapted by introducing the theta mixture measure and parameters were set in each algorithm as explained in this section. As the VAMP estimation is repeated for every iteration of the external loop, the running times related to grVAMP appear to be larger than those related to EP.

C. Correlated patterns generated by a recurrent neural network

As an example of a diluted network with correlated inputs, we consider a network of N randomly diluted perceptrons without self-loops. We will denote the ith row of the weight matrix $\mathbf{W} \in \mathbb{R}^{N\times(N-1)}$ as $\boldsymbol{w}_i$. Each entry of $\boldsymbol{w}_i$ is the weight of an incoming link of the ith perceptron. Each perceptron receives binary inputs $\boldsymbol{x}$ generated according to a Glauber dynamics at zero temperature. We considered both the case of synchronous update of the patterns at each time step and the case where the binary inputs are updated asynchronously.

In the case of synchronous update, starting from an initial random vector $\boldsymbol{x}_0 = \mathrm{sgn}(\boldsymbol{\xi}_0)$, where $\boldsymbol{\xi}_0 \sim \mathcal{N}(\boldsymbol{\xi}; \mathbf{0}, \mathbf{I})$ at discrete time $t = 0$ and given a pattern $\boldsymbol{x}^t$ at time t, each perceptron computes its output at time $t + 1$ according to

$$z_i^t = \boldsymbol{w}_i^T \boldsymbol{x}_{\setminus i}^t, \tag{56}$$

and to

$$x_i^{t+1} = \mathrm{sgn}\left(z_i^t\right), \tag{57}$$

where $\boldsymbol{x}_{\setminus i}^t$ denotes the vector of outputs produced by all perceptrons except the ith one at time t. The patterns at time $t + 1$ are given by the set of outputs resulting from Eqs. (56) and

TABLE V. Running time related to the EP and R1BCS based sparse perceptron learning from (a) i.i.d. Gaussian patterns and from (b) Gaussian patterns from multivariate normal distribution with covariance matrix given by Eq. (54) ($u = 1$) in the presence of label noise, with $\eta = 0.95$ and damping factor equal to 0.99 in the case of EP. The uncertainty on these values was estimated as $\sigma/\sqrt{N_{\mathrm{conv}}}$, where σ is the sample standard deviation over the set of converged trials and N_{conv} is the number of converged simulations.

(a) i.i.d. patterns				(b) Patterns from MVN			
α	t_{EP} (s)	t_{grVAMP} (s)	t_{R1BCS} (s)	α	t_{EP} (s)	t_{grVAMP} (s)	t_{R1BCS} (s)
0.5	2.7 ± 0.2	126.7 ± 0.8	6.5 ± 0.2	0.5	2.7 ± 0.2	139.2 ± 2.0	6.7 ± 0.2
1.0	2.6 ± 0.03	134.6 ± 0.9	12.0 ± 0.3	1.0	3.7 ± 0.1	155.7 ± 2.0	14.4 ± 0.4
1.5	3.74 ± 0.04	147.0 ± 1.0	22.2 ± 0.5	1.5	4.8 ± 0.1	206.8 ± 20.0	27.0 ± 0.7
2.0	5.65 ± 0.08	159.0 ± 1.0	47.6 ± 1.0	2.0	6.2 ± 0.2	289.0 ± 40.0	53.6 ± 1.0
2.5	6.97 ± 0.09	175.7 ± 1.0	94.9 ± 2.0	2.5	7.7 ± 0.2	236.3 ± 20.0	105.8 ± 3.0
3.0	8.3 ± 0.1	212.9 ± 20.0	150.9 ± 3.0	3.0	9.5 ± 0.3	294.3 ± 30.0	158.5 ± 4.0
4.0	10.2 ± 0.1	258.3 ± 10.0	373.4 ± 7.0	4.0	11.4 ± 0.3	366.9 ± 40.0	379.4 ± 8.0
5.0	12.2 ± 0.1	311.6 ± 10.0	628.0 ± 10.0	5.0	14.1 ± 0.4	423.5 ± 40.0	592.7 ± 10.0
6.0	15.6 ± 0.2	353.9 ± 10.0	1052.9 ± 20.0	6.0	15.9 ± 0.4	498.5 ± 50.0	1057.2 ± 20.0

