DEEP LEARNING VIA MESSAGE PASSING ALGORITHMS BASED ON BELIEF PROPAGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Message-passing algorithms based on the Belief Propagation (BP) equations constitute a well-known distributed computational scheme. They yield exact marginals on tree-like graphical models and have also proven to be effective in many problems defined on loopy graphs, from inference to optimization, from signal processing to clustering. The BP-based schemes are fundamentally different from stochastic gradient descent (SGD), on which the current success of deep networks is based. In this paper, we present and adapt to mini-batch training on GPUs a family of BP-based message-passing algorithms with a reinforcement term that biases distributions towards locally entropic solutions. These algorithms are capable of training multi-layer neural networks with performance comparable to SGD heuristics in a diverse set of experiments on natural datasets including multi-class image classification and continual learning, while being capable of yielding improved performances on sparse networks. Furthermore, they allow to make approximate Bayesian predictions that have higher accuracy than point-wise ones.

1 INTRODUCTION

Belief Propagation is a method for computing marginals and entropies in probabilistic inference problems (Bethe, 1935; Peierls, 1936; Gallager, 1962; Pearl, 1982). These include optimization problems as well once they are written as zero temperature limit of a Gibbs distribution that uses the cost function as energy. Learning is one particular case, in which one wants to minimize a cost which is a data dependent loss function. These problems are generally intractable and message-passing techniques have been particularly successful at providing principled approximations through efficient distributed computations.

A particularly compact representation of inference/optimization problems that is used to build massage-passing algorithms is provided by factor graphs. A factor graph is a bipartite graph composed of variables nodes and factor nodes expressing the interactions among variables. Belief Propagation is exact for tree-like factor graphs (Yedidia et al., 2003)), where the Gibbs distribution is naturally factorized, whereas it is approximate for graphs with loops. Still, loopy BP is routinely used with success in many real world applications ranging from error correcting codes, vision, clustering, just to mention a few. In all these problems, loops are indeed present in the factor graph and yet the variables are weakly correlated at long range and BP gives good results. A field in which BP has a long history is the statistical physics of disordered systems where it is known as Cavity Method (Mézard et al., 1987). It has been used to study the typical properties of spin glass models which represent binary variables interacting through random interactions over a given graph. It is very well known that in spin glass models defined on complete graphs and in locally tree-like random graphs, which are both loopy, the weak correlation conditions between variables may hold and BP give asymptotic exact results (Mézard & Montanari, 2009). Here we will mostly focus on neural networks ± 1 binary weights and sign activation functions, for which the messages and the marginals can be described simply by the difference between the probabilities associated with the +1 and -1 states, the so called *magnetizations*. The effectiveness of BP for deep learning has never been numerically tested in a systematic way, however there is clear evidence that the weak correlation decay condition does not hold and thus BP convergence and approximation quality is unpredictable.

In this paper we explore the effectiveness of a variant of BP that has shown excellent convergence properties in hard optimization problems and in non-convex shallow networks. It goes under the

name of focusing BP (fBP) and is based on a probability distribution, a likelihood, that focuses on highly entropic wide minima, neglecting the contribution to marginals from narrow minima even when they are the majority (and hence dominate the Gibbs distribution). This version of BP is thus expected to give good results only in models that have such wide entropic minima as part of their energy landscape. As discussed in (Baldassi et al., 2016a), a simple way to define fBP is to add a "reinforcement" term to the BP equations: an iteration-dependent local field is introduced for each variable, with an intensity proportional to its marginal probability computed in the previous iteration step. This field is gradually increased until the entire system becomes fully biased on a configuration. The first version of reinforced BP was introduced in (Braunstein & Zecchina, 2006) as a heuristic algorithm to solve the learning problem in shallow binary networks. Baldassi et al. (2016a) showed that this version of BP is a limiting case of fBP, i.e., BP equations written for a likelihood that uses the local entropy function instead of the error (energy) loss function. As discussed in depth in that study, one way to introduce a likelihood that focuses on highly entropic regions is to create y coupled replicas of the original system. fBP equations are obtained as BP equations for the replicated system. It turns out that the fBP equations are identical to the BP equations for the original system with the only addition of a self-reinforcing term in the message passing scheme. The fBP algorithm can be used as a solver by gradually increasing the effect of the reinforcement: one can control the size of the regions over which the fBP equations estimate the marginals by tuning the parameters that appear in the expression of the reinforcement, until the high entropy regions reduce to a single configuration. Interestingly, by keeping the size of the high entropy region fixed, the fBP fixed point allows one to estimate the marginals and entropy relative to the region.

