Journal of Machine Learning Research 24 (2023) 1-89 Submitted 7/20; Revised 12/21; Published 2/23

Tree-AMP: Compositional Inference with Tree Approximate Message Passing

Antoine Baker ANTOINE.BAKER@ENS.FR
Florent Krzakala FLORENT.KRZAKALA@ENS.FR
Laboratoire de Physique
CNRS, École Normale Supérieure, PSL University
Paris, France

Benjamin Aubin BENJAMIN.AUBIN@CEA.FR
Lenka Zdeborová LENKA.ZDEBOROVA@CEA.FR
Institut de Physique Théorique
CNRS, CEA, Université Paris-Saclay
Saclay, France

Editor: Frank Wood

Abstract

We introduce Tree-AMP, standing for *Tree Approximate Message Passing*, a python package for compositional inference in high-dimensional tree-structured models. The package provides a unifying framework to study several approximate message passing algorithms previously derived for a variety of machine learning tasks such as generalized linear models, inference in multi-layer networks, matrix factorization, and reconstruction using non-separable penalties. For some models, the asymptotic performance of the algorithm can be theoretically predicted by the state evolution, and the measurements entropy estimated by the free entropy formalism. The implementation is modular by design: each module, which implements a factor, can be composed at will with other modules to solve complex inference tasks. The user only needs to declare the factor graph of the model: the inference algorithm, state evolution and entropy estimation are fully automated. The source code is publicly available at https://github.com/sphinxteam/tramp and the documentation at https://sphinxteam.github.io/tramp.docs.

Keywords: probabilistic programming, graphical models, Bethe entropy, state evolution, expectation propagation

©2023 Antoine Baker, Benjamin Aubin, Florent Krzakala and Lenka Zdeborová.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v24/20-695.html`.

1. Introduction

Probabilistic models have been used in many applications, as diverse as scientific data analysis, coding, natural language and signal processing. They also offer a powerful framework (Bishop, 2013) for several challenges in machine learning: dealing with uncertainty, choosing hyper-parameters, causal reasoning and model selection. However, the difficulty of deriving and implementing approximate inference algorithms for each new model may have hindered the wider adoption of Bayesian methods. The probabilistic programming approach seeks to make Bayesian inference as user friendly and streamlined as possible: ideally the user would only need to declare the probabilistic model and run an inference engine. Several probabilistic programming frameworks have been proposed, well suited for different contexts and leveraging variational inference or sampling methods to automate inference. To give a few examples, pomegranate (Schreiber, 2018) fits probabilistic models using maximum likelihood. Church (Goodman et al., 2008) and successors are universal languages for representing generative models. Infer.NET (Minka et al., 2018) implements several message passing algorithms such as expectation propagation. Stan (Carpenter et al., 2017) uses Hamiltonian Monte Carlo, while Anglican (Wood et al., 2014) uses particle MCMC as the sampling method. Turing (Ge et al., 2018) offers a Julia implementation. Recently, Edward (Tran et al., 2016) and Pyro (Bingham et al., 2019) tackle deep probabilistic problems, scaling inference up to large data and complex models.

In this paper, we present Tree-AMP (Tree Approximate Message Passing). In the current rich software ecosystem, Tree-AMP aims to fill a particular niche: using message passing algorithms with theoretical guarantees of performance in specific asymptotic settings. As will be detailed in Section 2, Tree-AMP uses the expectation propagation (EP) algorithm (Minka, 2001a) as its inference engine, which is also implemented by Infer.NET. Application-wise, for inference in statistical models or Bayesian machine learning tasks, the scope of Tree-AMP is very limited compared to the Infer.NET package. Indeed Tree-AMP is restricted to models like the ones presented in Figure 1, that is tree-structured factor graphs connecting high-dimensional variables, while Infer.NET can be applied to generic factor graphs with variables of arbitrary dimensions and types. However for the models considered in Figure 1, under specific asymptotic settings, Tree-AMP offers an in-depth theoretical analysis of its performance as will be detailed in Section 3: its errors can be predicted using the state evolution formalism, the free entropy formalism further predicts when the algorithm achieves or not the Bayes-optimal performance and also allows to estimate information theoretic quantities. For this reason, we believe the Tree-AMP package should be of special interest to theoretical researchers seeking a better understanding of EP/AMP algorithms. As an alternative to Tree-AMP let us mention the Vampyre package that allows inference in multi-layer networks (Fletcher et al., 2018).

There is a long history behind message passing (Yedidia et al., 2003; Mézard and Montanari, 2009), approximate message passing (AMP) (Donoho et al., 2009) and vector approximate message passing (VAMP) (Schniter et al., 2017), that we shall discuss later on. As exemplified in the context of compressed sensing, the AMP algorithm has a fundamental property: its performance on random instances in the high-dimensional limit, measured by the mean squared error on the signals, can be rigorously predicted by the so-called state evolution (Donoho et al., 2009; Bayati and Montanari, 2011), a rigorous version of the

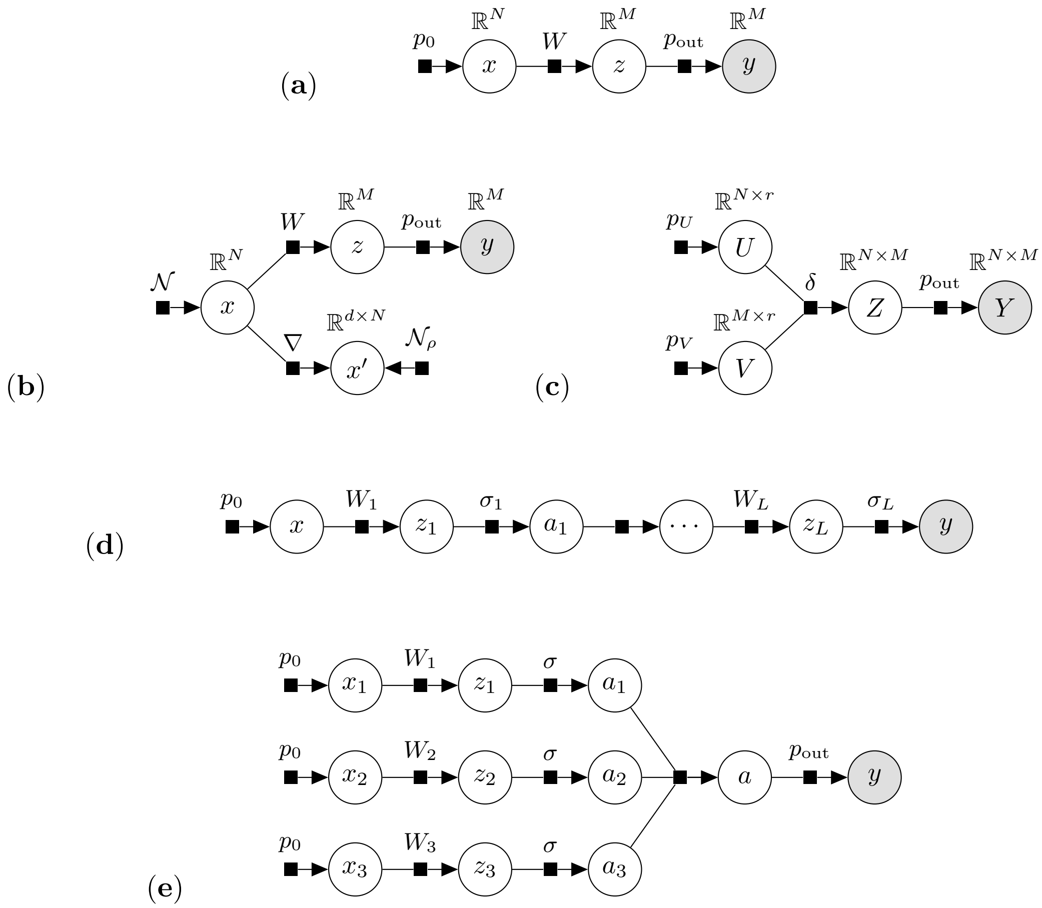


Figure 1: Tree-structured models. (**a**) Generalized linear model with separable prior $p_0(x)$, separable likelihood $p_{\text{out}}(y|z)$ and linear channel $z = Wx$. (**b**) Reconstruction using a sparse gradient prior where sparsity is enforced by the Gauss-Bernoulli prior $\mathcal{N}_\rho = [1-\rho]\delta + \rho\mathcal{N}$. (**c**) Low-rank matrix factorization with separable priors p_U and p_V, separable likelihood $p_{\text{out}}(Y|Z)$ and factorization $Z = \frac{UV^\intercal}{\sqrt{N}}$. (**d**) Multi-layer network with activation functions $a_l = \sigma_l(z_l)$ and linear channels $z_l = W_l a_{l-1}$. (**e**) Committee machine with three experts.

physicists "cavity method" (Mézard et al., 1987). These performances can be shown, in some cases, to reach the Bayes optimal one in polynomial time (Barbier et al., 2016; Reeves and Pfister, 2016), quite a remarkable feat! More recently, variant of the AMP approach has been developed with (some) correlated data and matrices (Schniter et al., 2017; Ma and Ping, 2017), again with guarantees of optimally in some cases (Barbier et al., 2018; Gerbelot et al., 2020). These approaches are intimately linked with the expectation propagation algorithm (Minka, 2001a) and the expectation consistency framework (Opper and Winther, 2005a).

The state evolution and the Bayes optimal guarantees were extended to a wide variety of models including for instance generalized linear models (GLM) (Rangan, 2011; Barbier et al., 2019), matrix factorization (Rangan and Fletcher, 2012; Deshpande and Montanari, 2014; Dia et al., 2016; Lesieur et al., 2017), committee machines (Aubin et al., 2018), optimization with non separable penalties (such as total variation) (Som and Schniter, 2012; Metzler et al., 2015; Tan et al., 2015; Manoel et al., 2018), inference in multi-layer networks (Manoel et al., 2017; Fletcher et al., 2018; Gabrié et al., 2018) and even arbitrary trees of GLMs (Reeves, 2017). In all these cases, the entropy of the system in the high dimensional limit can be obtained as the minimum of the so-called free entropy potential (Yedidia et al., 2003, 2005; Krzakala et al., 2014) and this allows the computation of interesting information theoretic quantities such as the mutual information between layers in a neural network (Gabrié et al., 2018). Furthermore, the global minimizer of the free entropy potential corresponds to the minimal mean squared error, which allows to determine fundamental limits to inference. Interestingly, the mean squared error achieved by AMP, predicted by the state evolution, is a stationary point of the same free entropy potential, which allows for an interesting interpretation of when the algorithm actually works (Zdeborová and Krzakala, 2016) in terms of phase transitions.

Unfortunately the development of AMP algorithms faced the same caveat as probabilistic modeling: for each new model, the AMP algorithm and the associated theory (free entropy and state evolution) had to be derived and implemented separately, which can be time-consuming. However, a key observation is that the factor graphs (Kschischang et al., 2001) for all the models mentioned above are tree-structured as illustrated in Figure 1. Each factor corresponds to an elementary inference problem that can be solved analytically or approximately. The Tree-AMP python package offers a unifying framework for all the models discussed above and extends to arbitrary tree-structured models. Similar to other probabilistic programming frameworks, the user only has to declare the model (here a tree-structured factor graph) then the inference, state evolution and entropy estimation are fully automated. The implementation is also completely modularized and extending Tree-AMP is in principle straightforward. If a new factor is needed, the user only has to solve (analytically or approximately) the elementary inference problem corresponding to this factor and implement it as a module in Tree-AMP.

Many of the AMP algorithms previously mentioned, especially the vectorized versions considered in (Schniter et al., 2017; Manoel et al., 2018; Fletcher et al., 2018), can be stated as particular instances of the expectation propagation (EP) algorithm (Minka, 2001a). The EP algorithm is equivalent to the expectation consistency framework (Opper and Winther, 2005a) which further yields an approximation for the log-evidence. Actually both approaches are solutions of the same relaxed Bethe variational problem (Heskes et al., 2005). In Section 2 we present the weak consistency derivation of EP by (Heskes et al., 2005) which offers a unifying framework that extends the previously mentioned AMP algorithms to tree-structured models. These are classic results but we hope that this pedagogical review will clarify the link between the various free energy formulations of the EP/AMP algorithms. Besides it allows us to introduce the key quantities (posterior moments and log-evidence) at the heart of the Tree-AMP implementation. Next in Section 3 we heuristically derive the replica free entropy (Mézard and Montanari, 2009) by using weak consistency on the overlaps and conjecture the state evolution. Even if the derivation is non-rigorous, we recover earlier results derived for specific models. Interestingly, the state evolution and the free entropy potential can be reinterpreted as simple ensemble average of the posterior variances and log-evidence estimated by EP. This allows us to extend the state evolution and free entropy formalism to tree-structured models and implement them in the Tree-AMP package. Finally, in Section 4 we illustrate the package on a few examples.

2. Expectation Propagation

In this section we review the derivation of EP as a relaxed variational problem. First we briefly recall the variational inference framework and the Bethe decomposition of the free energy (Yedidia et al., 2003) which is exact for tree-structured models. Then following Heskes et al. (2005), the Bethe variational problem can be approximately solved by enforcing moment-matching instead of full consistency of the marginals, which yields the EP algorithm. The EP solution consists of exponential family distributions which satisfy a duality between natural parameters and moments (Wainwright and Jordan, 2008). The EP free energy (Minka, 2001b) is shown to be equivalent to the AMP free energies and satisfies a tree decomposition which is at the heart of the modularization in the Tree-AMP package. Finally we expose the Tree-AMP implementation of EP and maximum a posteriori (MAP) estimation.

2.1 Model Settings

The Tree-AMP package is limited to high-dimensional tree-structured factor graphs, like the models presented in Figure 1. To define more precisely such models and set some notation, consider an inference problem $p(\mathbf{x},\mathbf{y})$ where $\mathbf{x}=(x_i)_{i\in V}$ are the signals to infer and $\mathbf{y}=(y_j)_{j\in O}$ the measurements. We emphasize that in our context each signal $x_i \in \mathbb{R}^{N_i}$ and measurement $y_j \in \mathbb{R}^{N_j}$ is itself a high dimensional object. The model is typically considered in the large N limit with ratios $\alpha_i = \frac{N_i}{N} = O(1)$ and $\alpha_j = \frac{N_j}{N} = O(1)$. We will assume that $p(\mathbf{x},\mathbf{y})$ can be factorized as a tree-structured probabilistic graphical model:

$$p(\mathbf{x},\mathbf{y}) = \frac{1}{Z_N}\prod_{k\in F} f_k(x_k;y_k), \quad Z_N = \int d\mathbf{x}d\mathbf{y}\prod_{k\in F} f_k(x_k;y_k), \tag{1}$$

with factors $(f_k)_{k\in F}$. The factorization structure can be conveniently represented as a factor graph (Kschischang et al., 2001): a bipartite graph $\mathcal{G}=(V,F,E)$, where each signal x_i is represented as a variable node $i\in V$ (circle), each factor f_k by a factor node $k\in F$ (square), with an edge $(i,k)\in E$ connecting the variable node i to the factor node k if and only if x_i is an argument of f_k. We will use the symbol ∂ to denote the neighbors nodes in the factor graph. Thus $i\in\partial k$ denotes the variable nodes neighboring the factor node k and the arguments of the factor f_k are $x_k=(x_i)_{i\in\partial k}$. Similarly $k\in\partial i$ denotes the factor nodes neighboring the variable node i.

Note that while we denote $\mathbb{R}^{N_i}$ the integration domain of the signal x_i our inference tasks are not limited to real high-dimensional variables. Indeed a high-dimensional binary, sparse, categorical or complex variable can always be embedded in some $\mathbb{R}^{N_i}$ and its type enforced by an appropriate factor. For instance a binary variable $x\in\pm^N$ can be enforced by a binary prior $p_0(x)=p_+\delta_+(x)+p_-\delta_-(x)$. As a consequence, we allow generic measures for the factors, including Dirac measures. This additionally allows us to represent hard constraints as factors, for example the linear channel $z = Wx$ will be represented by the factor $\delta(z - Wx)$. The high-dimensional factor f_k has to be simple enough to lead to a tractable inference problem as will be explained in Section 2.8. Some representative models are given in Examples 1-4.

The goal of the inference is then to get the posterior $p(\mathbf{x} \mid \mathbf{y})$ and evidence $p(\mathbf{y})$, or equivalently the negative log-evidence known as surprisal in information theory. For the factorization Eq. (1) the posterior is equal to:

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z_N(\mathbf{y})} \prod_{k \in F} f_k(x_k; y_k), \quad Z_N(\mathbf{y}) = \int d\mathbf{x} \prod_{k \in F} f_k(x_k; y_k). \tag{2}$$

The *Helmholtz free energy* is here defined as the negative log partition

$$F(\mathbf{y}) = -\ln Z_N(\mathbf{y}) = -\ln p(\mathbf{y}) - \ln Z_N \tag{3}$$

and gives the surprisal up to a constant.

Remark 1 (Bayesian network) *In a Bayesian network, the factors correspond to the conditional distributions $f_k(x_k; y_k) = p(x_k^+, y_k \mid x_k^-)$ where x_k^+ (resp. x_k^-) denote the outputs (resp. inputs) variables of the factor, besides $Z_N = 1$ so the Helmholtz free energy Eq.* (3) *directly gives the surprisal. All the models considered in Figure 1 are Bayesian networks, with the exception of **(b)**.*

Example 1 (GLM) *The factor graph for the generalized linear model (GLM) is shown in Figure 1 **(a)**. A high-dimensional signal $x \in \mathbb{R}^N$ is drawn from a separable prior p_0, a high-dimensional measurement $y \in \mathbb{R}^M$ is obtained through a separable likelihood p_{out} from $z = Wx \in \mathbb{R}^M$, where W is a given $M \times N$ matrix. The model is typically considered in the large N limit with $\alpha = \frac{M}{N} = O(1)$. The variables to infer are $\mathbf{x} = (x_i)_{i \in V} = (x, z)$, the observation $\mathbf{y} = (y_j)_{j \in O} = y$ and the factors $(f_k)_{k \in F} = (p_0(x), \delta(z - Wx), p_{out}(y|z))$.*

Example 2 (Low rank matrix factorization) *The factor graph for the low rank matrix factorization model considered in (Lesieur et al., 2017) is shown in Figure 1 **(c)**. Two matrices $U \in \mathbb{R}^{N \times r}$ and $V \in \mathbb{R}^{M \times r}$ are drawn from separable priors p_U and p_V, a high-dimensional measurement $Y \in \mathbb{R}^{N \times M}$ is obtained through a separable likelihood p_{out} from $Z = \frac{UV^\intercal}{\sqrt{N}} \in \mathbb{R}^{N \times M}$. The model is typically considered in the large N limit with $\alpha = \frac{M}{N} = O(1)$ and finite rank $r = O(1)$. The variables to infer are $\mathbf{x} = (x_i)_{i \in V} = (U, V, Z)$, the observation $\mathbf{y} = (y_j)_{j \in O} = Y$ and the factors $(f_k)_{k \in F} = (p_U(U), p_V(V), \delta(Z - \frac{UV^\intercal}{\sqrt{N}}), p_{out}(Y|Z))$.*

Example 3 (Extensive rank matrix factorization) *The factor graph for the extensive rank matrix factorization model considered in (Kabashima et al., 2016) is the same as the low rank case but considered in a different asymptotic regime. Two matrices $F \in \mathbb{R}^{M \times N}$ and $X \in \mathbb{R}^{N \times P}$ are drawn from separable priors p_F and p_X, a high-dimensional measurement $Y \in \mathbb{R}^{M \times P}$ is obtained through a separable likelihood p_{out} from $Z = \frac{FX}{\sqrt{N}} \in \mathbb{R}^{M \times P}$. The model is typically considered in the large N, M, P limit with fixed ratios $\alpha = \frac{M}{N} = O(1)$ and $\pi = \frac{P}{N} = O(1)$. The variables to infer are $\mathbf{x} = (x_i)_{i \in V} = (F, X, Z)$, the observation $\mathbf{y} = (y_j)_{j \in O} = Y$ and the factors $(f_k)_{k \in F} = \Big(p_F(F), p_X(X), \delta(Z - \frac{FX}{\sqrt{N}}), p_{out}(Y|Z)\Big)$.*

Example 4 (Sparse gradient regression) *The factor graph for this model is shown in Figure 1* ***(b)****. Compared to the GLM, we wish to infer a signal* $x \in \mathbb{R}^N$ *which gradient* $x' = \nabla x \in \mathbb{R}^{d \times N}$ *is sparse (we view the signal as having* d *axis, for instance* $d = 2$ *for an image, and consequently the gradient is taken along* d *directions). The sparsity is enforced by a Gauss-Bernoulli prior* $\mathcal{N}_\rho = [1-\rho]\delta + \rho\mathcal{N}$*. The variables to infer are* $\mathbf{x} = (x_i)_{i \in V} = (x, x', z)$*, the observation* $\mathbf{y} = (y_j)_{j \in O} = y$ *and the factors* $(f_k)_{k \in F} = (p_0(x), \delta(x' - \nabla x), \mathcal{N}_\rho(x'), \delta(z - Wx), p_{out}(y|z))$*. The factor graph is not a Bayesian network, in particular the partition function Eq.* (1) *is equal to:*

$$Z_N = \int_{\mathbb{R}^N} dx \mathcal{N}(x) \mathcal{N}_\rho(\nabla x) \neq 1. \tag{4}$$

2.2 Bethe Free Energy

We briefly recall the Bethe variational formulation of the belief propagation algorithm following (Yedidia et al., 2003) and refer the reader to this reference or (Wainwright and Jordan, 2008) for further details. We are interested in computationally hard inference problems and seek an approximation $\tilde{p}$ of the posterior distribution $\tilde{p}(\mathbf{x}) \simeq p(\mathbf{x} \mid \mathbf{y})$. Consider such an approximation $\tilde{p}$ and the following functional

$$\mathcal{F}[\tilde{p}] = \mathrm{KL}[\tilde{p}(\mathbf{x}) \| p(\mathbf{x} \mid \mathbf{y})] + F(\mathbf{y}) \,, \tag{5}$$

called the *variational free energy.* As the KL divergence is always positive and equal to zero only if the two distributions are equal, we can formally get the posterior and the Helmholtz free energy $F(\mathbf{y})$ as the solution of the variational problem:

$$F(\mathbf{y}) = \min_{\tilde{p}} \mathcal{F}[\tilde{p}] \quad \text{for} \quad \tilde{p}^\star(\mathbf{x}) = p(\mathbf{x} \mid \mathbf{y}). \tag{6}$$

However, for a tree-structured model it can be shown that the posterior factorizes as

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{\prod_{k \in F} p(x_k \mid \mathbf{y})}{\prod_{i \in V} p(x_i \mid \mathbf{y})^{n_i - 1}} \,, \tag{7}$$

where $p(x_i \mid \mathbf{y})$ is the marginal of the variable x_i, $p(x_k \mid \mathbf{y})$ is the joint marginal over $x_k = (x_i)_{i \in \partial k}$ and $n_i = |\partial i|$ is the number of neighbor factors of the variable x_i. Therefore we can restrict the variational problem Eq. (6) to distributions of the form

$$\tilde{p}(\mathbf{x}) = \frac{\prod_{k \in F} \tilde{p}_k(x_k)}{\prod_{i \in V} \tilde{p}_i(x_i)^{n_i - 1}} \,, \tag{8}$$

and minimize over the collection of variable marginals $\tilde{p}_V = (\tilde{p}_i)_{i \in V}$ and factor marginals $\tilde{p}_F = (\tilde{p}_k)_{k \in F}$. This collection however has to satisfy a strong self-consistency constraint: whenever the variable x_i is an argument of the factor f_k, the i-marginal of the factor marginal $\tilde{p}_k$ must give back the variable marginal $\tilde{p}_i$. In other words the collection of marginals $(\tilde{p}_V, \tilde{p}_F)$ must belong to the set:

$$\mathcal{M} = \left\{ (\tilde{p}_V, \tilde{p}_F) \quad : \quad \forall (i,k) \in E, \quad \tilde{p}_i(x_i) = \tilde{p}_k(x_i) = \int dx_{k \setminus i} \, \tilde{p}_k(x_k) \right\}. \tag{9}$$

For distributions of the type Eq. (8), the variational free energy Eq. (5) is equal to

$$\mathcal{F}_{\text{Bethe}}[\tilde{p}_V, \tilde{p}_F] = \sum_{k \in F} F_k[\tilde{p}_k] + \sum_{i \in V} (1 - n_i) F_i[\tilde{p}_i] ,$$
$$\text{with} \quad F_k[\tilde{p}_k] = \mathrm{KL}[\tilde{p}_k \| f_k] , \quad F_i[\tilde{p}_i] = -\mathrm{H}[\tilde{p}_i] , \tag{10}$$

called the *Bethe free energy* (Yedidia et al., 2003), where KL and H denote respectively the Kullback-Leibler divergence and the entropy. Therefore for a tree-structured model we have:

$$F(\mathbf{y}) = \min_{(\tilde{p}_V, \tilde{p}_F) \in \mathcal{M}} \mathcal{F}_{\text{Bethe}}[\tilde{p}_V, \tilde{p}_F] \quad \text{at} \quad \begin{cases} \tilde{p}_k^\star(x_k) = p(x_k \mid \mathbf{y}) & \text{for all } k \in F \\ \tilde{p}_i^\star(x_i) = p(x_i \mid \mathbf{y}) & \text{for all } i \in V \end{cases} \tag{11}$$

The solution of this Bethe variational problem actually leads to the belief propagation algorithm (Pearl, 1988; Yedidia et al., 2003).

2.3 Weak Consistency

For tree-structured models the Bethe variational problem Eq. (11) yields the exact posterior and Helmoltz free energy, and is solved by the belief propagation algorithm. Unfortunately for the models introduced in Figure 1, which involve high-dimensional vectors or matrices, the belief propagation algorithm is not tractable. Following Heskes et al. (2005) we consider instead a relaxed version of the Bethe variational problem, by replacing the strong consistency constraint Eq. (9) by the *weak consistency* constraint:

$$\mathcal{M}_\phi = \{ (\tilde{p}_V, \tilde{p}_F) \quad : \quad \forall (i,k) \in E, \quad \mathbb{E}_{\tilde{p}_i} \phi_i(x_i) = \mathbb{E}_{\tilde{p}_k} \phi_i(x_i) \} , \tag{12}$$

for a collection $\phi = (\phi_i)_{i \in V}$ of sufficient statistics $\phi_i : \mathbb{R}^{N_i} \to \mathbb{R}^{d_i}$ for each variable x_i. In other words instead of requiring the full consistency of the marginals we only require moment matching. The collection ϕ is a choice and each choice leads to a different approximation scheme. As the notation suggests, one can choose different sufficient statistics for each variable x_i. Following Wainwright and Jordan (2008) we use $\langle \lambda_i, \phi_i(x_i) \rangle$ to denote the Euclidian inner product in $\mathbb{R}^{d_i}$ of the so-called natural parameter $\lambda_i \in \mathbb{R}^{d_i}$ with $\phi_i(x_i) \in \mathbb{R}^{d_i}$. Currently the Tree-AMP package only supports isotropic Gaussian beliefs (Example 5) but we plan to include more generic Gaussian beliefs (Examples 6-8) in future versions of the package. The *relaxed Bethe variational problem*:

$$F_\phi(\mathbf{y}) = \min_{(\tilde{p}_V, \tilde{p}_F) \in \mathcal{M}_\phi} \mathcal{F}_{\text{Bethe}}[\tilde{p}_V, \tilde{p}_F] \simeq F(\mathbf{y}) \quad \text{at} \quad \begin{cases} \tilde{p}_k^\star(x_k) \simeq p(x_k \mid \mathbf{y}) & \text{for all } k \in F \\ \tilde{p}_i^\star(x_i) \simeq p(x_i \mid \mathbf{y}) & \text{for all } i \in V \end{cases} \tag{13}$$

leads to the following solution as proven by (Heskes et al., 2005). Let $\lambda_{i \to k} \in \mathbb{R}^{d_i}$ denotes the Lagrange multiplier associated to the moment matching constraint $\mathbb{E}_{\tilde{p}_i} \phi_i(x_i) = \mathbb{E}_{\tilde{p}_k} \phi_i(x_i)$. The (approximate) factor marginal $\tilde{p}_k^\star(x_k)$ belongs to the exponential family

$$p_k(x_k \mid \lambda_k) = \frac{1}{Z_k[\lambda_k]} f_k(x_k; y_k) e^{\langle \lambda_k, \phi_k(x_k) \rangle}, \quad Z_k[\lambda_k] = \int dx_k f_k(x_k; y_k) e^{\langle \lambda_k, \phi_k(x_k) \rangle} \tag{14}$$

with natural parameter and sufficient statistics

$$\lambda_k = (\lambda_{i\to k})_{i\in\partial k}, \quad \phi_k(x_k) = (\phi_i(x_i))_{i\in\partial k}, \quad \langle \lambda_k, \phi_k(x_k)\rangle = \sum_{i\in\partial k} \langle \lambda_{i\to k}, \phi_i(x_i)\rangle. \tag{15}$$

The (approximate) variable marginal $\tilde{p}_i^\star(x_i)$ belongs to the exponential family

$$p_i(x_i \mid \lambda_i) = \frac{1}{Z_i[\lambda_i]} e^{\langle \lambda_i, \phi_i(x_i)\rangle}, \quad Z_i[\lambda_i] = \int dx_i e^{\langle \lambda_i, \phi_i(x_i)\rangle} \tag{16}$$

with natural parameter given by

$$(n_i - 1)\lambda_i = \sum_{k\in\partial i} \lambda_{i\to k}, \tag{17}$$

or introducing the factor-to-variable messages $\lambda_{k\to i} = \lambda_i - \lambda_{i\to k}$:

$$\lambda_i = \sum_{k\in\partial i} \lambda_{k\to i}. \tag{18}$$

The moment matching condition can be written as:

$$\mu_i^\star = \mu_i[\lambda_i] = \mu_i^k[\lambda_k] \tag{19}$$

where $\mu_i[\lambda_i] = \mathbb{E}_{p_i(x_i|\lambda_i)}\phi_i(x_i)$ and $\mu_i^k[\lambda_k] = \mathbb{E}_{p_k(x_k|\lambda_k)}\phi_i(x_i)$ are the moments of the variable x_i as estimated by the variable and factor marginal respectively. This is the fixed point searched by the EP algorithm (Minka, 2001a).

Example 5 (Isotropic Gaussian belief) *It corresponds to the sufficient statistics* $\phi_i(x_i) = (x_i, -\frac{1}{2}\|x_i\|^2)$ *so* $d_i = N_i + 1$. *The associated natural parameters are* $\lambda_i = (b_i, a_i)$ *with* $b_i \in \mathbb{R}^{N_i}$ *and scalar precision* $a_i \in \mathbb{R}_+$. *The inner product leads to an isotropic Gaussian belief* $e^{\langle \lambda_i, \phi_i(x_i)\rangle} = e^{-\frac{a_i}{2}\|x_i\|^2 + b_i^\intercal x_i}$ *on* x_i.

Example 6 (Diagonal Gaussian belief) *It corresponds to the sufficient statistics* $\phi_i(x_i) = (x_i, -\frac{1}{2}x_i^2)$ *so* $d_i = 2N_i$ *and the precision* $a_i \in \mathbb{R}_+^{N_i}$ *is a vector.*

Example 7 (Full covariance Gaussian belief) *It corresponds to the sufficient statistics* $\phi_i(x_i) = (x_i, -\frac{1}{2}x_i x_i^\intercal)$ *so* $d_i = N_i + \frac{N_i(N_i+1)}{2}$ *and the precision* a_i *is a* $N_i \times N_i$ *positive symmetric matrix. Note however that such a belief will be computationally demanding as inverting the precision matrix will take* $O(N_i^3)$ *time.*

Example 8 (Structured Gaussian beliefs) *When the high dimensional variable* x_i *has an inner structure with multiple indices, a more complex covariance structure can be envisioned. For instance in the low rank matrix factorization problem (Example 2) Lesieur et al. (2017) consider for* $U \in \mathbb{R}^{N\times r}$ *a Gaussian belief with a full covariance in the second coordinate but diagonal in the first. In other words the low rank matrix* $U = (U_n)_{n=1}^N$ *is viewed as a collection of vectors* $U_n \in \mathbb{R}^r$ *with sufficient statistics* $\phi(U) = (U_n, -\frac{1}{2}U_n U_n^\intercal)_{n=1}^N$, *natural parameters* $\lambda_U = (b_n, a_n)_{n=1}^N$ *with* $b_n \in \mathbb{R}^r$ *and* a_n *is a* $r \times r$ *positive symmetric matrix. Then* $d_U = Nr + N\frac{r(r+1)}{2}$ *and* $\lambda_U = (b_U, a_U)$ *with* $b_U \in \mathbb{R}^{N\times r}$ *and* $a_U \in \mathbb{R}^{N\times r\times r}$.

2.4 Moments and Natural Parameters Duality

As the solution of the relaxed Bethe variational problem involves exponential family distributions, it is useful to recall some of their basic properties (Wainwright and Jordan, 2008). The log-partitions

$$A_k[\lambda_k] = \ln Z_k[\lambda_k] = \ln \int dx_k\, f_k(x_k; y_k)\, e^{\langle \lambda_k, \phi_k(x_k)\rangle}\,, \tag{20}$$

$$A_i[\lambda_i] = \ln Z_i[\lambda_i] = \ln \int dx_i\, e^{\langle \lambda_i, \phi_i(x_i)\rangle}\,, \tag{21}$$

are convex functions and provide the bijective mappings between the convex set of natural parameters and the convex set of moments:

$$\mu_k[\lambda_k] = \mathbb{E}_{p_k(x_k|\lambda_k)}\phi_k(x_k) = \partial_{\lambda_k} A_k[\lambda_k]\,, \tag{22}$$

$$\mu_i[\lambda_i] = \mathbb{E}_{p_i(x_i|\lambda_i)}\phi_i(x_i) \ = \partial_{\lambda_i} A_i[\lambda_i]\,, \tag{23}$$

and the inverse mappings are given by:

$$\lambda_k[\mu_k] = \partial_{\mu_k} G_k[\mu_k]\,, \tag{24}$$

$$\lambda_i[\mu_i] = \partial_{\mu_i} G_i[\mu_i]\,, \tag{25}$$

where $G_i[\mu_i]$ and $G_k[\mu_k]$ are the Legendre transformations (convex conjugates) of the log-partitions and are equal to the KL divergence and the negative entropy respectively:

$$G_k[\mu_k] = \max_{\lambda_k} \langle \lambda_k, \mu_k\rangle - A_k[\lambda_k] = \mathrm{KL}[p_k(x_k \mid \lambda_k) \| f_k(x_k; y_k)]\,, \tag{26}$$

$$G_i[\mu_i] = \max_{\lambda_i} \langle \lambda_i, \mu_i\rangle - A_i[\lambda_i] = -\mathrm{H}[p_i(x_i \mid \lambda_i)]\,. \tag{27}$$

We recall that for the factor marginal the natural parameter, sufficient statistics and inner product are given by Eq. (15). The corresponding moment is $\mu_k = (\mu_i^k)_{i\in\partial k}$ and the mapping Eq. (22) and the inner product in Eq. (26) are explicitly:

$$\mu_i^k[\lambda_k] = \mathbb{E}_{p_k(x_k|\lambda_k)}\phi_i(x_i) = \partial_{\lambda_{i\to k}} A_k[\lambda_k] \quad \text{for all } i \in \partial k, \tag{28}$$

$$\langle \lambda_k, \mu_k\rangle = \sum_{i\in\partial k} \langle \lambda_{i\to k}, \mu_i^k\rangle. \tag{29}$$

Example 9 (Duality for isotropic Gaussian beliefs) *Let us consider isotropic Gaussian beliefs (Example 5) with natural parameters $\lambda_i = (b_i, a_i) \in \mathbb{R}^{N_i} \times \mathbb{R}_+$. The corresponding moments $\mu_i = (r_i, -\frac{N_i}{2}\tau_i)$ are the mean $r_i = \mathbb{E}_{p_i} x_i \in \mathbb{R}^{N_i}$ and second moment $\tau_i = \mathbb{E}_{p_i} \frac{\|x_i\|^2}{N_i} \in \mathbb{R}_+$. The variable marginal Eq. (16) is the isotropic Gaussian*

$$p_i(x_i \mid a_i, b_i) = e^{-\frac{a_i}{2}\|x_i\|^2 + b_i^\intercal x_i - A_i[b_i, a_i]} = \mathcal{N}(x_i \mid r_i, v_i) \tag{30}$$

with mean $r_i \in \mathbb{R}^{N_i}$ and variance $v_i \in \mathbb{R}_+$. The mapping between the two parametrizations is particularly simple:

$$r_i = \frac{b_i}{a_i}, \quad v_i = \frac{1}{a_i} \quad \text{and} \quad b_i = \frac{r_i}{v_i}, \quad a_i = \frac{1}{v_i}. \tag{31}$$

The variable log-partition

$$A_i[a_i, b_i] = \frac{\|b_i\|^2}{2a_i} + \frac{N_i}{2} \ln \frac{2\pi}{a_i}, \tag{32}$$

gives consistently the forward mapping $r_i = \partial_{b_i} A_i = \frac{b_i}{a_i}$ *and* $-\frac{N_i}{2}\tau_i = \partial_{a_i} A_i$ *with* $\tau_i = \frac{\|r_i\|^2}{N_i} + v_i$ *and* $v_i = \frac{1}{a_i}$. *The variable negative entropy*

$$G_i[r_i, \tau_i] = -\frac{N_i}{2} \ln 2\pi e v_i \quad \textit{with} \quad v_i = \tau_i - \frac{\|r_i\|^2}{N_i} \tag{33}$$

gives consistently the inverse mapping $b_i = \partial_{r_i} G_i = \frac{r_i}{v_i}$ *and* $-\frac{N_i}{2} a_i = \partial_{\tau_i} G_i$ *with* $a_i = \frac{1}{v_i}$. *The moment matching condition Eq.* (19) *is equivalent to match the mean* $r_i = r_i^k$ *and isotropic variance* $v_i = v_i^k$. *The factor marginal Eq.* (14) *is the factor titled by Gaussian beliefs:*

$$p_k(x_k \mid a_k, b_k) = f(x_k; y_k) e^{-\frac{1}{2} a_k \|x_k\|^2 + b_k^\intercal x_k - A_k[a_k, b_k]} \tag{34}$$

where following Eq. (15) *we denote compactly the inner product by:*

$$-\frac{a_k}{2} \|x_k\|^2 + b_k^\intercal x_k = \sum_{i \in \partial k} -\frac{a_{i \to k}}{2} \|x_i\|^2 + b_{i \to k}^\intercal x_i \tag{35}$$

The factor log-partition, mean and isotropic variance are explicitly given by:

$$A_k[a_k, b_k] = \ln \int dx_k \, f(x_k; y_k) \, e^{-\frac{1}{2} a_k \|x_k\|^2 + b_k^\intercal x_k}, \tag{36}$$

$$r_i^k[a_k, b_k] = \mathbb{E}_{p_k(x_k|a_k, b_k)} x_i = \partial_{b_{i \to k}} A_k[a_k, b_k], \tag{37}$$

$$v_i^k[a_k, b_k] = \langle \mathrm{Var}_{p_k(x_k|a_k, b_k)}(x_i) \rangle = \langle \partial^2_{b_{i \to k}} A_k[a_k, b_k] \rangle \tag{38}$$

where $\langle \cdot \rangle$ *denotes the average over components. Several factor log-partitions with isotropic Gaussian beliefs are given in Appendix E.*

2.5 Tree Decomposition of the Free Energy

We now present several but equivalent free energy formulations of the EP and AMP algorithms. The expectation consistency (EC) Gibbs free energy (Opper and Winther, 2005b) is a function defined over the posterior moments $\mu_V = (\mu_i)_{i \in V}$:

$$G[\mu_V] = \sum_{k \in F} G_k[\mu_k] + \sum_{i \in V} (1 - n_i) G_i[\mu_i] \quad \text{where} \quad \mu_k = (\mu_i)_{i \in \partial k}. \tag{39}$$

The EP free energy (Minka, 2001b) is a function defined over the variable $\lambda_V = (\lambda_i)_{i \in V}$ and factor $\lambda_F = (\lambda_k)_{k \in F}$ natural parameters:

$$A[\lambda_V, \lambda_F] = \sum_{k \in F} A_k[\lambda_k] + \sum_{i \in V} (1 - n_i) A_i[\lambda_i]. \tag{40}$$

The Tree-AMP free energy is the same as the EP free energy but is parameterized in term of factor to variable messages and variable to factor messages $\lambda_E = (\lambda_{k\to i}, \lambda_{i\to k})_{(i,k)\in E}$:

$$A[\lambda_E] = \sum_{k\in F} A_k[\lambda_k] - \sum_{(i,k)\in E} A_i[\lambda_i^k] + \sum_{i\in V} A_i[\lambda_i], \tag{41}$$

$$\text{where} \quad \lambda_k = (\lambda_{i\to k})_{i\in\partial k} \quad \lambda_i^k = \lambda_{i\to k} + \lambda_{k\to i}, \quad \lambda_i = \sum_{k\in\partial i} \lambda_{k\to i}. \tag{42}$$

The parametrization Eq. (42) in term of messages has a nice interpretation: the variable natural parameter is the sum of the incoming messages (coming from the neighboring factors), the factor natural parameter is the set of the incoming messages (coming from the neighboring variables).