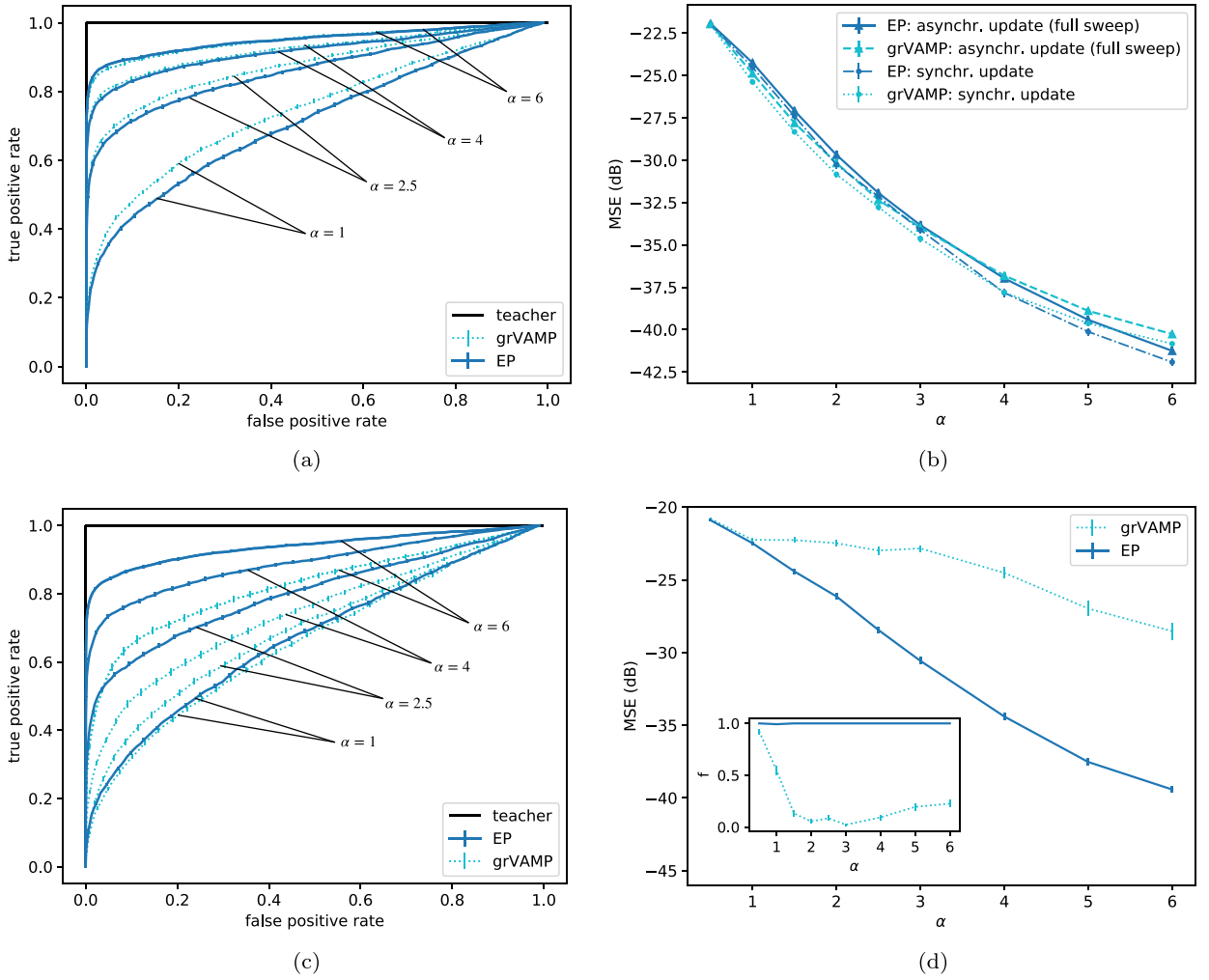

FIG. 6. Comparison between grVAMP and EP based sparse perceptron learning from correlated patterns generated by a recurrent network according to a Glauber dynamics. The case of *weakly correlated* patterns and that of *strongly correlated* patterns are shown in panels (a) and (b) and in panels (c) and (d), respectively. The number of perceptrons in the network is $N = 128$ and the density of the weights of each perceptron is $\rho = 0.25$. In each plot, 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 is the sample standard deviation computed over the set of N trained student perceptrons. (a) ROC curves related to the estimated topology of the learned weights of each perceptron for several values of α when patterns are synchronously updated. All N perceptrons achieved convergence during the training task. (b) MSE (in dB) associated with the synchronously updated patterns of (a) and with the asynchronous one in which patterns are included in the training set only after each perceptron was selected to yield the corresponding update ("full sweep"). (c) ROC curves related to the estimated topology of the weights learned from asynchronously updated patterns for several values of α. The Hamming distance between sets of patterns at consecutive times was chosen equal to 10. (d) MSE (in dB) corresponding to the case of the patterns considered in (c). The fraction of perceptrons of the recurrent network the training tasks of which achieved convergence is shown in the inset.

(57). While in principle such a recurrent network dynamics could become trapped in a limit cycle when coupled to a synchronous update rule for the patterns, in practice this never happened when generating such patterns from a recurrent network of $N = 128$ diluted perceptrons, each of which having dilution level $\rho = 0.25$.

In the case of asynchronous updates, at each time step one perceptron i is selected at random and, given a pattern $\boldsymbol{x}^t$ at time t, the ith component of the pattern vector at time $t + 1$ is computed according to Eqs. (56) and (57), while all other components are left unchanged. In order to tune the degree of correlation of the patterns, we generated them in such a way that two sets of patterns at consecutive times have a given Hamming distance. In order to do so, we ran the Glauber dynamics of the recurrent network as described and updated the patterns asynchronously, but only stored their configuration when the desired Hamming distance between the candidate set of patterns at time $t + 1$ and the set of patterns at time t was achieved.