In this work, we present and adapt to GPU computation a family of fBP inspired message passing algorithms that are capable of training multi-layer neural networks with generalization performance and computational speed comparable to SGD. This is the first work that shows that learning by message passing in deep neural networks 1) is possible and 2) is a viable alternative to SGD. Our version of fBP adds the reinforcement term at each mini-batch step in what we call the Posterior-as-Prior (PasP) rule. Furthermore, using the message-passing algorithm not as a solver but as an estimator of marginals allows us to make locally Bayesian predictions, averaging the predictions over the approximate posterior. The resulting generalization error is significantly better than those of the solver, showing that, although approximate, the marginals of the weights estimated by message-passing retain useful information. Consistently with the assumptions underlying fBP, we find that the solutions provided by the message passing algorithms belong to flat entropic regions of the loss landscape and have good performance in continual learning tasks and on sparse networks as well.

We also remark that our PasP update scheme is of independent interest and can be combined with different posterior approximation techniques.

The paper is structured as follows: in Sec. 2 we give a brief review of some related works. In Sec. 3 we provide a detailed description of the message-passing equations and of the high level structure of the algorithms. In Sec. 4 we compare the performance of the message passing algorithms versus SGD based approaches in different learning settings.

2 Related Works

The literature on message passing algorithms is extensive, we refer to (Mézard & Montanari, 2009; Zdeborová & Krzakala, 2016) for a general overview. More related to our work, multilayer message-passing algorithms have been developed in inference contexts (Manoel et al., 2017; Fletcher et al., 2018), where they have been shown to produce exact marginals under certain statistical assumptions on (unlearned) weight matrices.

The properties of message-passing for learning shallow neural networks have been extensively studied (see Baldassi et al. (2020) and reference therein). Barbier et al. (2019) rigorously show that message passing algorithms in generalized linear models perform asymptotically exact inference under some statistical assumptions. Dictionary learning and matrix factorization are harder problems closely related to deep network learning problems, in particular to the modelling of a single intermediate layer. They have been approached using message passing in Kabashima et al. (2016) and Parker et al. (2014), although the resulting predictions are found to be asymptotically inexact (Maillard et al., 2021). Same problem faced by the message passing algorithm recently proposed for a multi-layer matrix factorization scenario (Zou et al., 2021). Unfortunately, our framework as well doesn't yield

asymptotic exact predictions. Nonetheless, it gives a message passing heuristic that for the first time is able to train deep neural networks on natural datasets, therefore sets a reference for the algorithmic applications of this research line.

A few papers advocate the success of SGD to the geometrical structure (smoothness and flatness) of the loss landscape in neural networks (Baldassi et al., 2015; Chaudhari et al., 2017; Garipov et al., 2018; Li et al., 2018; Pittorino et al., 2021; Feng & Tu, 2021). These considerations do not depend on the particular form of the SGD dynamics and should extend also to other types of algorithms, although SGD is by far the most popular choice among NNs practitioners due to its simplicity, flexibility, speed, and generalization performance.

While our work focuses on message passing schemes, some of the ideas presented here, such as the PasP rule, can be combined with algorithms for Bayesian neural networks' training (Hernández-Lobato & Adams, 2015; Wu et al., 2018). Recent work extends BP by combining it with graph neural networks (Kuck et al., 2020; Satorras & Welling, 2021). Finally, some work in computational neuroscience shows similarities to our approach (Rao, 2007).

3 Learning by message passing

3.1 Posterior-as-Prior updates

We consider a multi-layer perceptron with L hidden neuron layers, having weight and bias parameters $\mathcal{W} = \{\boldsymbol{W}^\ell, \boldsymbol{b}^\ell\}_{\ell=0}^L$. We allow for stochastic activations $P^\ell(\boldsymbol{x}^{\ell+1}|\boldsymbol{z}^\ell)$, where $\boldsymbol{z}^\ell$ is the neuron's pre-activation vector for layer ℓ, and P^ℓ is assumed to be factorized over the neurons. If no stochasticity is present, P^ℓ just encodes an element-wise activation function. The probability of output y given an input $\boldsymbol{x}$ is then given by