Proposition 2 *The relaxed Bethe variational problem Eq.* (13) *can be formulated in term of the posterior moments using the EC Gibbs free energy, in term of the factor and variable natural parameters using the EP free energy, or in term of the natural parameters messages using the Tree-AMP free energy:*

$$F_\phi(\mathbf{y}) = \min_{\mu_V} G[\mu_V] \tag{43}$$

$$= \min_{\lambda_V} \max_{\lambda_F} -A[\lambda_V, \lambda_F] \quad s.t. \quad \forall i \in V : (n_i - 1)\lambda_i = \sum_{k\in\partial i} \lambda_{i\to k} \tag{44}$$

$$= \min \operatorname*{extr}_{\lambda_E} -A[\lambda_E]. \tag{45}$$

Besides, any stationary point of the free energies (not necessarily the global optima) is an EP fixed point:

$$\mu_i = \mu_i^k, \quad (n_i - 1)\lambda_i = \sum_{k\in\partial i} \lambda_{i\to k}. \tag{46}$$

The $\min \operatorname{extr}_{\lambda_E}$ notation in Eq. (45) means to search for stationary (in general saddle) points of $A[\lambda_E]$ and among these critical points select the minimizer of G.

2.5.1 Proofs

There are several but separate derivations of these equivalences in the literature. In (Minka, 2001b) the optimization problem Eq. (44) is shown to lead to the EP algorithm, where $A[\lambda_V, \lambda_F]$ is called the EP dual energy function while the relaxed Bethe variational problem is called the EP primal energy function. Opper and Winther (2005a,b) further show that the minimization Eq. (43), or equivalently the dual saddle point problem Eq. (44), yields the so-called EC approximation of $-\ln Z_N(\mathbf{y})$, here denoted by $F_\phi(\mathbf{y})$. The EC approximation is presented for a very simple factor graph with only one variable x and two factors $f_q(x)$ and $f_r(x)$ but can be straightforwardly extended to a tree-structured factor graph which yields $A[\lambda_V, \lambda_F]$ and $G[\mu_V]$. Finally, Heskes et al. (2005) unify the two formalisms as solutions to the relaxed Bethe variational problem, and a similar approach is presented in (Wainwright and Jordan, 2008) using an extended exponential family distribution. Proposition 2 is the straightforward application of these ideas to the tree-structured models considered in this manuscript. We present in Appendix A a condensed proof for the reader convenience using the duality between moments and natural parameters.

2.5.2 Exact Tree Decomposition

The Bethe, EC Gibbs, EP and Tree-AMP free energies all follow the same tree decomposition: a sum over factors − sum over edges + sum over variables. The same tree decomposition holds for the Helmholtz free energy $F(\mathbf{y})$ and the EC approximation $F_\phi(\mathbf{y})$. Indeed $F(\mathbf{y})$ is simply the Bethe free energy evaluated at the true marginals according to Eq. (11) and $F_\phi(\mathbf{y})$ is the EP/EC Gibbs/Tree-AMP free energy evaluated at the optimal EP fixed point according to Proposition 2. This tree decomposition is at the heart of the modularization (Section 2.8) in the Tree-AMP package. We emphasize that this tree decomposition is exact, the approximate part in the inference comes from relaxing the full consistency of the marginals to moment-matching, which effectively projects the marginals onto exponential family approximate marginals Eq. (14) and Eq. (16) indexed by a finite-dimensional natural parameter. It is this projection of the beliefs onto finite-dimensional exponential family distributions which makes the inference tractable. Note however that the factor log-partition Eq. (20) involves an integration over the high-dimensional factor $f_k(x_k; y_k)$ and can still be a challenge to compute or approximate. See Section 2.8 for the kind of factors that the Tree-AMP package can currently handle.

2.5.3 Iterative Schemes

The fixed point Eq. (46) consists of the moment matching constraint $\mu_i = \mu_i^k$ and the natural parameter constraint $(n_i - 1)\lambda_i = \sum_{k \in \partial i} \lambda_{i \to k}$. The three optimization problems in Proposition 2 suggest different iterative schemes to reach this fixed point. As shown by (Minka, 2001b; Heskes et al., 2005), the optimization Eq. (44) of the EP free energy naturally suggests the EP algorithm, where the natural parameter constraint is enforced at each iteration. The iterative scheme suggested by the Tree-AMP free energy will be presented in Section 2.6. The direct minimization Eq. (43) of the EC Gibbs free energy enforces the moment-matching constraint at each iteration. Note that the EC Gibbs free energy is in general not convex (Opper and Winther, 2005b) because:

$$G[\mu_V] = \sum_{k \in F} \underbrace{G_k[\mu_k]}_{\text{convex}} + \sum_{i \in V} \underbrace{(1 - n_i)}_{\leq 0} \underbrace{G_i[\mu_i]}_{\text{convex}} \tag{47}$$

Message passing procedures can only aim[1] at a local minimum. According to Eq. (43), the global minimum / minimizer is expected to give the best approximation $F_\phi(\mathbf{y})$ of the surprisal / posterior moments. The algorithm is said to be in a computational hard phase if on typical instances the message passing procedure converges towards a sub-optimal minimum.

2.5.4 Connection with AMP

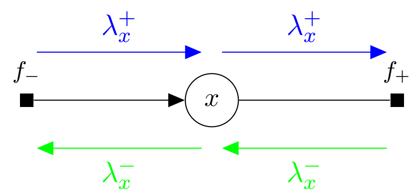
Finally we note that many AMP algorithms, such as LowRAMP (Lesieur et al., 2017) for the low-rank matrix factorization problem, GAMP (Zdeborová and Krzakala, 2016) for the GLM, or (Kabashima et al., 2016) for the extensive rank matrix factorization problem,

1. Actually the EP and Tree-AMP algorithms are not even guaranteed to converge, although damping the updates often works in practice. The double loop algorithm (Heskes and Zoeter, 2002) is guaranteed to converge but is usually very slow.

also follow a free energy formulation. The corresponding AMP free energies can be shown to be equivalent to Proposition 2 through appropriate Legendre transformations using the moment/natural parameter duality. They therefore seek the same fixed point Eq. (46) but yield another iterative schemes (the AMP algorithms) to find this fixed point. The EC Gibbs free energy is actually equal to the so-called variational Bethe energy in AMP literature (Examples 10-12). This is quite remarkable as the derivation of AMP algorithms and corresponding variational Bethe free energies follows a different path. In contrast with the approach presented here, it starts by unfolding the factor graph at the level of individual scalar components (as a result the factor graph is dense and far from being a tree) and then considers the relaxation obtained in the asymptotic limit (Zdeborová and Krzakala, 2016). The tree decomposition in Proposition 2 (at the level of the high-dimensional factors and variables) is usually recovered "after the fact".

Remark 3 (Variable neighbored by two factors) *When the variable, say x, has two factor neighbors, say f_- and f_+, the fixed point messages must satisfy:*

$$\lambda_x^+ \doteq \lambda_{f_-\to x} = \lambda_{x\to f_+}, \qquad \lambda_x^- \doteq \lambda_{f_+\to x} = \lambda_{x\to f_-}. \tag{48}$$



In many cases, for instance all the models in Figure 1 except ***(b)****, each variable has only two factor neighbors, the reparametrization Eq.* (48) *allows to reduce by half the number of messages. We use this reparametrization in the Examples 10-12 discussed below.*

Example 10 (GLM) *The Tree-AMP free energy for the GLM (Example 1), with isotropic Gaussian beliefs (Example 5), is given by:*

$$\begin{aligned} &A[a_x^\pm, b_x^\pm, a_z^\pm, b_z^\pm] \\ &= A_{p_0}[a_x^-, b_x^-] + A_W[a_x^+, b_x^+, a_z^-, b_z^-] + A_{p_{out}}[a_z^+, b_z^+] - A_x[a_x, b_x] - A_z[a_z, b_z] \end{aligned} \tag{49}$$

where $a_x = a_x^+ + a_x^-$, $b_x = b_x^+ + b_x^-$ (idem z). The factor log-partitions are given in Appendix E. The EC Gibbs free energy is given by:

$$G[r_x, \tau_x, r_z, \tau_z] = G_{p_0}[r_x, \tau_x] + G_{p_{out}}[r_z, \tau_z] - \tilde{G}_W[r_x, \tau_x, r_z, \tau_z] \tag{50}$$

with $\tilde{G}_W = G_x + G_z - G_W$ given in Appendix E.2.6 for a generic linear channel. In particular when the matrix W has iid entries (Appendix E.2.7) one recovers exactly the variational Bethe free energy of (Krzakala et al., 2014) which is shown to be equivalent to the GAMP free energy.

Example 11 (Low rank matrix factorization) *The Tree-AMP free energy for the low rank factorization (Example 2), with isotropic (Example 5) or structured (Example 8) Gaussian beliefs on U and V, and diagonal Gaussian beliefs (Example 6) on Z, is given by:*

$$\begin{aligned}&A[a_U^\pm, b_U^\pm, a_V^\pm, b_V^\pm]\\&= A_{p_U}[a_U^-, b_U^-] + A_{p_V}[a_V^-, b_V^-] + A_\delta[a_U^+, b_U^+, a_V^+, b_V^+; a_Z^-, b_Z^-] - A_U[a_U, b_U] - A_V[a_V, b_V] \quad (51)\end{aligned}$$

where $a_U = a_U^+ + a_U^-$, $b_U = b_U^+ + b_U^-$ (idem V). The factor log-partitions are given in Appendix E. In the large N limit, the output channel do not contribute to the free energy and its net effect is to send the constant messages:

$$a_Z^- = -\partial_z^2 \ln p_{out}(y \mid z)|_{z=0,y=Y}, \quad b_Z^- = \partial_z \ln p_{out}(y \mid z)|_{z=0,y=Y}, \tag{52}$$

a phenomenon known as channel universality in (Lesieur et al., 2017). The messages $a_Z^-, b_Z^- \in \mathbb{R}^{N\times M}$ are thus viewed as fixed parameters. The EC Gibbs free energy is given by:

$$G[r_U, \tau_U, r_V, \tau_V] = G_{p_U}[r_U, \tau_U] + G_{p_V}[r_V, \tau_V] - \tilde{G}_\delta[r_U, \tau_U, r_V, \tau_V; a_Z^-, b_Z^-] \tag{53}$$

where $\tilde{G}_\delta = G_U + G_V - G_\delta$ is derived in (Lesieur et al., 2017) using a Plefka-Georges-Yedidia expansion (Plefka, 1982; Georges and Yedidia, 1991), which is asymptotically exact in the large N limit:

$$\begin{aligned}&\tilde{G}_\delta[r_U, \tau_U, r_V, \tau_V; a_Z^-, b_Z^-]\\&= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M} \frac{(b_Z^-)^2_{nm}}{N} \operatorname{Tr}(\Sigma_U^n \Sigma_V^m) + 2\frac{(b_Z^-)_{nm}}{\sqrt{N}}(r_U r_V^T)_{nm} - \frac{(a_Z^-)_{nm}}{N} \operatorname{Tr} \tau_U^n \tau_V^m \quad (54)\end{aligned}$$

One recovers exactly[2] *the variational Bethe free energy of (Lesieur et al., 2017) which is shown to be equivalent to the LowRAMP free energy.*

Example 12 (Extensive rank matrix factorization) *The Tree-AMP free energy for the extensive rank factorization (Example 3), with diagonal Gaussian beliefs (Example 6), is given by:*

$$\begin{aligned}A[a_F^\pm, b_F^\pm, a_X^\pm, b_X^\pm, a_Z^\pm, b_Z^\pm,] &= A_{p_F}[a_F^-, b_F^-] + A_{p_X}[a_X^-, b_X^-] + A_\delta[a_F^+, b_F^+, a_X^+, b_X^+, a_Z^-, b_Z^-]\\&\quad - A_F[a_F, b_F] - A_X[a_X, b_X] - A_Z[a_Z, b_Z] \quad (55)\end{aligned}$$

where $a_F = a_F^+ + a_F^-$, $b_F = b_F^+ + b_F^-$ (idem X, Z). The factor log-partitions are given in Appendix E. Contrary to the low rank case, the output channel does contribute to the free energy and the messages a_Z^-, b_Z^- are no longer constant. The EC Gibbs free energy is given by:

$$G[r_F, \tau_F, r_X, \tau_X] = G_{p_F}[r_F, \tau_F] + G_{p_X}[r_X, \tau_X] + G_{p_Z}[r_Z, \tau_Z] - \tilde{G}_\delta[r_F, \tau_F, r_X, \tau_X, r_Z, \tau_Z] \tag{56}$$

2. We think there is a typo in Eq. (111) of (Lesieur et al., 2017) and that Eq. (54) is the correct expression.

One recovers exactly the variational Bethe free energy of (Kabashima et al., 2016), which is shown to be equivalent to the AMP free energy, with $\tilde{G}_\delta = G_F + G_X + G_Z - G_\delta$ given by:

$$\tilde{G}_\delta[r_F, \tau_F, r_X, \tau_X, r_Z, \tau_Z] = -\frac{1}{2}\sum_{m=1}^{M}\sum_{p=1}^{P}\frac{v_Z^{mp}}{V_{mp}} + \ln 2\pi V_{mp} + \frac{(v_F v_X)_{mp}}{N} g_{mp}^2$$

$$\textit{with} \quad g = \frac{r_Z - \frac{r_F r_X}{\sqrt{N}}}{\frac{v_F v_X}{N}}, \quad V = \frac{v_F v_X + r_F^2 v_X + v_F r_X^2}{N} = \frac{\tau_F \tau_X - r_F^2 r_X^2}{N}. \tag{57}$$

However we warn the reader that the derivation of (Kabashima et al., 2016) wrongly assumes that Z behaves as a multivariate Gaussian as pointed out by (Maillard et al., 2022). In consequence the free energy is not asymptotically exact and we expect the AMP algorithm of (Kabashima et al., 2016) to be sub-optimal. See (Maillard et al., 2022) for the corrections to the expression Eq. (57) *of $\tilde{G}_\delta$.*

2.6 Tree-AMP Implementation of Expectation Propagation

The Tree-AMP implementation of EP works with the full set of messages $(\lambda_{k\to i}, \lambda_{i\to k})_{(i,k)\in E}$. Due to the parametrization Eq. (42) and the moment functions Eqs (23) and (28), a stationary point of $A[\lambda_E]$ satisfies:

$$\partial_{\lambda_{k\to i}} A[\lambda_E] = 0 \implies \mu_i[\lambda_i^k] = \mu_i[\lambda_i] \implies \lambda_i^k = \lambda_i\,, \tag{58}$$

$$\partial_{\lambda_{i\to k}} A[\lambda_E] = 0 \implies \mu_i[\lambda_i^k] = \mu_i^k[\lambda_k] \implies \lambda_i^k = \lambda_i[\mu_i^k[\lambda_k]]\,, \tag{59}$$

which suggests the iterative procedure summarized in Algorithm 1, where E_+ denotes a topological ordering of the edges and E_- the reverse ordering.

The message-passing schedule seems the most natural: iterate over the edges in topological order (forward pass) then iterate in reverse topological order (backward pass) and repeat until convergence. In fact, if the exponential beliefs and the factors are conjugate[3] then the moment-matching is exact and Algorithm 1 is actually equivalent to exact belief propagation, where one forward pass and one backward pass yield the exact marginals (Bishop, 2006).

Algorithm 1: Generic Tree-AMP algorithm

```
initialize λ_{i→k}, λ_{k→i} = 0
repeat
    foreach edge e ∈ E_+ ∪ E_- do                  // forward and backward pass
        if e = k → i then
            λ_i^k = λ_i[μ_i^k[λ_k]]                // moment-matching
            λ_{k→i}^new = λ_i^k − λ_{i→k}          // message f_k → x_i
        if e = i → k then
            λ_i = Σ_{k'∈∂i} λ_{k'→i}               // posterior
            λ_{i→k}^new = λ_i − λ_{k→i}            // message x_i → f_k
until convergence
```

3. For example Gaussian beliefs and the factors are either linear transform or Gaussian noise, prior or likelihood

For isotropic or diagonal Gaussian beliefs (Examples 5 and 6) the moment-matching is equivalent to match the mean $r_i = r_i^k$ and variance $v_i = v_i^k$ which leads to Algorithm 2. When the variable x_i has only two neighbor factors, say f_k and f_l, the $x_i \to f_k$ update is particularly simple. The variable just passes through the corresponding messages:

$$a_{i\to k}^{\text{new}} = a_{l\to i}, \quad b_{i\to k}^{\text{new}} = b_{l\to i} \tag{60}$$

in compliance with Remark 3. Algorithm 2 can be straightforwardly extended to full covariance and structured Gaussian beliefs (Example 7 and 8) using the moment-matching update

$$a_i^k = v_i^k[a_k, b_k]^{-1}, \quad b_i^k = v_i^k[a_k, b_k]^{-1} r_i^k[a_k, b_k]$$

where v_i^k is now a covariance matrix.

Algorithm 2: Expectation propagation in Tree-AMP (Gaussian beliefs)

```
initialize a_{i→k}, b_{i→k}, a_{k→i}, b_{k→i} = 0
repeat
    foreach edge e ∈ E_+ ∪ E_- do                       // forward and backward pass
        if e = k → i then
            a_i^k = 1 / v_i^k[a_k,b_k],   b_i^k = r_i^k[a_k,b_k] / v_i^k[a_k,b_k]   // moment-matching
            a_{k→i}^new = a_i^k − a_{i→k},   b_{k→i}^new = b_i^k − b_{i→k}         // message f_k → x_i
        if e = i → k then
            a_i = Σ_{k'∈∂i} a_{k'→i},   b_i = Σ_{k'∈∂i} b_{k'→i}                    // posterior
            a_{i→k}^new = a_i − a_{k→i},   b_{i→k}^new = b_i − b_{k→i}             // message x_i → f_k
until convergence
```

2.7 MAP Estimation

The Tree-AMP Algorithm can also be used for maximum at posteriori (MAP) estimation, where the mode of the posterior Eq. (2) is the solution to the energy minimization problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} E(\mathbf{x}, \mathbf{y}), \quad E(\mathbf{x}, \mathbf{y}) = \sum_{k\in F} E_k(x_k; y_k) \tag{61}$$

where $E_k(x_k; y_k)$ is the energy associated to the factor $f_k(x_k; y_k) = e^{-E_k(x_k;y_k)}$. Conversely any optimization problem of the form Eq. (61), with a tree-structured graph of penalties/constraints, can be viewed as the MAP estimation of Eq. (2) with pseudo-factors $f_k(x_k; y_k) = e^{-E_k(x_k;y_k)}$. Following (Manoel et al., 2018) for the derivation of the TV-VAMP algorithm, that we generalize to tree-structured models in Appendix B, the MAP estimation / energy minimization can be derived by introducing an inverse temperature β in the posterior Eq. (2) and considering the zero temperature limit, as usually done in the statistical physics literature (Mézard and Montanari, 2009).

Proposition 4 (MAP estimation in Tree-AMP) *The energy minimization / MAP estimation problem can be formulated as the $\beta \to \infty$ limit of Proposition 2:*

$$\min_{\mathbf{x}} E(\mathbf{x}, \mathbf{y}) = \min \operatorname*{extr}_{a_E, b_E} -A[a_E, b_E] \tag{62}$$

with MAP variable log-partition, mean and variance:

$$A_i[a_i, b_i] = \frac{\|b_i\|^2}{2a_i}, \quad r_i[a_i, b_i] = \frac{b_i}{a_i}, \quad v_i[a_i, b_i] = \frac{1}{a_i}, \tag{63}$$

and MAP factor log-partition, mean and variance:

$$A_k[a_k, b_k] = \frac{\|b_k\|^2}{2a_k} - \mathcal{M}_{\frac{1}{a_k} E_k(\cdot\,;\, y_k)}\left(\frac{b_k}{a_k}\right), \tag{64}$$

$$r_k[a_k, b_k] = \operatorname{prox}_{\frac{1}{a_k} E_k(\cdot\,;\, y_k)}\left(\frac{b_k}{a_k}\right), \quad v_k[a_k, b_k] = \langle \partial_{b_k} r_k[a_k, b_k] \rangle, \tag{65}$$

where we introduce the Moreau envelop $\mathcal{M}_g(y) = \min_x\{g(x) + \frac{1}{2}\|x - y\|^2\}$ and the proximal operator $\operatorname{prox}_g(y) = \arg\min_x\{g(x) + \frac{1}{2}\|x - y\|^2\}$. A stationary point of $A[a_E, b_E]$ can be searched with the Tree-AMP Algorithm 2. At the optimal fixed point, the "means" $(r_i)_{i\in V}$ actually yield the MAP estimate / minimizer $\mathbf{x}^* = (x_i^*)_{i\in V}$.

The TV-VAMP algorithm is recovered as a special case (Section 2.9). See the discussion in (Manoel et al., 2018) for the close relationship to proximal methods in optimization (Parikh and Boyd, 2014), in particular the "variances" $(v_i)_{i\in V}$ can be viewed as adaptive stepsizes in the Peaceman-Rachford splitting.

2.8 Expectation Propagation Modules

In the weak consistency framework, we have the freedom to choose any kind of approximate beliefs, that is choose a set of sufficient statistics ϕ_i for each variable x_i. Each choice leads to a different approximate inference scheme, but all are implemented by the same message passing Algorithm 1. Of course the algorithm requires each relevant module to be implemented: in practice the variable x_i (resp. factor f_k) module should be able to compute the log-partition $A_i[\lambda_i]$ (resp. $A_k[\lambda_k]$) and its associated moment function $\mu_i[\lambda_i]$ (resp. $\mu_k[\lambda_k]$). Note that the definition of the module directly depends on the choice of sufficient statistics ϕ, so choosing a different kind of approximate beliefs actually leads to a distinct module. Currently the Tree-AMP package only supports isotropic Gaussian beliefs (Example 5) but we plan to include more generic Gaussian beliefs (Examples 6-8) in future versions of the package. The corresponding variable log-partition $A_i[a_i, b_i]$, mean $r_i[a_i, b_i]$ and variance $r_i[a_i, b_i]$ and factor log-partition $A_k[a_k, b_k]$, mean $r_k[a_k, b_k]$ and variance $v_k[a_k, b_k]$ are presented in Example 9. The implementation is detailed in the documentation[4]. We list below the modules considered in Tree-AMP.

2.8.1 Variable Modules

A list of approximate beliefs is presented in Appendix E.1. As shown, such beliefs can be defined over many types of variable: binary, sparse, real, constrained to an interval, or circular for instance. The associated variable modules correspond to well known exponential family distributions, including the Gauss-Bernoulli for a sparse variable. Even if the Tree-AMP package only implements isotropic Gaussian beliefs, the variable modules are useful to derive the factor modules.

4. See https://sphinxteam.github.io/tramp.docs/0.1/html/implementation.html

2.8.2 Analytical vs Approximate Factor Modules

Note that the factor log-partition Eq. (20) involves the high-dimensional factor $f_k(x_k; y_k)$ and can thus be a challenge to compute or approximate. Nonetheless, many factor modules implemented in the Tree-AMP package can be analytically derived, which means providing an explicit formula for the factor log-partition $A_k[a_k, b_k]$, mean $r_k[a_k, b_k]$ and variance $v_k[a_k, b_k]$. The following analytical modules are derived in Appendix E:

- linear channels which include the rotation channel, the discrete Fourier transform and convolutional filters as special cases;
- separable priors such as the Gaussian, binary, Gauss-Bernoulli, and positive priors;
- separable likelihoods such as the Gaussian or a deterministic likelihood like observing the sign, absolute value, modulus or phase;
- separable channels such as additive Gaussian noise or the piecewise linear activation channel.

For other modules, that we did not manage to obtain analytically, one resorts to an approximation or an algorithm to estimate the log-partition $A_k[a_k, b_k]$ and the associated mean $r_k[a_k, b_k]$ and variance $v_k[a_k, b_k]$. One such example in the Tree-AMP package is the low rank factorisation module $Z = \frac{UV^\intercal}{\sqrt{N}}$, for which we use the AMP algorithm developed in (Lesieur et al., 2017) to estimate $A_k[a_k, b_k]$, $r_k[a_k, b_k]$ and $v_k[a_k, b_k]$.

2.8.3 MAP Modules

The maximum a posteriori (MAP) modules are worth mentioning especially due to their connection to proximal methods in optimization (Parikh and Boyd, 2014). They are of course used in MAP estimation (Section 2.7) – where all modules are MAP modules – or can be used in isolation for a specific factor to approximate. Indeed, for any factor $f_k(x_f; y_k) = e^{-E_k(x_k;y_k)}$, one can use the Laplace method to obtain the MAP approximation Eqs (64)-(65) to the log-partition, mean and variance. Two such MAP modules are implemented in the Tree-AMP package for the penalties $E(x) = \lambda\|x\|_1$ and $E(x) = \lambda\|x\|_{2,1}$ associated to the ℓ_1 and $\ell_{2,1}$ norms. We recall that the $\ell_{2,1}$ norm is defined as:

$$\|x\|_{2,1} = \sum_{n=1}^{N} \|x_n\|_2 = \sum_{n=1}^{N} \sqrt{\sum_{l=1}^{d} x_{ln}^2} \quad \text{for} \quad x \in \mathbb{R}^{d\times N}. \tag{66}$$

The corresponding proximal operators are the soft thresholding and group soft thresholding operators.

2.9 Related Algorithms

We recover several algorithms as special cases of Algorithm 2. For instance the G-VAMP (Schniter et al., 2017), TV-VAMP (Manoel et al., 2018) and ML-VAMP (Fletcher et al., 2018) algorithms correspond respectively to the factor graphs Figure 1 (**a**), (**b**) and (**d**). In this subsection, we make explicit the equivalence with these algorithms and argue that the modularity of Tree-AMP allows to tackle a greater variety of inference tasks and optimization problems.

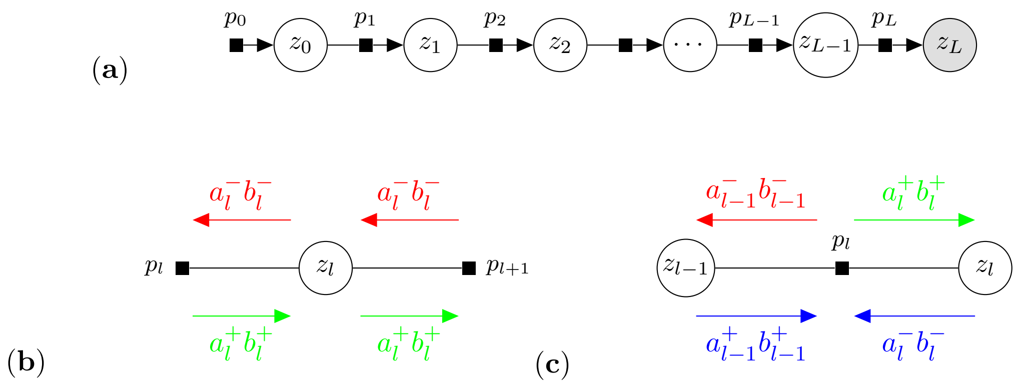


Figure 2: Message passing in ML-VAMP. (**a**) Multi-layer generalized linear model. (**b**) The variable z_l passes through the forward message $a_l^+ b_l^+$ (green) during the forward pass, and the backward message $a_l^- b_l^-$ (red) during the backward pass. (**c**) The factor p_l always takes as inputs the messages $a_{l-1}^+ b_{l-1}^+$ and $a_l^- b_l^-$ (blue). It outputs the message $a_l^+ b_l^+$ (green) during the forward pass, and the message $a_{l-1}^- b_{l-1}^-$ (red) during the backward pass.

2.9.1 Inference in Multi-Layer Networks

The ML-VAMP algorithm (Fletcher et al., 2018) performs inference in multi-layer networks such as Figure 1 (**d**). The one layer case reduces to the G-VAMP algorithm (Schniter et al., 2017) for GLM such as Figure 1 (**a**). Following Fletcher et al. (2018) let us consider the multi-layer model:

$$p(\mathbf{z}) = \prod_{l=0}^{L} p_l(z_l \mid z_{l-1}) \tag{67}$$

where $\mathbf{z} = \{z_l\}_{l=0}^{L}$ includes the measurement $y = z_L$ and the signals $\mathbf{x} = \{z_l\}_{l=0}^{L-1}$ to infer. The corresponding factor graph is displayed in Figure 2 (**a**). The model generally consists of a succession of linear channels (with possibly a bias and additive Gaussian noise) and separable non-linear activations; however, it is not necessary to specify further the architecture as all factors are treated on the same footing in both ML-VAMP and Algorithm 2.

We are interested in the isotropic Gaussian beliefs version of Algorithm 2. According to Eq. (60), the $z_l \to p_{l+1}$ update during the forward pass leads to

$$a_l^+ \doteq a_{z_l \to p_{l+1}} = a_{p_l \to z_l}, \quad b_l^+ \doteq b_{z_l \to p_{l+1}} = b_{p_l \to z_l}, \tag{68}$$

while the $z_l \to p_l$ update during the backward pass leads to

$$a_l^- \doteq a_{z_l \to p_l} = a_{p_{l+1} \to z_l}, \quad b_l^- \doteq b_{z_l \to p_l} = b_{p_{l+1} \to z_l}. \tag{69}$$

Each variable in Figure 2 (**a**) has exactly two neighbors, so each variable just passes through the corresponding messages as illustrated in Figure 2 (**b**).

The $p_l \to z_l$ update during the forward pass leads to

$$r_l = r^{p_l}_{z_l}[a^+_{l-1}, b^+_{l-1}, a^-_l, b^-_l], \quad v_l = v^{p_l}_{z_l}[a^+_{l-1}, b^+_{l-1}, a^-_l, b^-_l], \tag{70}$$

$$a_l = \frac{1}{v_l}, \quad b_l = \frac{r_l}{v_l}, \quad a^+_l = a_l - a^-_l, \quad b^+_l = b_l - b^-_l. \tag{71}$$

while the $p_l \to z_{l-1}$ update during the backward pass leads to

$$r_{l-1} = r^{p_l}_{z_{l-1}}[a^+_{l-1}, b^+_{l-1}, a^-_l, b^-_l], \quad v_{l-1} = v^{p_l}_{z_{l-1}}[a^+_{l-1}, b^+_{l-1}, a^-_l, b^-_l], \tag{72}$$

$$a_{l-1} = \frac{1}{v_{l-1}} \quad b_{l-1} = \frac{r_{l-1}}{v_{l-1}}, \quad a^-_{l-1} = a_{l-1} - a^+_{l-1}, \quad b^-_{l-1} = b_{l-1} - b^+_{l-1}. \tag{73}$$

as illustrated in Figure 2 (**c**). The ML-VAMP forward pass computes r_l and v_l using Eq. (70) and updates the message according to Eq. (71) where these quantities in (Fletcher et al., 2018) are denoted by:

$$a^\pm_l = \gamma^\pm_l, \quad b^\pm_l = \gamma^\pm_l r^\pm_l, \quad r_l = \hat{z}^+_l, \quad a_l = \eta^+_l. \tag{74}$$

Similarly the backward pass in the ML-VAMP algorithm is equivalent to Eqs (72) and (73). Finally the equivalence also holds for the prior $p_0(z_0)$ and the likelihood $p_L(y|z_{L-1})$. The prior is only used during the forward pass, it receives the backward message a^-_0, b^-_0 as input and outputs the forward message a^+_0, b^+_0. The likelihood is only used during the backward pass, it receives the forward message a^+_{L-1}, b^+_{L-1} as input and outputs the backward message a^-_{L-1}, b^-_{L-1}.

Therefore the EP Algorithm 2 with isotropic Gaussian beliefs is exactly equivalent to the ML-VAMP algorithm. It offers a direct generalization to any tree-structured model, for instance the tree network of GLMs considered in (Reeves, 2017).

2.9.2 Optimization with Non-Separable Penalties

We now turn to the TV-VAMP algorithm (Manoel et al., 2018) designed to solve optimization problem of the form:

$$x^* = \arg\min \frac{1}{2\Delta}\|y - Ax\|^2 + \lambda f(Kx). \tag{75}$$

This corresponds to the MAP estimate (Section 2.7) for the factor graph displayed in Figure 3. Of particular interest is the case $K = \nabla$ and $f(z) = \|z\|_{2,1}$ which is identical to the total variation penalty for x.

We are interested in the version of Algorithm 2 with isotropic Gaussian beliefs on all variables except x for which we consider a full covariance belief. The penalty term λf would correspond to a factor $e^{-\lambda f(z)}$ in a probabilistic setting, but here we are only considering the MAP module for which the mean and variance are given by Eq. (65):

$$r_f[a_f, b_f] = \eta_{\frac{\lambda}{a_f}}\left(\frac{b_f}{a_f}\right), \quad v_f[a_f, b_f] = \frac{1}{a_f}\left\langle \nabla\eta_{\frac{\lambda}{a_f}}\left(\frac{b_f}{a_f}\right)\right\rangle, \tag{76}$$

where we introduce the function $\eta_\lambda(x) = \mathrm{prox}_{\lambda f}(x)$ following Manoel et al. (2018).

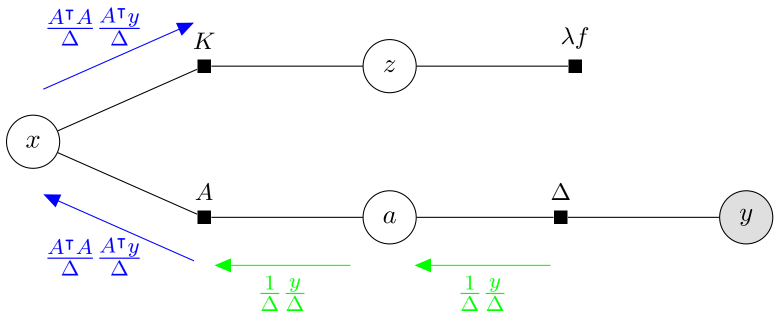


Figure 3: Message passing in TV-VAMP.

First note that each variable in Figure 3 has exactly two neighbors. According to Eq. (60) the message passing is then particularly simple: the variable just passes through the corresponding messages. The Gaussian likelihood Δ (Appendix E.4.5) leads to the messages:

$$a_{a\to A} = a_{\Delta\to a} = \frac{1}{\Delta}, \quad b_{a\to A} = b_{\Delta\to a} = \frac{y}{\Delta}, \tag{77}$$

and the linear channel A with full covariance belief on x (Appendix E.2.15) leads to the messages:

$$a_{x\to K} = a_{A\to x} = \frac{A^\intercal A}{\Delta}, \quad b_{x\to K} = b_{A\to x} = \frac{A^\intercal y}{\Delta}. \tag{78}$$

This stream of constant messages from the likelihood Δ up to factor K is displayed on Figure 3. The $K \to z$ update for the linear channel K with isotropic belief on z (Appendix E.2.1) leads to the messages:

$$r_x^K = \Sigma_x^K \left[\frac{A^\intercal y}{\Delta} + K^\intercal b_{z\to K}\right], \quad \Sigma_x^K = \left[\frac{A^\intercal A}{\Delta} + a_{z\to K} K^\intercal K\right]^{-1}, \tag{79}$$

$$r_z^K = K r_x^K, \quad v_z^K = \frac{1}{N_z} \mathrm{Tr}\left[K \Sigma_x^K K^\intercal\right], \tag{80}$$

$$a_{z\to f} = a_{K\to z} = \frac{1}{v_z^K} - a_{z\to K}, \quad b_{z\to f} = b_{K\to z} = \frac{r_z^K}{v_z^K} - b_{z\to K}, \tag{81}$$

and the $f \to z$ update for the MAP module λf leads to the messages:

$$r_z^f = \eta_{\frac{\lambda}{a_{z\to f}}}\left(\frac{b_{z\to f}}{a_{z\to f}}\right), \quad v_z^f = \frac{1}{a_{z\to f}}\left\langle \nabla\eta_{\frac{\lambda}{a_{z\to f}}}\left(\frac{b_{z\to f}}{a_{z\to f}}\right)\right\rangle, \tag{82}$$

$$a_{z\to K} = a_{f\to z} = \frac{1}{v_z^f} - a_{z\to f}, \quad b_{z\to K} = b_{f\to z} = \frac{r_z^f}{v_z^f} - b_{z\to f}. \tag{83}$$

In (Manoel et al., 2018) the following quantities at iteration t are denoted by:

$$a_{z\to K} = \rho^t, \quad b_{z\to K} = u^t, \quad r_x^K = x^t, \quad v_z^K = \sigma_x^t, \quad r_z^f = z^t, \quad v_z^f = \sigma_z^t \tag{84}$$

Then Eqs (79), (82) and (83) are exactly equivalent to the Eqs (24), (25) and (26) of (Manoel et al., 2018) defining the TV-VAMP algorithm.