We consider the noiseless case and analyze the performance of training all the perceptrons of a student network of the same type of the network that generated the patterns using both EP and grVAMP. We tested the two methods on a network with $N = 128$ perceptrons and density parameter of each perceptron given by $\rho = 0.25$. We considered both the case where the patterns are generated with the synchronous

TABLE VI. (a) Estimate of the parameter η of the theta mixture pseudoprior resulting from perceptron learning from the same patterns with label noise. The true unknown value of η is 0.9. (b) Estimated value of the density of the weights of single perceptrons in the network at large alphas, both in the case of synchronously updated patterns and in the case of asynchronous update with fixed Hamming distance $d_H = 10$ between patterns at consecutive time steps, without label noise. In both (a) and (b), $N = 128$ and $\rho = 0.25$.

(a)

α	η_L, synchr. update	$\Delta\eta/\eta$, synchr. update	η_L, asynchr. update ($d_H = 10$)	$\Delta\eta/\eta$, asynchr. update
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

(b)

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

update rule presented above and the case of asynchronous update. In both cases, 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 Gaussian part of the spike-and-slab prior was set to a standard normal distribution in both EP and grVAMP.

When considering patterns generated with a synchronous update, both algorithms achieved convergence during the training task for all perceptrons of the recurrent network. We evaluated the ROC curves and the reconstruction error associated with the correct or incorrect selection of the nonzero weights at the level of a single perceptron by considering the whole network and computing the average and the standard deviation of the relevant quantities (i.e., the MSE, the fraction of false positives, and the fraction of true positives) over all the perceptrons of the student network. Both algorithms were able to select the same fraction of relevant weights, as shown by the vertical portion of the corresponding ROC curves in Fig. 6(a), which were constructed by sorting the weights according to their absolute value, in decreasing order. As a consequence, both grVAMP and EP showed comparable values for the MSE between the weights estimated by the student perceptrons and the ones of the corresponding teachers as a function of α. In addition, similar values of the MSE were obtained for both algorithms by generating weakly correlated patterns according to an asynchronous generative process and including the generated vector of patterns in the training set only after a full sweep of N updates, i.e., one for each perceptron of the teacher recurrent network. In this situation, the convergence rate of grVAMP was still very large ($\geqslant$95%) at all values of α. We show the values of the MSE obtained from these two kinds of weakly correlated patterns in Fig. 6(b).

We repeated the same numerical experiments with patterns generated performing an asynchronous update of the perceptron outputs and fixing $d_H(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t) = 10$, corresponding to a Pearson correlation coefficient $r_{\text{Pearson}} = 0.84$ between pattern vectors at consecutive times. In this case, the observed convergence rate within the student network was significantly lower for grVAMP than for EP. Although the performances of EP and grVAMP were comparable when considering the subset of perceptrons the grVAMP guided learning tasks of which converged (not shown), the ROC curves [Fig. 6(c)] and MSE values [Fig. 6(d)] obtained from these experiments, as evaluated over the N perceptrons of the student network regardless of their convergence status at the end of the training task, show that grVAMP is not able to accurately estimate the weights of the teacher when it fails to converge. The fraction of perceptrons of the recurrent network the training task of which successfully converged is plotted in the inset of Fig. 6(d) for both EP and grVAMP. These results strongly suggest that the observed convergence failure of grVAMP is related to the degree of correlation of the patterns presented to the student.

We verified that the student perceptrons were able to estimate the density of the weights of the teacher perceptrons quite accurately while learning the classification rule of the teacher from this kind of patterns using EP and we checked that it is still possible to estimate the parameter η of the theta mixture pseudoprior during the training phase if a fraction of the labels are corrupted by noise, provided that the noise level $1 - \eta$ is not too large, similarly to what we observed in the noisy case with i.i.d. patterns and with the correlated patterns drawn from a multivariate normal distribution analyzed in Sec. III B of the paper. We provide the estimated noise level in the noisy scenario where $\eta = 0.9$ as well as some examples of the estimated values of the density at large values of α in Table VI. The estimate of the parameter η proves to be fairly accurate in the case of patterns generated with a synchronous update rule, but is overestimated in the case where the patterns are generated using an asynchronous update rule, as shown in Table VI.

IV. CONCLUSIONS

In this paper we have proposed an expectation propagation based strategy for efficient 1-bit compressed sensing reconstruction, the computational complexity of which is dominated by $O[(1+\alpha)N^3]$ elementary operations. We analyzed the behavior in the zero temperature case and assuming that the patterns are generated by a teacher having the same structure as the student.

