Lecture Notes on Complex Systems and Physical Models
BP-based algorithms in the perceptron model

Carlo Lucibello & Enrico M. Malatesta*

* enrico.malatesta@unibocconi.it

May 5, 2025

Contents

The goal of those lecture notes is to introduce an efficient implementation of the Belief Propagation (BP) equations, called relaxed Belief Propagation (rBP). We are going to see that rBP can efficiently compute the marginals in the binary teacher-student perceptron model. Moreover it can be employed as an algorithm able to infer the teacher if the size of the dataset is sufficiently large.

Finally we introduce an even more efficient implementation of rBP called Approximate Message Passing (AMP) algorithm, that is formerly known in the statistical physics literature as the Thouless-Anderson-Palmer (TAP) equations, from the names of the people that introduced it.

1 Factor Graph representation

Before writing down the BP equations for our perceptron models, we need to derive its factor graph representation. Recall that in our perceptron models we had a dataset $\mathcal{D}$ composed of P inputs with the corresponding label $\mathcal{D} = \left\{\xi^\mu, y^\mu\right\}_{\mu=1}^P$. In the teacher student setting, for example, the labels y^μ are generated by

$$y_\mu = \text{sign}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N w_i^T \xi_i^\mu\right) \tag{1}$$

We are given the posterior distribution $P(\boldsymbol{w}|\mathcal{D})$

$$P(\boldsymbol{w}|\mathcal{D}) = \frac{1}{Z_\mathcal{D}}\prod_{\mu=1}^P P(y_\mu|h_\mu)\prod_{i=1}^N P(w_i) \tag{2}$$

where $\prod_\mu P(y_\mu|h_\mu)$ and $\prod_i P(w_i)$ represent respectively the (factorized) likelihood and prior of the student; we have also defined h_μ to be the preactivation of the student

$$h_\mu \equiv \frac{1}{\sqrt{N}}\sum_i w_i \xi_i^\mu \tag{3}$$

What is the factor graph representation corresponding to this model? We have N variable nodes, corresponding to the weights w_i and P associated constraints to satisfy. As can be seen from the definition of the variable h_μ, each constraint μ has an impact on each variable node i, so each of those "pattern" factor nodes must be connected to each variable node. Moreover, for each of the N variable nodes we have a "prior" factor $P(w_i)$. The factor graph representation corresponding to this graphical model is given in Fig. 1.

In the following we will consider generic likelihood and prior distributions for our student model. Only later on we will specialize the analysis to the standard binary perceptron model that we have analyzed previously with the replica method; in that case we have a prior and a likelihood of the form

$$P(w_i) = \frac{1}{2}\delta(w_i - 1) + \frac{1}{2}\delta(w_i + 1) \tag{4a}$$

$$P(y_\mu|h_\mu) = \Theta\left(y_\mu h_\mu\right). \tag{4b}$$

2 Belief Propagation equations

We now want to write the Belief Propagation (BP) equations for the factor graph of Fig. 1. Note that this is quite a *dense* graph, and contains many short loops. Despite each variable

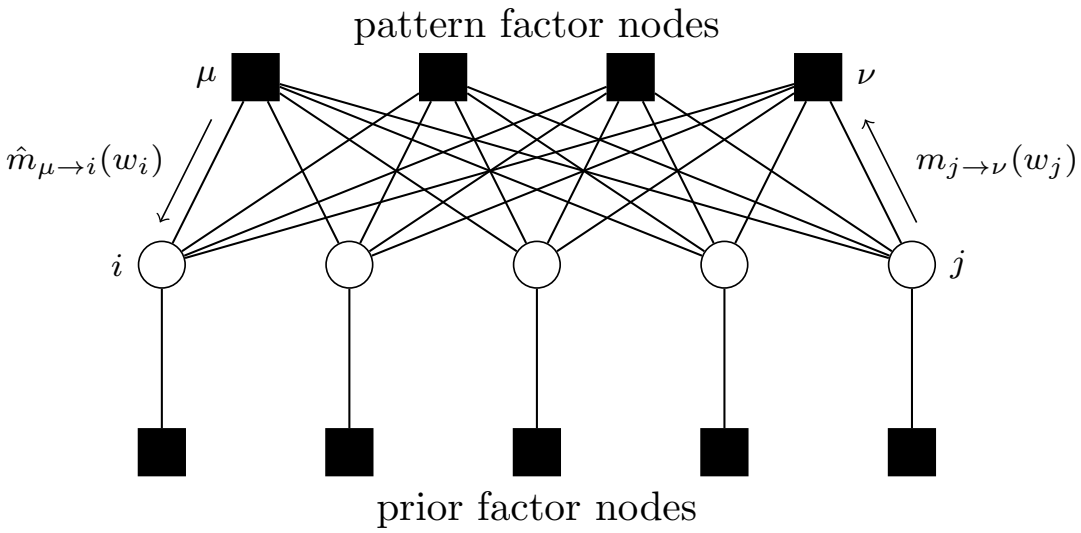

Figure 1: **Teacher-student problem**: Factor graph representation of the perceptron models. Each variable node is connected to each "pattern" factor node enforcing that. Each variable node is connected to a single variable factor node that gives the prior for that variable. Here $N = 5$ and $P = 4$.

node enters in in all interaction nodes, the individual contribution of a variable is negligible in the large system size as it is of order $1/\sqrt{N}$. We can write right away the Belief Propagation equations corresponding to the posterior distribution (2)

$$m_{i\to\mu}(w_i) = \frac{1}{z_{i\to\mu}} P(w_i) \prod_{\nu\neq\mu} \hat{m}_{\nu\to i}(w_i) \tag{5a}$$

$$\hat{m}_{\mu\to i}(w_i) = \frac{1}{\hat{z}_{\mu\to i}} \int \prod_{j\neq i} dw_j \prod_{j\neq i} m_{j\to\mu}(w_j)\, P\left(y_\mu | \frac{1}{\sqrt{N}} \sum_i w_i \xi_i^\mu\right) \tag{5b}$$

At the present stage, the previous equations are, however, computationally intractable since each factor node enforcing a pattern constraint involves N variables; the second of the BP equations requires the computation of $N-1$ integrals (sums if the weights are binary). When N is large, however, we can considerably simplify the previous equations, as we will see in the next section.

3 Relaxed Belief Propagation equations

Let's start analyzing the second of the BP equations (5). We start by isolating the variables of integration from w_i

$$\hat{m}_{\mu\to i}(w_i) = \frac{1}{\hat{z}_{\mu\to i}} \int \prod_{j\neq i} dw_j \prod_{j\neq i} m_{j\to\mu}(w_j)\, P\left(y_\mu | \frac{1}{\sqrt{N}} w_i \xi_i^\mu + \frac{1}{\sqrt{N}} \sum_{j\neq i} w_j \xi_j^\mu\right) \tag{6}$$

When N is large, due to the central limit theorem we expect the variable $h_i^\mu \equiv \frac{1}{\sqrt{N}} \sum_{j\neq i} w_j \xi_j^\mu$ to be Gaussian distributed. When N is sufficiently large, therefore, we expect only the mean and the variance of the messages to be relevant. This key observation will allow us to write the BP equations that can be efficiently solved numerically.