$$P(y \,|\, \boldsymbol{x}, \mathcal{W}) = \int d\boldsymbol{x}^{1:L} \prod_{\ell=0}^{L} P^{\ell+1}(\boldsymbol{x}^{\ell+1} \mid \boldsymbol{W}^\ell \boldsymbol{x}^\ell + \boldsymbol{b}^\ell), \tag{1}$$

where for convenience we defined $\boldsymbol{x}^0 = \boldsymbol{x}$ and $\boldsymbol{x}^{L+1} = y$. In a Bayesian framework, given a training set $D = \{(\boldsymbol{x}_n, y_n)\}_n$ and a prior distribution $q_\theta(\mathcal{W})$ over the weight in some parametric family, the posterior distribution is given by

$$P(\mathcal{W} \,|\, D, \theta) \propto \prod_n P(y_n \,|\, \boldsymbol{x}_n, \mathcal{W})\, q_\theta(\mathcal{W}). \tag{2}$$

Here $\propto$ denotes equality up to a normalization factor. Using the posterior one can compute the Bayesian prediction for a new data-point $\boldsymbol{x}$ through the average $P(y \,|\, \boldsymbol{x}, D, \theta) = \int d\mathcal{W}\; P(y \,|\, \boldsymbol{x}, \mathcal{W})\, P(\mathcal{W} \,|\, D, \theta)$. Unfortunately, the posterior is generically intractable due to the hard-to-compute normalization factor. On the other hand, we are mainly interested in training a distribution that covers wide minima of the loss landscape that generalize well (Baldassi et al., 2016a) and in recovering pointwise estimators within these regions. The Bayesian modeling becomes an auxiliary tool to set the stage for the message passing algorithms seeking flat minima. We also need a formalism that allows for mini-batch training to speed-up the computation and deal with large datasets. Therefore, we devise an update scheme that we call Posterior-as-Prior (PasP), where we evolve the parameters θ^t of a distribution $q_{\theta^t}(\mathcal{W})$ computed as an approximate mini-batch posterior, in such a way that the outcome of the previous iteration becomes the prior in the following step. In the PasP scheme, θ^t retains the memory of past observations. We also add an exponential factor ρ, that we typically set close to 1, tuning the forgetting rate and playing a role similar to the learning rate in SGD. Given a mini-batch $(\boldsymbol{X}^t, \boldsymbol{y}^t)$ sampled from the training set at time t and a scalar $\rho > 0$, the PasP update reads

$$q_{\theta^{t+1}}(\mathcal{W}) \approx \left[P(\mathcal{W} \,|\, \boldsymbol{y}^t, \boldsymbol{X}^t, \theta^t)\right]^\rho, \tag{3}$$

where $\approx$ denotes approximate equality and we do not account for the normalization factor. A first approximation may be needed in the computation of the posterior, a second to project the approximate posterior onto the distribution manifold spanned by θ (Minka, 2001). In practice, we will consider factorized approximate posterior in an exponential family and priors q_θ in the same family, although Eq. 3 generically allow for more refined approximations.

Notice that setting $\rho = 1$, the batch-size to 1, and taking a single pass over the dataset, we recover the Assumed Density Filtering algorithm (Minka, 2001). For large enough ρ (including $\rho = 1$), the iterations of q_{θ^t} will concentrate on a pointwise estimator. This mechanism mimics the reinforcement heuristic commonly used to turn Belief Propagation into a solver for constrained satisfaction problems (Braunstein & Zecchina, 2006) and related to flat-minima discovery (see focusing-BP in Baldassi et al. (2016a)). A different prior updating mechanism which can be understood as empirical Bayes has been used in Baldassi et al. (2016b).

3.2 Inner message passing loop

While the PasP rule takes care of the reinforcement heuristic across mini-batches, we compute the mini-batch posterior in Eq. 3 using message passing approaches derived from Belief Propagation. BP is an iterative scheme for computing marginals and entropies of statistical models Mézard & Montanari (2009). It is most conveniently expressed on factor graphs, that is bipartite graphs where the two sets of nodes are called variable nodes and factor nodes. They respectively represent the variables involved in the statistical model and their interactions. Message from factor nodes to variable nodes and viceversa are exchanged along the edges of the factor graph in the BP iterations.