The performance of the algorithm has been extensively tested under several conditions. For i.i.d. patterns generated by a Gaussian distribution of zero mean and unit variance, the algorithm performance is on par with two other state-of-the art algorithms: 1bitAMP [27] and grVAMP [30]. However, in the correlated Gaussian pattern case, where 1bitAMP fails to converge, EP outperforms grVAMP both in terms of accuracy and specificity. Moreover, both in the i.i.d. and correlated Gaussian case, EP is able to learn with remarkable accuracy the density ρ of the weights of the teacher during the retrieval task. This feature of variational method (EP, 1bitAMP, grVAMP, etc.) puts them at an advantage over other algorithms for 1-bit compressed sensing which require that the dilution level is known and provided among their inputs.

We then tested the robustness of EP reconstruction against noise. To do so, we mislabeled a fraction $(1-\eta)$ of the examples and compared EP with grVAMP and with the R1BCS algorithm [35] for 1-bit compressed sensing with sign flip. Again, as in the noiseless case discussed above, we first considered the i.i.d. pattern case and then the correlated Gaussian pattern case. The ROC curves and the sensitivity plots associated with EP, grVAMP, and R1BCS allow us to conclude that the variable selection properties displayed by the three algorithms are not very different in the case of i.i.d. patterns. However, in the case of correlated Gaussian patterns and as long as the noise level affecting the labels is small enough, while on the one hand EP appears to be mostly comparable to R1BCS, on the other hand it systematically and consistently outperforms grVAMP regardless of the specific value of α. In addition to the estimation of the density parameter, which we mentioned above in the case of sparse perceptron learning from noiseless examples, EP was able to successfully retrieve the noise level affecting the labels by including in the algorithm an analogous optimization strategy consisting in one step of gradient descent on the EP free energy at each iteration. One important limitation of R1BCS in comparison with EP is the computational complexity: while the computational complexity of EP is $O[(1+\alpha)N^3]$ (at least in the zero temperature case), R1BCS scales as $O[(1+\alpha^3)N^3]$. Therefore, from a computational point of view, EP turns out to be especially advantageous in the large α regime as compared to R1BCS, but does not scale as well as grVAMP (and of course 1bitAMP) in terms of simulation time as a function of N.

Finally, we explored a more realistic scenario for "temporally" correlated patterns generated by a recurrent network of N randomly diluted perceptrons both in the case of synchronous and asynchronous update schemes. First, we compared the performance of EP and grVAMP in the noiseless case. The first striking observation is the poor performance of grVAMP in terms of percentage of patterns for which the iterative strategy converges. While EP converges basically on all the presented sets of patterns, the convergence rate drops to less than 20% for a large interval of α values for grVAMP. EP in this regime shows a remarkable performance both in terms of accuracy and sensitivity and is still able to learn the parameters of the priors fairly well.

Taken together, these results show that EP is a competitive algorithmic scheme with very good variable selection properties, particularly when one cannot rely on the statistical independence of the entries of the pattern matrix. An important point of strength of the algorithm is the possibility to infer *online* the optimal dilution of the problem using a maximum likelihood data-driven iterative strategy, whereas, for other competitive strategies, the dilution is a fixed parameter to set on the basis of some prior knowledge about the problem. In addition, we have shown that the same maximum likelihood online strategy can be used to learn the consistency level of the labels when they are affected by noise. As for the limitations of EP, the main disadvantage is given by its cubic computational complexity, which makes it slower as compared with algorithms such as 1bitAMP and implementations of generalized VAMP. Furthermore, we found that, while EP was able to deal with the task of learning from noisy examples, a not too large level of noise was needed in order for the algorithm to learn the target classification rule with an acceptable accuracy. Finally, although in this paper convergence was generally achieved within the thresholds we specified, it is well known that EP can suffer from numerical instabilities that prevent it from converging to its true fixed points. As a consequence, we expect that the results presented in this paper may not be the best results obtainable by means of the EP approximation. Attempting to better understand the convergence properties of EP in the settings presented in this paper as well as in other scenarios will be the object of future work.

ACKNOWLEDGMENTS

A.B., A.P., and M.P. acknowledge funding from INFERNET, a European Union Horizon 2020 research and innovation program under Marie Skłodowska-Curie Grant No. 734439, and from the SmartData@Polito interdepartmental center for data science of Politecnico di Torino.

APPENDIX A: FINITE TEMPERATURE FORMULATION OF EXPECTATION PROPAGATION

We here recall the finite temperature EP scheme used in [39] and state the update equations for the sparse perceptron learning problem.

In this case, the posterior distribution is given by

$$P(\boldsymbol{h}) = \frac{1}{Z_P} e^{-\frac{\beta}{2}\boldsymbol{h}^T \mathbf{E}^{-1}\boldsymbol{h}} \prod_{i=1}^{N} \Gamma_i(h_i) \prod_{\tau=N+1}^{N+M} \Lambda_\tau(h_\tau), \tag{A1}$$

where we recall that

$$\mathbf{E}^{-1} = \begin{pmatrix} \mathbf{X}_\sigma^T \mathbf{X}_\sigma & -\mathbf{X}_\sigma^T \\ -\mathbf{X}_\sigma & \mathbf{I} \end{pmatrix}. \tag{A2}$$