3.1 The Central Limit theorem comes to the rescue

As we have anticipated, since N is large we can apply the central limit theorem to the quantity $h_i^\mu \equiv \frac{1}{\sqrt{N}}\sum_{j\neq i} w_j \xi_j^\mu$. Equivalently, we can derive the resulting expression by introducing a delta function that enforces the definition of this variable and using its Fourier representation

$$\hat{m}_{\mu\to i}(w_i) = \frac{1}{\hat{z}_{\mu\to i}} \int \frac{dh d\hat{h}}{2\pi} P\left(y_\mu | \frac{1}{\sqrt{N}} w_i \xi_i^\mu + h\right) e^{ih\hat{h}} \times \int \prod_{j\neq i} dw_j \prod_{j\neq i} m_{j\to\mu}(w_j) e^{-i\hat{h}\frac{1}{\sqrt{N}}\sum_{j\neq i} w_j \xi_j^\mu} \tag{7}$$

writing the last term as a product of $j \neq i$, and Taylor expanding the exponential

$$\begin{aligned} &\int \prod_{j\neq i} dw_j \prod_{j\neq i} m_{j\to\mu}(w_j) e^{-i\hat{h}\frac{1}{\sqrt{N}}\sum_{j\neq i} w_j \xi_j^\mu} = \prod_{j\neq i}\left[\int dw_j\, m_{j\to\mu}(w_j) e^{-i\hat{h}\frac{1}{\sqrt{N}} w_j \xi_j^\mu}\right] \\ &= \prod_{j\neq i}\left[\int dw_j\, m_{j\to\mu}(w_j)\left(1 - i\hat{h}\frac{1}{\sqrt{N}} w_j \xi_j^\mu - \frac{\hat{h}^2}{2N} w_j^2 (\xi_j^\mu)^2\right)\right] \end{aligned} \tag{8}$$

Recalling that the messages $m_{j\to\mu}(w_j)$ are properly normalized probability distributions we can proceed by integrating term by term and re-exponentiating we have

$$\int \prod_{j\neq i} dw_j \prod_{j\neq i} m_{j\to\mu}(w_j) e^{-i\hat{h}\frac{1}{\sqrt{N}}\sum_{j\neq i} w_j \xi_j^\mu} \simeq e^{i\hat{h}\left(h - M_{i\to\mu}\right) - \frac{\hat{h}^2}{2} V_{i\to\mu}} \tag{9}$$

where

$$M_{i\to\mu} \equiv \frac{1}{\sqrt{N}} \sum_{j\neq i} \xi_j^\mu a_{j\to\mu} \tag{10a}$$

$$a_{i\to\mu} \equiv \int dw_i\, m_{i\to\mu}(w_i)\, w_i \tag{10b}$$

$$V_{i\to\mu} \equiv \frac{1}{N} \sum_{j\neq i} (\xi_j^\mu)^2 b_{j\to\mu} \tag{10c}$$

$$b_{i\to\mu} \equiv \int dw_i\, m_{i\to\mu}(w_i)\, w_i^2 - a_{i\to\mu}^2 \tag{10d}$$

Plugging back (9) into (7) we have

$$\begin{aligned} \hat{m}_{\mu\to i}(w_i) &= \frac{1}{\hat{z}_{\mu\to i}} \int \frac{dh d\hat{h}}{2\pi} P\left(y_\mu | \frac{1}{\sqrt{N}} w_i \xi_i^\mu + h\right) e^{i\hat{h}\left(h - M_{i\to\mu}\right) - \frac{\hat{h}^2}{2} V_{i\to\mu}} \\ &= \frac{1}{\hat{z}_{\mu\to i}} \int Dh\, P\left(y_\mu | \frac{1}{\sqrt{N}} w_i \xi_i^\mu + M_{i\to\mu} + \sqrt{V_{i\to\mu}}\, h\right) \end{aligned} \tag{11}$$

where in the last equality we have integrated over $\hat{h}$ and performed a change of variable in order to recover the standard Gaussian measure over the variable v. Note that we have written the message $\hat{m}_{\mu\to i}(w_i)$ only in terms of the mean and variance of the probability distribution of the messages $m_{j\to\mu}(w_j)$ with $j \neq i$. This allowed us to significantly reduce the complexity of the second BP equation, that now requires the computation of one Gaussian integral only!

3.2 The energetic channel

The next step is to notice that the argument of the likelihood contains a small term of $O\left(\frac{1}{\sqrt{N}}\right)$. We can Taylor expand it up to second order

$$\begin{aligned}\hat{m}_{\mu\to i}(w_i) &\propto \int Dh\, P\left(y_\mu|\frac{1}{\sqrt{N}}w_i\xi_i + M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)\\ &\propto 1 + \frac{1}{\sqrt{N}}w_i\xi_i^\mu \frac{\int Dh\,\partial_{M_{i\to\mu}} P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)}{\int Dh\, P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)}\\ &+ \frac{1}{2N}w_i^2(\xi_i^\mu)^2 \frac{\int Dh\,\partial^2_{M_{i\to\mu}} P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)}{\int Dh\, P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)}\end{aligned} \tag{12}$$

notice that the zeroth order term $\int Dh\, P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)$ does not depend on w_i, so we can absorb it into the proportionality constant. We conveniently introduce the two quantities

$$B_{\mu\to i} \equiv \frac{\xi_i^\mu}{\sqrt{N}} \frac{\int Dh\,\partial_{M_{i\to\mu}} P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)}{\int Dh\, P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)} \tag{13a}$$

$$-A_{\mu\to i} \equiv \frac{(\xi_i^\mu)^2}{N} \frac{\int Dh\,\partial^2_{M_{i\to\mu}} P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)}{\int Dh\, P\left(y_\mu|M_{i\to\mu} + \sqrt{V_{i\to\mu}}h\right)} - B^2_{\mu\to i} \tag{13b}$$

which can be compactly written in terms of just one function

$$g_E(y, M, V) \equiv \frac{\int Dh\,\partial_M P\left(y|M + \sqrt{V}\,h\right)}{\int Dh\, P\left(y|M + \sqrt{V}\,h\right)} \tag{14}$$

as

$$B_{\mu\to i} = \frac{\xi_i^\mu}{\sqrt{N}} g_E(y_\mu, M_{i\to\mu}, V_{i\to\mu}) \tag{15a}$$

$$A_{\mu\to i} = -\frac{(\xi_i^\mu)^2}{N}\partial_{M_{i\to\mu}} g_E(y_\mu, M_{i\to\mu}, V_{i\to\mu}) \tag{15b}$$