The factor graph for $P(\mathcal{W} \,|\, \boldsymbol{X}^t, \boldsymbol{y}^t, \theta^t)$ can be derived from Eq. 2, with the following additional specifications. For simplicity, we will ignore the bias term in each layer. We assume factorized $q_{\theta^t}(\mathcal{W})$, each factor parameterized by its first two moments. In what follows, we drop the PasP iteration index t. For each example $(\boldsymbol{x}_n, y_n)$ in the mini-batch, we introduce the auxiliary variables $\boldsymbol{x}_n^\ell, \ell = 1, \dots, L$, representing the layers' activations. For each example, each neuron in the network contributes a factor node to the factor graph. The scalar components of the weight matrices and the activation vectors become variable nodes. This construction is presented in Appendix A, where we also derive the message update rules on the factor graph. The factor graph thus defined is extremely loopy and straightforward iteration of Belief Propagation has convergence issues. Moreover, in presence of a homogeneous prior over the weights, the neuron permutation symmetry in each hidden layer induces a strongly attractive symmetric fixed point that hinders learning. We work around these issues by breaking the symmetry at time $t = 0$ with a inhomogeneous prior $\theta^{t=0}$. In our experiments a little initial heterogeneity is sufficient to obtain specialized neurons at each following time step. Additionally, we do not require message passing convergence in the inner loop (see Algorithm 1) but perform one or a few iterations for each θ update. We also include an inertia term, commonly called damping factor, in the message updates (see B.2). As we shall discuss, these simple rules suffice to train deep networks by message passing.

For the inner loop we adapt to deep neural networks four different message passing algorithms, all of which are well known to the literature although derived in simpler settings: Belief Propagation (BP), BP-Inspired (BPI) message passing, mean-field (MF), and approximate message passing (AMP). The last three can be considered approximations of the first one. In the following paragraphs we will discuss their common traits, present the BP updates as an example, and refer to Appendix A for an in-depth exposition. For all algorithms, message updates can be divided in a forward pass and backward pass, as also done in (Fletcher et al., 2018) in a multi-layer inference setting. The BP algorithm is compactly reported in Algorithm 1.

Meaning of messages. All the messages involved in the message passing can be understood in terms of cavity marginals or full marginals (as mentioned in the introduction BP is also known as Cavity Method, see (Mézard & Montanari, 2009)). Of particular relevance are m_{ki}^ℓ and σ_{ki}^ℓ, denoting the mean and variance of the weights W_{ki}^ℓ. The quantities $\hat{x}_{in}^\ell$ and Δ_{in}^ℓ instead denote the mean and variance of the i-th neuron's activation in layer ℓ for a given input $\boldsymbol{x}_n$.

Scalar free energies. All message passing schemes are conveniently expressed in terms of two functions, corresponding respectively to single neuron and single weight effective free-energies:

$$\varphi^\ell(B, A, \omega, V) = \log \int \mathrm{d}x \, \mathrm{d}z \; e^{-\frac{1}{2}Ax^2 + Bx} \, P^\ell\left(x|z\right) e^{-\frac{(\omega - z)^2}{2V}} \qquad \ell = 1, \dots, L \tag{4}$$

$$\psi(H, G, \theta) = \log \int \mathrm{d}w \; e^{-\frac{1}{2}G^2 w^2 + Hw} \, q_\theta(w) \tag{5}$$

Notice that for common deterministic activations such as ReLU and sign, the function φ has analytic and smooth expressions (see Appendix A.8). The same holds for the function ψ when $q_\theta(w)$ is Gaussian (continuous weights) or a mixture of atoms (discrete weights). At the last layer we impose $P^{L+1}(y|z) = \mathbb{I}(y = \mathrm{sign}(z))$ in binary classification tasks and $P^{L+1}(y|\boldsymbol{z}) = \mathbb{I}(y = \arg\max(\boldsymbol{z}))$ in multi-class classification (see Appendix A.9). While in our experiments we use hard constraints for the final output, therefore solving a constraint satisfaction problem, it would be interesting to also consider soft constraints and introduce a temperature, but this is beyond the scope of our work.

Start and end of message passing. At the beginning of a new PasP iteration t, we reset the messages to zero (see Appendix A) and run message passing for $\tau_{\max}$ iterations. We then compute the new prior $q_{\theta^{t+1}}(\mathcal{W})$ from the posterior given by the message passing.