As in Sec. II, we introduce Gaussian approximating factors (14) and write a fully Gaussian approximation of the posterior distribution (A1):

$$\begin{aligned} Q(\boldsymbol{h}) &= \frac{1}{Z_Q} e^{-\frac{\beta}{2}\boldsymbol{h}^T \mathbf{E}^{-1}\boldsymbol{h}} \prod_{i\in W} \phi(h_i; a_i, d_i) \prod_{\tau\in Y} \phi(h_\tau; a_\tau, d_\tau) \\ &= \frac{1}{Z_Q} \exp\left(-\frac{1}{2}(\boldsymbol{h}-\bar{\boldsymbol{h}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{h}-\bar{\boldsymbol{h}})\right). \end{aligned} \tag{A3}$$

The covariance matrix $\boldsymbol{\Sigma}$ and the mean $\bar{\boldsymbol{h}}$ in Eq. (A3) are given, respectively, by

$$\boldsymbol{\Sigma}^{-1} = \beta \mathbf{E}^{-1} + \mathbf{D}, \tag{A4}$$

and by

$$\bar{\boldsymbol{h}} = \boldsymbol{\Sigma}\mathbf{D}\boldsymbol{a}. \tag{A5}$$

Moreover, we define the tilted distributions $Q^{(i)}$ for all $i = 1, \ldots, N+M$ as

$$\begin{aligned} Q^{(i)}(\boldsymbol{h}) &= \frac{1}{Z_Q} e^{-\frac{\beta}{2}\boldsymbol{h}^T \mathbf{E}^{-1}\boldsymbol{h}} \psi_i(h_i) \prod_{k\neq i} \phi(h_k; a_k, d_k) \\ &= \frac{1}{Z_{Q^{(i)}}} \psi_i(h_i) \exp\left(-\frac{1}{2}(\boldsymbol{h}-\bar{\boldsymbol{h}}^{(i)})^T \left(\boldsymbol{\Sigma}^{(i)}\right)^{-1}(\boldsymbol{h}-\bar{\boldsymbol{h}}^{(i)})\right), \end{aligned} \tag{A6}$$

where $\psi_i = \Gamma$ if $i \in W$ and $\psi_i = \Lambda$ if $i \in Y$.

As explained in Sec. II, the EP update equations for the means $\boldsymbol{a}_i$ and for the variances $\boldsymbol{d}_i$ are obtained by imposing the matching of the first and second moments of each variable h_i with respect to the tilted marginal distributions $Q^{(i)}(h_i)$ and the fully approximated marginal distributions $Q(h_i)$:

$$\langle h_i \rangle_Q = \langle h_i \rangle_{Q^{(i)}}, \qquad \langle h_i^2 \rangle_Q = \langle h_i^2 \rangle_{Q^{(i)}}, \tag{A7}$$

which leads, again, to Eqs. (31) and (32).

APPENDIX B: EP FREE ENERGY FOR THE DILUTED PERCEPTRON PROBLEM

Let us recall the definition of the EP free energy for a system with $N+M$ variables:

$$F_{\rm EP} = (N+M-1)\log Z_Q - \sum_{k=1}^{N+M} \log Z_{Q^{(k)}}, \tag{B1}$$

where

$$\log Z_Q = \frac{N+M}{2}\log(2\pi) + \frac{1}{2}\log(\det \boldsymbol{\Sigma}), \tag{B2}$$

and the expression of $\log Z_{Q^{(k)}}$ depends on the type of prior considered. If $k = 1, \ldots, N$, the prior is a spike-and-slab and one has

$$\log Z_{Q^{(k)}} = \log\left((1-\rho)\frac{1}{\sqrt{2\pi\,\Sigma_k}} e^{-\frac{\mu_k^2}{2\Sigma_k}} + \frac{\rho}{\sqrt{2\pi}}\sqrt{\frac{\lambda}{1+\lambda\,\Sigma_k}} e^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}\right), \quad k = 1, \ldots, N. \tag{B3}$$

For the remaining variables, the expression of $Z_{Q^{(k)}}$ either reads

$$Z_{Q^{(k)}} = \sqrt{\frac{\pi\,\Sigma_k}{2}}\left[1 + \mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right], \quad k = N+1, \ldots, N+M \tag{B4}$$

if $\Lambda(h_k) = \Theta(h_k)$, or it reads

$$\begin{aligned} Z_{Q^{(k)}} &= \sqrt{\frac{\pi\,\Sigma_k}{2}}\left[\frac{\eta}{2}\mathrm{erfc}\left(-\frac{\mu_k}{\sqrt{2\Sigma_k}}\right) + \frac{1-\eta}{2}\mathrm{erfc}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right] \\ &= \sqrt{\frac{\pi\,\Sigma_k}{2}}\left[\frac{1}{2} + \left(\eta - \frac{1}{2}\right)\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right], \quad k = N+1, \ldots, N+M, \end{aligned} \tag{B5}$$

if $\Lambda(h_k) = \eta\Theta(h_k) + (1-\eta)\Theta(-h_k)$.