The quantity $g_E(y, M, V)$ is called "*energetic channel*" as it depends on the likelihood of the model. We can therefore write equation (12) as

$$\hat{m}_{\mu\to i}(w_i) \propto 1 + w_i B_{\mu\to i} - \frac{1}{2}w_i^2(A_{\mu\to i} - B^2_{\mu\to i}) = e^{-\frac{w_i^2}{2}A_{\mu\to i} + B_{\mu\to i}w_i} \propto e^{-\frac{1}{2}\left(w_i - \frac{B_{\mu\to i}}{A_{\mu\to i}}\right)^2 A_{\mu\to i}} \tag{16}$$

where in the second equality we have written the previous expression in terms of an exponential, since $B_{\mu\to i}$ is of order $\frac{1}{\sqrt{N}}$ and $A_{\mu\to i}$ is of order $1/N$. We finally can retrieve the normalization constant

$$\hat{m}_{\mu\to i}(w_i) = \sqrt{\frac{A_{\mu\to i}}{2\pi}} e^{-\frac{1}{2}\left(w_i - \frac{B_{\mu\to i}}{A_{\mu\to i}}\right)^2 A_{\mu\to i}} \tag{17}$$

3.3 The entropic channel

We can now plug the second message passing equation back into the first one, see equation (5a) to get

$$m_{i\to\mu}(w_i) = \frac{1}{z_{i\to\mu}} P(w_i) \prod_{\nu\neq\mu} \hat{m}_{\nu\to i}(w_i) \propto P(w_i)\, e^{-\frac{\left(w_i - R_{i\to\mu}\right)^2}{2\Sigma_{i\to\mu}}} \tag{18}$$

where

$$R_{i\to\mu} \equiv \frac{\sum_{\nu\neq\mu} B_{\nu\to i}}{\sum_{\nu\neq\mu} A_{\nu\to i}} \tag{19a}$$

$$\Sigma_{i\to\mu} \equiv \frac{1}{\sum_{\nu\neq\mu} A_{\nu\to i}} \tag{19b}$$

$$\tag{19c}$$

From those identities we can close the message passing equations

$$a_{i\to\mu} \equiv \int dw_i\, m_{i\to\mu}(w_i)\, w_i = \frac{\int dw_i\, P(w_i)\, w_i\, e^{-\frac{\left(w_i - R_{i\to\mu}\right)^2}{2\Sigma_{i\to\mu}}}}{\int dw_i\, P(w_i)\, e^{-\frac{\left(w_i - R_{i\to\mu}\right)^2}{2\Sigma_{i\to\mu}}}} \tag{20a}$$

$$b_{i\to\mu} \equiv \frac{\int dw_i\, P(w_i)\, w_i^2\, e^{-\frac{\left(w_i - R_{i\to\mu}\right)^2}{2\Sigma_{i\to\mu}}}}{\int dw_i\, P(w_i)\, e^{-\frac{\left(w_i - R_{i\to\mu}\right)^2}{2\Sigma_{i\to\mu}}}} - a_{i\to\mu}^2 \tag{20b}$$

The previous equations can be compactly written in terms of an "*entropic channel*"

$$g_S(R,\Sigma) \equiv \frac{\int dw\, P(w)\, w\, e^{-\frac{(w-R)^2}{2\Sigma}}}{\int dw\, P(w)\, e^{-\frac{(w-R)^2}{2\Sigma}}} \tag{21}$$

as

$$a_{i\to\mu} = g_S(R_{i\to\mu}, \Sigma_{i\to\mu}) \tag{22a}$$

$$b_{i\to\mu} = \Sigma_{i\to\mu}\, \partial_{R_{i\to\mu}} g_S(R_{i\to\mu}, \Sigma_{i\to\mu}) \tag{22b}$$

The name "entropic" descends from the fact that (21) depends only on the prior distribution of the weights.

3.4 The relaxed Belief Propagation algorithm

The self-consistent equations for the mean $a_{i\to\mu}$ and the variance $b_{i\to\mu}$ of the message distribution $m_{i\to\mu}(w_i)$ that we have derived so far constitute what is called the *relaxed Belief Propagation algorithm*. A compact pseudo-code summarizing its fundamental steps is reported below. Note that the rBP algorithm needs to store in memory order $NP \propto N^2$ variables, corresponding to total number of messages living on the links of the factor graph. It is important to note that the algorithm can be run on a generic dataset instance, both when the labels are generated by a teacher network and when they are drawn completely at random, as in the storage setting.

Once convergence is reached for the messages $a_{i\to\mu}$ and $b_{i\to\mu}$, we can use them to estimate the site marginals. This can be done first by computing

$$R_i \equiv \frac{\sum_{\mu} B_{\mu\to i}}{\sum_{\mu} A_{\mu\to i}} \tag{23a}$$

$$\Sigma_i \equiv \frac{1}{\sum_{\mu} A_{\mu\to i}} \tag{23b}$$

and then

$$a_i = \langle w_i \rangle = g_S(R_i, \Sigma_i) \tag{24a}$$
$$b_i = \langle w_i^2 \rangle - \langle w_i \rangle^2 = \Sigma_i \, \partial_{R_i} g_S(R_i, \Sigma_i) \tag{24b}$$

having denoted with the brackets $\langle \bullet \rangle$ the average over the posterior distribution.

Algorithm 1: relaxed Belief Propagation (rBP)

Input: random guess for $a_{i\to\mu}^{t=0}$ and $b_{i\to\mu}^{t=0}$, $t = 1$

while $a_{i\to\mu}$ *and* $b_{i\to\mu}$ *have not converged* **do**

update $M_{i\to\mu}$ and $V_{i\to\mu}$ via (10a) (10c)

$$M_{i\to\mu}^t \leftarrow \frac{1}{\sqrt{N}} \sum_{j\neq i} \xi_j^\mu a_{j\to\mu}^{t-1} \tag{25a}$$
$$V_{i\to\mu}^t \leftarrow \frac{1}{N} \sum_{j\neq i} (\xi_j^\mu)^2 b_{j\to\mu}^{t-1} \tag{25b}$$

update $A_{\mu\to i}$ and $B_{\mu\to i}$ via (15a) (15b)

$$A_{\mu\to i}^t \leftarrow -\frac{(\xi_i^\mu)^2}{N} \partial_{M_{i\to\mu}^t} g_E(y_\mu, M_{i\to\mu}^t, V_{i\to\mu}^t) \tag{26a}$$
$$B_{\mu\to i}^t \leftarrow \frac{\xi_i^\mu}{\sqrt{N}} g_E(y_\mu, M_{i\to\mu}^t, V_{i\to\mu}^t) \tag{26b}$$

update $R_{i\to\mu}$ and $\Sigma_{i\to\mu}$ via (19a) (19b)

$$R_{i\to\mu}^t \leftarrow \frac{\sum_{\nu\neq\mu} B_{\nu\to i}^t}{\sum_{\nu\neq\mu} A_{\nu\to i}^t} \tag{27a}$$
$$\Sigma_{i\to\mu}^t \leftarrow \frac{1}{\sum_{\nu\neq\mu} A_{\nu\to i}^t} \tag{27b}$$

update $a_{i\to\mu}$ and $b_{i\to\mu}$ via (22a) (22b)

$$a_{i\to\mu}^t \leftarrow g_S(R_{i\to\mu}^t, \Sigma_{i\to\mu}^t) \tag{28a}$$
$$b_{i\to\mu}^t \leftarrow \Sigma_{i\to\mu}^t \, \partial_{R_{i\to\mu}^t} g_S(R_{i\to\mu}^t, \Sigma_{i\to\mu}^t) \tag{28b}$$