BP Forward pass. After initialization of the messages at time $\tau = 0$, at each following time we propagate a set of message from the first to the last layer and then another set from the last to the first. For an intermediate layer ℓ the forward pass reads

$$\hat{x}^{\ell,\tau}_{in\to k} = \partial_B \varphi^\ell \left(B^{\ell,\tau-1}_{in\to k}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in} \right) \tag{6}$$

$$\Delta^{\ell,\tau}_{in} = \partial^2_B \varphi^\ell \left(B^{\ell,\tau-1}_{in}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in} \right) \tag{7}$$

$$m^{\ell,\tau}_{ki\to n} = \partial_H \psi(H^{\ell,\tau-1}_{ki\to n}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) \tag{8}$$

$$\sigma^{\ell,\tau}_{ki} = \partial^2_H \psi(H^{\ell,\tau-1}_{ki}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) \tag{9}$$

$$V^{\ell,\tau}_{kn} = \sum_i \left(\left(m^{\ell,\tau}_{ki\to n} \right)^2 \Delta^{\ell,\tau}_{in} + \sigma^{\ell,\tau}_{ki} (\hat{x}^{\ell,\tau}_{in\to k})^2 + \sigma^{\ell,\tau}_{ki} \Delta^{\ell,\tau}_{in} \right) \tag{10}$$

$$\omega^{\ell,\tau}_{kn\to i} = \sum_{i'\neq i} m^{\ell,\tau}_{ki'\to n} \hat{x}^{\ell,\tau}_{i'n\to k} \tag{11}$$

The equations for the first and the last layer differ slightly and in an intuitive way from the ones above (see Appendix).

BP Backward pass. The backward pass updates a set of messages from the last to the first layer:

$$g^{\ell,\tau}_{kn\to i} = \partial_\omega \varphi^{\ell+1} \left(B^{\ell+1,\tau}_{kn}, A^{\ell+1,\tau}_{kn}, \omega^{\ell,\tau}_{kn\to i}, V^{\ell,\tau}_{kn} \right) \tag{12}$$

$$\Gamma^{\ell,\tau}_{kn} = -\partial^2_\omega \varphi^{\ell+1} \left(B^{\ell+1,\tau}_{kn}, A^{\ell+1,\tau}_{kn}, \omega^{\ell,\tau}_{kn}, V^{\ell,\tau}_{kn} \right) \tag{13}$$

$$A^{\ell,\tau}_{in} = \sum_k \left((m^{\ell,\tau}_{ki\to n})^2 + \sigma^{\ell,\tau}_{ki} \right) \Gamma^{\ell,\tau}_{kn} - \sigma^{\ell,\tau}_{ki} \left(g^{\ell,\tau}_{kn\to i} \right)^2 \tag{14}$$

$$B^{\ell,\tau}_{in\to k} = \sum_{k'\neq k} m^{\ell,\tau}_{k'i\to n} g^{\ell,\tau}_{k'n\to i} \tag{15}$$

$$G^{\ell,\tau}_{ki} = \sum_n \left((\hat{x}^{\ell,\tau}_{in\to k})^2 + \Delta^{\ell,\tau}_{in} \right) \Gamma^{\ell,\tau}_{kn} - \Delta^{\ell,\tau}_{in} \left(g^{\ell,\tau}_{kn\to i} \right)^2 \tag{16}$$

$$H^{\ell,\tau}_{ki\to n} = \sum_{n'\neq n} \hat{x}^{\ell,\tau}_{in'\to k} g^{\ell,\tau}_{kn'\to i} \tag{17}$$

Again, the equations for the first and the last layer have a slightly different form.

On the parameter $\tau_{\max}$ In our experiments we fix $\tau_{\max} = 1$. We typically do not observe an increase in perfomance taking more steps, except for some specific cases and in particular for MF layers. We remark that for $\tau_{\max} = 1$ the BP and the BPI equations are identical, so in most of the subsequent numerical results we will only investigate BP.

Computational complexity The message passing equations boil down to element-wise operations and tensor contractions that we easily implement using the GPU friendly julia library Tullio.jl (Abbott

et al., 2021). For a layer of input and output size N and considering a batch-size of B, the time complexity of a forth-and-back iteration is $O(N^2B)$ for all message passing algorithms (BP, BPI, MF, and AMP), the same as SGD. The prefactor varies and it is generally larger than SGD (see Appendix B.8). Also, time complexity for message passing is proportional to $\tau_{\max}$ (which we typically set to 1). We provide our implementation in the GitHub repo **anonymous**.