As a consequence, when $\Lambda(h_k) = \Theta(h_k)$, the EP free energy of the problem is given by

$$\begin{aligned} F_{\rm EP} = (N+M-1)&\left(\frac{N+M}{2}\log(2\pi) + \frac{1}{2}\log(\det\boldsymbol{\Sigma})\right) \\ &- \sum_{k=1}^{N}\log\left((1-\rho)\frac{1}{\sqrt{2\pi\,\Sigma_k}} e^{-\frac{\mu_k^2}{2\Sigma_k}} + \frac{\rho}{\sqrt{2\pi}}\sqrt{\frac{\lambda}{1+\lambda\,\Sigma_k}} e^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}\right) \\ &- \frac{M}{2}\log(\pi/2) - \frac{1}{2}\sum_{k=N+1}^{N+M}\log\Sigma_k - \sum_{k=N+1}^{N+M}\log\left[1 + \mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right], \end{aligned} \tag{B6}$$

whereas, when $\Lambda(h_k) = \eta\Theta(h_k) + (1-\eta)\Theta(-h_k)$, it is given by

$$\begin{aligned} F_{\rm EP} = (N+M-1)&\left(\frac{N+M}{2}\log(2\pi) + \frac{1}{2}\log(\det\mathbf{\Sigma})\right) \\ &- \sum_{k=1}^{N}\log\left((1-\rho)\frac{1}{\sqrt{2\pi\,\Sigma_k}}e^{-\frac{\mu_k^2}{2\Sigma_k}} + \frac{\rho}{\sqrt{2\pi}}\sqrt{\frac{\lambda}{1+\lambda\Sigma_k}}e^{-\frac{1}{2}\frac{\lambda\mu_k^2}{1+\lambda\Sigma_k}}\right) \\ &- \frac{M}{2}\log(\pi/2) - \frac{1}{2}\sum_{k=N+1}^{N+M}\log\Sigma_k - \sum_{k=N+1}^{N+M}\log\left[\frac{1}{2} + \left(\eta-\frac{1}{2}\right)\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)\right]. \end{aligned} \tag{B7}$$

As a practical remark, an efficient way to compute $F_{\rm EP}$ numerically involves using the Cholesky decomposition of the covariance matrix in $\log(\det\mathbf{\Sigma})$, that is $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular matrix with real positive diagonal entries L_{kk} for all $k = 1,\dots,N+M$. Then, the log determinant contribution to the EP free energy is efficiently computed as $\log(\det\mathbf{\Sigma}) = 2\sum_{k=1}^{N+M}\log(L_{kk})$.

APPENDIX C: LEARNING THE DENSITY LEVEL OF THE WEIGHTS OF THE TEACHER AND THE NOISE ON THE LABELS

The parameters of the prior distributions, such as the density ρ of the weights of the teacher signal and the fraction η of labels fulfilling the consistency constraints, can be iteratively learned by the student perceptron by minimizing the free energy associated with the EP algorithm. We follow the reasoning laid out in [42], which we here recall and adapt to the sparse perceptron learning problem.

Let $\boldsymbol{\theta}$ denote the set of parameters of the prior distribution $P_{\boldsymbol{\theta}}(\boldsymbol{h})$. For example, the density ρ appears in the prior distribution in the factors $\Gamma(w_i)$, for $i = 1,\dots,N$, whereas the consistency level η of the labels appears in the factors Λ in the noisy case. Such parameters can be estimated by the student perceptron by maximizing the following likelihood function:

$$\begin{aligned} P(\sigma_1,\dots,\sigma_M|\boldsymbol{\theta},\boldsymbol{x}_1,\dots,\boldsymbol{x}_M) &= \int d\boldsymbol{h}P(\sigma_1,\dots,\sigma_M,\boldsymbol{h}|\boldsymbol{\theta},\boldsymbol{x}_1,\dots,\boldsymbol{x}_M) = \int d\boldsymbol{h}P(\mathbf{X}_\sigma|\boldsymbol{h},\boldsymbol{x}_1,\dots,\boldsymbol{x}_M)P(\boldsymbol{h}|\boldsymbol{\theta}) \\ &= \int d\boldsymbol{h}P(\mathbf{X}_\sigma|\boldsymbol{h},\boldsymbol{x}_1,\dots,\boldsymbol{x}_M)P(h_1,\dots,h_N|\rho,\lambda)P(h_{N+1},\dots,h_{N+M}|\eta) = Z(\boldsymbol{\theta}), \end{aligned} \tag{C1}$$

which is nothing but the normalization of the posterior distribution in Eq. (11).

It is possible to associate a free energy to the partition function (C1) by using the definition $F = -\log Z(\boldsymbol{\theta})$. When EP reaches its fixed point, F is approximated by the EP free energy (B1) and the student perceptron can attempt to minimize the latter via gradient descent:

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \delta\theta_j\frac{\partial F_{\rm EP}}{\partial\theta_j}, \tag{C2}$$

where t denotes the current iteration, θ_j denotes the jth component of the parameter vector $\boldsymbol{\theta}$, and $\delta\theta_j$ is its corresponding learning rate, which was taken equal to 10^{-5} in our numerical experiments both when learning ρ and when learning η.