As discussed in greater detail by Manoel et al. (2018), the TV-VAMP algorithm is closely related to proximal methods: it can be viewed as the Peaceman-Rachford splitting where the step-size ρ^t is set adaptively. Then Algorithm 2 offers a generalization to any optimization problem for which the factor graph of penalty/constraints is tree-structured (Section 2.7). For instance, while the TV-VAMP can only solve linear regression with a TV penalty, Algorithm 2 can be easily applied to a classification setting: one just needs to replace the Gaussian likelihood Δ by the appropriate likelihood. Also Algorithm 2 offers more flexibility in designing the approximate inference scheme: for example one can choose isotropic or diagonal Gaussian belief for x to alleviate the computational burden of inverting a matrix in Eq. (79).

2.10 EP, EC, AdaTAP and Message-Passing

There is a long history behind the methods used in this section and the literature on statistical physics. In particular, broadening the class of matrices amenable to mean-field treatments was the motivation behind a decades long series of works.

Parisi and Potters (1995) were among the pioneers in this direction by deriving mean-field equations for orthogonal matrices. The adaTAP approach of Csató et al. (2001), and their reinterpretation as a particular case of the expectation propagation algorithm (Minka, 2001a) allowed for a generic reinterpretation of these ideas as an approximation of the log partition named expectation consistency (EC) (Heskes et al., 2005; Opper and Winther, 2005a,b). Many works then applied these ideas to problems such as the perceptron (Shinzato and Kabashima, 2008b,a; Kabashima, 2003).

All these ideas were behind the recent renewal of interest of message-passing algorithms with generic rotationally invariant matrices (Schniter et al., 2017; Ma and Ping, 2017; Çakmak et al., 2014, 2016). In a recent work, Maillard et al. (2019) showed the consistency and the equivalence of these approaches.

3. State Evolution and Free Entropy

In this section we present an heuristic derivation of the free entropy and state evolution formalisms for the tree-structured models considered in Section 2.1. There is now a very vast literature on the state evolution and free entropy formalisms applied to machine learning models (Zdeborová and Krzakala, 2016) so an exhaustive review is beyond the scope of this manuscript, however see Examples 13 and 14 for some representative prior works. Our primary goal in this section is to tie these results together in an unifying framework, extend them to tree-structured factor graphs and justify the modularization of the free entropy and state evolution as done in the Tree-AMP package.

We first heuristically derive the so-called *replica free entropy* (Mézard and Montanari, 2009) using weak consistency (Heskes et al., 2005) on the overlaps. We expect our formulas to be valid only when the overlaps are the relevant order parameters needed to describe the ensemble average. The replica symmetric solution is exposed in Section 3.3 and we present more specifically the Bayes-optimal setting in Section 3.4. The solution is easily interpreted as local ensemble averages defined for each factor and variable and allows us to conjecture the state evolution of the EP Algorithm 2. This effective ensemble average is at the heart of the modularization of the free entropy and state evolution formalisms in the Tree-AMP package. We express the free entropy potentials using information theoretic quantities in Section 3.5 and recover (Reeves, 2017) formalism as a special case. Finally we briefly list the state evolution modules currently implemented in the Tree-AMP package. We emphasize that the derivation is non-rigorous and largely conjectural, however we recover many replica free entropies and state evolutions previously derived for specific models, that are conjectured to be exact or even rigorously proven in some cases (Examples 13 and 14).

Example 13 (Multi-layer and tree network of GLMs) *A very general setting is the tree network of GLMs proposed by Reeves (2017), which includes the GLM and multi-layer network as special cases. A key assumption in such models is that the weight matrices in linear channels are drawn from an orthogonally invariant ensemble. The state evolution is rigorously proven in the multi-layer case (Fletcher et al., 2018) for the corresponding ML-VAMP algorithm. The replica free entropy for the multi-layer case is derived in (Gabrié et al., 2018). When the entries of the weight matrices are iid Gaussian (a special case of an orthogonally-invariant ensemble) the replica free entropy was further be shown to be rigorous in the compressed sensing (Reeves and Pfister, 2016) and GLM (Barbier et al., 2019) cases.*

Example 14 (Low rank matrix factorization) *The replica free entropy for the low-rank matrix factorization problem (Example 2) and the state evolution of the LowRAMP algorithm are derived in (Lesieur et al., 2017). The replica free entropy was further shown to be rigorous in (Miolane, 2017; Lelarge and Miolane, 2019). Stacking a multi-layer GLM with a low rank factorization model was considered in (Aubin et al., 2019).*

3.1 Model Settings

In this subsection we first present the teacher-student scenario and its high-dimensional limit. Then we give a brief introduction to the replica free entropy computation. In particular we define the overlaps, that are here assumed to be the relevant order parameters to describe the ensemble average.

3.1.1 Teacher-Student Scenario

We will consider a generic teacher-student scenario where the teacher generates the signals $\mathbf{x}^{(0)}$ and measurements $\mathbf{y}$. The student is only given the measurements $\mathbf{y}$ and must infer the signals. The teacher generative model is a tree-structured factor graph:

$$p^{(0)}(\mathbf{x}^{(0)}, \mathbf{y}) = \frac{1}{Z_N^{(0)}} \prod_{k \in F} f_k^{(0)}(x_k^{(0)}; y_k), \quad Z_N^{(0)} = \int d\mathbf{x}^{(0)} d\mathbf{y} \prod_{k \in F} f_k^{(0)}(x_k^{(0)}; y_k), \tag{85}$$

where $\mathbf{x}^{(0)} = (x_i^{(0)})_{i \in V}$ are the (ground truth) signals and $\mathbf{y} = (y_j)_{j \in O}$ the measurements, see Section 2.1 for more information on the factor graph notation. For the student, we will assume the same tree factorization as the teacher, however the student factors $f_k(x_k; y_k)$ can be mismatched. The student generative model is given by Eq. (1) and the student posterior by Eq. (2). The student posterior mean, variance and second moment are given by:

$$r_i(\mathbf{y}) = \mathbb{E}_{p(\mathbf{x}|\mathbf{y})} x_i, \quad v_i(\mathbf{y}) = \langle \mathrm{Var}_{p(\mathbf{x}|\mathbf{y})} x_i \rangle, \quad \tau_i(\mathbf{y}) = \mathbb{E}_{p(\mathbf{x}|\mathbf{y})} \frac{\|x_i\|^2}{N_i} = \frac{\|r_i(\mathbf{y}))\|^2}{N_i} + v_i(\mathbf{y}). \tag{86}$$

The *mismatched setting* corresponds to the case where at least one of the student factors is mismatched. On the opposite, the *Bayes-optimal setting* refers to the case where all the student factors match the teacher factors $f_k(x_k; y_k) = f_k^{(0)}(x_k; y_k)$, consequently the student and teacher generative models are identical $p(\mathbf{x}, \mathbf{y}) = p^{(0)}(\mathbf{x}, \mathbf{y})$ and the student posterior Eq. (2) is indeed Bayes-optimal, in particular the posterior mean $r_i(\mathbf{y})$ is the minimal mean-squared-error (MMSE) estimator and the posterior variance $v_i(\mathbf{y})$ the MMSE.

3.1.2 High-Dimensional Limit

We will consider the high-dimensional limit $N \to \infty$ where each signal $x_i \in \mathbb{R}^{N_i}$ is itself a high-dimensional object with scaling $\alpha_i = \frac{N_i}{N} = O(1)$. We will denote by N_k the dimension of the factor f_k (for instance we can choose the dimension of its inputs signals by convention) and $\alpha_k = \frac{N_k}{N} = O(1)$ the corresponding scaling. Finally we will denote $\alpha_i^k = \frac{N_i}{N_k} = O(1)$ the scaling of the variable x_i wrt the factor f_k. In the large N limit we expect the log-partition to self-average:

$$A_N(\mathbf{y}) = \frac{1}{N} \ln Z_N(\mathbf{y}) \simeq \bar{A} = \lim_{N\to\infty} \mathbb{E}_{p^{(0)}(\mathbf{y})} A_N(\mathbf{y}) \tag{87}$$

where the log-partition $A_N(\mathbf{y})$ is scaled by N in order to be $O(1)$. The ensemble average $\bar{A}$ is called the *free entropy* and gives the cross-entropy up to a constant:

$$-\bar{A} = \lim_{N\to\infty} \frac{1}{N} H[p^{(0)}(\mathbf{y}), p(\mathbf{y})] - A_N \tag{88}$$

where $A_N = \frac{1}{N} \ln Z_N$ is the scaled log-partition associated to the student generative model Eq. (1) and $H[p, q] = -\mathbb{E}_p \ln q$ denotes the differential cross-entropy. In particular when the model is a Bayesian network (Remark 1) $Z_N = 1$ and $A_N = 0$ so the free entropy directly gives the cross-entropy. The goal of the free entropy formalism is to provide an analytical expression for $\bar{A}$ and describe the limiting ensemble average.

3.1.3 Replica Free Entropy

The replica trick (Mézard et al., 1987) can be viewed as an heuristic method to compute

$$A(n) = \lim_{N\to\infty} \frac{1}{N} \ln \mathbb{E}_{p^{(0)}(\mathbf{y})} Z_N(\mathbf{y})^n = \lim_{N\to\infty} \frac{1}{N} \ln \mathbb{E}_{p^{(0)}(\mathbf{y})} e^{NnA_N(\mathbf{y})} \tag{89}$$

which is interpreted in large deviation theory (Touchette, 2009) as the scaled cumulant generating function (SCGF) of the log-partition $A_N(\mathbf{y}) = \frac{1}{N} \ln Z_N(\mathbf{y})$. If the SCGF is well defined we can get the ensemble average log-partition as:

$$\bar{A} = \frac{d}{dn} A(n)\Big|_{n=0} \tag{90}$$

We can formally decompose Eq. (89) at finite N, before talking the $N \to \infty$ limit:

$$A(n) = A_N^{(n)} - A_N^{(0)}, \quad A_N^{(n)} = \frac{1}{N} \ln Z_N^{(n)}, \quad A_N^{(0)} = \frac{1}{N} \ln Z_N^{(0)}, \tag{91}$$

where $Z_N^{(0)}$ is the partition function introduced in Eq. (85) and $Z_N^{(n)}$ the partition function of the replicated system:

$$p^{(n)}(\{\mathbf{x}^{(a)}\}_{a=0}^{n}, \mathbf{y}) = \frac{1}{Z_N^{(n)}} \prod_{k \in F} \left\{ f_k^{(0)}(x_k^{(0)}; y_f) \prod_{a=1}^{n} f_k(x_k^{(a)}; y_k) \right\} \tag{92}$$

where $\mathbf{x}^{(a)}$ for $a = 1 \cdots n$ denote the n replicas and $\mathbf{x}^{(0)}$ the ground truth. The replica free entropy is obtained by computing $Z_N^{(n)}$ as if $n \in \mathbb{N}$ in Eq. (92), but then letting $n \to 0$ in Eq. (90) as if n was real (Mézard and Montanari, 2009). The replica method is non-rigorous, however it has been successfully applied for decades and given numerous given exact results, some of which were later confirmed by rigorous methods. There is therefore a very high level of trust in the replica method by the statistical physics community.

3.1.4 Overlaps

In the statistical physics literature, the system is said to have a well defined thermodynamic limit $N \to \infty$ if the limiting ensemble average is fully characterized by a few scalar parameters (called order parameters). Here we will restrict our analysis to systems where these order parameters are the overlaps:

$$\phi(\mathbf{x}) = \left(\frac{x_i^{(a)} \cdot x_i^{(b)}}{N_i} \right)_{i \in V, 0 \le a \le b \le n} \tag{93}$$

which due to the definition of the replicated system Eq. (92) correspond to:

$$\tau_i^{(0)} = \mathbb{E}_{p^{(n)}} \frac{\|x_i^{(0)}\|^2}{N_i} = \mathbb{E}_{p^{(0)}(\mathbf{x}^{(0)})} \frac{\|x_i^{(0)}\|^2}{N_i} \tag{94}$$

$$\tau_i = \mathbb{E}_{p^{(n)}} \frac{\|x_i^{(a)}\|^2}{N_i} = \mathbb{E}_{p^{(0)}(\mathbf{y})} \tau_i(\mathbf{y}) \qquad \text{for all } 1 \le a \le n \tag{95}$$

$$m_i = \mathbb{E}_{p^{(n)}} \frac{x_i^{(a)} \cdot x_i^{(0)}}{N_i} = \mathbb{E}_{p^{(0)}(\mathbf{x}^{(0)}, \mathbf{y})} \frac{r_i(\mathbf{y}) \cdot x_i^{(0)}}{N_i} \qquad \text{for all } 1 \le a \le n \tag{96}$$

$$q_i^{(ab)} = \mathbb{E}_{p^{(n)}} \frac{x_i^{(a)} \cdot x_i^{(b)}}{N_i} \qquad \text{for all } 1 \le a \neq b \le n \tag{97}$$

In the ensemble average, m_i denotes the overlap with the ground truth, $\tau_i^{(0)}$ the teacher prior second moment, and τ_i is the student posterior second moment. It is difficult to tell under which conditions (on the high-dimensional factors and their arrangement in a tree graph) the ensemble average will be fully characterized by the overlaps, and we hope that future theoretical work could clarify this point. Examples 13 and 14 show however a few representative models and conditions. For instance in GLMs and network of GLMs the weight matrices in the linear channels must come from an orthogonally invariant ensemble. The core of our heuristic derivation of the replica free entropy is to assume weak consistency on the overlaps Eq. (93). We note that this weak consistency derivation could be extended by adapting the sufficient statistics Eq. (93) to include other order parameters, if those turn out to be relevant to describe the ensemble average.

3.1.5 Replica Symmetry

The system is said to be *replica symmetric* if the overlap $q_i^{(ab)}$ between two replicas concentrates to a single value q_i which is then equal to

$$q_i = \mathbb{E}_{p^{(0)}(\mathbf{y})} \frac{\|r_i(\mathbf{y}))\|^2}{N_i}. \tag{98}$$

In the Bayes-optimal setting, the system will always be replica symmetric (Nishimori, 2001; Mézard and Montanari, 2009). In the mismatched setting, we expect the system to sometimes exhibit *replica symmetry breaking*, where the overlap $q_i^{(ab)}$ between two replicas converges instead to a discrete distribution:

$$P(q_i) = \sum_{r=0}^{R} \pi_r \delta(q_i - q_i^r), \quad q_i^0 \leq \ldots \leq q_i^R. \tag{99}$$

This situation is called R-level symmetry breaking and the cumulative $s_r = \sum_{r'=0}^{r} \pi_{r'}$ which satisfy $0 \leq s_0 \leq \ldots \leq s_R = 1$ are called the Parisi parameters (Mézard and Montanari, 2009). The system can also undergo full replica symmetry breaking where the overlap distribution has a continuous part. In this manuscript, we focus on the replica symmetric solution, the replica symmetry breaking solution is deferred to a forthcoming publication.

3.2 Teacher Prior Second Moments

The replica symmetric solution requires the teacher prior second moments $\tau_V = (\tau_i^{(0)})_{i \in V}$ which we explain how to compute in this section. The weak consistency approximation of the log-partition $A_N^{(0)} = \frac{1}{N} Z_N^{(0)}$ using the sufficient statistics

$$\phi_i(x_i^{(0)}) = -\tfrac{1}{2}\|x_i^{(0)}\|^2 \quad \text{for all } i \in V \tag{100}$$

is summarized by Proposition 5. It yields an approximation for both the log-partition $A_N^{(0)}$ and second-moments $\tau_V^{(0)}$ that we assume to be asymptotically exact in the large N limit. The full derivation follows exactly the same steps as Section 2 for the student posterior Eq. (2) but here applied to the teacher prior

$$p^{(0)}(\mathbf{x}^{(0)}) = \frac{1}{Z_N^{(0)}} \prod_{k \in F} f_k^{(0)}(x_k^{(0)}) \quad \text{with} \quad f_k^{(0)}(x_k^{(0)}) = \int dy_k \, f_k^{(0)}(x_k^{(0)}; y_k) \tag{101}$$

obtained by marginalizing Eq. (85) over $\mathbf{y}$.

Proposition 5 (Weak consistency derivation of $A_N^{(0)}$) *Solving the relaxed Bethe variational problem, using the sufficient statistics Eq.* (100)*, leads to:*

$$-A_N^{(0)} = \min_{\tau_V^{(0)}} G^{(0)}[\tau_V^{(0)}] = \min_{\hat{\tau}_E^{(0)}} \operatorname{extr} -A^{(0)}[\hat{\tau}_E^{(0)}]. \tag{102}$$

where the minimizer corresponds to the teacher prior second moments $\tau_V^{(0)} = (\tau_i^{(0)})_{i\in V}$ *and* $\hat{\tau}_E^{(0)} = (\hat{\tau}_{i\to k}^{(0)}, \hat{\tau}_{k\to i}^{(0)})_{(i,k)\in E}$ *denotes the dual natural parameter messages. The potentials satisfy the tree decomposition:*

$$G^{(0)}[\tau_V^{(0)}] = \sum_{k\in F} \alpha_k G_k^{(0)}[\tau_k^{(0)}] + \sum_{i\in V} \alpha_i(1-n_i) G_i^{(0)}[\tau_i^{(0)}] \quad \text{with} \quad \tau_k^{(0)} = (\tau_i^{(0)})_{i\in\partial k} \tag{103}$$

$$A^{(0)}[\hat{\tau}_E^{(0)}] = \sum_{k\in F} \alpha_k A_k^{(0)}[\hat{\tau}_k^{(0)}] - \sum_{(i,k)\in E} \alpha_i A_i^{(0)}[\hat{\tau}_i^{k(0)}] + \sum_{i\in V} \alpha_i A_i^{(0)}[\hat{\tau}_i^{(0)}]$$

$$\text{with} \quad \hat{\tau}_k^{(0)} = (\hat{\tau}_{i\to k}^{(0)})_{i\in\partial k}, \quad \hat{\tau}_i^{k(0)} = \hat{\tau}_{i\to k}^{(0)} + \hat{\tau}_{k\to i}^{(0)}, \quad \hat{\tau}_i^{(0)} = \sum_{k\in\partial i} \hat{\tau}_{k\to i}^{(0)}. \tag{104}$$

The scaled factor and variable log-partitions are given by:

$$A_k^{(0)}[\hat{\tau}_k^{(0)}] = \frac{1}{N_k} \ln \int dy_k dx_k^{(0)}\, f_k^{(0)}(x_k^{(0)}; y_k)\, e^{-\frac{1}{2}\hat{\tau}_k^{(0)} \|x_k^{(0)}\|^2} \tag{105}$$

$$A_i^{(0)}[\hat{\tau}_i^{(0)}] = \frac{1}{N_i} \ln \int dx_i^{(0)}\, e^{-\frac{1}{2}\hat{\tau}_i^{(0)} \|x_i^{(0)}\|^2} = \frac{1}{2} \ln \frac{2\pi}{\hat{\tau}_i^{(0)}} \tag{106}$$

$A_k^{(0)} = \frac{1}{N_k} \ln Z_k^{(0)}$ *is the log-partition of the exponential family distribution that approximates the teacher factor marginal* $p^{(0)}(x_k^{(0)}, y_k)$

$$p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)}) = \frac{1}{Z_k^{(0)}[\hat{\tau}_k^{(0)}]} f_k^{(0)}(x_k^{(0)}; y_k)\, e^{-\frac{1}{2}\hat{\tau}_k^{(0)} \|x_k^{(0)}\|^2}. \tag{107}$$

Similarly $A_i^{(0)} = \frac{1}{N_i} \ln Z_i^{(0)}$ *is the log-partition of the zero-mean Normal that approximates the teacher variable marginal* $p^{(0)}(x_i^{(0)})$

$$p_i^{(0)}(x_i^{(0)} \mid \hat{\tau}_i^{(0)}) = \frac{1}{Z_i^{(0)}[\hat{\tau}_i^{(0)}]} e^{-\frac{1}{2}\hat{\tau}_i^{(0)} \|x_i^{(0)}\|^2} = \mathcal{N}(x_i^{(0)} \mid 0, \tau_i^{(0)}) \quad \text{with} \quad \tau_i^{(0)} = \frac{1}{\hat{\tau}_i^{(0)}}. \tag{108}$$

The gradients of the factor and variable log-partitions give the dual mapping to the second moments:

$$\tau_i^{k(0)}[\hat{\tau}_k^{(0)}] = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)})} \frac{\|x_i^{(0)}\|^2}{N_i}, \qquad -\frac{1}{2}\alpha_i^k \tau_i^{k(0)} = \partial_{\hat{\tau}_{i\to k}^{(0)}} A_k^{(0)}, \tag{109}$$

$$\tau_i^{(0)}[\hat{\tau}_i^{(0)}] = \mathbb{E}_{p_i^{(0)}(x_i^{(0)} \mid \hat{\tau}_i^{(0)})} \frac{\|x_i^{(0)}\|^2}{N_i} = \frac{1}{\hat{\tau}_i^{(0)}}, \qquad -\frac{1}{2}\tau_i^{(0)} = \partial_{\hat{\tau}_i^{(0)}} A_i^{(0)}. \tag{110}$$

$G_k^{(0)}$ *and* $G_i^{(0)}$ *are the corresponding Legendre transforms. Any stationary point of the potentials (not necessarily the global optima) is a fixed point:*

$$\tau_i^{k(0)} = \tau_i^{(0)}, \quad \hat{\tau}_i^{(0)} = \sum_{k\in\partial i} \hat{\tau}_{k\to i}^{(0)}. \tag{111}$$

3.2.1 Bayesian Network Teacher

When the teacher factor graph is a Bayesian network (Remark 1) the fixed point in Proposition 5 is particularly simple. If x_i is an input signal of the factor f_k:

$$\hat{\tau}_{i\to k}^{(0)} = \hat{\tau}_i^{(0)} = \frac{1}{\tau_i^{(0)}}, \qquad \hat{\tau}_{k\to i}^{(0)} = 0. \tag{112}$$

If x_i is an output signal of the factor f_k:

$$\hat{\tau}_{k\to i}^{(0)} = \hat{\tau}_i^{(0)} = \frac{1}{\tau_i^{(0)}}, \qquad \hat{\tau}_{i\to k}^{(0)} = 0. \tag{113}$$

In other words the forward messages are equal to the precisions while the backward messages are null. The factor log-partition is equal to:

$$A_k^{(0)}[\hat{\tau}_k^{(0)}] = \sum_{i\in(\partial k)^-} \alpha_i^k A_i^{(0)}[\hat{\tau}_i^{(0)}] \tag{114}$$

where $(\partial k)^-$ denotes the input signals nodes of the factor node k. In particular Proposition 5 gives consistently:

$$A_N^{(0)} = \sum_{k\in F} \alpha_k A_k^{(0)}[\hat{\tau}_k^{(0)}] + \sum_{i\in V} \alpha_i(1-n_i) A_i^{(0)}[\hat{\tau}_i^{(0)}] = 0 \tag{115}$$

as it should, because $Z_N^{(0)} = 1$ and $A_N^{(0)} = 0$ for a Bayesian network.

3.2.2 Computing the Fixed Point

In practice one can use the Tree-AMP Algorithm 3 to find the fixed point, which will compute the second-moments $\tau_V^{(0)}$ as well as the messages $\hat{\tau}_E^{(0)}$. Note that in the usual state evolution algorithms, for example in the multi-layer GLM case (Gabrié et al., 2018; Fletcher et al., 2018), a first step is always to compute the teacher prior second moments, often invoking the central limit theorem and using approximate isotropic Gaussian distributions along the way. These routines are exactly equivalent to Algorithm 3, which indeed yields the fixed point in a single forward pass when the teacher factor graph is a Bayesian network. By contrast in factor graphs that are not Bayesian network the Tree-AMP Algorithm 3 will be useful to find the fixed point which is no longer trivial and to compute the normalization constant $A_N^{(0)}$.

Algorithm 3: Tree-AMP algorithm for the teacher prior second moments

repeat
  foreach *edge* $e \in E_+ \cup E_-$ **do** // forward and backward pass
    if $e = k \to i$ **then**
      $\hat{\tau}_i^{k(0)} = 1/\tau_i^{k(0)}[\hat{\tau}_k^{(0)}]$
      $\hat{\tau}_{k\to i}^{(0)\text{new}} = \hat{\tau}_i^{k(0)} - \hat{\tau}_{i\to k}^{(0)}$ // message $f_k \to x_i$
    if $e = i \to k$ **then**
      $\hat{\tau}_i^{(0)} = \sum_{k'\in\partial i} \hat{\tau}_{k'\to i}^{(0)}$
      $\hat{\tau}_{i\to k}^{(0)\text{new}} = \hat{\tau}_i^{(0)} - \hat{\tau}_{k\to i}^{(0)}$ // message $x_i \to f_k$

until *convergence*

3.3 Replica Symmetric Free Entropy

The replica symmetric free entropy is derived by assuming weak consistency on the overlaps Eq. (93). The full derivation is presented in Appendix C and summarized in Proposition 6. We assume that the teacher prior second moments $\tau_V^{(0)}$ and dual messages $\hat{\tau}_E^{(0)}$ are known thanks to Proposition 5 and should be now considered as fixed parameters.

Proposition 6 (Replica symmetric $\bar{A}$) *The replica symmetric (RS) $\bar{A}$ is given by:*

$$-\bar{A} = \min_{m_V,q_V,\tau_V} \bar{A}^*[m_V, q_V, \tau_V] = \min \operatorname*{extr}_{\hat{m}_E,\hat{q}_E,\hat{\tau}_E} -\bar{A}[\hat{m}_E, \hat{q}_E, \hat{\tau}_E]. \tag{116}$$

where the minimizer corresponds to the overlaps $m_V = (m_i)_{i\in V}$ and $\hat{m}_E = (\hat{m}_{i\to k}, \hat{m}_{k\to i})_{(i,k)\in E}$ denotes the dual natural parameter messages (idem q, τ). The RS potentials satisfy the tree decomposition:

$$\bar{A}^*[m_V, q_V, \tau_V] = \sum_{k\in F} \alpha_k \bar{A}_k^*[m_k, q_k, \tau_k] + \sum_{i\in V} \alpha_i (1-n_i) \bar{A}_i^*[m_i, q_i, \tau_i]$$
$$\text{with} \quad m_k = (m_i)_{i\in\partial k} \quad (\text{idem } q, \tau), \tag{117}$$
$$\bar{A}[\hat{m}_E, \hat{q}_E, \hat{\tau}_E] = \sum_{k\in F} \alpha_k \bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] - \sum_{(i,k)\in E} \alpha_i \bar{A}_i[\hat{m}_i^k, \hat{q}_i^k, \hat{\tau}_i^k] + \sum_{i\in V} \alpha_i \bar{A}_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i]$$
$$\text{with} \quad \hat{m}_k = (\hat{m}_{i\to k})_{i\in\partial k}, \quad \hat{m}_i^k = \hat{m}_{i\to k} + \hat{m}_{k\to i}, \quad \hat{m}_i = \sum_{k\in\partial i} \hat{m}_{k\to i} \quad (\text{idem } \hat{q}, \hat{\tau}). \tag{118}$$

The factor and variable RS potentials are given by:

$$\bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] = \lim_{N_k\to\infty} \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} A_k[a_k, b_k; y_k] \tag{119}$$
$$\bar{A}_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \lim_{N_i\to\infty} \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} A_i[a_i, b_i] \tag{120}$$

where $A_k[a_k, b_k; y_k]$ and $A_i[a_i, b_i]$ are the scaled student EP log-partitions with isotropic Gaussian beliefs (Example 9):

$$A_k[a_k, b_k; y_k] = \frac{1}{N_k} \ln \int dx_k\, f(x_k; y_k)\, e^{-\frac{1}{2}a_k\|x_k\|^2 + b_k^\intercal x_k} \tag{121}$$
$$A_i[a_i, b_i] = \frac{1}{N_i} \ln \int dx_i\, e^{-\frac{1}{2}a_i\|x_i\|^2 + b_i^\intercal x_i} \tag{122}$$

and the factor and variable ensemble averages are taken with:

$$p_k^{(0)}(x_k^{(0)}, y_k, b_k) = \mathcal{N}(b_k \mid \hat{m}_k x_k^{(0)}, \hat{q}_k)\, p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)}) \qquad \text{and} \quad a_k = \hat{\tau}_k + \hat{q}_k, \tag{123}$$

$$p_i^{(0)}(x_i^{(0)}, b_i) = \mathcal{N}(b_i \mid \hat{m}_i x_i^{(0)}, \hat{q}_i)\, p_i^{(0)}(x_i^{(0)} \mid \hat{\tau}_i^{(0)}) \qquad \text{and} \quad a_i = \hat{\tau}_i + \hat{q}_i, \tag{124}$$

where $p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)})$ *and* $p_i^{(0)}(x_i^{(0)} \mid \hat{\tau}_i^{(0)})$ *are the approximate teacher marginals defined in Eqs* (107) *and* (108). *The gradient of the factor RS potential give the dual mapping to the overlaps:*

$$m_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} \frac{r_i^k[a_k, b_k; y_k] \cdot x_i^{(0)}}{N_i}, \qquad \alpha_i^k m_i^k = \partial_{\hat{m}_{i\to k}} \bar{A}_k, \tag{125}$$

$$q_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} \frac{\|r_i^k[a_k, b_k; y_k]\|^2}{N_i}, \qquad -\tfrac{1}{2}\alpha_i^k q_i^k = \partial_{\hat{q}_{i\to k}} \bar{A}_k, \tag{126}$$

$$\tau_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} \frac{\|r_i^k[a_k, b_k; y_k]\|^2}{N_i} + v_i^k[a_k, b_k; y_k], \quad -\tfrac{1}{2}\alpha_i^k \tau_i^k = \partial_{\hat{\tau}_{i\to k}} \bar{A}_k, \tag{127}$$

where $r_i^k[a_k, b_k; y_k]$ *and* $v_i^k[a_k, b_k; y_k]$ *are the posterior mean and isotropic variance Eqs* (37) *and* (38) *as estimated by the student EP factor marginal Eq.* (34). *The gradient of the variable RS potential give the dual mapping to the overlaps:*

$$m_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} \frac{r_i[a_i, b_i] \cdot x_i^{(0)}}{N_i}, \qquad m_i = \partial_{\hat{m}_i} \bar{A}_i, \tag{128}$$

$$q_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} \frac{\|r_i[a_i, b_i]\|^2}{N_i}, \qquad -\tfrac{1}{2} q_i = \partial_{\hat{q}_i} \bar{A}_i, \tag{129}$$

$$\tau_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} \frac{\|r_i[a_i, b_i]\|^2}{N_i} + v_i[a_i, b_i], \qquad -\tfrac{1}{2} \tau_i = \partial_{\hat{\tau}_i} \bar{A}_i, \tag{130}$$

where $r_i[a_i, b_i] = \frac{b_i}{a_i}$ *and* $v_i[a_i, b_i] = \frac{1}{a_i}$ *are the posterior mean and isotropic variance as estimated by the student EP variable marginal Eq.* (30). $\bar{A}_k^*$ *and* $\bar{A}_i^*$ *are the corresponding Legendre transforms. The ensemble average variances are given by:*

$$v_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} v_i^k[a_k, b_k; y_k] = \tau_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] - q_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k], \tag{131}$$

$$v_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} v_i[a_i, b_i] = \tau_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] - q_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i]. \tag{132}$$

Any stationary point of the potentials (not necessarily the global optima) is a fixed point:

$$m_i^k = m_i, \quad \hat{m}_i = \sum_{k \in \partial i} \hat{m}_{k\to i}, \quad (\textit{idem } q, \tau) \tag{133}$$

3.3.1 RS Potentials

The variable RS potential Eq. (120) is explicitly given by:

$$\bar{A}_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \frac{\hat{m}_i^2 \tau_i^{(0)} + \hat{q}_i}{2a_i} + \frac{1}{2}\ln\frac{2\pi}{a_i} \quad \text{with} \quad a_i = \hat{\tau}_i + \hat{q}_i, \tag{134}$$

which yields the dual mapping:

$$m_i = \frac{\hat{m}_i \tau_i^{(0)}}{a_i}, \quad q_i = \frac{\hat{m}_i^2 \tau_i^{(0)} + \hat{q}_i}{a_i^2}, \quad \tau_i = q_i + v_i \quad \text{with} \quad v_i = \frac{1}{a_i}. \tag{135}$$

Several factor RS potentials are given in Appendix E.

3.3.2 Cross-Entropy Estimation and State Evolution

The ensemble average $\bar{A}$ which gives access to the cross-entropy through Eq. (88) is given by the global minimum according to Proposition 6. The global minimizer gives access to the ensemble average overlaps m_i, q_i, τ_i (as well as $\text{mse}_i = \tau_i^{(0)} - 2m_i + q_i$ and $v_i = \tau_i - q_i$) for the student posterior Eq. (2). However the replica free entropy solution also appears as an ensemble average of the underlying EP Algorithm 2 where the effective ensemble average is defined locally for each factor Eq. (123) and variable Eq. (124). With this effective ensemble average interpretation in mind, we conjecture Algorithm 4 to give the state evolution of the EP Algorithm 2. Then the state evolution fixed point will give the overlaps m_i, q_i, τ_i (as well as $\text{mse}_i = \tau_i^{(0)} - 2m_i + q_i$ and $v_i = \tau_i - q_i$) corresponding to the student EP solution, which is in general only a local minimizer in Proposition 6.

Algorithm 4: Tree-AMP State evolution (replica symmetric mismatched setting)

initialize $\hat{m}_{i\to k}, \hat{q}_{i\to k}, \hat{\tau}_{i\to k}, \hat{m}_{k\to i}, \hat{q}_{k\to i}, \hat{\tau}_{k\to i} = 0$

repeat

  foreach *edge* $e \in E_+ \cup E_-$ **do** `// forward and backward pass`

    if $e = k \to i$ **then**

      $v_i^k = v_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k], \quad m_i^k = m_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k], \quad q_i^k = q_i^k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$

      $a_i^k = \frac{1}{v_i^k}, \quad \hat{m}_i^k = \frac{a_i^k m_i^k}{\tau_i^{(0)}}, \quad \hat{q}_i^k = (a_i^k)^2 q_i^k - (\hat{m}_i^k)^2 \tau_i^{(0)}, \quad \hat{\tau}_i^k = a_i^k - \hat{q}_i^k$

      $\hat{m}_{k\to i}^{\text{new}} = \hat{m}_i^k - \hat{m}_{i\to k}, \quad \hat{q}_{k\to i}^{\text{new}} = \hat{q}_i^k - \hat{q}_{i\to k}, \quad \hat{\tau}_{k\to i}^{\text{new}} = \hat{\tau}_i^k - \hat{\tau}_{i\to k}$

    if $e = i \to k$ **then**

      $\hat{m}_i = \sum_{k'\in\partial i} \hat{m}_{k'\to i}, \quad \hat{q}_i = \sum_{k'\in\partial i} \hat{q}_{k'\to i}, \quad \hat{\tau}_i = \sum_{k'\in\partial i} \hat{\tau}_{k'\to i}$

      $\hat{m}_{i\to k}^{\text{new}} = \hat{m}_i - \hat{m}_{k\to i}, \quad \hat{q}_{i\to k}^{\text{new}} = \hat{q}_i - \hat{q}_{k\to i}, \quad \hat{\tau}_{i\to k}^{\text{new}} = \hat{\tau}_i - \hat{\tau}_{k\to i}$

until *convergence*

3.3.3 Computational Hard Phase

When the SE fixed point happens to be the global minimizer, the EP algorithm is in a sense optimal as its solution achieves the same overlaps as the student posterior according to Proposition 6. By contrast, when the SE fixed point fails to be a global minimizer, the EP algorithm is sub-optimal and is said to be in a computational *hard phase*. Finally note that Algorithm 4 can be viewed more generally as an iterative routine to find stationary points of the replica free entropy potential. When initialized as in Algorithm 4 it leads to the SE fixed point, but if initialized in the right basin of attraction it leads to the global minimizer.

3.3.4 Connection with Previous Work

We recover several results as particular cases. The state evolution for the ML-VAMP algorithm in the mismatched setting rigorously proven in (Pandit et al., 2020) is equivalent to Algorithm 4 applied to the multi-layer network with orthogonally invariant weight matrices. We recover the replica symmetric free entropy in the mismatched setting derived for the GLM (with orthogonally invariant weight matrix) in (Kabashima, 2008) and the low rank factorization in (Lesieur et al., 2017).

3.4 Bayes-Optimal Setting

In the Bayes-optimal setting, where all the student factors match the teacher factors $f_k(x_k, y_k) = f_k^{(0)}(x_k, y_k)$, the solution should be replica symmetric (Nishimori, 2001) and furthermore the ground truth $\mathbf{x}^{(0)}$ should behave as one the replicas $\mathbf{x}^{(a)}$ $(a = 1 \cdots n)$ in Eq. (92). In particular:

$$\tau_i = \tau_i^{(0)}, \quad m_i = q_i, \quad \text{mse}_i = v_i = \tau_i^{(0)} - m_i. \tag{136}$$

As the student and teacher generative models are identical $p(\mathbf{x}, \mathbf{y}) = p^{(0)}(\mathbf{x}, \mathbf{y})$, the ensemble average $\bar{A}$ in Eq. (88) now gives access to the entropy

$$-\bar{A}^{(0)} = \lim_{N \to \infty} \frac{1}{N} H[p^{(0)}(\mathbf{y})] - A_N^{(0)} \tag{137}$$

up to the constant $A_N^{(0)}$ that can be estimated through Proposition 5. With these simplifications, we get the following free entropy.