Algorithm 1: BP for deep neural networks

```
// Message passing used in the PasP Eq.  3 to approximate.
// the mini-batch posterior.
// Here we specifically refer to BP updates.
// BPI, MF, and AMP updates take the same form but using
// the rules in Appendix A.4, A.5, and A.7 respectively
Initialize messages.
for τ = 1, ... τ_max do
    // Forward Pass
    for l = 0, ..., L do
        compute x̂^ℓ, Δ^ℓ using (6, 7)
        compute m^ℓ, σ^ℓ using (8, 9)
        compute V^ℓ, ω^ℓ using (10, 11)
    // Backward Pass
    for l = L, ..., 0 do
        compute g^ℓ, Γ^ℓ using (12, 13)
        compute A^ℓ, B^ℓ using (14, 15)
        compute G^ℓ, H^ℓ using (16, 17)
```

4 NUMERICAL RESULTS

We implement our message passing algorithms on neural networks with continuous and binary weights and with binary activations. We compare their performance with a SGD based algorithm adapted to binary architectures (Hubara et al., 2016) which we call BinaryNet along the paper (see Appendix B.6 for details). Comparison of Bayesian predictions are with the gradient-based Expectation Backpropagation (EBP) algorithm (Soudry et al., 2014), also able to deal with discrete weights and activation. In all architectures we avoid the use of bias terms and batch-normalization layers.

We find that message-passing algorithms are able to train generic MLP architectures (with varying numbers and sizes of hidden layers). As for the datasets, we are able to perform both binary classification and multi-class classification on standard computer vision datasets such as MNIST, Fashion-MNIST, and CIFAR-10. Since these datasets consist of 10 classes, for the binary classification task we divide each dataset in two classes (even vs odd).

We report that it is possible to find values of the hyper-parameters such that message passing algorithms are able to solve these optimization problems with generalization performance comparable to or better than competing algorithms. The training error is usually lower for optimized configurations from message passing schemes, suggesting that these algorithms are able to achieve higher capacity than SGD-based algorithms (given the same test error). Timings of our GPU implementations of the message passing algorithms are competitive with SGD (see Appendix B.8).

4.1 EXPERIMENTS ACROSS ARCHITECTURES

We select a specific task, namely we perform multi-class classification on Fashion-MNIST, and we compare the message passing algorithms with BinaryNet for different choices of the architecture (i.e. we vary the number and the size of the hidden layers). In Fig.1 (Left), we present the learning curves for a MLP with 3 hidden layers with 501 units with binary weights and activations. Similar results hold in our experiments with 2 or 3 hidden layers of 101, 501 or 1001 units each and with batch sizes

from 1 to from 1024. The parameters used in our simulations are reported in Appendix B.3. Results on networks with continuous weights can be found in Fig.2 (Right).

4.2 SPARSE LAYERS

Since the BP algorithm has notoriously been successful on sparse graphs, we perform a straightforward implementation of pruning at initialization, i.e. we impose a random boolean mask on the weights that we keep fixed along the training. We call *sparsity* the fraction of zeroed weights. This kind of non-adaptive pruning is known to largely hinder learning (Frankle et al., 2021; Sung et al., 2021). In the right panel of Fig. 1 (Right), we report results on sparse binary networks, in which we train a MLP with 2 hidden layers of 101 units on the MNIST dataset (see also (Han et al., 2016; Ardakani et al., 2017; Tung & Mori, 2018; Diffenderfer & Kailkhura, 2021) for results on pruning quantized/binary networks). In this setting BP and MF perform better than BinaryNet at high sparsity (up to 90%). AMP struggles with sparsity similarly to BinaryNet instead.