Notice that, while the only contributions to $F_{\rm EP}$ depending explicitly on the parameters $\boldsymbol{\theta}$ of the prior are given by the terms $F_{Q^{(k)}}$, $k = 1,\dots,N+M$, the components of the gradient include other terms as well. However, these terms depend on the derivatives of the free energy with respect to the cavity parameters, which vanish at the EP fixed point. The optimization strategy we employ consists in iteratively alternating an EP update step at fixed prior parameters and an update of $\boldsymbol{\theta}$ performed via gradient descent at fixed EP parameters, similarly to an EM scheme, where the optimization over the EP parameters corresponds to the expectation step and the minimization of the free energy with respect to $\boldsymbol{\theta}$ corresponds to the maximization step. A "proper" EM procedure would also be possible and would involve alternating 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 allows us to only consider the explicit dependence of the free energy on the prior parameters. In particular, we have that $\partial F_{\rm EP}/\partial\rho$ in Eq. (C2) reads

$$\frac{\partial F_{\rm EP}}{\partial\rho} = \sum_{k=1}^{N}\frac{\frac{1}{\sqrt{2\pi\,\Sigma_k}}e^{-\frac{\mu_k^2}{2\Sigma_k}} - \frac{1}{\sqrt{2\pi(\lambda+\Sigma_k)}}e^{-\frac{1}{2}\frac{\mu_k^2}{\lambda+\Sigma_k}}}{(1-\rho)\frac{1}{\sqrt{2\pi\,\Sigma_k}}e^{-\frac{\mu_k^2}{2\Sigma_k}} + \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma_k)}}e^{-\frac{1}{2}\frac{\mu_k^2}{\lambda+\Sigma_k}}}. \tag{C3}$$

Moreover, as highlighted in [42], the second derivative with respect to ρ is strictly positive if $\lambda > 0$, which guarantees the uniqueness of the value of the density value that minimizes $F_{\rm EP}$, provided that the EP parameters μ_k and Σ_k are fixed for all $k = 1,\dots,N$.

The same line of reasoning applies to the estimation of the parameter η. In this case, the only contributions to $F_{\rm EP}$ where η appears are the terms $F_{Q^{(k)}}$, $k = N+1,\dots,N+M$, and by taking the derivative with respect to η one obtains

$$\frac{\partial F_{\rm EP}}{\partial\eta} = \sum_{k=N+1}^{N+M}\frac{-2\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)}{1+(2\eta-1)\mathrm{erf}\left(\frac{\mu_k}{\sqrt{2\Sigma_k}}\right)}. \tag{C4}$$

[1] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*, Studies in Fuzziness and Soft Computing Vol. 207 (Springer, New York, 2008).

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2009).

[3] The cancer genome atlas web site, https://www.cancer.gov/tcga (accessed February 16, 2020).

[4] E. J. Candès and M. B. Wakin, An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition], IEEE Signal Process. Mag. **25**, 21 (2008).

[5] R. G. Baraniuk, More is less: Signal processing and the data deluge, Science **331**, 717 (2011).

[6] D. L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory **52**, 1289 (2006).

[7] E. J. Candès and T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, IEEE Trans. Inf. Theory **52**, 5406 (2006).

[8] D. L. Donoho, A. Maleki, and A. Montanari, Message-passing algorithms for compressed sensing, Proc. Natl. Acad. Sci. USA **106**, 18914 (2009).

[9] Y. Kabashima, T. Wadayama, and T. Tanaka, A typical reconstruction limit for compressed sensing based on l p -norm minimization, J. Stat. Mech.: Theory Exp. (2009) L09003.

[10] S. Ganguli and H. Sompolinsky, Statistical Mechanics of Compressed Sensing, Phys. Rev. Lett. **104**, 188701 (2010).

[11] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices, J. Stat. Mech.: Theory Exp. (2012) P08009.

[12] P. T. Boufounos and R. G. Baraniuk, 1-bit Compressive Sensing, in *2008 42nd Annual Conference on Information Sciences and Systems* (IEEE, New York, 2008), pp. 16–21.

[13] D. Lee, T. Sasaki, T. Yamada, K. Akabane, Y. Yamaguchi, and K. Uehara, Spectrum sensing for networked system using 1-bit compressed sensing with partial random circulant measurement matrices, in *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)* (IEEE, New York, 2012), pp. 1–5.

[14] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B (Methodol.) **58**, 267 (1996).

[15] A Braunstein, A Pagnani, M Weigt, and R Zecchina, Gene-network inference by message passing, J. Phys.: Conf. Series **95**, 012016 (2008).

[16] A Braunstein, A Pagnani, M Weigt, and R Zecchina, Inference algorithms for gene networks: A statistical mechanics analysis, J. Stat. Mech.: Theory Exp. (2008) P12001.

[17] A. Pagnani, F. Tria, and M. Weigt, Classification and sparse-signature extraction from gene-expression data, J. Stat. Mech.: Theory Exp. (2009) P05001.