$t \leftarrow t + 1$

end

Return: one site marginals:

$$R_i \equiv \frac{\sum_\mu B_{\mu\to i}}{\sum_\mu A_{\mu\to i}} \tag{29a}$$
$$\Sigma_i \equiv \frac{1}{\sum_\mu A_{\mu\to i}} \tag{29b}$$
$$a_i = g_S(R_i, \Sigma_i) \tag{29c}$$
$$b_i = \Sigma_i \, \partial_{R_i} g_S(R_i, \Sigma_i) \tag{29d}$$

3.4.1 Binary Perceptron

The rBP algorithm is quite general, as it permits to handle generic prior and likelihood distributions that appear respectively in the definitions of the entropic and energetic channels. In the particular case of the binary perceptron, the entropic and energetic channels can be explicitly computed analytically. Indeed, inserting (4) inside the definition of the entropic and energetic channels we have

$$g_S(R,\Sigma) \equiv \frac{\int dw\, P(w)\, w\, e^{-\frac{(w-R)^2}{2\Sigma}}}{\int dw\, P(w)\, e^{-\frac{(w-R)^2}{2\Sigma}}} = \tanh\left(\frac{R}{\Sigma}\right) \tag{30a}$$

$$g_E(y,M,V) \equiv \frac{\int Dh\, \partial_M P\left(y|M+\sqrt{V}\,h\right)}{\int Dh\, P\left(y|M+\sqrt{V}\,h\right)} = \frac{y\int Dh\, \delta\left(yM+\sqrt{V}\,h\right)}{\int Dh\, \Theta\left(yM+\sqrt{V}\,h\right)} = \frac{y}{\sqrt{V}} GH\left(-\frac{yM}{\sqrt{V}}\right) \tag{30b}$$

where $GH(x) = G(x)/H(x)$, with $G(x) = e^{-x^2/2}/\sqrt{2\pi}$ and $H(x) \equiv \frac{1}{2}\mathrm{Erfc}\left(\frac{x}{\sqrt{2}}\right)$.

Another simplification that occurs in the binary case is that the entire algorithm can be expressed in terms of averages over the distribution of messages, since $w_i^2 = 1$. Indeed from equation (10d) we see that

$$b_{i\to\mu} = 1 - a_{i\to\mu}^2 . \tag{31}$$

As the entropic channel g_S depends on the ratio of its arguments $R_{i\to\mu}$ and $\Sigma_{i\to\mu}$ only, the variable $A_{\nu\to i}$ is not playing any role and needs not to be computed anymore. The argument in the entropic channel in (30a) is called *local cavity field* and reads

$$h_{i\to\mu} \equiv \frac{R_{i\to\mu}}{\Sigma_{i\to\mu}} = \sum_{\nu\neq\mu} B^t_{\nu\to i} = \frac{1}{\sqrt{N}} \sum_{\nu\neq\mu} \xi_i^\nu\, g_E(y_\nu, M^t_{i\to\nu}, V^t_{i\to\nu}) . \tag{32}$$

We show below the rBP algorithm specialized to the binary perceptron.

Algorithm 2: relaxed Belief Propagation (rBP), binary perceptron

Input: random guess for $a_{i\to\mu}^{t=0}$, $t=1$

while $a_{i\to\mu}$ *has not converged* **do**

update $M_{i\to\mu}$ and $V_{i\to\mu}$

$$M_{i\to\mu}^{t} \leftarrow \frac{1}{\sqrt{N}} \sum_{j\neq i} \xi_j^\mu a_{j\to\mu}^{t-1} \tag{33a}$$

$$V_{i\to\mu}^{t} \leftarrow \frac{1}{N} \sum_{j\neq i} (\xi_j^\mu)^2 \left[1-(a_{j\to\mu}^{t-1})^2\right] \tag{33b}$$

update the *local cavity field* $h_{i\to\mu}$

$$h_{i\to\mu}^{t} \leftarrow \frac{1}{\sqrt{N}} \sum_{\nu\neq\mu} \xi_i^\nu \, g_E(y_\nu, M_{i\to\nu}^{t}, V_{i\to\nu}^{t}) \tag{34}$$

update $a_{i\to\mu}$ via (22a)

$$a_{i\to\mu}^{t} \leftarrow \tanh\left(h_{i\to\mu}^{t}\right) \tag{35}$$

$t \leftarrow t+1$

end

Return: one site marginals via the computation of the *local field* h_i:

$$h_i = \frac{1}{\sqrt{N}} \sum_{\mu} \xi_i^\mu \, g_E(y_\mu, M_{i\to\mu}, V_{i\to\mu}) \tag{36a}$$

$$a_i = \tanh(h_i) \tag{36b}$$

3.5 Inferring the teacher

As we saw before, when convergence of the rBP messages is reached, we can compute the $\langle w_i \rangle$, i.e the average over the posterior distribution for weight w_i, by evaluating a_i. These quantities provide estimates of the teacher's weights, as we now explain. It is important to emphasize, however, that in general the estimates a_i assume continuous values in the interval $[-1,1]$, and are not always binary ± 1 variables like the true teacher weights. This is because a_i represents an average over the posterior distribution; if the posterior is peaked around multiple binary weights, the corresponding average of w_i will not be binary.

A natural way to assess the quality of the inference is through the overlap with the teacher weights, i.e.

$$r = \frac{1}{N} \sum_{i=1}^{N} w_i^T a_i = \left\langle \frac{1}{N} \sum_{i=1}^{N} w_i^T w_i \right\rangle \tag{37}$$

The corresponding generalization error can be computed by $\epsilon_g = \frac{1}{\pi} \arccos r$. Inferring the teacher therefore is equivalent to obtaining $r=1$ from the rBP equations. We show the generalization error found by rBP as a function of α in Figure 2. As expected from the replica formalism, for $\alpha < \alpha_{\rm IT}$ it is impossible to infer the teacher. Indeed in this regime we learned that there exists an *exponential* number of weights $\boldsymbol{w} \neq \boldsymbol{w}^T$ that correctly classify the labeled data.