Proposition 7 (Bayes-optimal setting $\bar{A}^{(0)}$) *The Bayes-optimal (BO) setting $\bar{A}^{(0)}$ is given by:*

$$-\bar{A}^{(0)} = \min_{m_V} \bar{A}^{(0)*}[m_V] = \min \operatorname*{extr}_{\hat{m}_E} -\bar{A}^{(0)}[\hat{m}_E]. \tag{138}$$

where the minimizer corresponds to the overlaps $m_V = (m_i)_{i \in V}$ and $\hat{m}_E = (\hat{m}_{i \to k}, \hat{m}_{k \to i})_{(i,k) \in E}$ denotes the dual natural parameter messages. The BO potentials satisfy the tree decomposition:

$$\bar{A}^{(0)*}[m_V] = \sum_{k \in F} \alpha_k \bar{A}_k^{(0)*}[m_k] + \sum_{i \in V} \alpha_i (1 - n_i) \bar{A}_i^{(0)*}[m_i] \qquad \text{with} \quad m_k = (m_i)_{i \in \partial k}, \tag{139}$$

$$\bar{A}^{(0)}[\hat{m}_E] = \sum_{k \in F} \alpha_k \bar{A}_k^{(0)}[\hat{m}_k] - \sum_{(i,k) \in E} \alpha_i \bar{A}_i^{(0)}[\hat{m}_i^k] + \sum_{i \in V} \alpha_i \bar{A}_i^{(0)}[\hat{m}_i]$$
$$\text{with} \quad \hat{m}_k = (\hat{m}_{i \to k})_{i \in \partial k}, \quad \hat{m}_i^k = \hat{m}_{i \to k} + \hat{m}_{k \to i}, \quad \hat{m}_i = \sum_{k \in \partial i} \hat{m}_{k \to i}. \tag{140}$$

The factor and variable BO potentials are given by:

$$\bar{A}_k^{(0)}[\hat{m}_k] = \lim_{N_k \to \infty} \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} A_k^{(0)}[a_k, b_k; y_k] \tag{141}$$

$$\bar{A}_i^{(0)}[\hat{m}_i] = \lim_{N_i \to \infty} \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} A_i^{(0)}[a_i, b_i] \tag{142}$$

where $A_k^{(0)}[a_k, b_k; y_k]$ *and* $A_i^{(0)}[a_i, b_i]$ *are the scaled EP log-partitions with isotropic Gaussian beliefs (Example 9):*

$$A_k^{(0)}[a_k, b_k; y_k] = \frac{1}{N_k} \ln \int dx_k \, f_k^{(0)}(x_k; y_k) \, e^{-\frac{1}{2}a_k \|x_k\|^2 + b_k^\intercal x_k} \tag{143}$$

$$A_i^{(0)}[a_i, b_i] = \frac{1}{N_i} \ln \int dx_i \, e^{-\frac{1}{2}a_i \|x_i\|^2 + b_i^\intercal x_i} \tag{144}$$

and the factor and variable ensemble averages are taken with:

$$p_k^{(0)}(x_k^{(0)}, y_k, b_k) = \mathcal{N}(b_k \mid \hat{m}_k x_k^{(0)}, \hat{m}_k) \, p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)}) \qquad \text{and} \quad a_k = \hat{\tau}_k^{(0)} + \hat{m}_k, \tag{145}$$

$$p_i^{(0)}(x_i^{(0)}, b_i) = \mathcal{N}(b_i \mid \hat{m}_i x_i^{(0)}, \hat{m}_i) \, p_i^{(0)}(x_i^{(0)} \mid \hat{\tau}_i^{(0)}) \qquad \text{and} \quad a_i = \hat{\tau}_i^{(0)} + \hat{m}_i, \tag{146}$$

where $p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)})$ *and* $p_i^{(0)}(x_i^{(0)} \mid \hat{\tau}_i^{(0)})$ *are the approximate teacher marginals defined in Eqs* (107) *and* (108)*. The gradient of the factor RS potential give the dual mapping to the overlap:*

$$m_i^k[\hat{m}_k] = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} \frac{r_i^k[a_k, b_k; y_k] \cdot x_i^{(0)}}{N_i}, \quad \tfrac{1}{2}\alpha_i^k m_i^k = \partial_{\hat{m}_{i\to k}} \bar{A}_k^{(0)}, \tag{147}$$

where $r_i^k[a_k, b_k; y_k]$ *and* $v_i^k[a_k, b_k; y_k]$ *are the posterior mean and isotropic variance Eqs* (37) *and* (38) *as estimated by the EP factor marginal Eq.* (34)*. The gradient of the variable RS potential give the dual mapping to the overlap:*

$$m_i[\hat{m}_i] = \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} \frac{r_i[a_i, b_i] \cdot x_i^{(0)}}{N_i}, \quad \tfrac{1}{2} m_i = \partial_{\hat{m}_i} \bar{A}_i^{(0)}, \tag{148}$$

where $r_i[a_i, b_i] = \frac{b_i}{a_i}$ *and* $v_i[a_i, b_i] = \frac{1}{a_i}$ *are the posterior mean and isotropic variance as estimated by the EP variable marginal Eq.* (30)*.* $\bar{A}_k^{(0)*}$ *and* $\bar{A}_i^{(0)*}$ *are the corresponding Legendre transforms. The ensemble average variances are given by:*

$$v_i^k[\hat{m}_k] = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} v_i^k[a_k, b_k; y_k] = \tau_i^{(0)} - m_i^k[\hat{m}_k], \tag{149}$$

$$v_i[\hat{m}_i] = \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} v_i[a_i, b_i] = \tau_i^{(0)} - m_i[\hat{m}_i]. \tag{150}$$

Any stationary point of the potentials (not necessarily the global optima) is a fixed point:

$$m_i^k = m_i, \quad \hat{m}_i = \sum_{k \in \partial i} \hat{m}_{k \to i}. \tag{151}$$

3.4.1 BO Potentials

The variable BO potential Eq. (142) is explicitly given by:

$$\bar{A}_i^{(0)}[\hat{m}_i] = \frac{\hat{m}_i \tau_i^{(0)}}{2} + \frac{1}{2} \ln \frac{2\pi}{a_i} \quad \text{with} \quad a_i = \hat{\tau}_i^{(0)} + \hat{m}_i \tag{152}$$

which yields the dual mapping:

$$m_i = \tau_i^{(0)} - v_i, \quad v_i = \frac{1}{a_i}. \tag{153}$$

Several factor BO potentials are given in Appendix E.

3.4.2 Entropy Estimation and State Evolution

The ensemble average $\bar{A}^{(0)}$ which gives access to the entropy through Eq. (137) is given by the global minimum according to Proposition 7. The global minimizer gives access to the ensemble average overlap m_i as well as $\text{mse}_i = \tau_i^{(0)} - m_i = v_i$ for the student posterior Eq. (2). As the student posterior is Bayes-optimal $\text{mse}_\text{i} = \text{mmse}_i$ is the MMSE. However the replica free entropy solution also appears as an ensemble average of the underlying EP Algorithm 2 where the effective ensemble average is defined locally for each factor Eq. (145) and variable Eq. (146). With this effective ensemble average interpretation in mind we conjecture Algorithm 5 to give the state evolution of the EP Algorithm 2 in the Bayes-optimal setting. Then the state evolution fixed point will give the overlap m_i as well as $\text{mse}_i = \tau_i^{(0)} - m_i = v_i$ corresponding to the EP student solution, which is in general only a local minimizer in Proposition 7.

Algorithm 5: Tree-AMP State evolution (Bayes-optimal setting)

initialize $\hat{m}_{i\to k}, \hat{m}_{k\to i} = 0$
repeat
  foreach *edge* $e \in E_+ \cup E_-$ **do** // forward and backward pass
    if $e = k \to i$ **then**
      $a_i^k = 1/v_i^k[\hat{m}_k], \quad \hat{m}_i^k = a_i^k - \hat{\tau}_i^{k(0)}$ // variance-matching
      $\hat{m}_{k\to i}^{\text{new}} = \hat{m}_i^k - \hat{m}_{i\to k}$ // message $f_k \to x_i$
    if $e = i \to k$ **then**
      $\hat{m}_i = \sum_{k'\in\partial i} \hat{m}_{k'\to i}$ // precision
      $\hat{m}_{i\to k}^{\text{new}} = \hat{m}_i^k - \hat{m}_{k\to i}$ // message $x_i \to f_k$
until *convergence*

3.4.3 Computational Hard phase

When the SE fixed point happens to be the global minimizer, the EP algorithm is in a sense optimal as its solution achieves the same overlap as the Bayes-optimal posterior according to Proposition 7, in particular $\text{mse}_i = \text{mmse}_i$. By contrast, when the SE fixed point fails to be a global minimizer, the EP algorithm is sub-optimal $\text{mse}_i > \text{mmse}_i$ and is said to be in a computational *hard phase*. Finally note that Algorithm 5 can be viewed more generally as an iterative routine to find stationary points of the replica free entropy potential. When initialized as in Algorithm 5 it leads to the SE fixed point, but if initialized in the right basin of attraction it leads to the global minimizer.

3.4.4 Connection with Previous Work

We recover several results as particular cases. The state evolution for the ML-VAMP algorithm in the matched setting, rigorously proven in (Fletcher et al., 2018), is equivalent to Algorithm 5 applied to the multi-layer network with orthogonally invariant weight matrices. We recover the corresponding replica symmetric free entropy derived in (Gabrié et al., 2018). The Reeves (2017) formalism, developed for tree networks of GLMs and expressed in term of mutual information potentials, is shown to be equivalent to Proposition 7 in Section 3.5.

3.5 Information Theoretic Expressions

In this subsection we wish to express the RS and BO potentials using information theoretic quantities. These potentials are all given by a local ensemble average of the EP log-partition which can be interpreted as a local teacher-student scenario. We will find that the BO potential is related to a mutual information term, recovering (Reeves, 2017) formalism, while the RS potential differs from the BO potential by a KL divergence term.

3.5.1 Local Teacher-Student Scenario

The RS factor potential Eq. (119) is given by a local ensemble average of the EP log-partition. The effective ensemble average in Eq. (123) can be interpreted as a local teacher-student scenario where the local teacher generative model is given by:

$$p_k^{(0)}(x_k^{(0)}, y_k, b_k) = \mathcal{N}(b_k \mid \hat{m}_k x_k^{(0)}, \hat{q}_k) \underbrace{f_k^{(0)}(x_k^{(0)}; y_k)\, e^{-\frac{1}{2}\hat{\tau}_k^{(0)}\|x_k^{(0)}\|^2 - N_k A_k^{(0)}[\hat{\tau}_k^{(0)}]}}_{p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)})} \tag{154}$$

while the local student generative model is given by:

$$p_k(x_k, y_k, b_k) = \mathcal{N}(b_k \mid \hat{q}_k x_k, \hat{q}_k) \underbrace{f_k(x_k; y_k)\, e^{-\frac{1}{2}\hat{\tau}_k\|x_k\|^2 - N_k A_k[\hat{\tau}_k]}}_{p_k(x_k, y_k \mid \hat{\tau}_k)}. \tag{155}$$

The teacher generates the ground truth signals $x_k^{(0)}$, the measurements y_k and the messages b_k according to Eq. (154) and the goal of the student is to infer the signals x_k from (y_k, b_k) assuming Eq. (155). The local student posterior is then given by:

$$p_k(x_k \mid y_k, b_k) = f_k(x_k; y_k)\, e^{-\frac{1}{2}a_k\|x_k\|^2 + b_k^\intercal x_k - N_k A_k[a_k, b_k; y_k]} \quad \text{with} \quad a_k = \hat{\tau}_k + \hat{q}_k \tag{156}$$

which we recognize as the EP factor marginal Eq. (34). Note that in this local teacher-student scenario, the messages b_k acts as pseudo-measurements of the signals x_k corrupted by Gaussian noise. We see that the local teacher Eq. (154) actually generates the messages b_k with signal-to-noise ratios (SNR) $\hat{m}_k^{(0)} = \hat{m}_k^2/\hat{q}_k$ while the student Eq. (155) believes that they are generated with SNR $\hat{q}_k$.

3.5.2 Bayes-Optimal Setting

The teacher and student factors are matched $f_k = f_k^{(0)}$ and the local teacher and student generative models are identical:

$$p_k^{(0)}(x_k^{(0)}, y_k, b_k) = \mathcal{N}(b_k \mid \hat{m}_k x_k^{(0)}, \hat{m}_k) \underbrace{f_k^{(0)}(x_k^{(0)}; y_k)\, e^{-\frac{1}{2}\hat{\tau}_k^{(0)}\|x_k^{(0)}\|^2 - A_k^{(0)}[\hat{\tau}_k^{(0)}]}}_{p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)})} \tag{157}$$

We recall that in the Bayes-optimal setting $\hat{m}_k = \hat{q}_k$ and $\hat{\tau}_k = \hat{\tau}_k^{(0)}$ and consistently the teacher and student SNR are matched $\hat{m}_k^{(0)} = \hat{q}_k = \hat{m}_k$. The local student posterior is the EP factor marginal Eq. (34)

$$p_k^{(0)}(x_k \mid y_k, b_k) = f_k^{(0)}(x_k; y_k) e^{-\frac{1}{2}a_k\|x_k\|^2 + b_k^\intercal x_k - N_k A_k^{(0)}[a_k, b_k; y_k]} \quad \text{with} \quad a_k = \hat{\tau}_k^{(0)} + \hat{m}_k \tag{158}$$

and is Bayes-optimal. In particular, for all $i \in \partial k$, the posterior mean $r_i^k[a_k, b_k; y_k]$ is the MMSE estimator of the signal x_i and the posterior variance $v_i^k[a_k, b_k; y_k]$ gives the MMSE.

3.5.3 Decomposition of the RS and BO Factor Potentials

With these local teacher and student generative models, we can give the following information theoretic interpretation of the RS factor potential Eq. (119) and BO factor potential Eq. (141), see Appendix D for a proof.

Proposition 8 *The RS potential $\bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$ differs from the BO potential $\bar{A}_k^{(0)}[\hat{m}_k^{(0)}]$ by a KL divergence term:*

$$\bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] - A_k[\hat{\tau}_k] = \bar{A}_k^{(0)}[\hat{m}_k^{(0)}] - A_k^{(0)}[\hat{\tau}_k^{(0)}] - K_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$$
$$\text{with} \quad K_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] = \lim_{N_k \to \infty} \frac{1}{N_k} \mathrm{KL}[p_k^{(0)}(y_k, b_k) \| p_k(y_k, b_k)] \quad \text{and} \quad \hat{m}_k^{(0)} = \frac{\hat{m}_k^2}{\hat{q}_k}, \tag{159}$$

up to the student prior log-partition $A_k[\hat{\tau}_k]$ and teacher prior log-partition $A_k^{(0)}[\hat{\tau}_k^{(0)}]$, and $K_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$ is the KL divergence between the local teacher evidence $p_k^{(0)}(y_k, b_k)$ in Eq. (154) and the local student evidence $p_k(y_k, b_k)$ in Eq. (155). The BO potential $\bar{A}_k^{(0)}[\hat{m}_k]$ is directly related to an entropic term:

$$\bar{A}_k^{(0)}[\hat{m}_k] - A_k^{(0)}[\hat{\tau}_k^{(0)}] = \sum_{i \in \partial k} \frac{\alpha_i^k \hat{m}_{i \to k} \tau_i^{(0)}}{2} - H_k[\hat{m}_k]$$
$$\text{with} \quad H_k[\hat{m}_k] = \lim_{N_k \to \infty} \frac{1}{N_k} H[y_k, b_k] - \frac{1}{N_k} H[b_k \mid x_k^{(0)}], \tag{160}$$

where the entropies are defined over the random variables $x_k^{(0)}, y_k, b_k$ distributed according to Eq. (157).

3.5.4 Reeves Formalism

The entropic expression Eq. (160) in the Bayes-optimal setting allows to recover the (Reeves, 2017) formalism developed for a tree network of GLMs (Example 13) and extends it to other tree-structured factor graphs. For a non-likelihood factor, that is $y_k = \emptyset$, the entropic potential H_k reduced to the mutual information between the signals $x_k^{(0)}$ and the messages b_k:

$$H_k[\hat{m}_k] = I_k[\hat{m}_k] = \lim_{N_k \to \infty} \frac{1}{N_k} I[x_k^{(0)}; b_k]. \tag{161}$$

For a likelihood factor $y_k \neq \emptyset$, the entropic potential H_k reduced to the mutual information between the signals $x_k^{(0)}$ and the pair of measurements y_k and messages b_k plus an entropic noise term:

$$H_k[\hat{m}_k] = I_k[\hat{m}_k] + E_k,$$
$$I_k[\hat{m}_k] = \lim_{N_k \to \infty} \frac{1}{N_k} I[x_k^{(0)}; (y_k, b_k)], \quad E_k = \lim_{N_k \to \infty} \frac{1}{N_k} H[y_k \mid x_k^{(0)}]. \tag{162}$$

For a noiseless output channel (deterministic relationship between y_k and $x_k^{(0)}$) the mutual information $I_k[\hat{m}_k]$ and the entropic noise E_k can be ill-defined and the more general relation

Eq. (160) should be preferred. The mutual information potential $I_k[\hat{m}_k]$ and the entropic potential $H_k[\hat{m}_k]$ are functions of the SNR $\hat{m}_k$. Their gradients give the dual mapping with the variances:

$$\frac{1}{2}\alpha_i^k v_i^k = \partial_{\hat{m}_{i\to k}} I_k[\hat{m}_k] = \partial_{\hat{m}_{i\to k}} H_k[\hat{m}_k] \quad \text{for all } i \in \partial k \tag{163}$$

known as the I-MMSE theorem (Guo et al., 2005) as v_i^k gives the MMSE of the signal x_i. Then Proposition 7 is exactly equivalent to (Reeves, 2017) formalism, where the BO potentials $\bar{A}_k^{(0)}[\hat{m}_k]$ are replaced by the mutual information potentials $I_k[\hat{m}_k]$ and the overlaps m_k by the variances v_k.

3.5.5 RS and BO Variable Potentials

The RS variable potential Eq. (120) is given by the local ensemble average Eq. (124) of the EP log-partition, which can again be interpreted as a local teacher-student scenario. The local teacher and student generative models are the Gaussians:

$$p_i^{(0)}(x_i^{(0)}, b_i) = \mathcal{N}(b_i \mid \hat{m}_i x_i^{(0)}, \hat{q}_i)\,\mathcal{N}(x_i^{(0)} \mid 0, 1/\hat{\tau}_i^{(0)}), \tag{164}$$

$$p_i(x_i, b_i) = \mathcal{N}(b_i \mid \hat{q}_i x_i, \hat{q}_i)\,\mathcal{N}(x_i \mid 0, 1/\hat{\tau}_i). \tag{165}$$

In the Bayes-optimal setting $\hat{m}_i = \hat{q}_i$ and $\hat{\tau}_i = \hat{\tau}_i^{(0)}$ so the local teacher and student generative models are identical. The RS and BO variable potentials follows the same decomposition Eqs (159)-(160) as the factor potentials. But in that case the decomposition can be straightforwardly checked from the explicit expressions:

$$\bar{A}_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \frac{\hat{m}_i^2 \tau_i^{(0)} + \hat{q}_i}{2a_i} + \frac{1}{2}\ln\frac{2\pi}{a_i} \quad \text{with} \quad a_i = \hat{\tau}_i + \hat{q}_i, \quad A_i[\hat{\tau}_i] = \frac{1}{2}\ln\frac{2\pi}{\hat{\tau}_i}, \tag{166}$$

$$K_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \frac{1}{2}\left(\ln\frac{\Delta_S}{\Delta_T} + \frac{\Delta_T}{\Delta_S} - 1\right) \quad \text{with} \quad \Delta_T = \hat{q}_i + \frac{\hat{m}_i^2}{\hat{\tau}_i^{(0)}}, \quad \Delta_S = \hat{q}_i + \frac{\hat{q}_i^2}{\hat{\tau}_i}, \tag{167}$$

$$\bar{A}_i^{(0)}[\hat{m}_i^{(0)}] = \frac{\hat{m}_i^{(0)}\tau_i^{(0)}}{2} + \frac{1}{2}\ln\frac{2\pi}{a_i^{(0)}} \quad \text{with} \quad a_i^{(0)} = \hat{\tau}_i^{(0)} + \hat{m}_i^{(0)}, \; A_i^{(0)}[\hat{\tau}_i^{(0)}] = \frac{1}{2}\ln\frac{2\pi}{\hat{\tau}_i^{(0)}}, \tag{168}$$

$$H_i[\hat{m}_i^{(0)}] = I_i[\hat{m}_i^{(0)}] = \frac{1}{2}\ln a_i^{(0)}\tau_i^{(0)}. \tag{169}$$

K_i is the KL divergence between the local teacher evidence $p_i^{(0)}(b_i) = \mathcal{N}(b_i \mid 0, \Delta_T)$ in Eq. (164) and the local student evidence $p_i(b_i) = \mathcal{N}(b_i \mid 0, \Delta_S)$ in Eq. (165).

3.6 State Evolution Modules

For each EP factor module with isotropic Gaussian beliefs (Section 2.8) one can easily implement the corresponding free entropy / state evolution module by taking the ensemble average Eq. (123) in the replica symmetric (RS) setting or Eq. (145) in the Bayes-optimal (BO) setting. In practice the RS module must be able to compute the RS potential $\bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$ given in Eq. (119) and the associated overlaps $m_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$, $q_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$ and $\tau_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$ given in Eqs (125)-(127). Similarly the BO module must be able to compute the BO potential $\bar{A}_k^{(0)}[\hat{m}_k]$ given in Eq. (141) and the associated overlap $m_k[\hat{m}_k]$ given in Eq. (147).

We have closed-form expressions for the linear channel in the RS (Appendix E.2.3) and BO (Appendix E.2.4) cases. For separable factors the RS and BO potentials and associated overlaps can be analytically obtained through a low dimensional integration: see Appendix E.3.3-E.3.4 for a separable prior, Appendix E.4.3-E.4.4 for a separable likelihood, and Appendix E.5.3-E.5.4 for a separable channel.

4. Examples

This section is dedicated to illustrating the Tree-AMP package. We first point out that its reconstruction performances asymptotically reaches the Bayes optimal limit out of the hard phase and that its fast execution speed often exceeds competing algorithms. Moreover we stress that the cornerstone of Tree-AMP is its modularity, which allows it to handle a wide range of inference tasks. To appreciate its great flexibility, we illustrate its performance on various tree-structured models. Finally, the last section depicts the ability of Tree-AMP to predict its own state evolution performance on two simple GLMs: compressed sensing and sparse phase retrieval. The codes corresponding to the examples presented in this section can be found in the documentation gallery[5] or in the repository[6].

4.1 Benchmark on Sparse Linear Regression

Let us consider a sparse signal $x \in \mathbb{R}^N$, iid drawn according to $x \sim \prod_{n=1}^N \mathcal{N}_\rho(x_n)$ where $\mathcal{N}_\rho = [1-\rho]\delta + \rho\mathcal{N}$ is the Gauss-Bernoulli prior and $\mathcal{N}$ the normal distribution. The inference task is to reconstruct the signal x from noisy observations $y \in \mathbb{R}^M$ generated according to

$$y = Ax + \xi \tag{170}$$

where $A \in \mathbb{R}^{M\times N}$ is the sensing matrix with iid Gaussian entries $A_{mn} \sim \mathcal{N}(0, 1/N)$ and ξ is a iid Gaussian noise $\xi_m \sim \mathcal{N}(0, \Delta)$. We define $\alpha = M/N$ the aspect ratio of the matrix A. The corresponding factor graph is depicted in Figure 4.

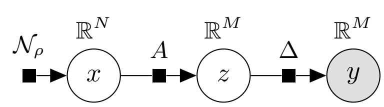


Figure 4: Sparse linear regression factor graph.

The sparse linear regression problem can be easily solved with the Tree-AMP package. We simply need to import the necessary modules, declare the model, and run the expectation propagation algorithm:

```
# import modules
from tramp.base import Variable as V
from tramp.priors import GaussBernoulliPrior
from tramp.likelihoods import GaussianLikelihood
from tramp.channels import LinearChannel
```

5. https://sphinxteam.github.io/tramp.docs/0.1/html/gallery/index.html
6. https://github.com/sphinxteam/tramp/tree/master/examples/figures

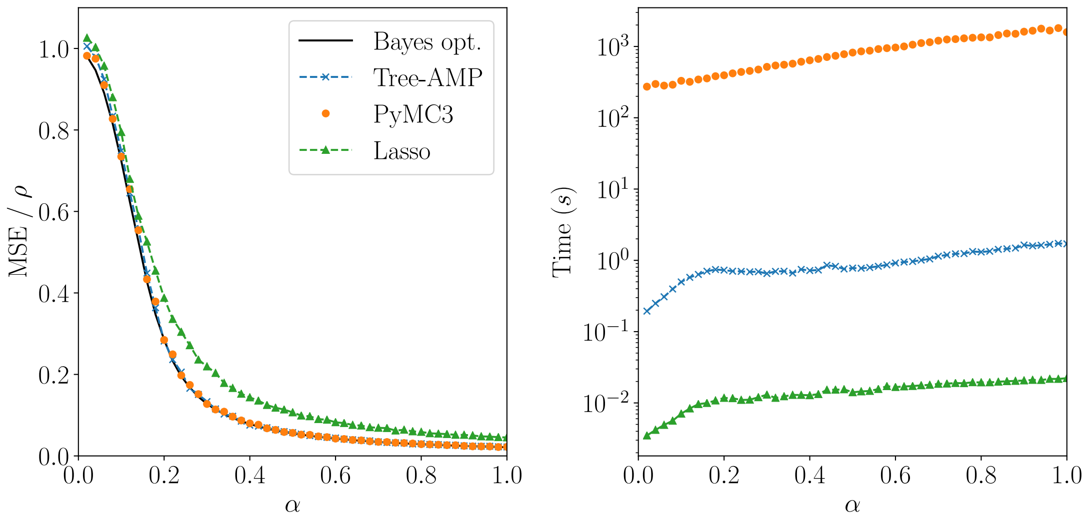


Figure 5: Benchmark on a sparse linear regression task: rescaled MSE as a function of $\alpha = M/N$. The MSE achieved by Tree-AMP (blue) is compared to the Bayes-optimal MMSE (black), Hamiltonian Monte-Carlo (orange) from PyMC3 (with $n_{\mathrm{s}} = 1000$ distribution samples and NUTS sampler) and Lasso (green) from Scikit-Learn (with the optimal regularization parameter obtained beforehand by simulation). The above experiments have been performed with parameters $(N, \rho, \Delta) = (1000, 0.05, 0.01)$ and have been averaged over 100 samples.

```
# declare sparse linear regression model
model = (
    GaussBernoulliPrior(rho=rho, size=N) @ V('x') @
    LinearChannel(A) @ V('z') @
    GaussianLikelihood(var=Delta, y=y)
).to_model()
# run EP
from tramp.algos import ExpectationPropagation
ep = ExpectationPropagation(model)
ep.iterate(max_iter=200)
```

We compare the Tree-AMP performance on this inference task to the Bayes optimal theoretical prediction from (Barbier et al., 2019) to two state of the art algorithms for this task: Hamiltonian Monte-Carlo from the PyMC3 package (Salvatier et al., 2016) and Lasso (L1-regularized linear regression) from the Scikit-Learn package (Pedregosa et al., 2011). Note that to perform our experimental benchmark in Figure 5 the Tree-AMP and PyMC3 algorithms had access to the ground-truth parameters (ρ, Δ) used to generate the observations. In order to make the benchmark as fair as possible, we use the optimal regularization parameter for the Lasso, obtained beforehand by simulation.

We observe in Figure 5 (**left**) that for this model Tree-AMP is Bayes-optimal and reaches the MMSE, up to finite size fluctuations, just as PyMC3. They naturally both outperform Lasso from Scikit-Learn that never achieves the Bayes-optimal MMSE for the full range of aspect ratio α under investigation. This is expected and unfair to Lasso as the two Bayesian methods have full knowledge of the exact generating distribution in our toy model, but this is rarely the case in real applications.

Whereas the Hamiltonian Monte-Carlo algorithm requires to draw a large number of samples ($n_s = 10^3$) to reach a given threshold of precision, Tree-AMP is an iterative algorithm that converges in a few iterations varying broadly speaking between $[10^0; 10^2]$. It leads interestingly to an execution time smaller by two orders of magnitude with respect to PyMC3 as illustrated in Figure 5 (**right**). Hence the fast convergence and execution time of Tree-AMP is certainly a deep asset over PyMC3, or similar Markov Chain Monte-Carlo packages.

4.2 Depicting Tree-AMP Modularity

In order to show the adaptability and modularity of Tree-AMP to handle various inference tasks, we present here different examples where the prior distributions are modified flexibly. In particular, we consider first Gaussian denoising of synthetic data with either sparse discrete Fourier transform (DFT) or sparse gradient, and second the denoising and inpainting of real images drawn from the MNIST data set, using a trained Variational Auto-Encoder (VAE) as a prior.

4.2.1 Sparse DFT/Gradient Denoising

Let us consider a signal $x \in \mathbb{R}^N$ corrupted by a Gaussian noise $\xi \sim \mathcal{N}(0, \Delta)$, that leads to the observation $y = x + \xi \in \mathbb{R}^N$. In contrast to the first section in which we considered the signal x to be sparse, we assume here that the signal is dense but that a linear transformation of the signal is sparse. In other words let us define the variable $z = \Omega x$ that we assume to be sparse, where Ω denotes a linear operator acting on the signal. The factor graph associated to this model is depicted in Figure 6.

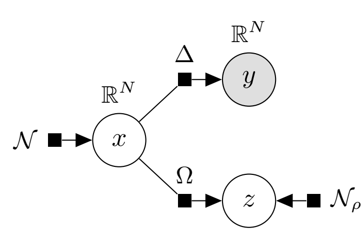


Figure 6: Factor graph for sparse Ω denoising, where Ω represents either the DFT or the gradient operator.

As a matter of clarity, we focus on two toy one-dimensional signals:

1. $x \in \mathbb{R}^N$ such that $\forall n \in [1 : N], x_n = \cos(t_n) + \sin(2t_n)$, with $t_n = 2\pi(-1 + \frac{2n}{N})$. The signal is sparse in the Fourier basis with only two spikes, that leads us to consider a sparse DFT prior: Ω is the discrete Fourier transform,

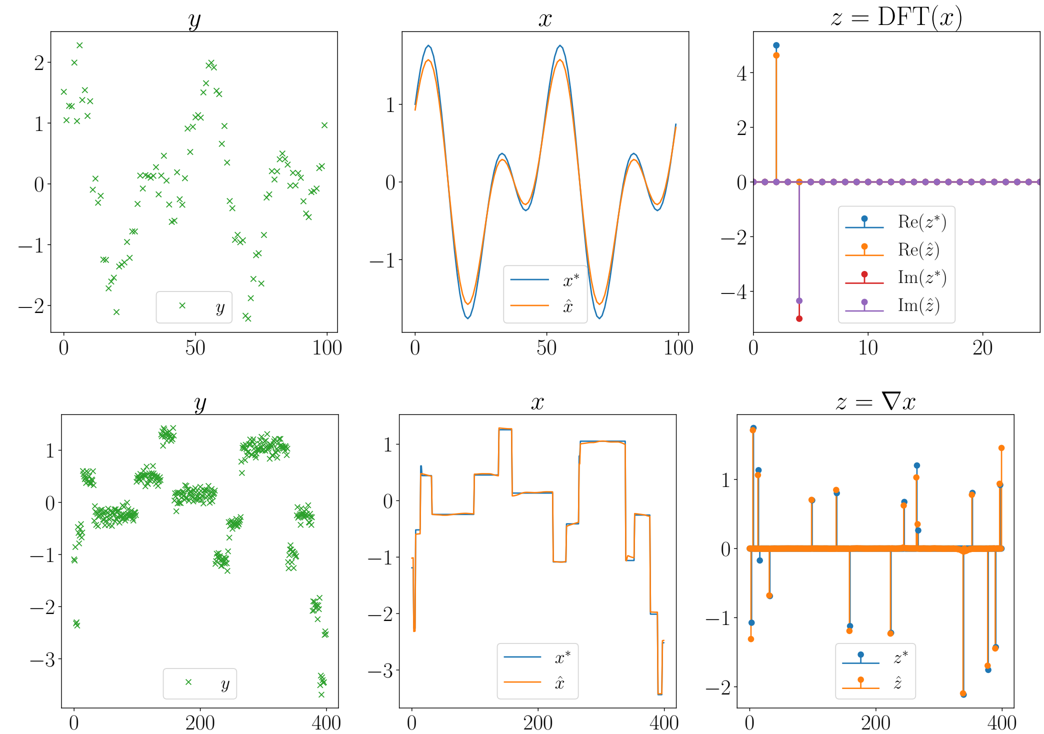


Figure 7: Sparse FFT/gradient denoising: **(left)** noisy observation y, **(middle)** ground truth signal x^* and predicted $\hat{x}$ and **(right)** ground truth linear transform z^* and predicted $\hat{z}$ for **(upper)** sparse DFT denoising with $(N, \rho, \Delta) = (100, 0.02, 0.1)$ and **(lower)** sparse gradient denoising with $(N, \rho, \Delta) = (400, 0.04, 0.01)$.

2. $x \in \mathbb{R}^N$ such that it is randomly drawn constant by pieces. Its gradient contains a lot of zeros and therefore inference with a sparse gradient prior is appropriate: Ω is the gradient operator.

After importing the relevant modules, declaring the model in the Tree-AMP package is simple, for instance for the sparse gradient model:

```
# sparse gradient denoising
model = (
  GaussianPrior(size=N) @ V('x', n_prev=1, n_next=2) @ (
    GaussianLikelihood(var=Delta, y=y) + (
      GradientChannel() +
      GaussBernoulliPrior(rho=rho, size=(1,N))
    ) @ V('z', n_prev=2, n_next=0)
  )
).to_model()
```

For the sparse DFT model, one just needs to replace `GradientChannel` by `DFTChannel`. Numerical experiments are shown in Figure 7. The left panel shows the observation y, while the middle and right panels illustrate the Tree-AMP reconstruction of the signal $\hat{x}$

and of its linear transform $\hat{z}$ compared to the ground truth x^* and z^*. The Tree-AMP reconstruction approaches closely the ground truth signal and leads to MSE $\sim 10^{-2}/10^{-3}$ for signals 1/2.

4.2.2 Variational Auto-Encoder on MNIST

Let us consider a signal $x \in \mathbb{R}^N$ (with $N = 784$) drawn from the MNIST data set. We want to reconstruct the original image from a corrupted observation $y = \varphi(x) \in \mathbb{R}^N$, where $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ represents a noisy channel. In the following the noisy channel represents either a Gaussian additive channel or an inpainting channel, that erases some pixels of the input image.

In order to reconstruct correctly the MNIST image, we investigated the possibility of using a generative prior such as a Variational Auto-Encoder (VAE) along the lines of (Bora et al., 2017; Fletcher et al., 2018). The information theoretical and approximate message passing properties of reconstruction of a low rank or GLM channel, using a dense feed-forward neural network generative prior with iid weights has been studied in particular in (Aubin et al., 2019, 2020). The VAE architecture is summarized in Figure 8 and the training procedure on the MNIST data set follows closely the canonical one detailed in (Keras-VAE). We considered two common inference tasks: denoising and inpainting.

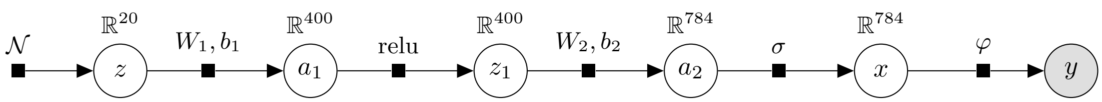


Figure 8: Denoising/inpaiting a MNIST image with a VAE prior. The weights W_1, W_2 and biases b_1, b_2 were learned beforehand on the MNIST data set.

Denoising: In that case, the corrupted channel $\varphi_{\text{den},\Delta}$ adds a Gaussian noise and corresponds to the noisy channel

$$\varphi_{\text{den},\Delta}(x) = x + \xi \quad \text{with} \quad \xi \sim \mathcal{N}(0, \Delta)\,.$$

Inpaiting: The corrupted channel erases a few pixels of the input image and corresponds formally to

$$\varphi_{\text{inp},\text{I}_\alpha}(x) = \begin{cases} 0 \text{ if } i \in \text{I}_\alpha\,, \\ x_i \text{ otherwise}\,. \end{cases}$$

where I_α denotes the set of erased indexes of size $\lfloor \alpha N \rfloor$ for some $\alpha \in [0; 1]$. As an illustration, we consider two different manners of generating the erased interval I_α:

1. A central horizontal band of width $\lfloor \alpha N \rfloor$: $\text{I}_\alpha^{\text{band}} = [\lfloor \frac{\text{N}}{2}(1-\alpha) \rfloor; \lfloor \frac{\text{N}}{2}(1+\alpha) \rfloor]$
2. Indices drawn uniformly at random $\lfloor \alpha N \rfloor$: $\text{I}_\alpha^{\text{uni}} \sim \text{U}([1, \text{N}]; \lfloor \alpha \text{N} \rfloor)$

Solving these inference tasks in Tree-AMP is straightforward: first declare the model Figure 8 and then run expectation propagation as exemplified in Section 4.1 for the sparse regression case. A few MNIST samples x^* compared to the noisy observations y and Tree-AMP reconstructions $\hat{x}$ are presented in Figure 9, that suggest that Tree-AMP is able to use

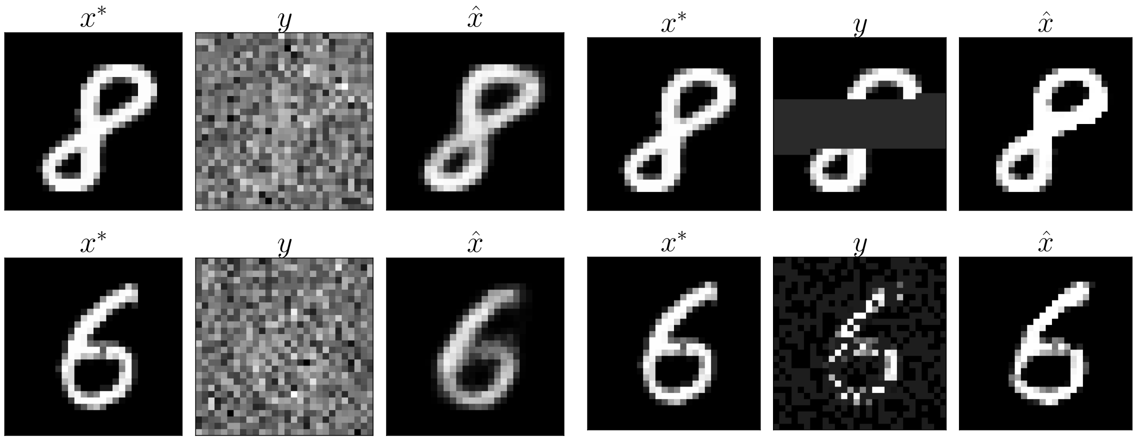


Figure 9: Illustration of the Tree-AMP prediction $\hat{x}$ using a VAE prior from observation $y = \varphi(x^*)$ with x^* a MNIST sample. **(left)** Denoising $\varphi = \varphi_{\text{den},\Delta}$ with $\Delta = 4$. **(right-upper)** Band-inpainting $\varphi_{\text{inp},\text{I}_\alpha^{\text{band}}}$ with $\alpha = 0.3$ **(right-lower)** Uniform-inpainting $\varphi_{\text{inp},\text{I}_\alpha^{\text{uni}}}$ with $\alpha = 0.5$.

the trained VAE prior information to either denoise very noisy observations or reconstruct missing pixels.