[18] M. Bailly-Bechet, A. Braunstein, A. Pagnani, M. Weigt, and R. Zecchina, Inference of sparse combinatorial-control networks from gene-expression data: A message passing approach, BMC Bioinf. **11**, 355 (2010).

[19] E. J. Molinelli, A. Korkut, W. Wang, M. L. Miller, N. P. Gauthier, X. Jing, P. Kaushik, Q. He, G. Mills, D. B. Solit, C. A. Pratilas, M. Weigt, A. Braunstein, A. Pagnani, R. Zecchina, and C. Sander, Perturbation biology: Inferring signaling networks in cellular systems, PLoS Comput. Biol. **9**, e1003290 (2013).

[20] Y. Kabashima, A CDMA multiuser detection algorithm on the basis of belief propagation, J. Phys. A: Math. Gen. **36**, 11111 (2003).

[21] S. Uda and Y. Kabashima, Statistical mechanical development of a sparse Bayesian classifier, J. Phys. Soc. Jpn. **74**, 2233 (2005).

[22] Y Kabashima, Inference from correlated patterns: A unified theory for perceptron learning and linear vector channels, J. Phys.: Conf. Ser. **95**, 012001 (2008).

[23] Y. Xu and Y. Kabashima, Statistical mechanics approach to 1-bit compressed sensing, J. Stat. Mech.: Theory Exp. (2013) P02041.

[24] A. Lage-Castellanos, A. Pagnani, and M. Weigt, Statistical mechanics of sparse generalization and graphical model selection, J. Stat. Mech.: Theory Exp. (2009) P10009.

[25] A. Lage-Castellanos, A. Pagnani, and M. Weigt, The importance of dilution in the inference of biological networks, in Proceedings of the 2009 47th Annual Allerton Conference on Communication, Control, and Computing, 2009 (unpublished), pp. 531–538.

[26] A. Lage-Castellanos, A. Pagnani, and G. Q. Angulo, Stability of the replica symmetric solution in diluted perceptron learning, J. Stat. Mech.: Theory Exp. (2013) L02002.

[27] Y. Xu, Y. Kabashima, and L. Zdeborová, Bayesian signal reconstruction for 1-bit compressed sensing, J. Stat. Mech.: Theory Exp. (2014) P11015.

[28] S. Rangan, P. Schniter, and A. K. Fletcher, Vector approximate message passing, IEEE Trans. Inf. Theory **65**, 6664 (2019).

[29] P. Schniter, S. Rangan, and A. K. Fletcher, Vector approximate message passing for the generalized linear model, arXiv:1612.01186 (2016).

[30] X. Meng, S. Wu, and J. Zhu, A unified Bayesian inference framework for generalized linear models, IEEE Signal Process. Lett. **25**, 398 (2018).

[31] P. T. Boufounos, Greedy sparse signal reconstruction from sign measurements, in Proceedings of the 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, 2009 (unpublished), pp. 1305–1309.

[32] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors, IEEE Trans. Inf. Theory **59**, 2082 (2013).

[33] M. Yan, Y. Yang, and S. Osher, Robust 1-bit compressive sensing using adaptive outlier pursuit, IEEE Trans. Signal Process. **60**, 3868 (2012).

[34] A. Movahed, A. Panahi, and G. Durisi, A robust rfpi-based 1-bit compressive sensing reconstruction algorithm, in Proceedings of the 2012 IEEE Information Theory Workshop, 2012 (unpublished), pp. 567–571.

[35] F. Li, J. Fang, H. Li, and L. Huang, Robust one-bit Bayesian compressed sensing with sign-flip errors, IEEE Signal Process. Lett. **22**, 857 (2015).

[36] M. Opper and O. Winther, Gaussian processes for classification: Mean-field algorithms, Neural Comput. **12**, 2655 (2000).

[37] M. Opper and O. Winther, Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling, Phys. Rev. E **64**, 056131 (2001).

[38] T. P. Minka, Expectation propagation for approximate Bayesian inference, in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (San Francisco, CA, 2001), pp. 362–369.

[39] A. Braunstein, A. P. Muntoni, and A. Pagnani, An analytic approximation of the feasible space of metabolic networks, Nat. Commun. **8**, 14915 (2017).

[40] T. J. Mitchell and J. J. Beauchamp, Bayesian variable selection in linear regression, J. Am. Stat. Assoc. **83**, 1023 (1988).

[41] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*, International Series of Monographs on Physics (Oxford University, New York, 2001).

[42] A. Braunstein, A. P. Muntoni, A. Pagnani, and M. Pieropan, Compressed sensing reconstruction using expectation propagation, J. Phys. A: Math. Theor. **53**, 184001 (2020).

[43] See https://github.com/abraunst/GaussianEP for EP, https://github.com/mengxiangming/glmcode for grVAMP, and https://github.com/livey/R1BCS for R1BCS. For reproducibility of the results of this paper, a repository is available at https://github.com/mpieropan/EP-diluted-classifier.