In the regime $\alpha > \alpha_{\rm IT}$ the inference of the teacher becomes information theoretically possible. However rBP is able to recover the teacher only for $\alpha \geq \alpha_{\rm AT}$. In the region $\alpha_{\rm IT} < \alpha < \alpha_{\rm AT}$

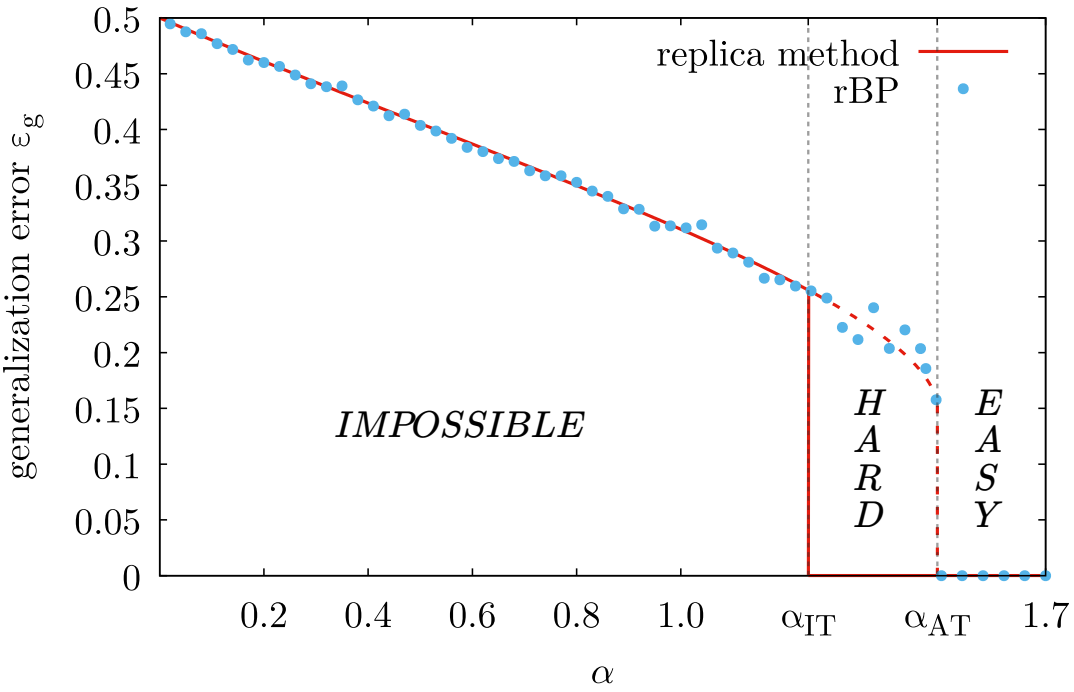

Figure 2: **Binary perceptron, teacher-student case**. Generalization error of the relaxed Belief Propagation algorithm as a function of α (light blue dots). For each value of α we have solved an instance of the problem, using $N = 4000$. In red we show also the generalization error of a typical student sampled from the posterior distribution as obtained from the replica method. For $\alpha < \alpha_{IT}$ inference is impossible, as there are exponentially many configurations fitting the training set together with the teacher. When $\alpha_{IT} < \alpha < \alpha_{AT}$ despite inference becomes information theoretically possible, rBP gets stuck in poor generalizing fixed points. For $\alpha > \alpha_{AT}$ rBP is able to correctly infer the teacher.

rBP is stuck into a "metastable" poor generalizing fixed point. In this window of α the problem of inferring the teacher is *hard on average* for rBP. At present no polynomial-time algorithm is known that is able to infer the teacher in $\alpha_{IT} < \alpha < \alpha_{AT}$.

3.6 Bridging with replicas: State Evolution

The agreement between a single instance of rBP and the replica results is by no means a coincidence. Similar agreement can be obtained in the storage case; in Fig. 3 we show the overlap between two students sampled from the posterior distribution. This can be estimated using rBP by

$$q_N = \frac{1}{N}\sum_{i=1}^{N} a_i^2 = \frac{1}{N}\sum_{i=1}^{N} \langle w_i \rangle^2 \tag{38}$$

The (small) deviations from the prediction of the replica theory are due finite size (i.e. N) effects.

This agreement can be justified rigorously in the large N limit. A nice property of the rBP equations, is indeed that its dynamical behavior can be analytically tracked in the large N limit via a set of equations called *density* or *state evolution*. As we are going to see, these are equivalent to the Replica-Symmetric saddle point equations that we have obtained through the replica method.

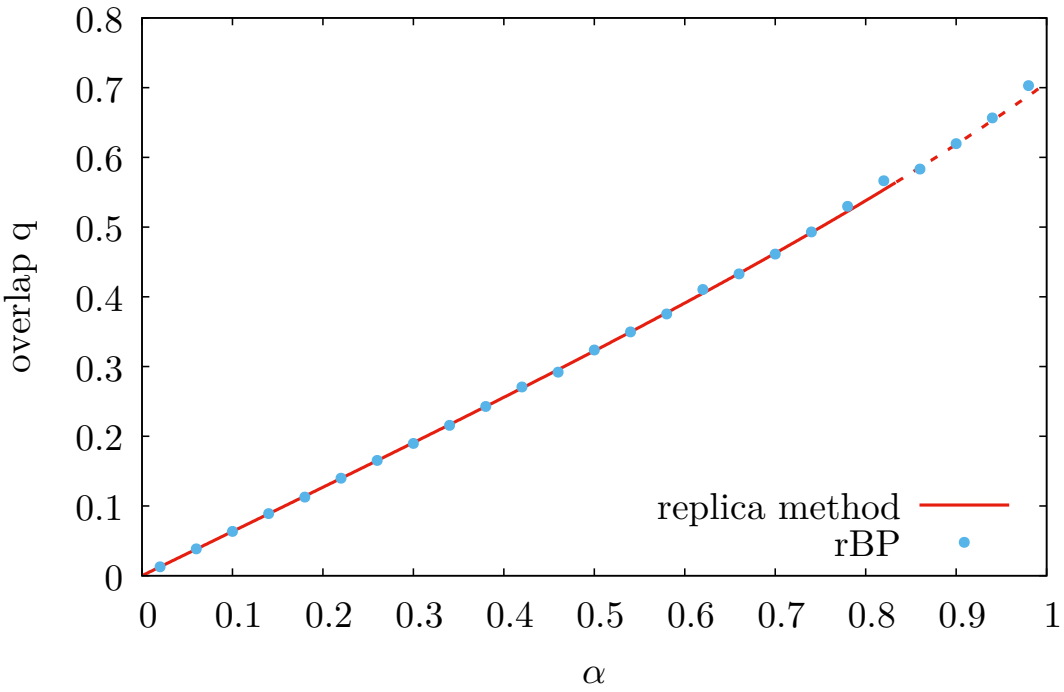

Figure 3: **Binary perceptron, storage case**. Overlap between students sampled from the posterior distribution estimated by the rBP algorithm using equation (38) (light blue dots). For each value of α we have solved an instance of the problem, using $N = 4000$. In red we show also the corresponding prediction using the replica method.