4.3 Theoretical Prediction of Performance

Previous sections were devoted to applications of the expectation propagation (EP) Algorithm 2 implemented in Tree-AMP. Moreover the state evolution (SE) Algorithm 5 has also been implemented in the package. This two-in-one package makes it easier to obtain performances of the EP algorithm on finite size instances as well as the infinite size limit behavior predicted by the state evolution.

We illustrate this on two generalized linear models: *compressed sensing* and *sparse phase retrieval*, whose common factor graph is represented in Figure 10. Briefly, we consider a sparse $x \in \mathbb{R}^N$ iid drawn from a Gauss-Bernoulli distribution $\mathcal{N}_\rho$. We observe $y \in \mathbb{R}^M = \varphi(Ax)$ with $A \in \mathbb{R}^{M\times N}$ a Gaussian iid matrix, and the noiseless channel is $\varphi(x) = x$ in the compressed sensing case and $\varphi(x) = |x|$ in the phase retrieval one.

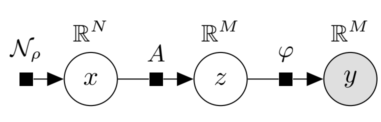


Figure 10: Graphical model representing the compressed sensing ($\varphi(x) = x$)) and phase retrieval ($\varphi(x) = |x|$)). We denote $\alpha = M/N$ the aspect ratio of the sensing matrix A.

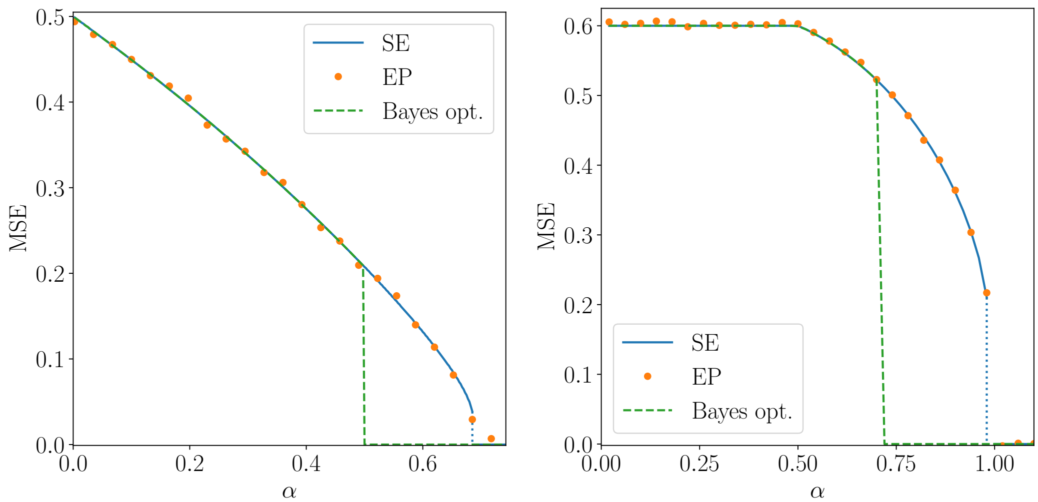


Figure 11: MSE as a function of $\alpha = M/N$ for (**left**) Compressed sensing ($\rho = 0.5$), (**right**) Sparse phase retrieval ($\rho = 0.6$).

Getting the MSE predicted by state evolution is straightforward in the Tree-AMP package. After importing the relevant modules, one just needs to declare the model and run the SE algorithm. For instance for the sparse phase retrieval model:

```
# declare sparse phase retrieval model
model = (
    GaussBernoulliPrior(rho=rho, size=N) @ V('x') @
    LinearChannel(A) @ V('z') @
    AbsLikelihood(y=y)
).to_model()
# run SE
from tramp.algos import StateEvolution
se = StateEvolution(model)
se.iterate(max_iter=200)
```

In Figure 11, we compare the MSE theoretically predicted by state evolution and the MSE obtained on high-dimensional ($N = 2000$) instances of EP. Notably up to finite size effects, the MSE averaged over 25 instances of EP match perfectly the MSE predicted by SE. Also the MSE is equal to the Bayes optimal MMSE proven in (Barbier et al., 2019), except for a region of α values known as the hard phase. In that phase, there is a significant gap between the MMSE that is information-theoretically achievable and the MSE actually achieved by EP.

Note that the MMSE is also a state evolution fixed point (Section 3.4) and can thus be obtained by initializing the SE Algorithm 5 in the right basin of attraction. For the two models discussed here, we found that initializing the incoming prior message $\hat{m}_{x\to\mathcal{N}_\rho} \gg 1$ was sufficient to converge towards the MMSE and obtain the Bayes optimal curve.

5. Discussion

The Tree-AMP package aims to solve compositional inference tasks, which can be broken down into local inference problems. As long as the underlying factor graph is tree-structured, the global inference task can be solved by message passing using Algorithm 2, which is just a particular instance of EP. Note that Algorithm 2 is generic, meaning that it can be implemented independently of the probabilistic graphical model under consideration. The main strength of the presented approach is therefore its modularity. In the Tree-AMP package, each module corresponds to a local inference problem given by a factor and associated beliefs on its variables. As long as the module is implemented (which means computing the log-partition $A_f[\lambda_f]$ and the moment function $\mu_f[\lambda_f]$), it can be composed at will with other modules to solve complex inference tasks. Several popular machine learning tasks can be reformulated that way as illustrated in Figure 1. We hope that the Tree-AMP package offers a unifying framework to run these models, as well as study them theoretically using the state evolution and free entropy formalism. Below, we review some shortcomings of the Tree-AMP package and possible ways to overcome them.

5.1 Hyper-Parameter Learning

In principle, it should be straightforward to learn hyper-parameters. As usually done in hierarchical Bayesian modelling, one simply needs to add the hyper-parameters as scalar variables in the graphical model with associated hyper priors. In term of the Tree-AMP package, one would simply need to implement the corresponding module (where the set of variables of the factor now includes the hyper-parameters to learn). In the typical use case, where the signals are high dimensional but the hyper-parameters are just scalars, Algorithm 2 will likely be equivalent to the expectation-maximization (Dempster et al., 1977) learning of hyper-parameters, as usually done in AMP algorithms (Krzakala et al., 2012).

5.2 Generic Belief

While the message passing Algorithm 2 is formulated for any kind of beliefs, the current Tree-AMP implementation only supports isotropic Gaussian beliefs. However we could consider more generic beliefs to deal with more complicated types of variable, such as Gaussian process beliefs (Rasmussen and Williams, 2006) for functions or harmonic exponential family beliefs (Cohen and Welling, 2015) for elements of compact groups. Maybe one can recover algorithms similar to (Opper and Winther, 2000) for Gaussian process classification or (Perry et al., 2018) for synchronisation problems over compact groups, and reformulate them in a more modular way. Even if we restrict ourselves to Gaussian beliefs, it may be beneficial to go beyond the isotropic case and consider diagonal or full covariance beliefs (Opper and Winther, 2005a), or any kind of prescribed covariance structure. As exemplified in the committee machine keeping a covariance between experts leads to a more accurate algorithm (Aubin et al., 2018).

5.3 Beyond Trees

By design, the Tree-AMP package can only handle tree-structured factor graphs. To overcome this fundamental limitation and extend to generic factor graphs, one could use the Kikuchi free energy (Yedidia et al., 2003) in place of the Bethe free energy as a starting point. Similar to the Bethe free energy which is exact for tree-structured factor graphs, the Kikuchi free energy will be exact if the graphical model admits a hyper-tree factorization (Wainwright and Jordan, 2008). The minimization of the Kikuchi free energy under weak consistency constraints (Zoeter and Heskes, 2005) could be used to implement a generalization of Algorithm 2, however the message passing will be more challenging than in the tree case. An equivalent of Proposition 2 will likely hold, where the tree decomposition Eqs (39)-(40) found for the EC Gibbs and EP free energies will be replaced by the hyper-tree factorization.

5.4 Convergence

The message passing Algorithm 2, like other EP algorithms, is not guaranteed to converge. This is a major drawback, and indeed on some instances the naive application of Algorithm 2 will diverge. Double loop algorithms like (Heskes and Zoeter, 2002) will ensure convergence, but are unfortunately very slow. In practice, damping the updates is often sufficient to converge towards a fixed point. In the Tree-AMP package the amount of damping has to be chosen by the user. It will be therefore interesting to generalize the adaptive damping scheme (Vila et al., 2015) in order to tune this damping automatically.

5.5 Proofs

In Section 3, we heuristically derive the free entropy using weak consistency on the overlaps and conjecture the corresponding state evolution, but did not provide any rigorous proof. We nonetheless recovered earlier derivations of these results for specific models. For instance in the multi-layer model with orthogonally invariant weight matrices, the state evolution was rigorously proven by Fletcher et al. (2018) while the free potential was heuristically derived using the replica method (Gabrié et al., 2018). When the weight matrices are Gaussian, the replica free entropy can further be shown to be rigorous (Reeves and Pfister, 2016; Barbier et al., 2019). One extension of our work would be to generalize existing proofs to arbitrary tree-structured models. In particular it would be beneficial to determine under which conditions the overlaps are the relevant order parameters and if our weak consistency derivation is indeed asymptotically exact.

Acknowledgments

This work is supported by the ERC under the European Union's Horizon 2020 Research and Innovation Program 714608-SMiLe, as well as by the French Agence Nationale de la Recherche under grant ANR-17-CE23-0023-01 PAIL. Additional funding is acknowledged by AB from 'Chaire de recherche sur les modèles et sciences des données', Fondation CFM pour la Recherche-ENS.

Appendix A. Proof of Proposition 2

A.1 Minimization of the EC Gibbs Free Energy

Let's first minimize the Bethe free energy at fixed moments $\mu_V = (\mu_i)_{i \in V}$:

$$F_{\mu_V} = \min_{(\tilde{p}_F, \tilde{p}_V) \in \mathcal{M}_{\mu_V}} \mathcal{F}_{\text{Bethe}}[\tilde{p}_F, \tilde{p}_V] \tag{171}$$

where $\mathcal{M}_{\mu_V}$ is the set of factor and variable marginals at fixed moment:

$$\mathcal{M}_{\mu_V} = \{(\tilde{p}_F, \tilde{p}_V) \quad : \quad \forall (i,k) \in E, \quad \mathbb{E}_{\tilde{p}_i} \phi_i(x_i) = \mathbb{E}_{\tilde{p}_k} \phi_i(x_i) = \mu_i\}. \tag{172}$$

The solution will be a stationary point of the Lagrangian

$$\begin{aligned} &\mathcal{L}[\tilde{p}_F, \tilde{p}_V, \lambda_F, \lambda_V] = \mathcal{F}_{\text{Bethe}}[\tilde{p}_F, \tilde{p}_V] \\ &+ \sum_{k \in F} \sum_{i \in \partial k} \langle \lambda_{i \to k}, \mu_i - \mathbb{E}_{\tilde{p}_k} \phi_i(x_i) \rangle + \sum_{i \in V} (1 - n_i) \langle \lambda_i, \mu_i - \mathbb{E}_{\tilde{p}_i} \phi_i(x_i) \rangle \end{aligned} \tag{173}$$

with Lagrange multipliers $\lambda_{i \to k}$ and λ_i associated to the moment constraint $\mathbb{E}_{\tilde{p}_k} \phi_i(x_i) = \mu_i$ and $\mathbb{E}_{\tilde{p}_i} \phi_i(x_i) = \mu_i$. Then $0 = \delta_{\tilde{p}_i} \mathcal{L} = \delta_{\tilde{p}_k} \mathcal{L} = \partial_{\lambda_{i \to k}} \mathcal{L} = \partial_{\lambda_i} \mathcal{L}$ leads to the solution:

$$\tilde{p}_k(x_k) = p_k(x_k \mid \lambda_k) = f(x_k; y_k)\, e^{\langle \lambda_k, \phi_k(x_k) \rangle - A_k[\lambda_k]}, \tag{174}$$

$$\tilde{p}_i(x_i) = p_i(x_i \mid \lambda_i) = e^{\langle \lambda_i, \phi_i(x_i) \rangle - A_i[\lambda_i]}, \tag{175}$$

$$\mu_i = \mu_i[\lambda_i] = \mu_i^k[\lambda_k]. \tag{176}$$

Besides the minimal F_{μ_V} is equal to:

$$\begin{aligned} F_{\mu_V} &= \sum_{k \in F} F_k[\tilde{p}_k] + \sum_{i \in V} (1 - n_i) F_i[\tilde{p}_i] \\ &= \sum_{k \in F} \text{KL}[\tilde{p}_k \| f_k] + \sum_{i \in V} (1 - n_i)(-\text{H}[\tilde{p}_i]) \\ &= \sum_{k \in F} \text{KL}[p_k(x_k \mid \lambda_k) \| f_k(x_k; y_k)] + \sum_{i \in V} (1 - n_i)(-\text{H}[p_i(x_i | \lambda_i)]) \\ &= \sum_{k \in F} G[\mu_k] + \sum_{i \in V} (1 - n_i) G_i[\mu_i] \\ &= G[\mu_V], \end{aligned}$$

due to the definition Eq. (10) of the Bethe free energy, the fact that the solutions $\tilde{p}_k$ and $\tilde{p}_i$ belong to the exponential families Eqs (174)-(175), and the definitions Eqs (26)-(27) of G_k and G_i. But then the relaxed Bethe variational problem Eq. (13) can be written as:

$$F_\phi(\mathbf{y}) = \min_{\mu_V} \min_{(\tilde{p}_F, \tilde{p}_V) \in \mathcal{M}_{\mu_V}} \mathcal{F}_{\text{Bethe}}[\tilde{p}_F, \tilde{p}_V] = \min_{\mu_V} F_{\mu_V} = \min_{\mu_V} G[\mu_V]. \tag{177}$$

Besides minimizing $G[\mu_V]$ leads to:

$$0 = \partial_{\mu_i} G[\mu_V] = \sum_{k \in \partial i} \lambda_{i \to k} + (1 - n_i) \lambda_i \tag{178}$$

which is the natural parameter constraint. The solution is therefore the same as the EP fixed point Eq. (46).

A.2 Stationary Point of the EP Free Energy

Using the duality (Section 2.4) between natural parameters and moments:

$$
\begin{aligned}
\min_{\mu_V} G[\mu_V] &= \min_{\mu_V} \sum_{k\in F} G_k[\mu_k] + \sum_{i\in V}(1-n_i)G_i[\mu_i] \\
&= \min_{\mu_V} \sum_{k\in F} \max_{\lambda_k} \{\langle\lambda_k,\mu_k\rangle - A_k[\lambda_k]\} + \sum_{i\in V} \underbrace{(1-n_i)}_{\leq 0} \max_{\lambda_i} \{\langle\lambda_i,\mu_i\rangle - A_i[\lambda_i]\} \\
&= \min_{\mu_V} \min_{\lambda_V} \max_{\lambda_F} - \sum_{k\in F} A_k[\lambda_k] - \sum_{i\in V}(1-n_i)A_i[\lambda_i] \\
&+ \sum_{i\in V}\langle \sum_{k\in\partial i} \lambda_{i\to k} + (1-n_i)\lambda_i,\ \mu_i\rangle \\
&= \min_{\lambda_V} \max_{\lambda_F} -A[\lambda_V,\lambda_F] \quad \text{s.t} \quad \forall i\in V : (n_i-1)\lambda_i = \sum_{k\in\partial i}\lambda_{i\to k}.
\end{aligned} \tag{179}
$$

As a consistency check, let's directly derive the solution to Eq. (179). Consider the Lagrangian

$$
\mathcal{L}[\lambda_V,\lambda_F,\mu_V] = A[\lambda_V,\lambda_F] + \sum_{i\in V}\langle (n_i-1)\lambda_i - \sum_{k\in\partial i}\lambda_{i\to k},\ \mu_i\rangle \tag{180}
$$

with Lagrangian multiplier μ_i associated to the constraint $(n_i-1)\lambda_i = \sum_{k\in\partial i}\lambda_{i\to k}$. At a stationary point:

$$
\partial_{\mu_i}\mathcal{L} = 0 \implies (n_i-1)\lambda_i = \sum_{k\in\partial i}\lambda_{i\to k} \tag{181}
$$

$$
\partial_{\lambda_{i\to k}}\mathcal{L} = 0 \implies \mu_i = \mu_i^k[\lambda_k] \tag{182}
$$

$$
\partial_{\lambda_i}\mathcal{L} = 0 \implies \mu_i = \mu_i[\lambda_i] \tag{183}
$$

which is exactly the same as the EP fixed point Eq. (46). Furthermore the Lagrangian multiplier is the posterior moment μ_i.

A.3 Stationary Point of the Tree-AMP Free Energy

Finally, the optimization under constraint of $A[\lambda_V,\lambda_F]$ is equivalent to finding a stationary point of $A[\lambda_E]$ without any constraint. Indeed:

$$
\partial_{\lambda_{k\to i}} A[\lambda_E] = 0 \implies \mu_i[\lambda_i^k] = \mu_i[\lambda_i] \implies \lambda_i^k = \lambda_i \tag{184}
$$

$$
\partial_{\lambda_{i\to k}} A[\lambda_E] = 0 \implies \mu_i^k[\lambda_k] = \mu_i[\lambda_i^k] \tag{185}
$$

which implies the natural parameter constraint

$$
\sum_{k\in\partial i}\lambda_{i\to k} = \sum_{k\in\partial i}(\lambda_i^k - \lambda_{k\to i}) = n_i\lambda_i - \lambda_i = (n_i-1)\lambda_i \tag{186}
$$

and the moment matching $\mu_i^k[\lambda_k] = \mu_i[\lambda_i]$ defining the EP fixed point Eq. (46).

Appendix B. MAP Estimation

To yield the maximum at posterior (MAP) estimate, we follow the derivation given in Appendix A.1 of (Manoel et al., 2018) in the context of the TV-VAMP algorithm and generalize it to tree-structured models. As usual in the statistical physics literature (Mézard and Montanari, 2009) we introduce an inverse temperature β and consider the distribution:

$$p^{(\beta)}(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z^{(\beta)}(\mathbf{y})} \prod_{k \in F} f_k(x_k; y_k)^\beta = \frac{1}{Z^{(\beta)}(\mathbf{y})} e^{-\beta \sum_{k\in F} E_k(x_k; y_k)} \tag{187}$$

and take the zero temperature limit $\beta \to \infty$ to make the distribution concentrate around its mode. Indeed by the Laplace method:

$$F^{(\beta)}(\mathbf{y}) = -\frac{1}{\beta} \ln Z^{(\beta)}(\mathbf{y}) = -\frac{1}{\beta} \int d\mathbf{x} e^{-\beta E(\mathbf{y})} \xrightarrow{\beta\to\infty} F^{(\infty)}(\mathbf{y}) = \min_{\mathbf{x}} E(\mathbf{x}, \mathbf{y}) \tag{188}$$

where E is the total energy of the system:

$$E(\mathbf{x}, \mathbf{y}) = \sum_{k\in F} E_k(x_k; y_k). \tag{189}$$

The inference problem becomes an energy minimization problem, the minimizer $\mathbf{x}^* = \arg\min_{\mathbf{x}} E(\mathbf{x}, \mathbf{y}) = \arg\max_{\mathbf{x}} p(\mathbf{x} \mid \mathbf{y})$ is the MAP estimate and called the ground state in statistical physics (Mézard and Montanari, 2009).

B.1 Zero Temperature Limit of the Weak Consistency Derivation

At fixed β, the weak consistency approximation Proposition 2, here expressed for isotopic Gaussian beliefs, gives:

$$F_\phi^{(\beta)}(\mathbf{y}) = \min_{r_V^{(\beta)}, \tau_V^{(\beta)}} G^{(\beta)}[r_V^{(\beta)}, \tau_V^{(\beta)}] = \min \operatorname*{extr}_{a_E^{(\beta)}, b_E^{(\beta)}} -A^{(\beta)}[a_E^{(\beta)}, b_E^{(\beta)}] \tag{190}$$

The limiting behavior can easily be found by scaling the natural parameters $a^{(\beta)} = \beta a$ and $b^{(\beta)} = \beta b$. By the Laplace method, the factor log-partitions are given by:

$$\begin{aligned} A_k^{(\beta)}[a_k^{(\beta)}, b_k^{(\beta)}] &= \frac{1}{\beta} \ln \int dx_k \, f_k(x_k; y_k)^\beta e^{-\frac{a_k^{(\beta)}}{2}\|x_k\|^2 + b_k^{(\beta)\intercal} x_k} \\ &= \frac{1}{\beta} \ln \int dx_k \, e^{-\beta\{E_k(x_k;y_k) + \frac{a_k}{2}\|x_k\|^2 - b_k^\intercal x_k\}} \\ &\xrightarrow{\beta\to\infty} A_k^{(\infty)}[a_k, b_k] = -\min_{x_k} E_k(x_k; y_k, a_k, b_k) \end{aligned}$$

where we introduce the tilted energy function:

$$E_k(x_k; y_k, a_k, b_k) = E_k(x_k; y_k) + \frac{a_k}{2}\|x_k\|^2 - b_k^\intercal x_k. \tag{191}$$

The limiting factor log-partition is thus given by:

$$A_k^{(\infty)}[a_k, b_k] = -E_k(x_k^*; y_k, a_k, b_k) \quad \text{with} \quad x_k^* = \arg\min_{x_k} E_k(x_k; y_k, a_k, b_k) \tag{192}$$

which can be alternatively expressed as:

$$A_k^{(\infty)}[a_k, b_k] = \frac{\|b_k\|^2}{2a_k} - \mathcal{M}_{\frac{1}{a_k}E_k(.;y_k)}\left(\frac{b_k}{a_k}\right), \quad x_k^* = \text{prox}_{\frac{1}{a_k}E_k(.;y_k)}\left(\frac{b_k}{a_k}\right) \tag{193}$$

using the Moreau envelop $\mathcal{M}_g(y) = \min_x\{g(x)+\frac{1}{2}\|x-y\|^2\}$ and proximal operator $\text{prox}_g(y) = \arg\min_x\{g(x) + \frac{1}{2}\|x - y\|^2\}$. The moment / natural parameter duality becomes:

$$r_i^{k(\beta)}[a_k^{(\beta)}, b_k^{(\beta)}] = \partial_{b_{i\to k}^{(\beta)}} \beta A_k^{(\beta)}[a_k^{(\beta)}, b_k^{(\beta)}] \xrightarrow{\beta\to\infty} r_i^{k(\infty)}[a_k, b_k] = \partial_{b_{i\to k}} A_k^{(\infty)}[a_k, b_k]$$

$$-\frac{N_i}{2}\tau_i^{k(\beta)}[a_k^{(\beta)}, b_k^{(\beta)}] = \partial_{a_{i\to k}^{(\beta)}} \beta A_k^{(\beta)}[a_k^{(\beta)}, b_k^{(\beta)}] \xrightarrow{\beta\to\infty} -\frac{N_i}{2}\tau_i^{k(\infty)}[a_k, b_k] = \partial_{a_{i\to k}} A_k^{(\beta)}[a_k, b_k]$$

As the precisions are scaled as $a^{(\beta)} = \beta a$, the variances must be scaled as $\beta v^{(\beta)} = v^{(\infty)}$ and:

$$\beta v_i^{k(\beta)}[a_k^{(\beta)}, b_k^{(\beta)}] = \beta\langle\partial^2_{b_{i\to k}^{(\beta)}} \beta A_k^{(\beta)}[a_k^{(\beta)}, b_k^{(\beta)}]\rangle \xrightarrow{\beta\to\infty} v_i^{k\infty}[a_k, b_k] = \langle\partial^2_{b_{i\to k}} A_k^{(\infty)}[a_k, b_k]\rangle.$$

where $\langle\cdot\rangle$ denotes the average over components of x_i. From Eq. (192), we then get the posterior mean and second moment:

$$r_i^{k(\infty)}[a_k, b_k] = \partial_{b_{i\to k}} A_k^{(\infty)}[a_k, b_k] = x_i^{k*}, \quad \tau_i^{k(\infty)}[a_k, b_k] = -\frac{2}{N_i}\partial_{a_{i\to k}} A_k^{(\infty)}[a_k, b_k] = \frac{\|x_i^{k*}\|^2}{N_i}$$

in agreement with the concentration at the mode. Note that in the zero temperature limit the second-moment is $\tau_i^{(\infty)} = \frac{\|r_i^{(\infty)}\|^2}{N_i} \neq \frac{\|r_i^{(\infty)}\|^2}{N_i} + v_i^{(\infty)}$. Indeed $\beta v_i^{(\beta)} \xrightarrow{\beta\to\infty} v_i^{(\infty)}$ is the *scaled* variance, the actual variance $v_i^{(\beta)} \xrightarrow{\beta\to\infty} 0$. The variable log-partition corresponds to the $E_i(x_i) = 0$ special case. The variable log-partition, mean and (scaled) variance are explicitly given by:

$$A_i^{(\infty)}[a_i, b_i] = \frac{\|b_i\|^2}{2a_i}, \quad r_i^{(\infty)}[a_i, b_i] = \partial_{b_i} A_i^{(\infty)} = \frac{b_i}{a_i}, \quad v_i^{(\infty)}[a_i, b_i] = \langle\partial^2_{b_i} A_i^{(\infty)}\rangle = \frac{1}{a_i}. \tag{194}$$

B.2 Zero Temperature Limit of the EP Algorithm

Let us denote $\text{EP}^{(\beta)}$ the instance of the EP Algorithm 2 running with $a^{(\beta)}, b^{(\beta)}$ and $r^{(\beta)}, v^{(\beta)}$, which can be used to search a stationary point of $A^{(\beta)}[a_E^{(\beta)}, b_E^{(\beta)}]$. Then we have the well defined limit $\text{EP}^{(\beta)} \xrightarrow{\beta\to\infty} \text{EP}^{(\infty)}$ where $\text{EP}^{(\infty)}$ is the instance of the EP Algorithm 2 running with a, b and $r^{(\infty)}, v^{(\infty)}$, which can be used to search a stationary point of $A^{(\infty)}[a_E, b_E]$. At an EP fixed point we will have

$$x_i^* = r_i = r_i^k, \quad v_i = v_i^k, \quad (n_i - 1)a_i = \sum_{k\in\partial i} a_{i\to k}, \quad (n_i - 1)b_i = \sum_{k\in\partial i} b_{i\to k}, \tag{195}$$

and therefore:

$$\begin{aligned}
-A^{(\infty)}[a_E, b_E] &= -\sum_{k\in F} A_k^{(\infty)}[a_k, b_k] - \sum_{i\in V}(1-n_i)A_i^{(\infty)}[a_i, b_i] \\
&= \sum_{k\in F} E_k(x_k^*; y_k, a_k, b_k) + \sum_{i\in V}(1-n_i)E_i(x_i^*; a_i, b_i) \\
&= \sum_{k\in F} E_k(x_k^*; y_k) \\
&\quad - \sum_{i\in V} \frac{\|x_i^*\|^2}{2} \underbrace{\left(\sum_{k\in\partial i} a_{i\to k} + (1-n_i)a_i\right)}_{0} + \sum_{i\in V} x_i^{*\intercal} \underbrace{\left(\sum_{k\in\partial i} b_{i\to k} + (1-n_i)b_i\right)}_{0} \\
&= E(\mathbf{x}^*, \mathbf{y}).
\end{aligned}$$

B.3 Summary

To recap, the energy minimization / MAP estimation problem can be formulated as:

$$F_\phi^{(\infty)}(\mathbf{y}) = \min_{\mathbf{x}} E(\mathbf{x}, \mathbf{y}) = \min \operatorname*{extr}_{a_E, b_E} -A^{(\infty)}[a_E, b_E] \tag{196}$$

which is the $\beta \to \infty$ limit of Proposition 2. A stationary point of $A^{(\infty)}[a_E, b_E]$ can be searched by the EP$^{(\infty)}$ Algorithm, which is just Algorithm 2 using the MAP modules:

$$A_k^{(\infty)}[a_k, b_k] = \frac{\|b_k\|^2}{2a_k} - \mathcal{M}_{\frac{1}{a_k}E_k(.;y_k)}\left(\frac{b_k}{a_k}\right),$$

$$r_k^{(\infty)}[a_k, b_k] = \operatorname{prox}_{\frac{1}{a_k}E_k(.;y_k)}\left(\frac{b_k}{a_k}\right), \quad v_k^{(\infty)}[a_k, b_k] = \langle \partial_{b_k} r_k^{(\infty)}[a_k, b_k]\rangle, \tag{197}$$

$$A_i^{(\infty)}[a_i, b_i] = \frac{\|b_i\|^2}{2a_i}, \quad r_i^{(\infty)} = \partial_{b_i} A_i^{(\infty)} = \frac{b_i}{a_i}, \quad v_i^{(\infty)} = \partial_{b_i}^2 A_i^{(\infty)} = \frac{1}{a_i}. \tag{198}$$

Appendix C. Proof of Proposition 6

C.1 Decomposition of the SCGF

We recall from Eq. (91) that we can formally decompose the SCGF $A(n)$ as:

$$A(n) = A_N^{(n)} - A_N^{(0)}, \quad A_N^{(n)} = \frac{1}{N}\ln Z_N^{(n)}, \quad A_N^{(0)} = \frac{1}{N}\ln Z_N^{(0)}, \tag{199}$$

where $Z_N^{(0)}$ is the partition function of the teacher generative model:

$$p^{(0)}(\mathbf{x}^{(0)}, \mathbf{y}) = \frac{1}{Z_N^{(0)}} \prod_{k\in F} f_k^{(0)}(x_k^{(0)}; y_k), \tag{200}$$

and $Z_N^{(n)}$ is the partition function of the replicated system:

$$p^{(n)}(\{\mathbf{x}^{(a)}\}_{a=0}^n, \mathbf{y}) = \frac{1}{Z_N^{(n)}} \prod_{k\in F} \left\{ f_k^{(0)}(x_k^{(0)}; y_k) \prod_{a=1}^n f_k(x_k^{(a)}; y_k) \right\} \tag{201}$$

where $x^{(a)}$ for $a = 1 \cdots n$ denote the n replicas and $x^{(0)}$ the ground truth. Now the key observation is that both $Z_N^{(0)}$ and $Z_N^{(n)}$ are partition functions associated to a tree-structured model (and this is actually the same tree). In that case we know that the log-partition can be obtained exactly from the Bethe variational problem which is unfortunately intractable. However following Heskes et al. (2005) we can solve the relaxed Bethe variational problem by enforcing weak consistency (moment matching) instead of the full consistency of the marginals, as reviewed in Section 2 for the derivation of the EP algorithm.

We must nonetheless specify for which sufficient statistics we want to enforce moment-matching, which in that setting we interpret as identifying the relevant order parameters for the thermodynamic limit. In the following, to make the connection with previously derived replica formulas, we will assume that the relevant order parameters are the overlaps:

$$\phi(\mathbf{x}) = \left(\frac{x_i^{(a)} \cdot x_i^{(b)}}{N_i} \right)_{i \in V, 0 \le a \le b \le n} \tag{202}$$

In some settings other order parameters could be relevant, fortunately the following derivation could be easily extended to these cases.

C.2 Weak Consistency Derivation (Replicated System)

Solving the relaxed Bethe variational problem applied to the log-partition $A_N^{(n)} = \frac{1}{N} \ln Z_N^{(n)}$ of the replicated system, using the overlaps Eq. (202) as sufficient statistics, leads to:

$$-A_N^{(n)} = \min_{Q_V} G^{(n)}[Q_V] = \min \operatorname*{extr}_{\hat{Q}_E} -A^{(n)}[\hat{Q}_E]. \tag{203}$$

where the minimizer corresponds to the overlaps:

$$Q_V = (Q_i^{(ab)})_{i \in V, 0 \le a \le b \le n}, \quad Q_i^{(ab)} = \mathbb{E} \frac{x_i^{(a)} \cdot x_i^{(b)}}{N_i} \tag{204}$$

with corresponding dual natural parameter messages:

$$\hat{Q}_E = (\hat{Q}_{i \to k}^{(ab)}, \hat{Q}_{k \to i}^{(ab)})_{(i,k) \in E, 0 \le a \le b \le n} \tag{205}$$

The potentials satisfy the tree decomposition:

$$G^{(n)}[Q_V] = \sum_{k \in F} \alpha_k G_k^{(n)}[Q_k] + \sum_{i \in V} \alpha_i (1 - n_i) G_i^{(n)}[Q_i] \quad \text{with} \quad Q_k = (Q_i)_{i \in \partial k}, \tag{206}$$

$$A^{(n)}[\hat{Q}_E] = \sum_{k \in F} \alpha_k A_k^{(n)}[\hat{Q}_k] - \sum_{(i,k) \in E} \alpha_i A_i^{(n)}[\hat{Q}_i^k] + \sum_{i \in V} \alpha_i A_i^{(n)}[\hat{Q}_i]$$

$$\text{with} \quad \hat{Q}_k = (\hat{Q}_{i \to k})_{i \in \partial k} \quad \hat{Q}_i^k = \hat{Q}_{i \to k} + \hat{Q}_{k \to i}, \quad \hat{Q}_i = \sum_{k \in \partial i} \hat{Q}_{k \to i}. \tag{207}$$

The factor and variable log-partition are given by:

$$A_k^{(n)}[\hat{Q}_k] = \frac{1}{N_k} \ln \int dy_k dx_k^{(0)} f_k^{(0)}(x_k^{(0)}; y_k) \left[\prod_{a=1}^{n} dx_k^{(a)} f_k(x_k^{(a)}; y_k) \right] e^{\sum_{i \in \partial k} \sum_{0 \le a \le b \le n} \hat{Q}_{i \to k}^{(ab)} x_i^{(a)} \cdot x_i^{(b)}} \quad (208)$$

$$A_i^{(n)}[\hat{Q}_i] = \frac{1}{N_i} \ln \int dx_i^{(0)} \left[\prod_{a=1}^{n} dx_i^{(a)} \right] e^{\sum_{0 \le a \le b \le n} \hat{Q}_i^{(ab)} x_i^{(a)} \cdot x_i^{(b)}} \quad (209)$$

and their gradients give the dual mapping to the overlaps:

$$\alpha_i^k Q_i^{k(ab)} = \partial_{\hat{Q}_{i \to k}^{(ab)}} A_k^{(n)}, \quad Q_i^{(ab)} = \partial_{\hat{Q}_i^{(ab)}} A_i^{(n)}. \quad (210)$$

$G_k^{(n)}$ and $G_i^{(n)}$ are the corresponding Legendre transforms. Any stationary point of the potentials (not necessarily the global optima) is a fixed point:

$$Q_i^{k(ab)} = Q_i^{(ab)}, \quad \hat{Q}_i^{(ab)} = \sum_{k \in \partial i} \hat{Q}_{k \to i}^{(ab)}. \quad (211)$$

C.3 Weak Consistency Derivation (Teacher)

The weak consistency derivation of the log-partition $A_N^{(0)}$ is given in Proposition 5 and consistently corresponds to the $n = 0$ case (ignoring the replicas and only keeping the ground truth) of $A_N^{(n)}$. The teacher second moments and dual messages in $A_N^{(0)}$ correspond to the $a = b = 0$ overlaps and dual messages in $A_N^{(n)}$:

$$\tau_i^{(0)} = Q_i^{(00)} = \mathbb{E} \frac{\|x_i^{(0)}\|^2}{N_i}, \quad -\frac{1}{2} \hat{\tau}_{i \to k}^{(0)} = \hat{Q}_{i \to k}^{(00)}, \quad -\frac{1}{2} \hat{\tau}_{k \to i}^{(0)} = \hat{Q}_{k \to i}^{(00)}. \quad (212)$$

C.4 Weak Consistency Derivation of the SCGF

Now we can finally express the SCGF $A(n)$ in Eq. (199) by subtracting the weak consistency derivation of $A_N^{(0)}$ to the weak consistency derivation of $A_N^{(n)}$. We assume that the teacher second moments $\tau_V^{(0)} = Q_V^{(00)}$ are known (Section 3.2) and are now considered as fixed parameters. The SCGF obtained by weak consistency on the overlaps is given by:

$$-A(n) = \min_{Q_V^*} G(n)[Q_V^*] = \min \operatorname*{extr}_{\hat{Q}_E^*} -A(n)[\hat{Q}_E^*]. \quad (213)$$

where the minimizer corresponds to the overlaps (with the $ab = 00$ overlap omitted):

$$Q_V^* = (Q_i^{(ab)})_{i \in V, 0 \le a \le b \le n, ab \neq 00}, \quad Q_i^{(ab)} = \mathbb{E} \frac{x_i^{(a)} \cdot x_i^{(b)}}{N_i} \quad (214)$$

with corresponding dual natural parameter messages:

$$\hat{Q}_E^* = (\hat{Q}_{i \to k}^{(ab)}, \hat{Q}_{k \to i}^{(ab)})_{(i,k) \in E, 0 \le a \le b \le n, ab \neq 00} \quad (215)$$