We restrict here for simplicity to the binary perceptron case in the storage setting, where the label is totally uncorrelated to the input. An analogous analysis can be performed in the teacher student scenario, and generic prior/likelihood distributions, but it is a little bit more involved. Our goal is to write self-consistent equations for the average overlap q_N in the large N limit

$$q = \mathbb{E}_\xi \left[\frac{1}{N} \sum_i a_i^2 \right] = \mathbb{E}_\xi \left[\frac{1}{N} \sum_i \tanh^2(h_i) \right]. \tag{39}$$

having used the explicit expression of the entropic channel in the binary case. For each i, h_i is the sum of P uncorrelated random variables. In the large N limit h_i therefore converges to a Gaussian random variable. All we have to do is to compute its mean and variance with respect to the input distribution ξ_i^μ. In order to do that, notice that for each variable i, the energetic channel g_E only contains terms depending on ξ_j^μ with $j \neq i$. We therefore obtain that the mean vanishes and the variance is given by

$$\mathbb{E}_\xi h_i^2 = \mathbb{E}_\xi \left[\frac{1}{N} \sum_\mu (\xi_i^\mu)^2 \, g_E^2(y_\mu, M_{i\to\mu}, V_{i\to\mu}) \right] = \mathbb{E}_\xi \left[\frac{1}{N} \sum_\mu g_E^2(y_\mu, M_{i\to\mu}, V_{i\to\mu}) \right]. \tag{40}$$

Now we use the fact that $M_{i\to\mu}$ and $V_{i\to\mu}$ from their very definitions in (10a) and (10c) are

quite close to their corresponding *non-cavity* quantities, i.e. respectively to M_μ and V_μ

$$M_{i\to\mu} = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\xi_j^\mu a_{j\to\mu} - \frac{1}{\sqrt{N}}\xi_i^\mu a_{i\to\mu} = M_\mu + O\left(\frac{1}{\sqrt{N}}\right), \tag{41a}$$

$$V_{i\to\mu} = \frac{1}{N}\sum_{j=1}^{N}(\xi_j^\mu)^2 b_{j\to\mu} - \frac{1}{N}(\xi_i^\mu)^2 b_{j\to\mu} = V_\mu + O\left(\frac{1}{N}\right). \tag{41b}$$

Discarding such small factors effectively tells us that $\mathbb{E}_\xi h_i^2$ is independent of the index i

$$\hat{q} \equiv \mathbb{E}_\xi h_i^2 = \mathbb{E}_\xi\left[\frac{1}{N}\sum_\mu g_E^2(y_\mu, M_\mu, V_\mu)\right] \tag{42}$$

Therefore equation (39) can be written as

$$q = \int Dh\, \tanh^2\left(\sqrt{\hat{q}}\, h\right) \tag{43}$$

In order to close the equations we have to perform the average in equation (42) examining the variables M_μ, V_μ. Let's start from M_μ. From its definition M_μ is a sum over uncorrelated random variables; in the large N limit it is therefore expected to converge, thanks to the central limit theorem, to a Gaussian variable with mean zero and variance

$$\mathbb{E}_\xi M_\mu^2 = \mathbb{E}_\xi\left[\frac{1}{N}\sum_j a_{j\to\mu}^2\right] \tag{44}$$

having exploited the fact that the average over ξ_j^μ can be performed explicitly, since $a_{j\to\mu}$ is independent of ξ_j^μ, as it only depends on ξ_j^ν with $\nu \neq \mu$. Next we use the fact that the message $a_{i\to\mu}$ is close, when N is large to a_i, i.e. $a_{i\to\mu} = a_i + O\left(\frac{1}{\sqrt{N}}\right)$, so that

$$\mathbb{E}_\xi M_\mu^2 \simeq \mathbb{E}_\xi\left[\frac{1}{N}\sum_i a_i^2\right] = q \tag{45}$$

Concerning V_μ we only need its mean, as its variance is vanishingly small in the large N limit

$$\mathbb{E}_\xi V_\mu \simeq \mathbb{E}_\xi\left[\frac{1}{N}\sum_{i=1}^{N} b_i\right] = 1 - \mathbb{E}_\xi\left[\frac{1}{N}\sum_i a_i^2\right] = 1 - q \tag{46}$$

Therefore equation (42) can be written as

$$\hat{q} = \alpha\int Dx\, g_E^2(y, \sqrt{q}x, 1-q) = \frac{\alpha}{1-q}\int Dx\left[GH\left(\sqrt{\frac{q}{1-q}}x\right)\right]^2 \tag{47}$$

Equations (43) and (47) describe the average evolution of the rBP algorithm. Note that those expressions exactly coincide with the Replica Symmetric saddle point equations obtained with the replica method. This way of deriving the results obtained by the replica method, using the BP equations is what is called in statistical physics as the *cavity method*.

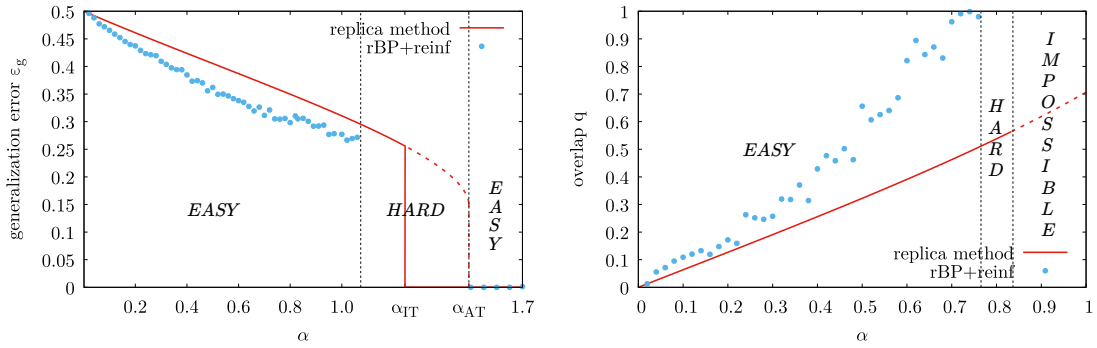

Figure 4: **Left:** Binary perceptron, teacher student setting. Generalization error found by the rBP+reinforcement algorithm. Both for small α and large α (where the teacher is the only solution) finding solutions is easy. In the interval $\alpha_{\text{alg}} < \alpha\alpha_{\text{AT}}$ finding solutions is hard on average. Here $\alpha_{\text{alg}} \simeq 1.1$. **Right:** Binary perceptron, storage setting. Overlap between students sampled from the posterior distribution estimated by the rBP algorithm using equation (38) (light blue dots). When this quantity is not near 1, it means that the one site marginals a_i are far from being binary variables; however their sign is a solution to the problem. Above the SAT/UNSAT transition $\alpha > \alpha_c \simeq 0.833$ finding solutions is impossible as they cease to exist. For $\alpha < \alpha_{\text{alg}}$ the algorithm finds solutions in polynomial time. When $\alpha_{\text{alg}} < \alpha < \alpha_c$ (with $\alpha_{\text{alg}} \simeq 0.77$) an exponential number of solutions exists but the algorithm is not able to find them. In both panels we have used $N = 4000$ and fixed $\delta r = 10^{-3}$. In red we have also shown the corresponding prediction using the replica method.