The potentials satisfy the tree decomposition:

$$G(n)[Q^*_V] = \sum_{k\in F} \alpha_k G_k(n)[Q^*_k] + \sum_{i\in V} \alpha_i(1-n_i)G_i(n)[Q^*_i] \quad \text{with} \quad Q^*_k = (Q^*_i)_{i\in\partial k}, \tag{216}$$

$$A(n)[\hat{Q}^*_E] = \sum_{k\in F} \alpha_k A_k(n)[\hat{Q}^*_k] - \sum_{(i,k)\in E} \alpha_i A_i(n)[\hat{Q}^{k*}_i] + \sum_{i\in V} \alpha_i A_i(n)[\hat{Q}^*_i]$$

$$\text{with} \quad \hat{Q}^*_k = (\hat{Q}^*_{i\to k})_{i\in\partial k} \quad \hat{Q}^{k*}_i = \hat{Q}^*_{i\to k} + \hat{Q}^*_{k\to i}, \quad \hat{Q}^*_i = \sum_{k\in\partial i} \hat{Q}^*_{k\to i}. \tag{217}$$

The factor and variable log-partition are now given by:

$$A_k(n)[\hat{Q}^*_k] = \frac{1}{N_k} \ln \int dy_k dx^{(0)}_k$$

$$p^{(0)}_k(x^{(0)}_k, y_k \mid \hat{\tau}^{(0)}_k) \left[\prod_{a=1}^n dx^{(a)}_k f_k(x^{(a)}_k; y_k)\right] e^{\sum_{i\in\partial k}\sum_{0\le a\le b\le n, ab\neq 00} \hat{Q}^{(ab)}_{i\to k} x^{(a)}_i \cdot x^{(b)}_i} \tag{218}$$

$$A_i(n)[\hat{Q}^*_i] = \frac{1}{N_i} \ln \int dx^{(0)}_i p^{(0)}_i(x^{(0)}_i \mid \hat{\tau}^{(0)}_i) \left[\prod_{a=1}^n dx^{(a)}_i\right] e^{\sum_{0\le a\le b\le n, ab\neq 00} \hat{Q}^{(ab)}_i x^{(a)}_i \cdot x^{(b)}_i} \tag{219}$$

and their gradients give the dual mapping to the overlaps:

$$\alpha^k_i Q^{k(ab)}_i = \partial_{\hat{Q}^{(ab)}_{i\to k}} A^{(n)}_k, \quad Q^{(ab)}_i = \partial_{\hat{Q}^{(ab)}_i} A^{(n)}_i. \tag{220}$$

$G_k(n)$ and $G_i(n)$ are the corresponding Legendre transforms. Any stationary point of the potentials (not necessarily the global optima) is a fixed point:

$$Q^{k(ab)}_i = Q^{(ab)}_i, \quad \hat{Q}^{(ab)}_i = \sum_{k\in\partial i} \hat{Q}^{(ab)}_{k\to i}. \tag{221}$$

Proof It follows straightforwardly from $\hat{Q}^{(00)}_k = -\frac{1}{2}\hat{\tau}^{(0)}_k$ and Eq. (107):

$$A_k(n)[\hat{Q}^*_k] = A^{(n)}_k[\hat{Q}_k] - A^{(0)}_k[\hat{\tau}^{(0)}_k]$$

$$= \frac{1}{N_k} \ln \int dy_k dx^{(0)}_k \underbrace{\frac{1}{Z^{(0)}_k[\hat{\tau}^{(0)}_k]} f^{(0)}_k(x^{(0)}_k; y_k) e^{-\frac{1}{2}\sum_{i\in\partial k} \hat{\tau}^0_{i\to k}\|x^{(0)}_i\|^2}}_{p^{(0)}_k(x^{(0)}_k, y_k|\hat{\tau}^{(0)}_k)} \left[\prod_{a=1}^n dx^{(a)}_k f_k(x^{(a)}_k; y_k)\right]$$

$$e^{\sum_{i\in\partial k}\sum_{0\le a\le b\le n, ab\neq 00} \hat{Q}^{(ab)}_{i\to k} x^{(a)}_i \cdot x^{(b)}_i}$$

Similarly it follows straightforwardly from $\hat{Q}^{(00)}_i = -\frac{1}{2}\hat{\tau}^{(0)}_i$ and Eq. (108):

$$A_i(n)[\hat{Q}^*_i] = A^{(n)}_i[\hat{Q}_i] - A^{(0)}_i[\hat{\tau}^{(0)}_i]$$

$$= \frac{1}{N_i} \ln \int dx^{(0)}_i \underbrace{\frac{1}{Z^{(0)}_i[\hat{\tau}^{(0)}_i]} e^{-\frac{1}{2}\hat{\tau}^0_i\|x^{(0)}_i\|^2}}_{p^{(0)}_i(x^{(0)}_i|\hat{\tau}^0_i)} \left[\prod_{a=1}^n dx^{(a)}_i\right] e^{\sum_{0\le a\le b\le n, ab\neq 00} \hat{Q}^{(ab)}_i x^{(a)}_i \cdot x^{(b)}_i}$$

■

C.5 Replica Symmetric SCGF

We take the replica symmetric ansatz (Mézard et al., 1987) for the overlaps ($1 \leq a < b \leq n$):

$$\tau = Q^{(aa)},\ m = Q^{(0a)},\ q = Q^{(ab)}, \quad -\frac{1}{2}\hat{\tau} = \hat{Q}^{(aa)},\ \hat{m} = \hat{Q}^{(0a)},\ \hat{q} = \hat{Q}^{(ab)}. \tag{222}$$

The replica symmetric SCGF $A(n)$ is given by:

$$-A(n) = \min_{m_V, q_V, \tau_V} G(n)[m_V, q_V, \tau_V] = \min \operatorname*{extr}_{\hat{m}_E, \hat{q}_E, \hat{\tau}_E} -A(n)[\hat{m}_E, \hat{q}_E, \hat{\tau}_E]. \tag{223}$$

The potentials satisfy the tree decomposition:

$$\begin{aligned}
&G(n)[m_V, q_V, \tau_V] = \sum_{k \in F} \alpha_k G_k(n)[m_k, q_k, \tau_k] + \sum_{i \in V} \alpha_i (1 - n_i) G_i(n)[m_i, q_i, \tau_i] \\
&\text{with} \quad m_k = (m_i)_{i \in \partial k} \quad (\text{idem } q, \tau),
\end{aligned} \tag{224}$$

$$\begin{aligned}
&A(n)[\hat{m}_E, \hat{q}_E, \hat{\tau}_E] \\
&= \sum_{k \in F} \alpha_k A_k(n)[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] - \sum_{(i,k) \in E} \alpha_i A_i(n)[\hat{m}_i^k, \hat{q}_i^k, \hat{\tau}_i^k] + \sum_{i \in V} \alpha_i A_i(n)[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] \\
&\text{with} \quad \hat{m}_k = (\hat{m}_{i \to k})_{i \in \partial k}, \quad \hat{m}_i^k = \hat{m}_{i \to k} + \hat{m}_{k \to i}, \quad \hat{m}_i = \sum_{k \in \partial i} \hat{m}_{k \to i} \quad (\text{idem } q, \tau).
\end{aligned} \tag{225}$$

The factor and variable log-partition are given by:

$$A_k(n)[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] = \frac{1}{N_k} \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} e^{N_k n A_k[a_k, b_k; y_k]} \tag{226}$$

$$A_i(n)[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] = \frac{1}{N_i} \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} e^{N_i n A_i[a_i, b_i]} \tag{227}$$

taken with:

$$p_k^{(0)}(x_k^{(0)}, y_k, b_k) = \mathcal{N}(b_k \mid \hat{m}_k x_k^{(0)}, \hat{q}_k)\ p_k^{(0)}(x_k^{(0)}, y_k; \hat{\tau}_k^{(0)}) \qquad \text{and} \quad a_k = \hat{\tau}_k + \hat{q}_k, \tag{228}$$

$$p_i^{(0)}(x_i^{(0)}, b_i) = \mathcal{N}(b_i \mid \hat{m}_i x_i^{(0)}, \hat{q}_i)\ p_i^{(0)}(x_i^{(0)}; \hat{\tau}_i^{(0)}) \qquad \text{and} \quad a_i = \hat{\tau}_i + \hat{q}_i, \tag{229}$$

and where $A_k[a_k, b_k; y_k]$ and $A_i[a_i, b_i]$ are the scaled EP log-partitions:

$$A_k[a_k, b_k; y_k] = \frac{1}{N_k} \ln \int dx_k\, f(x_k; y_k)\, e^{-\frac{1}{2} a_k \|x_k\|^2 + b_k^\intercal x_k} \tag{230}$$

$$A_i[a_i, b_i] = \frac{1}{N_i} \ln \int dx_i\, e^{-\frac{1}{2} a_i \|x_i\|^2 + b_i^\intercal x_i}. \tag{231}$$

We have therefore the following large deviation theory interpretation:

$$A(n) = \text{SCGF of } A_N(\mathbf{y}) \tag{232}$$

$$A_k(n) = \text{SCGF of } A_k[a_k, b_k; y_k] \tag{233}$$

$$A_i(n) = \text{SCGF of } A_i[a_i, b_i] \tag{234}$$

Proof The replica symmetric ansatz gives:

$$\sum_{0\leq a\leq b\leq n, ab\neq 00} \hat{Q}_{i\to k}^{(ab)} x_i^{(a)} \cdot x_i^{(b)} \Big|_{\mathrm{RS}}$$
$$= \hat{m}_{i\to k} x_i^{(0)} \cdot \sum_{a=1}^{n} x_i^{(a)} + \hat{q}_{i\to k} \sum_{1\leq a<b\leq n} x_i^{(a)} \cdot x_i^{(b)} - \frac{\hat{\tau}_{i\to k}}{2} \sum_{a=1}^{n} \|x_i^{(a)}\|^2$$
$$= \hat{m}_{i\to k} x_i^{(0)} \cdot \sum_{a=1}^{n} x_i^{(a)} + \frac{\hat{q}_{i\to k}}{2} \|\sum_{a=1}^{n} x_i^{(a)}\|^2 - \frac{\hat{\tau}_{i\to k} + \hat{q}_{i\to k}}{2} \sum_{a=1}^{n} \|x_i^{(a)}\|^2$$

Using the Gaussian identity:

$$e^{\frac{\hat{q}_{i\to k}}{2}\|\sum_{a=1}^{n} x_i^{(a)}\|^2} = \int_{\mathbb{R}^{N_i}} d\xi_{i\to k} \mathcal{N}(\xi_{i\to k}) e^{\sqrt{\hat{q}_{i\to k}}\xi_{i\to k}\cdot\sum_{a=1}^{n} x_i^{(a)}}$$

we can express the replica symmetric $A_k(n)[\hat{Q}_k^*]$ as:

$$A_k(n)[\hat{Q}_k^*]\Big|_{\mathrm{RS}}$$
$$= \frac{1}{N_k} \ln \int dy_k dx_k^{(0)} p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)}) \left[\prod_{i\in\partial k} d\xi_{i\to k}\mathcal{N}(\xi_{i\to k})\right] \left[\prod_{a=1}^{n} dx_k^{(a)} f_k(x_k^{(a)}; y_k)\right]$$
$$\prod_{a=1}^{n} e^{\sum_{i\in\partial k}[\hat{m}_{i\to k} x_i^{(0)} + \sqrt{\hat{q}_{i\to k}}\xi_{i\to k}]\cdot x_i^{(a)} - \frac{\hat{\tau}_{i\to k}+\hat{q}_{i\to k}}{2}\|x_i^{(a)}\|^2}$$
$$= \frac{1}{N_k} \ln \int dy_k dx_k^{(0)} p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)}) \left[\prod_{i\in\partial k} db_{i\to k}\mathcal{N}(b_{i\to k} \mid \hat{m}_{i\to k} x_i^{(0)}, \hat{q}_{i\to k})\right]$$
$$\left[\prod_{a=1}^{n} dx_k^{(a)} f_k(x_k^{(a)}; y_k) e^{\sum_{i\in\partial k} -\frac{a_{i\to k}}{2}\|x_i^{(a)}\|^2 + b_{i\to k}\cdot x_i^{(a)}}\right]$$
$$\text{with} \quad b_{i\to k} = \hat{m}_{i\to k} x_i^{(0)} + \sqrt{\hat{q}_{i\to k}}\xi_{i\to k}, \quad a_{i\to k} = \hat{\tau}_{i\to k} + \hat{q}_{i\to k}$$
$$= \frac{1}{N_k} \ln \int dy_k dx_k^{(0)} db_k \mathcal{N}(b_k \mid \hat{m}_k x_k^{(0)}, \hat{q}_k) p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)}) \prod_{a=1}^{n} Z_k[a_k, b_k; y_k]$$
$$\text{with} \quad Z_k[a_k, b_k; y_k] = \int dx_k\, f(x_k; y_k)\, e^{\sum_{i\in\partial k} -\frac{1}{2}a_{i\to k}\|x_i\|^2 + b_{i\to k}\cdot x_k}$$
$$= \frac{1}{N_k} \ln \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} Z_k[a_k, b_k; y_k]^n$$
$$= \frac{1}{N_k} \ln \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} e^{N_k n A_k[a_k, b_k; y_k]} \quad \text{with} \quad A_k[a_k, b_k; y_k] = \frac{1}{N_k} \ln Z_k[a_k, b_k; y_k]$$
$$= A_k(n)[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$$

The proof is identical for the replica symmetric $A_i(n)[\hat{Q}_i^*]$. ■

C.6 Replica Symmetric Free Entropy

We recall from Eq. (90) that the ensemble average $\bar{A}$ can be obtained by:

$$\bar{A} = \frac{d}{dn} A(n) \Big|_{n=0} \tag{235}$$

Taking the derivative Eq. (235) of the replica symmetric $A(n)$ we get Proposition 6 with:

$$\frac{d}{dn} A_k(n)[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] \Big|_{n=0} = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} A_k[a_k, b_k; y_k] = \bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k]$$
$$\frac{d}{dn} A_i(n)[\hat{m}_i, \hat{q}_i, \hat{\tau}_i] \Big|_{n=0} = \mathbb{E}_{p_i^{(0)}(x_i^{(0)}, b_i)} A_i[a_i, b_i] = \bar{A}_i[\hat{m}_i, \hat{q}_i, \hat{\tau}_i]$$

as expected from the large deviation theory interpretation Eqs (233)-(234).

Appendix D. Proof of Proposition 8

We recall that the local teacher generative model Eq. (154) is given by:

$$p_k^{(0)}(x_k^{(0)}, y_k, b_k) = \mathcal{N}(b_k \mid \hat{m}_k x_k^{(0)}, \hat{q}_k)\, p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)}),$$

while the local student generative model Eq. (155) is given by:

$$p_k(x_k, y_k, b_k) = \mathcal{N}(b_k \mid \hat{q}_k x_k, \hat{q}_k)\, p_k(x_k, y_k \mid \hat{\tau}_k).$$

Interestingly by rescaling $b_k^{(0)} = \frac{\hat{m}_k}{\hat{q}_k} b_k$ the local teacher generative model is equal to the Bayes-optimal setting generative model Eq. (157):

$$p_k^{(0)}(x_k^{(0)}, y_k, b_k^{(0)}) = \mathcal{N}(b_k^{(0)} \mid \hat{m}_k^{(0)} x_k^{(0)}, \hat{m}_k^{(0)})\, p_k^{(0)}(x_k^{(0)}, y_k \mid \hat{\tau}_k^{(0)}) \quad \text{with} \quad \hat{m}_k^{(0)} = \frac{\hat{m}_k^2}{\hat{q}_k}.$$

Lemma 9 *The RS potential Eq.* (119) *is related to the cross-entropy between the local teacher evidence* $p_k^{(0)}(y_k, b_k)$ *and the local student evidence* $p_k(y_k, b_k)$*:*

$$\begin{aligned}&\bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] - A_k[\hat{\tau}_k] \\ &= -\frac{1}{N_k} H[p_k^{(0)}(y_k, b_k), p_k(y_k, b_k)] + \sum_{i \in \partial k} \frac{\alpha_i^k \hat{m}_{i\to k}^{(0)} \tau_i^{(0)}}{2} + \frac{\alpha_i^k}{2} \ln 2\pi e \hat{q}_{i\to k}.\end{aligned} \tag{236}$$

Similarly, the BO potential Eq. (141) *is related to the entropy of the local Bayes-optimal evidence* $p_k^{(0)}(y_k, b_k^{(0)})$*:*

$$\bar{A}_k^{(0)}[\hat{m}_k^{(0)}] - A_k^{(0)}[\hat{\tau}_k^{(0)}] = -\frac{1}{N_k} H[p_k^{(0)}(y_k, b_k^{(0)})] + \sum_{i \in \partial k} \frac{\alpha_i^k \hat{m}_{i\to k}^{(0)} \tau_i^{(0)}}{2} + \frac{\alpha_i^k}{2} \ln 2\pi e \hat{m}_{i\to k}^{(0)}. \tag{237}$$

Proof The local student evidence is equal to:

$$
\begin{aligned}
p_k(y_k, b_k) &= \int dx_k p_k(x_k, y_k, b_k) = \int dx_k p_k(x_k, y_k \mid \hat{\tau}_k) \mathcal{N}(b_k \mid \hat{q}_k x_k, \hat{q}_k) \\
&= \int dx_k f_k(x_k; y_k) e^{-\frac{1}{2}\hat{\tau}_k \|x_k\|^2 - N_k A_k[\hat{\tau}_k]} e^{-\frac{1}{2}\hat{q}_k \|x_k\|^2 + b_k^\intercal x_k} \mathcal{N}(b_k \mid 0, \hat{q}_k) \\
&= \int dx_k f_k(x_k; y_k) e^{-\frac{1}{2} a_k \|x_k\|^2 + b_k^\intercal x_k - N_k A_k[\hat{\tau}_k]} \mathcal{N}(b_k \mid 0, \hat{q}_k) \quad \text{with} \quad a_k = \hat{\tau}_k + \hat{q}_k \\
&= e^{N_k A_k[a_k, b_k; y_k] - N_k A_k[\hat{\tau}_k]} \mathcal{N}(b_k \mid 0, \hat{q}_k)
\end{aligned}
$$

Then:

$$
\begin{aligned}
&\bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] - A_k[\hat{\tau}_k] = \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} A_k[a_k, b_k; y_k] - A_k[\hat{\tau}_k] \\
&= \frac{1}{N_k} \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} \ln \frac{p_k(y_k, b_k)}{\mathcal{N}(b_k \mid 0, \hat{q}_k)} \\
&= -\frac{1}{N_k} H[p_k^{(0)}(y_k, b_k), p_k(y_k, b_k)] + \frac{1}{N_k} \mathbb{E}_{p_k^{(0)}(x_k^{(0)}, y_k, b_k)} \sum_{i \in \partial k} \frac{\|b_{i\to k}\|^2}{2\hat{q}_{i\to k}} + \frac{N_i}{2} \ln 2\pi \hat{q}_{i\to k} \\
&= -\frac{1}{N_k} H[p_k^{(0)}(y_k, b_k), p_k(y_k, b_k)] + \frac{1}{N_k} \sum_{i \in \partial k} \frac{N_i(\hat{m}_{i\to k}^2 \tau_i^{(0)} + \hat{q}_{i\to k})}{2\hat{q}_{i\to k}} + \frac{N_i}{2} \ln 2\pi \hat{q}_{i\to k} \\
&= -\frac{1}{N_k} H[p_k^{(0)}(y_k, b_k), p_k(y_k, b_k)] + \sum_{i \in \partial k} \frac{\alpha_i^k \hat{m}_{i\to k}^{(0)} \tau_i^{(0)}}{2} + \frac{\alpha_i^k}{2} \ln 2\pi e \hat{q}_{i\to k}
\end{aligned}
$$

with $\hat{m}_{i\to k}^{(0)} = \frac{\hat{m}_{i\to k}^2}{\hat{q}_{i\to k}}$ and $\alpha_i^k = \frac{N_i}{N_k}$ which proves Eq. (236). The proof for the BO potential Eq. (237) is identical to the RS case with the simplifications $\hat{m}_k^{(0)} = \hat{q}_k = \hat{m}_k$ and $\hat{\tau}_k = \hat{\tau}_k^{(0)}$. ■

From Lemma 9 we easily recover the decomposition Eq. (159) for the RS potential:

$$
\bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] - A_k[\hat{\tau}_k] = \bar{A}_k^{(0)}[\hat{m}_k^{(0)}] - A_k^{(0)}[\hat{\tau}_k^{(0)}] - \frac{1}{N_k} \mathrm{KL}[p_k^{(0)}(y_k, b_k) \| p_k(y_k, b_k)]
$$

and Eq. (160) for the BO potential:

$$
\bar{A}_k^{(0)}[\hat{m}_k^{(0)}] - A_k^{(0)}[\hat{\tau}_k^{(0)}] = \sum_{i \in \partial k} \frac{\alpha_i^k \hat{m}_{i\to k}^{(0)} \tau_i^{(0)}}{2} + \frac{1}{N_k} H[p_k^{(0)}(b_k^{(0)} \mid x_k^{(0)})] - \frac{1}{N_k} H[p_k^{(0)}(y_k, b_k^{(0)})].
$$

Proof The proof for the BO case is straightforward:

$$
p_k^{(0)}(b_k^{(0)} \mid x_k^{(0)}) = \mathcal{N}(b_k^{(0)} \mid \hat{m}_k^{(0)} x_k^{(0)}, \hat{m}_k^{(0)}) \implies \frac{1}{N_k} H[p_k^{(0)}(b_k^{(0)} \mid x_k^{(0)})] = \sum_{i \in \partial k} \frac{\alpha_i^k}{2} \ln 2\pi e \hat{m}_{i\to k}^{(0)}.
$$

In the RS case, due to the rescaling $b_k^{(0)} = \frac{\hat{m}_k}{\hat{q}_k} b_k$ we have for the densities:

$$
\frac{1}{N_k} \ln p_k^{(0)}(y_k, b_k^{(0)}) = \frac{1}{N_k} \ln p_k^{(0)}(y_k, b_k) + \sum_{i \in \partial k} \alpha_i^k \ln \frac{\hat{q}_{i\to k}}{\hat{m}_{i\to k}}
$$

which gives for the entropies:

$$-\frac{1}{N_k}H[p_k^{(0)}(y_k, b_k^{(0)})] = -\frac{1}{N_k}H[p_k^{(0)}(y_k, b_k)] + \sum_{i\in\partial k} \alpha_i^k \ln \frac{\hat{q}_{i\to k}}{\hat{m}_{i\to k}}$$

Then Eq. (237) at $\hat{m}_k^{(0)} = \frac{\hat{m}_k^2}{\hat{q}_k}$ can be written:

$$\begin{aligned}
&\bar{A}_k^{(0)}[\hat{m}_k^{(0)}] - A_k^{(0)}[\hat{\tau}_k^{(0)}] \\
&= -\frac{1}{N_k}H[p_k^{(0)}(y_k, b_k)] + \sum_{i\in\partial k} \alpha_i^k \ln \frac{\hat{q}_{i\to k}}{\hat{m}_{i\to k}} + \sum_{i\in\partial k} \frac{\alpha_i^k \hat{m}_{i\to k}^{(0)} \tau_i^{(0)}}{2} + \frac{\alpha_i^k}{2} \ln 2\pi e \frac{\hat{m}_{i\to k}^2}{\hat{q}_{i\to k}} \\
&= -\frac{1}{N_k}H[p_k^{(0)}(y_k, b_k)] + \sum_{i\in\partial k} \frac{\alpha_i^k \hat{m}_{i\to k}^{(0)} \tau_i^{(0)}}{2} + \frac{\alpha_i^k}{2} \ln 2\pi e \hat{q}_{i\to k}
\end{aligned}$$

By using Eq. (236) we finally get:

$$\begin{aligned}
&\bar{A}_k[\hat{m}_k, \hat{q}_k, \hat{\tau}_k] - A_k[\hat{\tau}_k] \\
&= \bar{A}_k^{(0)}[\hat{m}_k^{(0)}] - A_k^{(0)}[\hat{\tau}_k^{(0)}] - \frac{1}{N_k}H[p_k^{(0)}(y_k, b_k), p_k(y_k, b_k)] + \frac{1}{N_k}H[p_k^{(0)}(y_k, b_k)] \\
&= \bar{A}_k^{(0)}[\hat{m}_k^{(0)}] - A_k^{(0)}[\hat{\tau}_k^{(0)}] - \frac{1}{N_k}\mathrm{KL}[p_k^{(0)}(y_k, b_k) \| p_k(y_k, b_k)]
\end{aligned}$$

■

Appendix E. Tree-AMP Modules

E.1 Variable

The Tree-AMP package only implements isotropic Gaussian beliefs, but the variable log-partitions presented here will be useful to derive the factor modules.

E.1.1 General Variable

An approximate belief, which we may as well call a variable type, is specified by the base space X as well as the chosen set of sufficient statistics $\phi(x)$. Any variable type defines an exponential family distribution

$$p(x \mid \lambda) = e^{\langle \lambda, \phi(x) \rangle - A[\lambda]} \tag{238}$$

indexed by the natural parameter λ. The family can be alternatively indexed by the moments $\mu = \mathbb{E}_{p(x|\lambda)}\phi(x)$. The log-partition

$$A[\lambda] = \ln \int_X dx e^{\lambda^\intercal \phi(x)} \tag{239}$$

provides the bijective mapping between the natural parameters and the moments:

$$\mu[\lambda] = \partial_\lambda A[\lambda]. \tag{240}$$

For all the variable types considered below, we will always have $x \in \phi(x)$ in the set of sufficient statistics. Its associated natural parameter $b \in \lambda$ is thus dual to the mean $r \in \mu$. The mean and variance are given by:

$$r[\lambda] = \partial_b A[\lambda], \qquad v[\lambda] = \partial_b^2 A[\lambda], \tag{241}$$

We list below the log-partition, mean and variance for several variable types, which correspond to well known exponential family distributions.

E.1.2 Isotropic Gaussian Variable

For $x \in \mathbb{R}^N$, sufficient statistics $\phi(x) = (x, -\frac{1}{2}x^\intercal x)$, natural parameters $\lambda = (b, a)$ with $b \in \mathbb{R}^N$ and scalar precision $a \in \mathbb{R}$.

$$A[a,b] = \ln \int dx\, e^{-\frac{1}{2}ax^\intercal x + b^\intercal x} = \frac{\|b\|^2}{2a} + \frac{N}{2}\ln\frac{2\pi}{a}, \tag{242}$$

$$r[a,b] = \frac{b}{a}, \qquad v[a,b] = \frac{1}{a} \in \mathbb{R}. \tag{243}$$

The corresponding exponential family is the isotropic multivariate Normal:

$$p(x \mid a,b) = \mathcal{N}(x \mid r, v) \quad \text{with} \quad r = \frac{b}{a}, \quad v = \frac{1}{a}. \tag{244}$$

E.1.3 Diagonal Gaussian Variable

For $x \in \mathbb{R}^N$, sufficient statistics $\phi(x) = (x, -\frac{1}{2}x^2)$, natural parameters $\lambda = (b, a)$ with $b \in \mathbb{R}^N$ and diagonal precision $a \in \mathbb{R}^N$.

$$A[a,b] = \ln \int dx\, e^{-\frac{1}{2}x^\intercal ax + b^\intercal x} = \sum_{n=1}^{N} \frac{b_n^2}{2a_n} + \frac{1}{2}\ln\frac{2\pi}{a_n}, \tag{245}$$

$$r[a,b] = \frac{b}{a}, \qquad v[a,b] = \frac{1}{a} \in \mathbb{R}^N. \tag{246}$$

The corresponding exponential family is the diagonal multivariate Normal:

$$p(x \mid a,b) = \mathcal{N}(x \mid r, v) \quad \text{with} \quad r = \frac{b}{a}, \quad v = \frac{1}{a}. \tag{247}$$

E.1.4 Full Covariance Gaussian Variable

For $x \in \mathbb{R}^N$, sufficient statistics $\phi(x) = (x, -\frac{1}{2}xx^\intercal)$, natural parameters $\lambda = (b, a)$ with $b \in \mathbb{R}^N$ and matrix precision $a \in \mathbb{R}^{N\times N}$.

$$A[a,b] = \ln \int dx\, e^{-\frac{1}{2}x^\intercal ax + b^\intercal x} = \frac{1}{2}b^\intercal a^{-1} b + \frac{1}{2}\ln\det 2\pi a^{-1}, \tag{248}$$

$$r[a,b] = \frac{b}{a}, \qquad \Sigma[a,b] = a^{-1} \in \mathbb{R}^{N\times N}, \tag{249}$$

The corresponding exponential family is the full covariance multivariate Normal:

$$p(x \mid a,b) = \mathcal{N}(x \mid r, \Sigma) \quad \text{with} \quad r = \frac{b}{a}, \quad \Sigma = a^{-1}. \tag{250}$$

E.1.5 Real Variable

For $x \in \mathbb{R}$, sufficient statistics $\phi(x) = (x, -\frac{1}{2}x^2)$, natural parameters $\lambda = (b, a)$.

$$A[a,b] = \ln \int dx\, e^{-\frac{1}{2}ax^2+bx} = \frac{b^2}{2a} + \frac{1}{2}\ln\frac{2\pi}{a}, \tag{251}$$

$$r[a,b] = \frac{b}{a}, \qquad v[a,b] = \frac{1}{a}. \tag{252}$$

The corresponding exponential family is the Normal:

$$p(x \mid a,b) = \mathcal{N}(x \mid r, v) \quad \text{with} \quad r = \frac{b}{a}, \quad v = \frac{1}{a}. \tag{253}$$

E.1.6 Binary Variable

For $x \in \pm$, sufficient statistics $\phi(x) = x$, natural parameter $\lambda = b$.

$$A[b] = \ln \sum_{x=\pm} e^{bx} = \ln(e^{+b} + e^{-b}), \tag{254}$$

$$r[b] = \tanh(b), \qquad v[b] = \frac{1}{\cosh(b)^2}. \tag{255}$$

The corresponding exponential family is the Bernoulli (over $\pm$):

$$p(x \mid b) = p_+\delta_{+1}(x) + p_-\delta_{-1}(x) \tag{256}$$

where the natural parameter $b = \frac{1}{2}\ln\frac{p_+}{p_-}$ is the log-odds.

E.1.7 Sparse Variable

For $x \in \mathbb{R} \cup \{0\}$, sufficient statistics $\phi(x) = (x, -\frac{1}{2}x^2, \delta(x))$, natural parameters $\lambda = (b, a, \eta)$. There is a finite probability that $x = 0$. The natural parameter η corresponding to the sufficient statistic $\delta(x)$ is dual to the fraction of zero elements $\kappa = \mathbb{E}\delta(x) = p(x = 0)$. The sparsity $\rho = 1 - \kappa = p(x \neq 0)$ is the fraction of non-zero elements.

$$A[a,b,\eta] = \ln\left[e^{\eta} + \int dx e^{-\frac{1}{2}ax^2+bx}\right] = \eta + \ln(1 + e^{\xi}) \quad \text{with} \quad \xi = A[a,b] - \eta, \tag{257}$$

$$r[a,b,\eta] = \frac{b}{a}\sigma(\xi), \qquad v[a,b,\eta] = \frac{1}{a}\sigma(\xi) + \frac{b^2}{a^2}\sigma(\xi)\sigma(-\xi), \qquad \rho[a,b,\eta] = \sigma(\xi), \tag{258}$$

where σ is the sigmoid function and the parameter ξ is the sparsity log-odds:

$$\xi = A[a,b] - \eta = \ln\frac{\sigma(\xi)}{\sigma(-\xi)} = \ln\frac{\rho}{1-\rho}. \tag{259}$$

The corresponding exponential family is the Gauss-Bernoulli:

$$p(x \mid a,b,\eta) = [1-\rho]\delta(x) + \rho\mathcal{N}(x \mid r,v) \quad \text{with} \quad r = \frac{b}{a}, \quad v = \frac{1}{a}, \quad \rho = \rho[a,b,\eta]. \tag{260}$$

E.1.8 Interval Variable

For $x \in X$, sufficient statistics $\phi(x) = (x, -\frac{1}{2}x^2)$, natural parameters $\lambda = (b, a)$, where $X = [x_{\min}, x_{\max}] \subset \mathbb{R}$ is a real interval. The probability that x belongs to X is equal to:

$$p_X[a,b] = \int_X dx \mathcal{N}(x \mid r, v) = \Phi(z_{\max}) - \Phi(z_{\min}), \tag{261}$$

$$z_{\min} = \frac{x_{\min} - r}{\sqrt{v}} = \frac{a x_{\min} - b}{\sqrt{a}}, \quad z_{\max} = \frac{x_{\max} - r}{\sqrt{v}} = \frac{a x_{\max} - b}{\sqrt{a}}, \tag{262}$$

where Φ is the cumulative Normal distribution and $z_{\min}$ and $z_{\max}$ are the z-scores of $x_{\min}$ and $x_{\max}$ for the Normal of mean $r = \frac{b}{a}$ and variance $v = \frac{1}{a}$. Then:

$$A_X[a,b] = \ln \int_X dx \, e^{-\frac{1}{2}ax^2 + bx} = A[a,b] + \ln p_X[a,b], \tag{263}$$

$$r_X[a,b] = \frac{b}{a} - \frac{1}{\sqrt{a}} \frac{\mathcal{N}(z_{\max}) - \mathcal{N}(z_{\min})}{\Phi(z_{\max}) - \Phi(z_{\min})}, \tag{264}$$

$$v_X[a,b] = \frac{1}{a} \left\{ 1 - \frac{z_{\max}\mathcal{N}(z_{\max}) - z_{\min}\mathcal{N}(z_{\min})}{\Phi(z_{\max}) - \Phi(z_{\min})} - \left[\frac{\mathcal{N}(z_{\max}) - \mathcal{N}(z_{\min})}{\Phi(z_{\max}) - \Phi(z_{\min})} \right]^2 \right\}. \tag{265}$$

The corresponding exponential family is the truncated Normal distribution:

$$p_X(x \mid a, b) = \frac{1}{p_X[a,b]} \mathcal{N}(x \mid r, v) \mathbf{1}_X(x) \quad \text{with} \quad r = \frac{b}{a}, \quad v = \frac{1}{a}. \tag{266}$$

E.1.9 Positive/Negative Variable

For $x \in \mathbb{R}_{\pm}$, sufficient statistics $\phi(x) = (x, -\frac{1}{2}x^2)$, natural parameters $\lambda = (b, a)$. It's a particular case of the interval variable with $X = \mathbb{R}_{\pm}$.

$$p_{\pm}[a,b] = \int_{\mathbb{R}_{\pm}} dx \mathcal{N}(x \mid a, b) = \Phi(z_{\pm}) \quad \text{with} \quad z_{\pm} = \pm \frac{b}{\sqrt{a}}, \tag{267}$$

$$A_{\pm}[a,b] = \ln \int_{\mathbb{R}_{\pm}} dx \, e^{-\frac{1}{2}ax^2 + bx} = A[a,b] + \ln p_{\pm}[a,b], \tag{268}$$

$$r_{\pm}[a,b] = \pm \frac{1}{\sqrt{a}} \left\{ z_{\pm} + \frac{\mathcal{N}(z_{\pm})}{\Phi(z_{\pm})} \right\}, \tag{269}$$

$$v_{\pm}[a,b] = \frac{1}{a} \left\{ 1 - \frac{z_{\pm}\mathcal{N}(z_{\pm})}{\Phi(z_{\pm})} - \frac{\mathcal{N}(z_{\pm})^2}{\Phi(z_{\pm})^2} \right\}. \tag{270}$$

The corresponding exponential family is the half Normal:

$$p_{\pm}(x \mid a, b) = \frac{1}{p_{\pm}[a,b]} \mathcal{N}(x \mid r, v) \mathbf{1}_{\mathbb{R}_{\pm}}(x) \quad \text{with} \quad r = \frac{b}{a}, \quad v = \frac{1}{a}. \tag{271}$$

E.1.10 Phase (Circular) Variable

For $x = e^{i\theta_x} \in \mathbb{S}^1$, sufficient statistics $\phi(x) = x$, natural parameter $\lambda = b$. Generally the von Mises distribution on the circle is defined over the angle $\theta_x \in [0, 2\pi[$ but we find it more convenient to define it over the phase $x = e^{i\theta_x} \in \mathbb{S}^1$. Then the natural parameter $b = |b|e^{i\theta_b} \in \mathbb{C}$ and:

$$A_{\mathbb{S}^1}[b] = \ln \int_{\mathbb{S}^1} dx\, e^{b^\intercal x} = \ln 2\pi I_0(|b|) \tag{272}$$

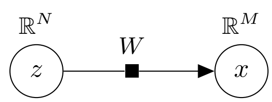
$$r_{\mathbb{S}^1}[b] = \frac{b}{|b|}\frac{I_1(|b|)}{I_0(|b|)}, \qquad v_{\mathbb{S}^1}[b] = \frac{1}{2}\left[1 - \frac{I_1(|b|)^2}{I_0(|b|)^2}\right], \tag{273}$$

where I_0 is the modified Bessel function of the first kind. For the natural parameter b, its modulus $|b|$ is called the concentration parameter and is analogous to the precision for a Gaussian, and its angle $\theta_b = \theta_r$ is the circular mean. The corresponding exponential family is the von Mises:

$$p(x \mid b) = \frac{e^{\kappa \cos(\theta_x - \mu)}}{2\pi I_0(\kappa)} \quad \text{with} \quad \kappa = |b|, \quad \mu = \theta_b. \tag{274}$$

E.2 Linear Channels

E.2.1 Generic Linear Channel



The factor $f(x,z) = p(x \mid z) = \delta(x - Wz)$ is the deterministic channel $x = Wz$. Unless explicitly specified, we will always consider isotropic Gaussian beliefs on both x and z. The log-partition is given by:

$$A_f[a_{z\to f}, b_{z\to f}, a_{x\to f}, b_{x\to f}] = \frac{1}{2}b^\intercal \Sigma b + \frac{1}{2}\ln\det 2\pi\Sigma \tag{275}$$

$$b = b_{z\to f} + W^\intercal b_{x\to f}, \quad a = a_{z\to f} + a_{x\to f}W^\intercal W, \quad \Sigma = a^{-1}. \tag{276}$$

The posterior means and variances are given by:

$$r_z^f = \Sigma b, \qquad v_z^f = \mathbb{E}_\lambda \frac{1}{a_{z\to f} + a_{x\to f}\lambda}, \tag{277}$$

$$r_x^f = W r_z^f, \qquad \alpha v_x^f = \mathbb{E}_\lambda \frac{\lambda}{a_{z\to f} + a_{x\to f}\lambda}, \tag{278}$$

where $\lambda = \mathrm{Spec}\, W^\intercal W$ denotes the spectrum of $W^\intercal W$ and $\alpha = \frac{M}{N}$ the aspect ratio of W. There is actually no need to explicitly compute the matrix inverse $\Sigma = a^{-1}$ at each update; it is more numerically efficient to use the SVD decomposition (see Section E.2.10). The variances satisfy:

$$a_{z\to f}v_z^f + \alpha a_{x\to f}v_x^f = 1, \tag{279}$$

$$\alpha a_{x\to f}v_x^f = 1 - a_{z\to f}v_z^f = n_{\text{eff}}, \tag{280}$$

where $n_{\text{eff}} = \mathbb{E}_\lambda \frac{a_{x\to f}\lambda}{a_{z\to f} + a_{x\to f}\lambda}$ is known as the effective number of parameters in Bayesian linear regression (Bishop, 2006).