4 Reinforcement

So far, we have only focused on estimating the mean and variance of the marginals of w_i over the posterior distribution; only in the teacher-student setting were we able (in the large data regime) to extract actual binary configurations that match the teacher weights.

But what if we are simply interested in finding a solution to the learning problem, without requiring it to coincide with the teacher? A simple method to extract solutions from the relaxed Belief Propagation (rBP) algorithm is called reinforcement, which can be interpreted as a form of smooth decimation. The idea is to gradually encourage the system to commit to a certain configuration by "reinforcing" the local fields.

Concretely, at each iteration we add an inertia term to the cavity local field in equation (34), proportional to the local field computed at the previous step:

$$h^t_{i\to\mu} = \frac{1}{\sqrt{N}} \sum_{\nu\neq\mu} \xi^\nu_i \, g_E(y_\nu, M^t_{i\to\nu}, V^t_{i\to\nu}) + r^{t-1} h^{t-1}_i \,. \tag{48}$$

The parameter r^{t-1} is the reinforcement term, which is updated incrementally:

$$r^t = r^{t-1} + \delta r, \,. \tag{49}$$

Over time, the marginals become increasingly polarized, allowing us to extract a candidate solution by simply taking the sign of the marginals:

$$w_i = \text{sign}(a_i), \,. \tag{50}$$

The pseudo-code is reported below.

Algorithm 3: relaxed Belief Propagation + reinforcement, binary perceptron

Input: random guess for $a_{i\to\mu}^{t=0}$, $t=1$, $r^{t=0}=0$, δr

while $w_i = \text{sign}(a_i)$ *is not a solution* **do**

update $M_{i\to\mu}$ and $V_{i\to\mu}$

$$M_{i\to\mu}^{t} \leftarrow \frac{1}{\sqrt{N}} \sum_{j\neq i} \xi_j^\mu a_{j\to\mu}^{t-1} \tag{51a}$$

$$V_{i\to\mu}^{t} \leftarrow \frac{1}{N} \sum_{j\neq i} (\xi_j^\mu)^2 \left[1-(a_{j\to\mu}^{t-1})^2\right] \tag{51b}$$

update the *local cavity field* $h_{i\to\mu}$

$$h_{i\to\mu}^{t} \leftarrow \frac{1}{\sqrt{N}} \sum_{\nu\neq\mu} \xi_i^\nu \, g_E(y_\nu, M_{i\to\nu}^{t}, V_{i\to\nu}^{t}) + r^t h^{t-1} \tag{52}$$

update $a_{i\to\mu}$ via (22a) and a_i

$$a_{i\to\mu}^{t} \leftarrow \tanh\left(h_{i\to\mu}^{t}\right) \tag{53}$$

update the local field h_i

$$h_i^t = h_{i\to\mu}^{t} + \frac{1}{\sqrt{N}} \xi_i^\mu \, g_E(y_\mu, M_{i\to\mu}^{t}, V_{i\to\mu}^{t}) \tag{54a}$$

$$a_i^t = \tanh(h_i^t) \tag{54b}$$

update reinforcement

$$r^t \leftarrow r^{t-1} + \delta r \tag{55}$$

$t \leftarrow t+1$

end

Return: solution $w_i = \text{sign}\, a_i$

In the left panel of Figure 4 we show the generalization error of the solutions found by rBP+reinforcement. We see that the algorithm is able to find solutions up to $\alpha \simeq \alpha_{\text{alg}} \simeq 1.1$. The corresponding generalization error is better than typical solutions (i.e. those extracted with the flat measure over all possible solutions). In the region $\alpha_{\text{alg}} < \alpha < \alpha_{\text{AT}}$ the problem is on average hard to solve for rBP+reinforcement. Above α_{AT}, rBP+reinforcement infers the teacher as rBP.

Similar conclusions can be drawn in the storage setting (right panel of Figure 4). The main difference with the teacher-student case is that above the SAT/UNSAT transition it is impossible to find solutions as they do not exist. In the region $\alpha_{\text{alg}} < \alpha < \alpha_c$ with $\alpha_{\text{alg}} \simeq 0.77$ the algorithm is not able to find solutions in polynomial time, as the optimization becomes hard on average.

5 Approximate Message Passing

From the computational perspective, the rBP algorithm needs to store in memory order N^2 variables corresponding to the messages living on the links of the factor graph. The goal of this section is introduce an even more efficient algorithm, called Approximate Message Passing

(AMP), that needs to store only order N variables.

The key observation that will allow this simplification, is noticing (again!) that each interaction in our fully connected factor graph give a small, order $O\left(\frac{1}{\sqrt{N}}\right)$ contribution. Therefore, each of the message in the rBP algorithm is very close the corresponding one-site messages. For example, as argued before

$$M^t_{i\to\mu} = M^t_\mu - \frac{1}{\sqrt{N}}\xi^\mu_i a^{t-1}_{i\to\mu} \tag{56a}$$

$$V^t_{i\to\mu} = V^t_\mu - \frac{1}{N}(\xi^\mu_i)^2 b^{t-1}_{i\to\mu} = V^t_\mu + O\left(\frac{1}{N}\right) \tag{56b}$$

Our goal is therefore to rewrite the rBP equations only in terms of the site marginals. As we will see, in order to get those equations, we only need to get the correction to $a_{i\to\mu}$ but not on $b_{i\to\mu} > 0$, since the last one will not contribute to the large N limit.

Inserting (56) into (26) and keeping terms up to order $O\left(\frac{1}{N}\right)$, we have

$$A^t_{\mu\to i} = -\frac{(\xi^\mu_i)^2}{N}\partial_{M^t_{i\to\mu}} g_E(y_\mu, M^t_{i\to\mu}, V^t_{i\to\mu}) \simeq -\frac{(\xi^\mu_i)^2}{N}\partial_{M^t_\mu} g_E(y_\mu, M^t_\mu, V^t_\mu) \tag{57a}$$

$$B^t_{\mu\to i} = \frac{\xi^\mu_i}{\sqrt{N}}\, g_E(y_\mu, M^t_{i\to\mu}, V^t_{i\to\mu}) \simeq \frac{\xi^\mu_i}{\sqrt{N}}\, g_E(y_\mu, M^t_\mu, V^t_\mu) - \frac{(\xi^\mu_i)^2}{N}\, a^{t-1}_{i\to\mu}\partial_{M^t_i} g_E(y_\mu, M^t_\mu, V^t_\mu) \tag{57b}$$

$$\simeq \frac{\xi^\mu_i}{\sqrt{N}}\, g_E(y_\mu, M^t_\mu, V^t_\mu) + a^{t-1}_i A^t_{\mu\to i}$$

Note that in the last step we have used that at first order $a_{i\to\mu} \simeq a_i$; we are allowed to use this approximation since the the second term is already of order $1/N$. Using those equations we can write the expressions of R_i and Σ_i as

$$\Sigma^t_i \equiv \frac{1}{\sum_\mu A^t_{\mu\to i}} = \left[-\sum_\mu \frac{(\xi^\mu_i)^2}{N}\partial_{M^t_\mu} g_E(y_\mu, M^t_\mu, V^t_\mu)\right]^{-1} \tag{58a}$$

$$R^t_i \equiv \frac{\sum_\mu B^t_{\mu\to i}}{\sum_\mu A^t_{\mu\to i}} \simeq \frac{\sum_\mu \left[\frac{\xi^\mu_i}{\sqrt{N}}\, g_E(y_\mu, M^t_\mu, V^t_\mu) + a^{t-1}_i A^t_{\mu\to i}\right]}{\sum_\mu A^t_{\mu\to i}} = a^{t-1}_i + \frac{\Sigma^t_i}{\sqrt{N}}\sum_\mu \xi^\mu_i\, g_E(y_\mu, M^t_\mu, V^t_\mu) \tag{58b}$$

In order to close the equations we need to get the small correction of $a_{i\to\mu}$ to a_i, so that we can write the update equations for the M_μ and V_μ in terms of a_i and b_i only. In order to do that we need to compute the small correction of $R_{i\to\mu}$ to R_i; from the previous equation one gets

$$\Sigma^t_{i\to\mu} \equiv \frac{1}{\sum_{\nu\neq\mu} A^t_{\nu\to i}} \simeq \frac{1}{\sum_\nu A^t_{\nu\to i}} = \Sigma^t_i \tag{59a}$$

$$R^t_{i\to\mu} \equiv \frac{\sum_{\nu\neq\mu} B^t_{\nu\to i}}{\sum_{\nu\neq\mu} A^t_{\nu\to i}} = \frac{\sum_\nu B^t_{\nu\to i} - B^t_{\mu\to i}}{\sum_\nu A^t_{\nu\to i} - A^t_{\mu\to i}} \simeq R^t_i - \frac{B^t_{\mu\to i}}{\sum_\nu A^t_{\nu\to i}} = R^t_i - \Sigma^t_i B^t_{\mu\to i} \tag{59b}$$

Note that, since $A_{\nu\to i}$ is already a $O(\frac{1}{N})$ quantity, we can safely approximate $\Sigma_{i\to\mu}$ with its corresponding one site quantity Σ_i. Then, we insert the relations (59) into (28) we have

$$a^t_{i\to\mu} = g_S(R^t_{i\to\mu}, \Sigma^t_{i\to\mu}) \simeq g_S(R^t_i, \Sigma^t_i) - \Sigma^t_i B^t_{\mu\to i}\partial_{R^t_i} g_S(R^t_i, \Sigma^t_i) = a^t_i - B^t_{\mu\to i} b^t_i \tag{60a}$$

$$b^t_{i\to\mu} = \Sigma^t_{i\to\mu}\,\partial_{R^t_{i\to\mu}} g_S(R^t_{i\to\mu}, \Sigma^t_{i\to\mu}) \simeq \Sigma^t_i\,\partial_{R^t_{i\to\mu}} g_S(R^t_{i\to\mu}, \Sigma^t_i) \simeq b^t_i \tag{60b}$$

so that (notice that here is the only part where time indexes are different!)

$$V_\mu^t \equiv \frac{1}{N}\sum_j (\xi_j^\mu)^2 b_{j\to\mu}^{t-1} \simeq \frac{1}{N}\sum_j (\xi_j^\mu)^2 b_j^{t-1} \tag{61a}$$

$$\begin{aligned} M_\mu^t &\equiv \frac{1}{\sqrt{N}}\sum_j \xi_j^\mu a_{j\to\mu}^{t-1} \simeq \frac{1}{\sqrt{N}}\sum_j \xi_j^\mu a_j^{t-1} - \frac{1}{\sqrt{N}}\sum_j \xi_j^\mu\, b_j^{t-1} B_{\mu\to j}^{t-1} \\ &= \frac{1}{\sqrt{N}}\sum_j \xi_j^\mu a_j^{t-1} - \frac{1}{N}\sum_j (\xi_j^\mu)^2 b_j^{t-1}\, g_E(y_\mu, M_\mu^{t-1}, V_\mu^{t-1}) \\ &= \frac{1}{\sqrt{N}}\sum_j \xi_j^\mu a_j^{t-1} - V_\mu^t\, g_E(y_\mu, M_\mu^{t-1}, V_\mu^{t-1}) \end{aligned} \tag{61b}$$

Algorithm 4: Approximate Message Passing (AMP)

Input: random guess for $a_i^{t=0}$, $b_i^{t=0}$ and $g_E(y_\mu, M_\mu^{t=0}, V_\mu^{t=0}) \sim \mathcal{N}(0,1)$

while *a_i and b_i have converged* **do**

update M_μ and V_μ via (61b) (62a)

$$V_\mu^t \leftarrow \frac{1}{N}\sum_j (\xi_j^\mu)^2 b_j^{t-1} \tag{62a}$$

$$M_\mu^t \leftarrow \frac{1}{\sqrt{N}}\sum_j \xi_j^\mu a_j^{t-1} - V_\mu^t\, g_E(y_\mu, M_\mu^{t-1}, V_\mu^{t-1}) \tag{62b}$$

update Σ_i and R_i via (58a) (58b)

$$\Sigma_i^t \leftarrow \left[-\frac{1}{N}\sum_\mu (\xi_i^\mu)^2 \partial_{M_\mu^t} g_E(y_\mu, M_\mu^t, V_\mu^t)\right]^{-1} \tag{63a}$$

$$R_i^t \leftarrow a_i^{t-1} + \frac{\Sigma_i^t}{\sqrt{N}}\sum_\mu \xi_i^\mu\, g_E(y_\mu, M_\mu^t, V_\mu^t) \tag{63b}$$

update estimation of the site marginals

$$a_i^t \leftarrow g_S(R_i^t, \Sigma_i^t) \tag{64a}$$

$$b_i^t \leftarrow \Sigma_i^t\, \partial_{R_i^t} g_S(R_i^t, \Sigma_i^t) \tag{64b}$$

end

Return: site marginals: a_i, b_i.