E.2.2 Teacher Prior Second Moment

The log-partition for the teacher prior second moment is equal to:

$$A_f^{(0)}[\hat{\tau}_{z\to f}^{(0)}, \hat{\tau}_{x\to f}^{(0)}] = \frac{1}{2}\mathbb{E}_\lambda \ln \frac{2\pi}{\hat{\tau}_\lambda^{(0)}} \quad \text{with} \quad \hat{\tau}_\lambda^{(0)} = \hat{\tau}_{z\to f}^{(0)} + \lambda \hat{\tau}_{x\to f}^{(0)} \tag{281}$$

which yields the dual mapping:

$$\tau_z^{(0)} = \mathbb{E}_\lambda \tau_\lambda^{(0)}, \quad \alpha\tau_x^{(0)} = \mathbb{E}_\lambda \lambda \tau_\lambda^{(0)} \quad \text{with} \quad \tau_\lambda^{(0)} = \frac{1}{\hat{\tau}_\lambda^{(0)}}. \tag{282}$$

When the teacher factor graph is a Bayesian network (Section 3.2) we have

$$\hat{\tau}_{z\to f}^{(0)} = \frac{1}{\tau_z^{(0)}}, \quad \hat{\tau}_{x\to f}^{(0)} = 0, \quad \tau_\lambda^{(0)} = \tau_z^{(0)}, \quad A_f^{(0)} = \frac{1}{2}\ln 2\pi\tau_z^{(0)}. \tag{283}$$

E.2.3 Replica Symmetric

The RS potential is given by:

$$\begin{aligned}
&\bar{A}_f[\hat{m}_{z\to f}, \hat{q}_{z\to f}, \hat{\tau}_{z\to f}, \hat{m}_{x\to f}, \hat{q}_{x\to f}, \hat{\tau}_{x\to f}] = \mathbb{E}_\lambda \bar{A}_\lambda[\hat{m}_\lambda, \hat{q}_\lambda, \hat{\tau}_\lambda], \\
&\bar{A}_\lambda[\hat{m}_\lambda, \hat{q}_\lambda, \hat{\tau}_\lambda] = \frac{\hat{m}_\lambda^2 \tau_\lambda^{(0)} + \hat{q}_\lambda}{2a_\lambda} + \frac{1}{2}\ln\frac{2\pi}{a_\lambda}, \\
&\text{with} \quad a_\lambda = \hat{\tau}_\lambda + \hat{q}_\lambda \quad \text{and} \quad \hat{m}_\lambda = \hat{m}_{z\to f} + \lambda \hat{m}_{x\to f} \quad (\text{idem } \hat{q}, \hat{\tau}, a, \hat{\tau}^{(0)})
\end{aligned} \tag{284}$$

We recognize $\bar{A}_\lambda$ as the variable RS potential Eq. (134) which yields the dual mapping:

$$\begin{aligned}
&m_z^f = \mathbb{E}_\lambda m_\lambda, \quad \alpha m_x^f = \mathbb{E}_\lambda \lambda m_\lambda \quad (\text{idem } q, \tau, v) \\
&\text{with} \quad m_\lambda = \frac{\hat{m}_\lambda \tau_\lambda^{(0)}}{a_\lambda}, \quad q_\lambda = \frac{\hat{m}_\lambda^2 \tau_\lambda^{(0)} + \hat{q}_\lambda}{a_\lambda^2}, \quad \tau_\lambda = q_\lambda + v_\lambda, \quad v_\lambda = \frac{1}{a_\lambda}.
\end{aligned} \tag{285}$$

In particular we recover Eqs (277)-(278) for the variances.

E.2.4 Bayes-Optimal

The BO potential is given by:

$$\begin{aligned}
&\bar{A}_f^{(0)}[\hat{m}_{z\to f}, \hat{m}_{x\to f}] = \mathbb{E}_\lambda \bar{A}_\lambda[\hat{m}_\lambda], \quad \bar{A}_\lambda^{(0)}[\hat{m}_\lambda] = \frac{\hat{m}_\lambda \tau_\lambda^{(0)}}{2} + \frac{1}{2}\ln\frac{2\pi}{a_\lambda}, \\
&\text{with} \quad a_\lambda = \hat{\tau}_\lambda^{(0)} + \hat{m}_\lambda \quad \text{and} \quad \hat{m}_\lambda = \hat{m}_{z\to f} + \lambda \hat{m}_{x\to f} \quad (\text{idem } a, \hat{\tau}^{(0)})
\end{aligned} \tag{286}$$

We recognize $\bar{A}_\lambda^{(0)}$ as the variable BO potential Eq. (152) which yields the dual mapping:

$$m_z^f = \mathbb{E}_\lambda m_\lambda, \quad \alpha m_x^f = \mathbb{E}_\lambda \lambda m_\lambda \quad (\text{idem } \tau^{(0)}, v) \quad \text{with} \quad m_\lambda = \tau_\lambda^{(0)} - v_\lambda, \quad v_\lambda = \frac{1}{a_\lambda}. \tag{287}$$

In particular we recover Eqs (277)-(278) for the variances. The decomposition Eq. (160) for the BO potential reads:

$$\bar{A}_f^{(0)}[\hat{m}_{z\to f}, \hat{m}_{x\to f}] = \frac{\hat{m}_{z\to f}\tau_z^{(0)} + \alpha\hat{m}_{x\to f}\tau_x^{(0)}}{2} - I_f[\hat{m}_{z\to f}, \hat{m}_{x\to f}] + A_f^{(0)}[\hat{\tau}_{z\to f}^{(0)}, \hat{\tau}_{z\to f}^{(0)}] \quad (288)$$

where $A_f^{(0)}$ is given by Eq. (281) and the mutual information by:

$$I_f[\hat{m}_{z\to f}, \hat{m}_{x\to f}] = \frac{1}{2}\mathbb{E}_\lambda \ln a_\lambda \tau_\lambda^{(0)}. \quad (289)$$

From these expressions, it is straightforward to check the dual mapping to the variances:

$$\tfrac{1}{2}v_z^f = \partial_{\hat{m}_{z\to f}} I_f, \qquad \tfrac{1}{2}\alpha v_x^f = \partial_{\hat{m}_{x\to f}} I_f, \quad (290)$$

as well as the dual mapping to the overlaps:

$$\tfrac{1}{2}m_z^f = \partial_{\hat{m}_{z\to f}} \bar{A}_f^{(0)}, \qquad \tfrac{1}{2}\alpha m_x^f = \partial_{\hat{m}_{x\to f}} \bar{A}_f^{(0)}. \quad (291)$$

E.2.5 Random Matrix Theory Expressions

The posterior variances and the mutual information are closely related to the following transforms in random matrix theory (Tulino and Verdú, 2004):

$$\text{Shannon transform} \quad \mathcal{V}(\gamma) = \mathbb{E}_\lambda \ln(1+\gamma\lambda) \quad (292)$$

$$\eta \text{ transform} \quad \eta(\gamma) = \mathbb{E}_\lambda \frac{1}{1+\gamma\lambda} \quad (293)$$

$$\text{Stieltjes transform} \quad \mathcal{S}(z) = \mathbb{E}_\lambda \frac{1}{\lambda - z} \quad (294)$$

$$\text{R transform} \quad R(s) = \mathcal{S}^{-1}(-s) - \frac{1}{s} \quad (295)$$

where $\mathcal{S}^{-1}$ denotes the functional inverse of $\mathcal{S}$. Following Reeves (2017) let's introduce the integrated R-transform and its Legendre transform

$$J(t) = \frac{1}{2}\int_0^t dz R(-z), \qquad J^*(u) = \sup_t J(t) - \frac{1}{2}ut. \quad (296)$$

Using the identities (Tulino and Verdú, 2004)

$$\gamma\frac{d}{d\gamma}\mathcal{V}(\gamma) = 1 - \eta(\gamma) = -\phi R(\phi) \quad \text{with } \phi = -\gamma\eta(\gamma), \quad (297)$$

it can be shown that:

$$a_{z\to f}v_z^f = \eta(\gamma) \quad \text{with } \gamma = \frac{a_{x\to f}}{a_{z\to f}}, \quad u = \frac{\alpha v_x^f}{v_z^f} = R(\phi), \quad J^*(u) = \frac{1}{2}\mathbb{E}_\lambda \ln a_\lambda v_z. \quad (298)$$

The Legendre transforms of the BO potential and mutual information

$$\bar{A}_f^{(0*}[m_z, m_x] = \sup_{\hat{m}_{z\to f}, \hat{m}_{x\to f}} \frac{\hat{m}_{z\to f}m_z + \alpha\hat{m}_{x\to f}m_x}{2} - \bar{A}_f^{(0)}[\hat{m}_{z\to f}, \hat{m}_{x\to f}], \quad (299)$$

$$I_f^*[v_z, v_x] = \sup_{\hat{m}_{z\to f}, \hat{m}_{x\to f}} I_f[\hat{m}_{z\to f}, \hat{m}_{x\to f}] - \frac{\hat{m}_{z\to f}v_z + \alpha\hat{m}_{x\to f}v_x}{2}, \quad (300)$$

are equal to

$$\bar{A}_f^{(0)*}[m_z, m_x] = J^*(u) - \frac{1}{2}\ln 2\pi e v_z + \frac{\hat{\tau}^{(0)}_{z\to f} v_z + \alpha \hat{\tau}^{(0)}_{x\to f} v_x}{2}, \tag{301}$$

$$I_f^*[v_z, v_x] = \bar{A}_f^{(0)*}[m_z, m_x] + A_f^{(0)}[\hat{\tau}^{(0)}_{z\to f}, \hat{\tau}^{(0)}_{x\to f}]. \tag{302}$$

In particular when the teacher factor graph is a Bayesian network we recover (Reeves, 2017):

$$I_f^*[v_z, v_x] = J^*(u) + I_z^*[v_z], \quad I_z^*[v_z] = \frac{1}{2}\left(\ln \frac{\tau_z^{(0)}}{v_z} + \frac{v_z}{\tau_z^{(0)}} - 1\right). \tag{303}$$

When the matrix W belongs to an ensemble for which the limiting spectral density of $W^\intercal W$ is known (for example Section E.2.7) the transforms above can be derived analytically, leading to the S-AMP approach (Çakmak et al., 2014, 2016).

E.2.6 Gibbs Free Energy

The Gibbs free energy for the linear channel can be expressed as a function of the posterior variances:

$$G_f[v_z, v_x] = G_z[v_z] + NJ^*(u), \quad u = \frac{\alpha v_x}{v_z}, \tag{304}$$

where $G_z[v_z] = -\frac{N}{2}\ln 2\pi e v_z$ is the variable negative entropy and $J^*(u)$ the dual integrated R-transform Eq. (296). Viewed as a function of the posterior variances, G_f gives the dual mapping to the incoming precisions:

$$-\frac{N}{2} a_{z\to f} = \partial_{v_z} G_f[v_z, v_x], \quad -\frac{M}{2} a_{x\to f} = \partial_{v_x} G_f[v_z, v_x], \tag{305}$$

Similarly the variable negative entropy G_z gives the dual mapping to the precision

$$-\frac{N}{2} a_z = \partial_{v_z} G_z[v_z] \implies a_z = \frac{1}{v_z} \quad \text{(idem x)}. \tag{306}$$

Then the potential

$$\tilde{G}_f[v_x, v_z] = G_x[v_x] + G_z[v_z] - G_f[v_z, v_x] = G_x[v_x] - NJ^*(u) \tag{307}$$

gives the dual mapping to the *outgoing* precisions:

$$-\frac{N}{2} a_{f\to z} = \partial_{v_z} \tilde{G}_f[v_z, v_z], \quad -\frac{M}{2} a_{f\to x} = \partial_{v_x} \tilde{G}_f[v_z, v_x]. \tag{308}$$

E.2.7 Matrix with iid Entries

Let W be a random matrix with iid entries of mean 0 and variance $\frac{1}{N}$. Then the limiting spectral density $\rho(\lambda)$ of $\lambda = \operatorname{Spec} W^\intercal W$ as $N \to \infty$ with $\alpha = \frac{M}{N} = O(1)$ follows the Marchenko-Pastur law:

$$\rho(\lambda) = \max(0, 1-\alpha)\delta(\lambda) + \frac{1}{2\pi\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}\mathbf{1}_{[\lambda_-, \lambda_+]}(\lambda) \quad \lambda_\pm = (1 \pm \sqrt{\alpha})^2 \tag{309}$$

Then the dual integrated R-transform is (Reeves, 2017):

$$J^*(u) = \frac{\alpha}{2}\left(\ln\frac{\alpha}{u} + \frac{u}{\alpha} - 1\right). \tag{310}$$

From Eq. (307) the potential $\tilde{G}_f$ is then equal to:

$$\tilde{G}_f[v_z, v_x] = -\frac{M}{2}\left(\ln 2\pi v_z + \frac{v_x}{v_z}\right) \tag{311}$$

which gives the dual mapping Eq. (308) to the *outgoing* precisions:

$$a_{f\to z} = \frac{\alpha}{v_z}\left(1 - \frac{v_x}{v_z}\right), \quad a_{f\to x} = \frac{1}{v_z}. \tag{312}$$

The fact that the outgoing precision $a_{f\to x}$ towards x is equal to the precision $a_z = \frac{1}{v_z}$ of z is only true for the iid entries case. For a generic linear channel $a_{f\to x}$ will depend on both v_x and v_z.

E.2.8 Rotation Channel

If $W = R$ is a rotation, then $v_z^f = v_x^f = \frac{1}{a}$ with $a = a_{z\to f} + a_{x\to f}$. Besides the forward $f \to x$ and backward $f \to z$ updates are simple rotations in parameter space:

$$a^{\text{new}}_{f\to x} = a_{z\to f}, \qquad b^{\text{new}}_{f\to x} = R b_{z\to f}, \tag{313}$$

$$a^{\text{new}}_{f\to z} = a_{x\to f}, \qquad b^{\text{new}}_{f\to z} = R^\intercal b_{x\to f}. \tag{314}$$

E.2.9 Scaling Channel

When the weight matrix $W = S$ is a diagonal $M \times N$ matrix, the eigenvalue distribution is equal to $\lambda = S^\intercal S$ and the posterior mean and variances are given by:

$$r_z^f = \frac{b_{z\to f} + S^\intercal b_{x\to f}}{a_{z\to f} + a_{x\to f}\lambda}, \qquad v_z^f = \mathbb{E}_\lambda \frac{1}{a_{z\to f} + a_{x\to f}\lambda}, \tag{315}$$

$$r_x^f = S r_z^f, \qquad \alpha v_x^f = \mathbb{E}_\lambda \frac{\lambda}{a_{z\to f} + a_{x\to f}\lambda}. \tag{316}$$

In particular, for out-of-rank components, the posterior mean r_z^f is set to the prior and the posterior mean r_x^f is set to zero:

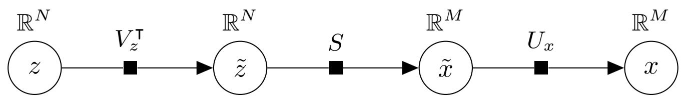
$$r_z^{f(n)} = \frac{b^{(n)}_{z\to f}}{a_{z\to f}} \quad \text{for} \quad R < n \le N, \quad r_x^{f(m)} = 0 \quad \text{for} \quad R < m \le M. \tag{317}$$

E.2.10 SVD Decomposition

As proposed by Rangan et al. (2017) for VAMP, it is more efficient to precompute the SVD decomposition:

$$W = U_x S V_z^\intercal,\ U_x \in O(M),\ V_z \in O(N),\ S \in \mathbb{R}^{M\times N} \text{ diagonal}. \tag{318}$$

The eigenvalue distribution of $W^\intercal W$ is equal to $\lambda = S^\intercal S$. Then the EP updates for W are equivalent to the composition of a rotation $V_z^\intercal$ in z-space, a scaling S that projects z into the x space and a rotation U_x in x-space.



These updates are only rotations or element-wise scaling and are thus considerably faster than solving $ar_z^f = b$ or even worse computing the inverse $\Sigma = a^{-1}$ at each update. It comes at the expense of computing the SVD decomposition of W, but this only needs to be done once.

E.2.11 Complex Linear Channel

The real linear channel can be easily extended to the complex linear channel $x = Wz$ with $x \in \mathbb{C}^M$, $z \in \mathbb{C}^N$ and $W \in \mathbb{C}^{M\times N}$ and $\lambda = \text{Spec } W^\dagger W$.

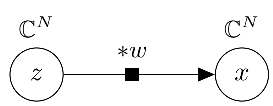
E.2.12 Unitary Channel

When $W = U$ is unitary, for instance when $W = \mathcal{F}$ is the discrete Fourier transform (DFT), then $v_z^f = v_x^f = \frac{1}{a}$ with $a = a_{z\to f} + a_{x\to f}$. Besides the forward $f \to x$ and backward $f \to z$ updates are simple unitary transforms in parameter space:

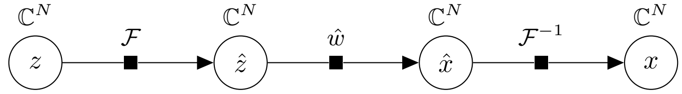
$$a_{f\to x}^{\text{new}} = a_{z\to f}, \qquad b_{f\to x}^{\text{new}} = U b_{z\to f}, \tag{319}$$

$$a_{f\to z}^{\text{new}} = a_{x\to f}, \qquad b_{f\to z}^{\text{new}} = U^\dagger b_{x\to f}. \tag{320}$$

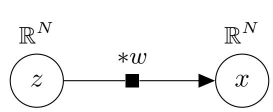
E.2.13 Convolution Channel (Complex)



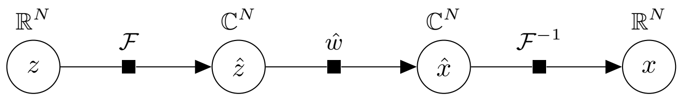
The convolution channel $x = w * z$ with convolution weights $w \in \mathbb{C}^N$ is a complex linear channel $x = Wz$ with $M = N$. It is equivalent to the composition of a discrete Fourier transform (DFT) $\mathcal{F}$ for z, a multiplication by $\hat{w} = \mathcal{F}w \in \mathbb{C}^N$, and an inverse DFT $\mathcal{F}^{-1}$ for x. The eigenvalue distribution of $W^\dagger W$ is equal to $\lambda = \hat{w}^\dagger \hat{w} = |\hat{w}|^2$.



E.2.14 Convolution Channel (Real)



The convolution channel $x = w * z$ with convolution weights $w \in \mathbb{R}^N$ is a real linear channel $x = Wz$ with $M = N$. It is equivalent to the composition of a discrete Fourier transform (DFT) $\mathcal{F}$ for z, a multiplication by $\hat{w} = \mathcal{F}w \in \mathbb{C}^N$, and an inverse DFT $\mathcal{F}^{-1}$ for x. The eigenvalue distribution of $W^\intercal W$ is equal to $\lambda = \hat{w}^\dagger \hat{w} = |\hat{w}|^2$.



E.2.15 Full Covariance Beliefs

In this subsection we will consider the linear channel $x = Wz$ with full covariance beliefs on x and z, meaning that the precisions $a_{z\to f}$ and $a_{x\to f}$ are $N \times N$ and $M \times M$ matrices. The log-partition is given by:

$$A_f[a_{z\to f}, b_{z\to f}, a_{x\to f}, b_{x\to f}] = \frac{1}{2} b^\intercal \Sigma b + \frac{1}{2} \ln \det 2\pi\Sigma \tag{321}$$

$$b = b_{z\to f} + W^\intercal b_{x\to f}, \quad a = a_{z\to f} + W^\intercal a_{x\to f} W, \quad \Sigma = a^{-1}. \tag{322}$$

The posterior mean and covariance are given by:

$$r_z^f = \Sigma b, \qquad \Sigma_z^f = \Sigma, \tag{323}$$

$$r_x^f = W r_z^f, \qquad \Sigma_x^f = W \Sigma W^\intercal, \tag{324}$$

We have $a_z^f = a$ and $b_z^f = b$ so the backward $f \to z$ update is simply given by:

$$a_{f\to z}^{\text{new}} = W^\intercal a_{x\to f} W, \qquad b_{f\to z}^{\text{new}} = W^\intercal b_{x\to f} \tag{325}$$

We have:

$$a_x^f = (W\Sigma W^\intercal)^{-1} = a_{x\to f} + (W a_{z\to f}^{-1} W^\intercal)^{-1} \tag{326}$$

$$b_x^f = (W\Sigma W^\intercal)^{-1} W \Sigma b = b_{x\to f} + (W\Sigma W^\intercal)^{-1} W \Sigma b_{z\to f} \tag{327}$$

We can obtain the RHS of Eq (326) by two applications of the Woodbury identity. The RHS of Eq (327) follows directly from the definition of b. Using the mean and covariance of the $z \to f$ and $f \to x$ messages:

$$\Sigma_{z\to f} = a_{z\to f}^{-1}, \qquad r_{z\to f} = a_{z\to f}^{-1} b_{z\to f} \tag{328}$$

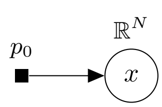
$$\Sigma_{f\to x} = a_{f\to x}^{-1}, \qquad r_{f\to x} = a_{f\to x}^{-1} b_{f\to x} \tag{329}$$

we can write the forward $f \to x$ update as:

$$\Sigma_{f\to x}^{\text{new}} = W \Sigma_{z\to f} W^\intercal, \qquad r_{f\to x}^{\text{new}} = W r_{z\to f}. \tag{330}$$

E.3 Separable Priors

E.3.1 Generic Separable Prior



Let $f(x) = p_0(x) = \prod_{n=1}^N p_0(x^{(n)})$ be a separable prior over $x \in \mathbb{R}^N$. The log-partition, posterior mean and variance are given by:

$$A_f[a_{x\to f}, b_{x\to f}] = \sum_{n=1}^N A_f[a_{x\to f}, b_{x\to f}^{(n)}], \tag{331}$$

$$r_x^f[a_{x\to f}, b_{x\to f}]^{(n)} = r_x^f[a_{x\to f}, b_{x\to f}^{(n)}], \quad v_x^f[a_{x\to f}, b_{x\to f}] = \frac{1}{N}\sum_{n=1}^N v_x^f[a_{x\to f}, b_{x\to f}^{(n)}]. \tag{332}$$

where on the RHS the same quantities are defined over scalar $b_{x\to f}^{(n)}$ in $\mathbb{R}$:

$$A_f[a_{x\to f}, b_{x\to f}] = \ln \int_{\mathbb{R}} dx\, p_0(x)\, e^{-\frac{1}{2}a_{x\to f}x^2 + b_{x\to f}x}, \tag{333}$$

$$r_x^f[a_{x\to f}, b_{x\to f}] = \partial_{b_{x\to f}} A_f[a_{x\to f}, b_{x\to f}], \quad v_x^f[a_{x\to f}, b_{x\to f}] = \partial^2_{b_{x\to f}} A_f[a_{x\to f}, b_{x\to f}]. \tag{334}$$

In the remainder we will only derive the scalar case, as it can be straightforwardly extended to the high dimensional counterpart through Eqs (331)-(332). For a large class of priors that we call natural priors we can derive closed-form expressions for the log-partition, mean and variance as shown in Section E.3.5; familiar examples include the Gaussian, binary, Gauss-Bernoulli, and positive priors.

E.3.2 Teacher Prior Second Moment

The approximate teacher marginal (Proposition 5) is:

$$p_f^{(0)}(x^{(0)} \mid \hat{\tau}_{x\to f}^{(0)}) = p_0^{(0)}(x^{(0)})\, e^{-\frac{1}{2}\hat{\tau}_{x\to f}^{(0)} x^{(0)2} - A_f^{(0)}[\hat{\tau}_{x\to f}^{(0)}]}, \tag{335}$$

with log-partition:

$$A_f^{(0)}[\hat{\tau}_{x\to f}^{(0)}] = \ln \int_{\mathbb{R}} dx\, p_0^{(0)}(x)\, e^{-\frac{1}{2}\hat{\tau}_{x\to f}^{(0)} x^2} \tag{336}$$

which yields the dual mapping:

$$\tau_x^{(0)} = -2\partial_{\hat{\tau}_{x\to f}^{(0)}} A_f^{(0)}[\hat{\tau}_{x\to f}^{(0)}]. \tag{337}$$

When the teacher factor graph is a Bayesian network (Section 3.2) we have

$$\hat{\tau}_{x\to f}^{(0)} = 0, \quad p_f^{(0)}(x^{(0)} \mid \hat{\tau}_{x\to f}^{(0)}) = p_0^{(0)}(x^{(0)}). \tag{338}$$

E.3.3 Replica Symmetric

The RS potential and overlaps are given by a low-dimensional integration of the corresponding scalar EP quantities Eqs (333)-(334):

$$\bar{A}_f[\hat{m}_{x\to f}, \hat{q}_{x\to f}, \hat{\tau}_{x\to f}] = \int_{\mathbb{R}} db_{x\to f}\, p_f^{(0)}(b_{x\to f})\, A_f[a_{x\to f}, b_{x\to f}], \tag{339}$$

$$m_x^f[\hat{m}_{x\to f}, \hat{q}_{x\to f}, \hat{\tau}_{x\to f}] = \int_{\mathbb{R}^2} db_{x\to f} dx^{(0)}\, p_f^{(0)}(b_{x\to f}, x^{(0)})\, x^{(0)} r_x^f[a_{x\to f}, b_{x\to f}] \tag{340}$$

$$q_x^f[\hat{m}_{x\to f}, \hat{q}_{x\to f}, \hat{\tau}_{x\to f}] = \int_{\mathbb{R}} db_{x\to f}\, p_f^{(0)}(b_{x\to f})\, r_x^f[a_{x\to f}, b_{x\to f}]^2, \tag{341}$$

$$v_x^f[\hat{m}_{x\to f}, \hat{q}_{x\to f}, \hat{\tau}_{x\to f}] = \int_{\mathbb{R}} db_{x\to f}\, p_f^{(0)}(b_{x\to f})\, v_x^f[a_{x\to f}, b_{x\to f}], \tag{342}$$

$$\tau_x^f[\hat{m}_{x\to f}, \hat{q}_{x\to f}, \hat{\tau}_{x\to f}] = q_x^f[\hat{m}_{x\to f}, \hat{q}_{x\to f}, \hat{\tau}_{x\to f}] + v_x^f[\hat{m}_{x\to f}, \hat{q}_{x\to f}, \hat{\tau}_{x\to f}], \tag{343}$$

where the ensemble average Eq. (123) is now over scalar $b_{x\to f}$ and $x^{(0)}$ in $\mathbb{R}$:

$$p_f^{(0)}(b_{x\to f}, x^{(0)}) = \mathcal{N}(b_{x\to f} \mid \hat{m}_{x\to f} x^{(0)}, \hat{q}_{x\to f})\, p_f^{(0)}(x^{(0)} \mid \hat{\tau}_{x\to f}^{(0)}) \tag{344}$$

with $a_{x\to f} = \hat{\tau}_{x\to f} + \hat{q}_{x\to f}$ and $p_f^{(0)}(x^{(0)} \mid \hat{\tau}_{x\to f}^{(0)})$ given by Eq. (335). The dual mapping to the overlaps now simply reads:

$$m_x^f = \partial_{\hat{m}_{x\to f}} \bar{A}_f, \quad -\tfrac{1}{2} q_x^f = \partial_{\hat{q}_{x\to f}} \bar{A}_f, \quad -\tfrac{1}{2} \tau_x^f = \partial_{\hat{\tau}_{x\to f}} \bar{A}_f. \tag{345}$$

E.3.4 Bayes-Optimal

The BO potential and overlap are given by a low-dimensional integration of the corresponding scalar EP quantities:

$$\bar{A}_f^{(0)}[\hat{m}_{x\to f}] = \int_{\mathbb{R}} db_{x\to f}\, p_f^{(0)}(b_{x\to f})\, A_f^{(0)}[a_{x\to f}, b_{x\to f}], \tag{346}$$

$$v_x^f[\hat{m}_{x\to f}] = \int_{\mathbb{R}} db_{x\to f}\, p_f^{(0)}(b_{x\to f})\, v_x^f[a_{x\to f}, b_{x\to f}], \tag{347}$$

$$m_x^f[\hat{m}_{x\to f}] = \tau_x^{(0)} - v_x^f[\hat{m}_{x\to f}], \tag{348}$$

where the ensemble average Eq. (145) is now over scalar $b_{x\to f}$ and $x^{(0)}$ in $\mathbb{R}$:

$$p_f^{(0)}(b_{x\to f}, x^{(0)}) = \mathcal{N}(b_{x\to f} \mid \hat{m}_{x\to f} x^{(0)}, \hat{m}_{x\to f})\, p_f^{(0)}(x^{(0)} \mid \hat{\tau}_{x\to f}^{(0)}) \tag{349}$$

with $a_{x\to f} = \hat{\tau}_{x\to f}^{(0)} + \hat{m}_{x\to f}$ and $p_f^{(0)}(x^{(0)} \mid \hat{\tau}_{x\to f}^{(0)})$ given by Eq. (335). The relationship Eq. (160) between the mutual information and the BO potential now reads:

$$I_f[\hat{m}_{x\to f}] = \tfrac{1}{2} \hat{m}_{x\to f} \tau_x^{(0)} - \bar{A}_f^{(0)}[\hat{m}_{x\to f}] + A_f^{(0)}[\hat{\tau}_{x\to f}^{(0)}]. \tag{350}$$

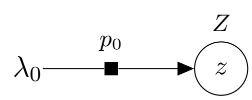
The dual mapping to the variance and overlap simply reads:

$$\tfrac{1}{2} v_x^f = \partial_{\hat{m}_{x\to f}} I_f, \qquad \tfrac{1}{2} m_x^f = \partial_{\hat{m}_{x\to f}} \bar{A}_f^{(0)}. \tag{351}$$

E.3.5 Natural Prior

Let z be a variable of base space Z, with sufficient statistics $\phi(z)$ and associated natural parameters λ_z. Several examples of variable types are presented in Section E.1 such as the real, binary, sparse and positive variable. A natural prior over the variable z is an exponential family distribution $p_0(z) = p(z \mid \lambda_0) = e^{\lambda_0^\intercal \phi(z) - A_z[\lambda_0]}$ with a given natural parameter λ_0.

Variable z belief Let the factor $f(z) = p_0(z) = p(z \mid \lambda_0)$ be a natural prior. Let us first consider the corresponding module with variable z belief.



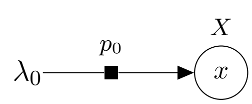
The log-partition and moment function are given by:

$$A_f[\lambda_{z\to f}] = A_z[\lambda_{z\to f} + \lambda_0] - A_z[\lambda_0], \tag{352}$$
$$\mu_z^f[\lambda_{z\to f}] = \mu_z[\lambda_{z\to f} + \lambda_0], \tag{353}$$

where $A_z[\lambda]$ and $\mu_z[\lambda]$ denotes the log-partition and moment function of the variable z. The $f \to z$ update is the constant message:

$$\lambda_{f\to z}^{\text{new}} = \lambda_0. \tag{354}$$

Variable x belief Let x be a variable of different type than z, with base space X and sufficient statistics $\phi(x)$ and associated natural parameters λ_x. We still consider the same factor $f(x) = p_0(x) = p(x \mid \lambda_0)$ which is a natural prior in the z variable, but we derive the corresponding module using a variable x belief.



This is meaningful only if we can inject $Z \hookrightarrow X$. We denote by $\phi^{(0)}$, $\phi^{(1)}$ and $\phi^{(2)}$ the set of sufficient statistics common to x and z, specific to z and specific to x respectively. We will assume that the sufficient statistics $\phi^{(2)}$ specific to x are constant on Z:

$$\phi^{(2)}(z) = \mu_Z^{(2)} \quad \text{for all} \quad z \in Z. \tag{355}$$

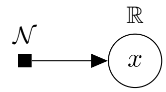
Then the log-partition and moment function are given by:

$$A_f[\lambda_{x\to f}] = A_z[\lambda_{z\to f} + \lambda_0] - A_z[\lambda_0] + \lambda_{x\to f}^{(2)T} \mu_Z^{(2)}, \tag{356}$$
$$\mu_x^{f(0)}[\lambda_{x\to f}] = \mu_z^{(0)}[\lambda_{z\to f} + \lambda_0], \qquad \mu_x^{f(2)}[\lambda_{x\to f}] = \mu_Z^{(2)} \tag{357}$$

with $\lambda_{z\to f}^{(0)} = \lambda_{x\to f}^{(0)}$ and $\lambda_{z\to f}^{(1)} = 0$.

Isotropic Gaussian belief To derive the isotropic Gaussian belief modules, we take x to be a real variable that is $X = \mathbb{R}$ and $\phi(x) = (x, -\frac{1}{2}x^2)$ and associated natural parameters $\lambda_x = (b_x, a_x)$. When z is a real, binary, sparse, and interval variable we obtain respectively the Gaussian, binary, Gauss-Bernoulli and truncated Normal prior as detailed in the next subsections.

E.3.6 Gaussian Prior



The factor $f(x) = p(x \mid a_0, b_0) = \mathcal{N}(x \mid r_0, v_0)$ is the Normal prior with natural parameters $a_0 = \frac{1}{v_0}$ and $b_0 = \frac{r_0}{v_0}$. The Normal prior corresponds to the natural prior for the real variable $x \in \mathbb{R}$. According to Section E.3.5 the log-partition is given by:

$$A_f[a_{x\to f}, b_{x\to f}] = A[a, b] - A[a_0, b_0] \quad \text{with} \quad a = a_{x\to f} + a_0, \quad b = b_{x\to f} + b_0, \tag{358}$$

where $A[a, b]$ is the log-partition of a real variable, see Section E.1.5. The posterior mean and variance are given by:

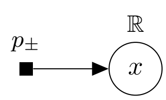
$$r_x^f = \frac{b}{a}, \qquad v_x^f = \frac{1}{a}, \tag{359}$$

leading to the constant $f \to x$ update:

$$a_{f\to x}^{\text{new}} = a_0, \qquad b_{f\to x}^{\text{new}} = b_0. \tag{360}$$

The ensemble average variance is directly given by $v_x^f = \frac{1}{a}$.

E.3.7 Binary Prior



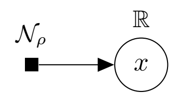
The factor $f(x) = p(x \mid b_0) = p_+\delta_{+1}(x) + p_-\delta_{-1}(x)$ is the binary prior with natural parameter $b_0 = \frac{1}{2}\ln\frac{p_+}{p_-}$. The binary prior corresponds to the natural prior for the binary variable $z \in \pm$. According to Section E.3.5 the log-partition, posterior mean and variance are given by:

$$A_f[a_{x\to f}, b_{x\to f}] = A[b] - A[b_0] - \frac{a_{x\to f}}{2} \quad \text{with} \quad b = b_0 + b_{x\to f}, \tag{361}$$

$$r_x^f = r[b], \qquad v_x^f = v[b], \tag{362}$$

where $A[b]$, $r[b]$ and $v[b]$ denote the log-partition, mean and variance of a binary variable, see Section E.1.6.

E.3.8 Sparse Prior



The factor $f(x) = p(x \mid a_0, b_0\eta_0) = [1-\rho_0]\delta_0(x) + \rho_0\mathcal{N}(x \mid r_0, v_0)$ is the Gauss-Bernoulli prior with natural parameters $a_0 = \frac{1}{v_0}$, $b_0 = \frac{r_0}{v_0}$ and $\eta_0 = A[a_0, b_0] - \ln\frac{\rho_0}{1-\rho_0}$ where $A[a,b] = \frac{b^2}{2a} + \frac{1}{2}\ln\frac{2\pi}{a}$ is the log-partition of a real variable. The Gauss-Bernoulli prior corresponds to the natural prior for the sparse variable $z \in \mathbb{R} \cup \{0\}$. According to Section E.3.5 the log-partition, posterior mean and variance are given by:

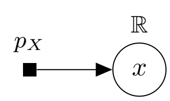
$$A_f[a_{x\to f}, b_{x\to f}] = A[a, b, \eta_0] - A[a_0, b_0, \eta_0] \quad \text{with} \quad a = a_0 + a_{x\to f}, \quad b = b_0 + b_{x\to f}, \tag{363}$$

$$r_x^f = r[a, b, \eta_0], \qquad v_x^f = v[a, b, \eta_0], \tag{364}$$

where $A[a, b, \eta]$, $r[a, b, \eta]$ and $v[a, b, \eta]$ denote the log-partition, mean and variance of a sparse variable, see Section E.1.7. Also the sparsity of x is equal to $\rho_x^f = \rho[a, b, \eta_0]$.

E.3.9 Interval Prior

Let $X \subset \mathbb{R}$ denotes any real interval, for example $X = \mathbb{R}_+$ for a positive prior.



The factor $f(x) = p_X(x \mid a_0, b_0) = \frac{1}{p_X[a_0, b_0]}\mathcal{N}(x \mid r_0, v_0)\delta_X(x)$ is the truncated Normal prior with natural parameters $a_0 = \frac{1}{v_0}$ and $b_0 = \frac{r_0}{v_0}$. The truncated Normal corresponds to the natural prior for the interval variable $z \in X$. According to Section E.3.5 the log-partition, posterior mean and variance are given by:

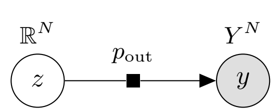
$$A_f[a_{x\to f}, b_{x\to f}] = A_X[a, b] - A_X[a_0, b_0] \quad \text{with} \quad a = a_0 + a_{x\to f}, \quad b = b_0 + b_{x\to f}, \tag{365}$$

$$r_x^f = r_X[a, b], \qquad v_x^f = v_X[a, b] \tag{366}$$

where $A_X[a, b]$, $r_X[a, b]$ and $v_X[a, b]$ denote the log-partition, mean and variance of a interval X variable, see Section E.1.8.

E.4 Separable Likelihoods

E.4.1 Generic Separable Likelihood



Let $f(z) = p_{\text{out}}(y \mid z) = \prod_{n=1}^{N} p_{\text{out}}(y^{(n)} \mid z^{(n)})$ be a separable likelihood over $z \in \mathbb{R}^N$ with observed $y \in Y^N$. The log-partition, posterior means and variances are given by:

$$A_f[a_{z\to f}, b_{z\to f}; y] = \sum_{n=1}^{N} A_f[a_{z\to f}, b_{z\to f}^{(n)}; y^{(n)}], \tag{367}$$

$$r_z^f[a_{z\to f}, b_{z\to f}; y]^{(n)} = r_z^f[a_{z\to f}, b_{z\to f}^{(n)}; y^{(n)}], \tag{368}$$

$$v_z^f[a_{z\to f}, b_{z\to f}; y] = \frac{1}{N}\sum_{n=1}^{N} v_z^f[a_{z\to f}, b_{z\to f}^{(n)}; y^{(n)}]. \tag{369}$$

where on the RHS the same quantities are defined over scalar $b_{z\to f}^{(n)}$ and $y^{(n)}$ in $\mathbb{R}$:

$$A_f[a_{z\to f}, b_{z\to f}; y] = \ln \int_{\mathbb{R}} dz\, p_{\text{out}}(y \mid z)\, e^{-\frac{1}{2}a_{z\to f}z^2 + b_{z\to f}z}, \tag{370}$$

$$r_z^f[a_{z\to f}, b_{z\to f}; y] = \partial_{b_{z\to f}} A_f[a_{z\to f}, b_{z\to f}; y], \tag{371}$$

$$v_z^f[a_{z\to f}, b_{z\to f}; y] = \partial^2_{b_{z\to f}} A_f[a_{z\to f}, b_{z\to f}; y]. \tag{372}$$

In the remainder we will only derive the scalar case, as it can be straightforwardly extended to the high dimensional counterpart through Eqs (367)-(369).

E.4.2 Teacher Prior Second Moment

The approximate teacher marginal (Proposition 5) is:

$$p_f^{(0)}(y, z^{(0)} \mid \hat{\tau}_{z\to f}^{(0)}) = p_{\text{out}}^{(0)}(y \mid z^{(0)})\, e^{-\frac{1}{2}\hat{\tau}_{z\to f}^{(0)} z^{(0)2} - A_z^{(0)}[\hat{\tau}_{z\to f}^{(0)}]}, \tag{373}$$

with log-partition:

$$A_f^{(0)}[\hat{\tau}_{z\to f}^{(0)}] = \ln \int_{\mathbb{R}} dydz\, p_{\text{out}}^{(0)}(y \mid z)\, e^{-\frac{1}{2}\hat{\tau}_{z\to f}^{(0)} z^2} = \frac{1}{2}\ln \frac{2\pi}{\hat{\tau}_{z\to f}^{(0)}} \tag{374}$$

which yields the dual mapping:

$$\tau_z^{(0)} = -2\partial_{\hat{\tau}_{z\to f}^{(0)}} A_f^{(0)}[\hat{\tau}_{z\to f}^{(0)}] = \frac{1}{\hat{\tau}_{z\to f}^{(0)}} \tag{375}$$

The approximate teacher marginal is therefore equal to:

$$p_f^{(0)}(y, z^{(0)} \mid \hat{\tau}_{z\to f}^{(0)}) = p_{\text{out}}^{(0)}(y \mid z^{(0)})\, \mathcal{N}(z^{(0)} \mid 0, \tau_z^{(0)}). \tag{376}$$

E.4.3 Replica Symmetric

The RS potential and overlaps are given by a low-dimensional integration of the corresponding scalar EP quantities Eqs (370)-(372):

$$\bar{A}_f[\hat{m}_{z\to f}, \hat{q}_{z\to f}, \hat{\tau}_{z\to f}] = \int_{\mathbb{R}^2} db_{z\to f} dy \, p_f^{(0)}(b_{z\to f}, y) \, A_f[a_{z\to f}, b_{z\to f}; y], \tag{377}$$

$$m_z^f[\hat{m}_{z\to f}, \hat{q}_{z\to f}, \hat{\tau}_{z\to f}] = \int_{\mathbb{R}^3} db_{z\to f} dy dz^{(0)} \, p_f^{(0)}(b_{z\to f}, y, z^{(0)}) \, z^{(0)} r_z^f[a_{z\to f}, b_{z\to f}; y] \tag{378}$$

$$q_z^f[\hat{m}_{z\to f}, \hat{q}_{z\to f}, \hat{\tau}_{z\to f}] = \int_{\mathbb{R}^2} db_{z\to f} dy \, p_f^{(0)}(b_{z\to f}, y) \, r_z^f[a_{z\to f}, b_{z\to f}; y]^2, \tag{379}$$

$$v_z^f[\hat{m}_{z\to f}, \hat{q}_{z\to f}, \hat{\tau}_{z\to f}] = \int_{\mathbb{R}^2} db_{z\to f} dy \, p_f^{(0)}(b_{z\to f}, y) \, v_z^f[a_{z\to f}, b_{z\to f}; y], \tag{380}$$

$$\tau_z^f[\hat{m}_{z\to f}, \hat{q}_{z\to f}, \hat{\tau}_{z\to f}] = q_z^f[\hat{m}_{z\to f}, \hat{q}_{z\to f}, \hat{\tau}_{z\to f}] + v_z^f[\hat{m}_{z\to f}, \hat{q}_{z\to f}, \hat{\tau}_{z\to f}], \tag{381}$$

where the ensemble average Eq. (123) is now over scalar $b_{z\to f}$, y and $z^{(0)}$ in $\mathbb{R}$:

$$p_f^{(0)}(b_{z\to f}, y, z^{(0)}) = \mathcal{N}(b_{z\to f} \mid \hat{m}_{z\to f} z^{(0)}, \hat{q}_{z\to f}) \, p_f^{(0)}(y, z^{(0)} \mid \hat{\tau}_{z\to f}^{(0)}) \tag{382}$$

with $a_{z\to f} = \hat{\tau}_{z\to f} + \hat{q}_{z\to f}$ and $p_f^{(0)}(y, z^{(0)} \mid \hat{\tau}_{z\to f}^{(0)})$ given by Eq. (373). The dual mapping to the overlaps now simply reads:

$$m_z^f = \partial_{\hat{m}_{z\to f}} \bar{A}_f, \quad -\tfrac{1}{2} q_z^f = \partial_{\hat{q}_{z\to f}} \bar{A}_f, \quad -\tfrac{1}{2} \tau_z^f = \partial_{\hat{\tau}_{z\to f}} \bar{A}_f. \tag{383}$$

E.4.4 Bayes-Optimal

The BO potential and overlap are given by a low-dimensional integration of the corresponding scalar EP quantities:

$$\bar{A}_f^{(0)}[\hat{m}_{z\to f}] = \int_{\mathbb{R}^2} db_{z\to f} dy \, p_f^{(0)}(b_{z\to f}, y) \, A_f^{(0)}[a_{z\to f}, b_{z\to f}; y], \tag{384}$$

$$v_z^f[\hat{m}_{z\to f}] = \int_{\mathbb{R}^2} db_{z\to f} dy \, p_f^{(0)}(b_{z\to f}, y) \, v_z^f[a_{z\to f}, b_{z\to f}; y], \tag{385}$$

$$m_z^f[\hat{m}_{z\to f}] = \tau_z^{(0)} - v_z^f[\hat{m}_{z\to f}], \tag{386}$$

where the ensemble average Eq. (145) is now over scalar $b_{z\to f}$, y and $z^{(0)}$ in $\mathbb{R}$:

$$p_f^{(0)}(b_{z\to f}, y, z^{(0)}) = \mathcal{N}(b_{z\to f} \mid \hat{m}_{z\to f} z^{(0)}, \hat{m}_{z\to f}) \, p_f^{(0)}(y, z^{(0)} \mid \hat{\tau}_{z\to f}^{(0)}) \tag{387}$$

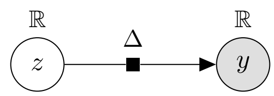
with $a_{z\to f} = \hat{\tau}_{z\to f}^{(0)} + \hat{m}_{z\to f}$ and $p_f^{(0)}(z^{(0)} \mid \hat{\tau}_{z\to f}^{(0)})$ given by Eq. (373). The relationship Eq. (160) between the mutual information and the BO potential now reads:

$$I_f[\hat{m}_{z\to f}] + E_f = \tfrac{1}{2} \hat{m}_{z\to f} \tau_z^{(0)} - \bar{A}_f^{(0)}[\hat{m}_{z\to f}] + A_f^{(0)}[\hat{\tau}_{z\to f}^{(0)}]. \tag{388}$$

The dual mapping to the variance and overlap simply reads:

$$\tfrac{1}{2} v_z^f = \partial_{\hat{m}_{z\to f}} I_f, \qquad \tfrac{1}{2} m_z^f = \partial_{\hat{m}_{z\to f}} \bar{A}_f^{(0)}. \tag{389}$$

E.4.5 Gaussian Likelihood



The factor $f(z) = p_{\mathrm{out}}(y \mid z) = \mathcal{N}(y \mid z, \Delta)$ is the Gaussian likelihood with noise variance Δ and observed y. The log-partition is given by:

$$A_f[a_{z\to f}, b_{z\to f}; y] = A[a,b] - A[a_y, b_y], \tag{390}$$

$$a_y = \frac{1}{\Delta}, \quad b_y = \frac{y}{\Delta}, \quad a = a_{z\to f} + a_y, \quad b = b_{z\to f} + b_y, \tag{391}$$

where $A[a,b]$ denotes the log-partition of a real variable, see Section E.1.5. The posterior mean and variance are given by:

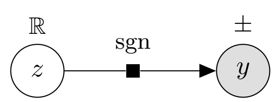
$$r_z^f = \frac{b}{a}, \qquad v_z^f = \frac{1}{a}, \tag{392}$$

leading to the constant $f \to z$ update:

$$a_{f\to z}^{\mathrm{new}} = a_y, \qquad b_{f\to z}^{\mathrm{new}} = b_y. \tag{393}$$

The ensemble average variance is directly given by $v_z^f = \frac{1}{a}$.

E.4.6 Sgn Likelihood



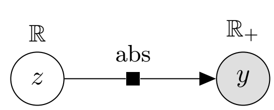
The factor $f(z) = p_{\mathrm{out}}(y \mid z) = \delta(y - \mathrm{sgn}(z))$ is the deterministic likelihood $y = \mathrm{sgn}(z) \in \pm$. The log-partition, posterior mean and variance are given by:

$$A_f[a_{z\to f}, b_{z\to f}; y] = A_y[a_{z\to f}, b_{z\to f}], \tag{394}$$

$$r_z^f = r_y[a_{z\to f}, b_{z\to f}], \qquad v_z^f = v_y[a_{z\to f}, b_{z\to f}], \tag{395}$$

where $A_\pm[a,b]$, $r_\pm[a,b]$ and $v_\pm[a,b]$ denote the log-partition, mean and variance of a positive/negative variable, see Section E.1.9.

E.4.7 Abs Likelihood



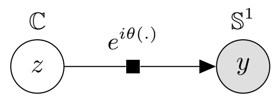
The factor $f(z) = p_{\text{out}}(y \mid z) = \delta(y - \text{abs}(z))$ is the deterministic likelihood $y = \text{abs}(z) \in \mathbb{R}_+$. The log-partition, posterior mean and variance are given by:

$$A_f[a_{z\to f}, b_{z\to f}; y] = -\frac{a_{z\to f} y^2}{2} + A[b] \quad \text{with} \quad b = y b_{z\to f}, \tag{396}$$

$$r_z^f = y\, r[b], \qquad v_z^f = y^2 v[b], \tag{397}$$

where $A[b]$, $r[b]$ and $v[b]$ denote the log-partition, mean and variance of a binary variable, see Section E.1.6.

E.4.8 Phase Likelihood



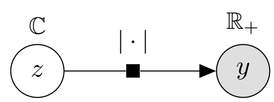
The factor $f(z) = p_{\text{out}}(y \mid z) = \delta(y - e^{i\theta(z)})$ is the deterministic likelihood $y = e^{i\theta(z)} \in \mathbb{S}^1$. The log-partition, posterior mean and variance are given by:

$$A_f[a_{z\to f}, b_{z\to f}; y] = A_+[a_{z\to f}, b] \quad \text{with} \quad b = y^\intercal b_{z\to f}, \tag{398}$$

$$r_z^f = y\, r_+[a_{z\to f}, b], \qquad v_z^f = \tfrac{1}{2} v_+[a_{z\to f}, b], \tag{399}$$

where $A_+[a, b]$, $r_+[a, b]$ and $v_+[a, b]$ denote the log-partition, mean and variance of a positive variable, see Section E.1.9. The $\frac{1}{2}$ factor in the variance comes from the average over the real and imaginary parts.

E.4.9 Modulus Likelihood



The factor $f(z) = p_{\text{out}}(y \mid z) = \delta(y - |z|)$ is the deterministic likelihood $y = |z| \in \mathbb{R}_+$. The log-partition, posterior mean and variance are given by:

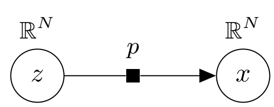
$$A_f[a_{z\to f}, b_{z\to f}; y] = -\frac{a_{z\to f} y^2}{2} + \ln y + A_{\mathbb{S}^1}[b] \quad \text{with} \quad b = y b_{z\to f}, \tag{400}$$

$$r_z^f = y\, r_{\mathbb{S}^1}[b], \qquad v_z^f = y^2 v_{\mathbb{S}^1}[b], \tag{401}$$

where $A_{\mathbb{S}^1}[b]$, $r_{\mathbb{S}^1}[b]$ and $v_{\mathbb{S}^1}[b]$ denote the log-partition, mean and variance of a phase variable, see Section E.1.10.

E.5 Separable Channels

E.5.1 Generic Separable Channel



Let $f(x,z) = p(x \mid z) = \prod_{n=1}^{N} p(x^{(n)} \mid z^{(n)})$ be a separable channel with input $z \in \mathbb{R}^N$ and output $x \in \mathbb{R}^N$. The log-partition, posterior means and variances are given by:

$$A_f[a_f, b_f] = \sum_{n=1}^{N} A_f[a_f, b_f^{(n)}], \tag{402}$$

$$r_x^{f(n)}[a_f, b_f] = r_x^f[a_f, b_f^{(n)}], \quad v_x^f[a_f, b_f] = \frac{1}{N}\sum_{n=1}^{N} v_x^f[a_f, b_f^{(n)}], \tag{403}$$

$$r_z^{f(n)}[a_f, b_f] = r_z^f[a_f, b_f^{(n)}], \quad v_z^f[a_f, b_f] = \frac{1}{N}\sum_{n=1}^{N} v_z^f[a_f, b_f^{(n)}]. \tag{404}$$

where on the RHS the same quantities are defined over scalar $b_{x\to f}^{(n)}$ and $b_{z\to f}^{(n)}$ in $\mathbb{R}$:

$$A_f[a_f, b_f] = \ln \int_{\mathbb{R}^2} dxdz\, p(x \mid z)\, e^{-\frac{1}{2}a_{x\to f}x^2 + b_{x\to f}x - \frac{1}{2}a_{z\to f}z^2 + b_{z\to f}z}, \tag{405}$$

$$r_x^f[a_f, b_f] = \partial_{b_{x\to f}} A_f[a_{x\to f}, b_{x\to f}], \quad v_x^f[a_f, b_f] = \partial^2_{b_{x\to f}} A_f[a_f, b_f], \tag{406}$$

$$r_z^f[a_f, b_f] = \partial_{b_{z\to f}} A_f[a_{x\to f}, b_{x\to f}], \quad v_z^f[a_f, b_f] = \partial^2_{b_{z\to f}} A_f[a_f, b_f]. \tag{407}$$

In the remainder we will only derive the scalar case, as it can be straightforwardly extended to the high dimensional counterpart through Eqs (402)-(404).

E.5.2 Teacher Prior Second Moment

The approximate teacher marginal (Proposition 5) is:

$$p_f^{(0)}(x^{(0)}, z^{(0)} \mid \hat{\tau}_f^{(0)}) = p^{(0)}(x^{(0)} \mid z^{(0)})\, e^{-\frac{1}{2}\hat{\tau}_{x\to f}^{(0)} x^{(0)2} - \frac{1}{2}\hat{\tau}_{z\to f}^{(0)} z^{(0)2} - A_f^{(0)}[\hat{\tau}_f^{(0)}]}, \tag{408}$$

with log-partition:

$$A_f^{(0)}[\hat{\tau}_f^{(0)}] = \ln \int_{\mathbb{R}^2} dxdz\, p^{(0)}(x \mid z)\, e^{-\frac{1}{2}\hat{\tau}_{x\to f}^{(0)} x^2 - \frac{1}{2}\hat{\tau}_{z\to f}^{(0)} z^2} \tag{409}$$

which yields the dual mapping:

$$\tau_x^{(0)} = -2\partial_{\hat{\tau}_{x\to f}^{(0)}} A_f^{(0)}[\hat{\tau}_f^{(0)}], \quad \tau_z^{(0)} = -2\partial_{\hat{\tau}_{z\to f}^{(0)}} A_f^{(0)}[\hat{\tau}_f^{(0)}], \tag{410}$$

When the teacher factor graph is a Bayesian network (Section 3.2) we have

$$\hat{\tau}_{x\to f}^{(0)} = 0, \quad \hat{\tau}_{z\to f}^{(0)} = \frac{1}{\tau_z^{(0)}}, \quad p_f^{(0)}(x^{(0)}, z^{(0)} \mid \hat{\tau}_f^{(0)}) = p^{(0)}(x^{(0)} \mid z^{(0)})\,\mathcal{N}(z^{(0)} \mid 0, \tau_z^{(0)}). \tag{411}$$

E.5.3 Replica Symmetric

The RS potential and overlaps are given by a low-dimensional integration of the corresponding scalar EP quantities Eqs (405)-(407):

$$\bar{A}_f[\hat{m}_f, \hat{q}_f, \hat{\tau}_f] = \int_{\mathbb{R}^2} db_{x\to f} db_{z\to f}\, p_f^{(0)}(b_{x\to f}, b_{z\to f})\, A_f[a_f, b_f], \tag{412}$$

$$m_x^f[\hat{m}_f, \hat{q}_f, \hat{\tau}_f] = \int_{\mathbb{R}^3} db_{x\to f} db_{z\to f} dx^{(0)}\, p_f^{(0)}(b_{x\to f}, b_{z\to f}, x^{(0)})\, x^{(0)} r_x^f[a_f, b_f], \tag{413}$$

$$q_x^f[\hat{m}_f, \hat{q}_f, \hat{\tau}_f] = \int_{\mathbb{R}^2} db_{x\to f} db_{z\to f}\, p_f^{(0)}(b_{x\to f}, b_{z\to f})\, r_x^f[a_f, b_f]^2, \tag{414}$$

$$v_x^f[\hat{m}_f, \hat{q}_f, \hat{\tau}_f] = \int_{\mathbb{R}^2} db_{x\to f} db_{z\to f}\, p_f^{(0)}(b_{x\to f}, b_{z\to f})\, v_x^f[a_f, b_f], \tag{415}$$

$$\tau_x^f[\hat{m}_f, \hat{q}_f, \hat{\tau}_f] = q_x^f[\hat{m}_f, \hat{q}_f, \hat{\tau}_f] + v_x^f[\hat{m}_f, \hat{q}_f, \hat{\tau}_f], \quad \text{(idem } z) \tag{416}$$

where the ensemble average Eq. (123) is now over scalar $b_{x\to f}$, $b_{z\to f}$, $x^{(0)}$ and $z^{(0)}$ in $\mathbb{R}$:

$$p_f^{(0)}(b_{x\to f}, b_{z\to f}, x^{(0)}, z^{(0)}) =$$
$$\mathcal{N}(b_{x\to f} \mid \hat{m}_{x\to f} x^{(0)}, \hat{q}_{x\to f})\, \mathcal{N}(b_{z\to f} \mid \hat{m}_{z\to f} x^{(0)}, \hat{q}_{z\to f})\, p_f^{(0)}(x^{(0)}, z^{(0)} \mid \hat{\tau}_f^{(0)}) \tag{417}$$

with $a_{x\to f} = \hat{\tau}_{x\to f} + \hat{q}_{x\to f}$, $a_{z\to f} = \hat{\tau}_{z\to f} + \hat{q}_{z\to f}$ and $p_f^{(0)}(x^{(0)}, z^{(0)} \mid \hat{\tau}_f^{(0)})$ given by Eq. (408). The dual mapping to the overlaps now simply reads:

$$m_x^f = \partial_{\hat{m}_{x\to f}} \bar{A}_f, \quad -\tfrac{1}{2} q_x^f = \partial_{\hat{q}_{x\to f}} \bar{A}_f, \quad -\tfrac{1}{2}\tau_x^f = \partial_{\hat{\tau}_{x\to f}} \bar{A}_f, \tag{418}$$

$$m_z^f = \partial_{\hat{m}_{z\to f}} \bar{A}_f, \quad -\tfrac{1}{2} q_z^f = \partial_{\hat{q}_{z\to f}} \bar{A}_f, \quad -\tfrac{1}{2}\tau_z^f = \partial_{\hat{\tau}_{z\to f}} \bar{A}_f. \tag{419}$$

E.5.4 Bayes-Optimal

The BO potential and overlap are given by a low-dimensional integration of the corresponding scalar EP quantities:

$$\bar{A}_f^{(0)}[\hat{m}_f] = \int_{\mathbb{R}^2} db_{x\to f} db_{z\to f}\, p_f^{(0)}(b_{x\to f}, b_{z\to f})\, A_f^{(0)}[a_f, b_f], \tag{420}$$

$$v_x^f[\hat{m}_f] = \int_{\mathbb{R}^2} db_{x\to f} db_{z\to f}\, p_f^{(0)}(b_{x\to f}, b_{z\to f})\, v_x^f[a_f, b_f], \tag{421}$$

$$m_x^f[\hat{m}_f] = \tau_x^{(0)} - v_x^f[\hat{m}_f], \quad \text{(idem } z) \tag{422}$$

where the ensemble average Eq. (145) is now over scalar $b_{x\to f}$, $b_{z\to f}$, $x^{(0)}$ and $z^{(0)}$ in $\mathbb{R}$:

$$p_f^{(0)}(b_{x\to f}, b_{z\to f}, x^{(0)}, z^{(0)}) = \tag{423}$$

$$\mathcal{N}(b_{x\to f} \mid \hat{m}_{x\to f} x^{(0)}, \hat{m}_{x\to f})\, \mathcal{N}(b_{z\to f} \mid \hat{m}_{z\to f} x^{(0)}, \hat{m}_{z\to f})\, p_f^{(0)}(x^{(0)}, z^{(0)} \mid \hat{\tau}_f^{(0)}) \tag{424}$$

with $a_{x\to f} = \hat{\tau}_{x\to f}^{(0)} + \hat{m}_{x\to f}$, $a_{z\to f} = \hat{\tau}_{z\to f}^{(0)} + \hat{m}_{z\to f}$ and $p_f^{(0)}(x^{(0)}, z^{(0)} \mid \hat{\tau}_f^{(0)})$ given by Eq. (408). The relationship Eq. (160) between the mutual information and the BO potential now reads:

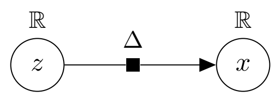
$$I_f[\hat{m}_f] = \tfrac{1}{2}\hat{m}_{x\to f}\tau_x^{(0)} + \tfrac{1}{2}\hat{m}_{z\to f}\tau_z^{(0)} - \bar{A}_f^{(0)}[\hat{m}_f] + A_f^{(0)}[\hat{\tau}_f^{(0)}]. \tag{425}$$

The dual mapping to the variances and overlaps simply reads:

$$\tfrac{1}{2}v_x^f = \partial_{\hat{m}_{x\to f}} I_f, \qquad \tfrac{1}{2}m_x^f = \partial_{\hat{m}_{x\to f}} \bar{A}_f^{(0)}, \tag{426}$$

$$\tfrac{1}{2}v_z^f = \partial_{\hat{m}_{z\to f}} I_f, \qquad \tfrac{1}{2}m_z^f = \partial_{\hat{m}_{z\to f}} \bar{A}_f^{(0)}. \tag{427}$$

E.5.5 Gaussian Noise Channel



The factor $f(x,z) = p(x \mid z) = \mathcal{N}(x \mid z, \Delta)$ is the additive Gaussian noise channel $x = z + \sqrt{\Delta}\xi$ with variance Δ and precision $a_\Delta = \Delta^{-1}$. The log-partition is given by:

$$A_f[a_f, b_f] = \frac{1}{2}b^\intercal \Sigma b + \frac{1}{2}\ln\det 2\pi\Sigma - \frac{1}{2}\ln 2\pi\Delta, \tag{428}$$

$$b = \begin{bmatrix} b_{z\to f} \\ b_{x\to f} \end{bmatrix}, \quad A = \begin{bmatrix} a_\Delta + a_{z\to f} & -a_\Delta \\ -a_\Delta & a_\Delta + a_{x\to f} \end{bmatrix}, \quad \Sigma = A^{-1}. \tag{429}$$

The posterior mean and variance are given by:

$$v_z^f = \frac{a_\Delta + a_{x\to\Delta}}{a_\Delta a}, \qquad r_z^f = v_z^f\left[b_{z\to f} + \frac{a_\Delta b_{x\to f}}{a_\Delta + a_{x\to f}}\right], \tag{430}$$

$$v_x^f = \frac{a_\Delta + a_{z\to\Delta}}{a_\Delta a}, \qquad r_x^f = v_x^f\left[b_{x\to f} + \frac{a_\Delta b_{z\to f}}{a_\Delta + a_{z\to f}}\right], \tag{431}$$

$$\text{with} \quad a = a_{x\to f} + a_{z\to f} + \frac{a_{x\to f}a_{z\to f}}{a_\Delta}. \tag{432}$$

We therefore have:

$$a_z^f = a_{z\to f} + \frac{a_\Delta a_{x\to f}}{a_\Delta + a_{x\to f}}, \qquad b_z^f = \left[b_{z\to f} + \frac{a_\Delta b_{x\to f}}{a_\Delta + a_{x\to f}}\right], \tag{433}$$

$$a_x^f = a_{x\to f} + \frac{a_\Delta a_{z\to f}}{a_\Delta + a_{z\to f}}, \qquad b_x^f = \left[b_{x\to f} + \frac{a_\Delta b_{z\to f}}{a_\Delta + a_{z\to f}}\right]. \tag{434}$$

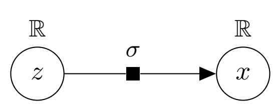
leading to the backward $f \to z$ and forward $f \to x$ updates:

$$a_{f\to z}^{\text{new}} = \frac{a_\Delta}{a_\Delta + a_{x\to f}} a_{x\to f}, \qquad b_{f\to z}^{\text{new}} = \frac{a_\Delta}{a_\Delta + a_{x\to f}} b_{x\to f}, \tag{435}$$

$$a_{f\to x}^{\text{new}} = \frac{a_\Delta}{a_\Delta + a_{z\to f}} a_{z\to f}, \qquad b_{f\to x}^{\text{new}} = \frac{a_\Delta}{a_\Delta + a_{z\to f}} b_{z\to f}. \tag{436}$$

The ensemble average variances are still given by Eqs (430)-(431).

E.5.6 Piecewise Linear Activation



Factor $f(x, z) = p(x \mid z) = \delta(x - \sigma(z))$ is the deterministic channel $x = \sigma(z)$, where we assume the activation to be piecewise linear[7]

$$\sigma(z) = \sum_{i \in I} \mathbf{1}_{R_i}(z)[x_i + \gamma_i z], \quad \mathbb{R} = \bigsqcup_{i \in I} R_i. \tag{437}$$

As the real line $\mathbb{R}$ is the disjoint union of the regions R_i the factor partition function is simply the sum of the region partition functions:

$$Z_f[a_f, b_f] = \sum_{i \in I} Z_f^i[a_f, b_f], \tag{438}$$

$$Z_f^i[a_f, b_f] = \int_{R_i} dz\, e^{-\frac{1}{2}a_{x\to f}[x_i+\gamma_i z]^2 + b_{x\to f}[x_i+\gamma_i z] - \frac{1}{2}a_{z\to f}z^2 + b_{z\to f}z}. \tag{439}$$

Thus at the factor level, the log-partition, posterior means and variances are given by:

$$A_f[a_f, b_f] = \ln \sum_{i \in I} e^{A_f^i[a_f, b_f]}, \tag{440}$$

$$r_z^f = \sum_{i \in I} p_i\, r_z^i, \quad v_z^f = \sum_{i \in I} p_i\, v_z^i + \sum_{i<j \in I} p_i\, p_j [r_z^i - r_z^j]^2, \tag{441}$$

$$r_x^f = \sum_{i \in I} p_i\, r_x^i, \quad v_x^f = \sum_{i \in I} p_i\, v_x^i + \sum_{i<j \in I} p_i\, p_j [r_x^i - r_x^j]^2. \tag{442}$$

where the region probabilities are given by

$$p_i = \frac{\exp(A_f^i)}{\sum_{j \in I} \exp(A_f^j)} \quad ie \quad p = \mathrm{softmax}\,[(A_f^i)_{i \in I}]. \tag{443}$$

The corresponding quantities at the linear region level are given by:

$$A_f^i[a_f, b_f] = -\frac{a_{x\to f} x_i^2}{2} + b_{x\to f} x_i + A_{R_i}[a_i, b_i]$$

$$\text{with} \quad a_i = a_{z\to f} + \gamma_i^2 a_{x\to f}, \quad b_i = b_{z\to f} + \gamma_i [b_{x\to f} - a_{x\to f} x_i], \tag{444}$$

$$r_z^i = r_{R_i}[a_i, b_i], \qquad r_x^i = x_i + \gamma_i r_z^i, \tag{445}$$

$$v_z^i = v_{R_i}[a_i, b_i], \qquad v_x^i = \gamma_i^2 v_z^i, \tag{446}$$

where $A_R[a, b]$, $r_R[a, b]$ and $v_R[a, b]$ denote the log-partition, mean and variance of an interval R variable, see Section E.1.8.

7. ReLU, leaky ReLU, hard tanh, hard sigmoid, sgn and abs are popular examples.

References

Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Benjamin Aubin, Bruno Loureiro, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová. The spiked matrix model with generative priors. *Advances in Neural Information Processing Systems*, 32, 2019.

Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. In *Mathematical and Scientific Machine Learning*, pages 55–73, 2020.

Jean Barbier, Mohamad Dia, Nicolas Macris, and Florent Krzakala. The mutual information in random linear estimation. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 625–632, 2016.

Jean Barbier, Nicolas Macris, Antoine Maillard, and Florent Krzakala. The mutual information in random linear estimation beyond iid matrices. In *IEEE International Symposium on Information Theory*, pages 1390–1394, 2018.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Christopher M Bishop. Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20120222, 2013.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546, 2017.

Burak Çakmak, Ole Winther, and Bernard H Fleury. S-AMP: Approximate message passing for general matrix ensembles. In *IEEE Information Theory Workshop*, pages 192–196, 2014.

Burak Çakmak, Manfred Opper, Bernard H Fleury, and Ole Winther. Self-averaging expectation propagation. *arXiv preprint arXiv:1608.06602*, 2016.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Taco Cohen and Max Welling. Harmonic exponential families on manifolds. In *International Conference on Machine Learning*, pages 1757–1765, 2015.

Lehel Csató, Manfred Opper, and Ole Winther. TAP Gibbs free energy, belief propagation and sparsity. *Advances in Neural Information Processing Systems*, 14, 2001.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory*, pages 2197–2201, 2014.

Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. *Advances in Neural Information Processing Systems*, 29, 2016.

David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Alyson K Fletcher, Sundeep Rangan, and Philip Schniter. Inference in deep networks in high dimensions. In *IEEE International Symposium on Information Theory*, 2018.

Marylou Gabrié, Andre Manoel, Clément Luneau, Jean Barbier, Nicolas Macris, Florent Krzakala, and Lenka Zdeborová. Entropy and mutual information in models of deep neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: A language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics*, volume 84, 2018.

Antoine Georges and Jonathan S Yedidia. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9): 2173, 1991.

Cédric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for convex penalized linear regression beyond Gaussian matrices. *arXiv preprint arXiv:2002.04372*, 2020.

Noah D Goodman, Vikash K Mansinghka, Daniel Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: A language for generative models. In *Uncertainty in Artificial Intelligence*, pages 220–229, 2008.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51 (4):1261–1282, 2005.

Tom Heskes and Onno Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 216–233, 2002.

Tom Heskes, Manfred Opper, Wim Wiegerinck, Ole Winther, and Onno Zoeter. Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*, 11:277–300, 2005.

Yoshiyuki Kabashima. A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111, 2003.

Yoshiyuki Kabashima. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. *Journal of Physics: Conference Series*, 95(1):012001, 2008.

Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborova. Phase transitions and sample complexity in Bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265, 2016.

Keras-VAE. Example of VAE on MNIST dataset using MLP. `https://keras.io/examples/variational_autoencoder/`.

Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(8), 2012.

Florent Krzakala, Andre Manoel, Eric W Tramel, and Lenka Zdeborová. Variational free energies for compressed sensing. In *IEEE International Symposium on Information Theory*, pages 1499–1503, 2014.

Frank R Kschischang, Brendan J Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929, 2019.

Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7), 2017.

Junjie Ma and Li Ping. Orthogonal AMP. *IEEE Access*, 5:2020–2033, 2017.

Antoine Maillard, Laura Foini, Alejandro Lage Castellanos, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. High-temperature expansions and message passing algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(11):113301, 2019.

Antoine Maillard, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083301, 2022.

Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborova. Multi-layer generalized linear estimation. In *IEEE International Symposium on Information Theory*, 2017.

Andre Manoel, Florent Krzakala, Gaël Varoquaux, Bertrand Thirion, and Lenka Zdeborová. Approximate message-passing for convex optimization with non-separable penalties. *arXiv preprint arXiv:1809.06304*, 2018.

Christopher A Metzler, Arian Maleki, and Richard G Baraniuk. Optimal recovery from compressive measurements via denoising-based approximate message passing. In *International Conference on Sampling Theory and Applications*, pages 508–512, 2015.

Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

T. Minka, J.M. Winn, J.P. Guiver, Y. Zaykov, D. Fabian, and J. Bronskill. /Infer.NET 0.3, 2018. Microsoft Research Cambridge. `http://dotnet.github.io/infer`.

Thomas P Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, 2001a.

Thomas P Minka. The EP energy function and minimization schemes, 2001b. URL `https://tminka.github.io/papers/ep/minka-ep-energy.pdf`.

Léo Miolane. Fundamental limits of low-rank matrix estimation: the non-symmetric case. *arXiv preprint arXiv:1702.00473*, 2017.

Marc Mézard and Andrea Montanari. *Information, physics, and computation.* Oxford University Press, 2009.

Hidetoshi Nishimori. *Statistical Physics of Spin Glasses and Information Processing*. Oxford University Press, 2001.

Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12:2655–2684, 2000.

Manfred Opper and Ole Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2206, 2005a.

Manfred Opper and Ole Winther. Expectation consistent free energies for approximate inference. *Advances in Neural Information Processing Systems*, 17, 2005b.

Parthe Pandit, Mojtaba Sahraee Ardakan, Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Matrix inference and estimation in multi-layer models. *Advances in Neural Information Processing Systems*, 33, 2020.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and trends in Optimization*, 1(3):127–239, 2014.

Giorgio Parisi and Marc Potters. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A: Mathematical and General*, 28(18):5267, 1995.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics*, 71(11):2275–2322, 2018.

Timm Plefka. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971, 1982.

Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symposium on Information Theory*, pages 2168–2172, 2011.

Sundeep Rangan and Alyson K Fletcher. Iterative estimation of constrained rank-one matrices in noise. In *IEEE International Symposium on Information Theory*, pages 1246–1250, 2012.

Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Vector approximate message passing. In *IEEE International Symposium on Information Theory*, pages 1588–1592, 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Galen Reeves. Additivity of information in multilayer networks via additive Gaussian noise transforms. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1064–1070, 2017.

Galen Reeves and Henry D Pfister. The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact. In *IEEE International Symposium on Information Theory*, 2016.

John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

Philip Schniter, Sundeep Rangan, and Alyson K Fletcher. Vector approximate message passing for the generalized linear model. In *Asilomar Conference on Signals, Systems and Computers*, pages 1525–1529, 2017.

Jacob Schreiber. Pomegranate: Fast and flexible probabilistic modeling in Python. *Journal of Machine Learning Research*, 18:1–6, 2018.

Takashi Shinzato and Yoshiyuki Kabashima. Learning from correlated patterns by simple perceptrons. *Journal of Physics A: Mathematical and Theoretical*, 42(1):015005, 2008a.

Takashi Shinzato and Yoshiyuki Kabashima. Perceptron capacity revisited: classification ability for correlated patterns. *Journal of Physics A: Mathematical and Theoretical*, 41 (32):324013, 2008b.

Subhojit Som and Philip Schniter. Compressive imaging using approximate message passing and a markov-tree prior. *IEEE transactions on signal processing*, 60(7):3439–3448, 2012.

Jin Tan, Yanting Ma, and Dror Baron. Compressive imaging via approximate message passing with image denoising. *IEEE Transactions on Signal Processing*, 63(8):2085–2092, 2015.

Hugo Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478(1-3):1–69, 2009.

Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

Antonia M Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1(1):1–182, 2004.

Jeremy Vila, Philip Schniter, Sundeep Rangan, Florent Krzakala, and Lenka Zdeborova. Adaptive damping and mean removal for the generalized approximate message passing algorithm. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2021–2025, 2015.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Frank Wood, Jan Willem Van De Meent, and Vikash Mansinghka. A new approach to probabilistic programming inference. *Journal of Machine Learning Research*, 33:1024–1032, 2014.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, pages 236–239, 2003.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Onno Zoeter and Tom Heskes. Change point problems in linear dynamical systems. *Journal of Machine Learning Research*, 6:1999–2026, 2005.