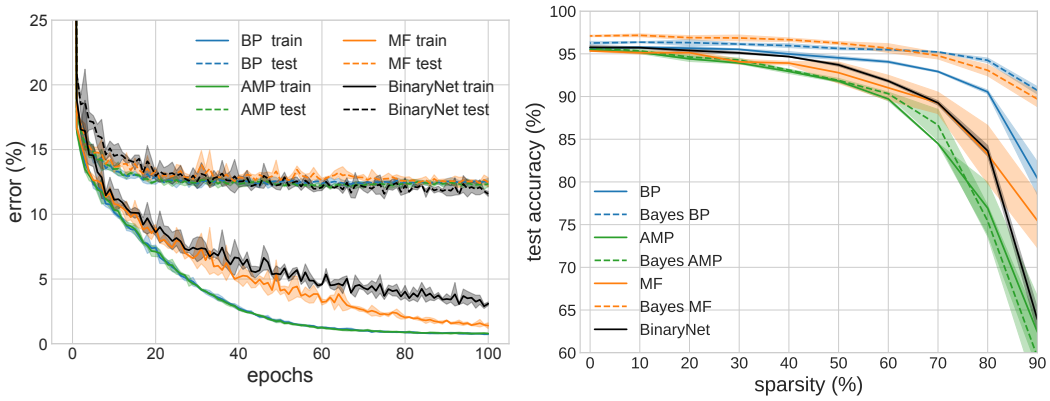

Figure 1: (Left) Test and training error curves of the message passing algorithms compared with BinaryNet on the Fashion-MNIST dataset (multi-class classification) with a MLP with 3 hidden layers of 501 units. (Right) Final test accuracy in function of the sparsity (i.e. percentage of pruned weights with a fixed mask imposed at initialization) of a MLP with 2 hidden layers of 101 units on the MNIST dataset (multi-class). In both panels the batch-size is 128 and curves are averaged over 5 realizations of the initial conditions (and sparsity pattern in the right panel).

4.3 EXPERIMENTS ACROSS DATASETS

We now fix the architecture, a MLP with 2 hidden layers of 501 neurons each with binary weights and activation, and vary the dataset, i.e. we test the BP-based algorithms on standard computer vision benchmark datasets such as MNIST, Fashion-MNIST and CIFAR-10, in both the multi-class and binary classification tasks. In Tab. 1 we report the final test errors (averaged over 5 realizations of the initial conditions) obtained by the message passing algorithms (BP, AMP, MF) compared to the BinaryNet baseline. See Appendix B.4 for the corresponding training errors and the parameters used in the simulations. We mention that while the test performance is mostly comparable, the train error tends to be lower for the message passing algorithms.

4.4 LOCALLY BAYESIAN ERROR

The message passing framework, used as an estimator of the mini-batch posterior marginals, allows us to perform approximate Bayesian prediction, i.e. averaging the pointwise predictions over the approximate posterior. We observe better generalization error from Bayesian predictions compared to point-wise ones, showing that the marginals retain useful information. However, we roughly estimate the marginals with the PasP mini-batch procedure (the exact ones should be computed with a

Dataset	BinaryNet	BP	AMP	MF
MNIST (2 classes)	1.3 ± 0.1	1.4 ± 0.2	1.4 ± 0.1	1.3 ± 0.2
Fashion-MNIST (2 classes)	2.4 ± 0.1	2.3 ± 0.1	2.4 ± 0.1	2.3 ± 0.1
CIFAR-10 (2 classes)	30.0 ± 0.3	31.4 ± 0.1	31.1 ± 0.3	31.1 ± 0.4
MNIST	2.2 ± 0.1	2.6 ± 0.1	2.6 ± 0.1	2.3 ± 0.1
Fashion-MNIST	12.0 ± 0.6	11.8 ± 0.3	11.9 ± 0.2	12.1 ± 0.2
CIFAR-10	59.0 ± 0.7	58.7 ± 0.3	58.5 ± 0.2	60.4 ± 1.1

Table 1: Test error (%) on Fashion-MNIST of various algorithms on a MLP with 2 hidden layers of 501 units, binary weights and activations. All algorithms are trained with batch-size 128 and for 100 epochs. Mean and standard deviations are calculated over 5 random initializations.

full-batch procedure, but this converges with difficulty in our tests). Since BP-based algorithms tend to focus on dense states (as also confirmed by the local energy measure performed in Appendix B.5), the Bayesian error we compute can be considered as a local approximation of the full one. We report results for binary classification on the MNIST dataset in Fig. 2, and we observe the same performance increase on different datasets and architectures. We obtain the Bayesian prediction from the output marginal given by a single forward pass of the message passing. To obtain good Bayesian estimates, it is important that the posterior distribution does not concentrate too much (otherwise, the Bayesian prediction will converge to the prediction of a single configuration).

In Fig.2 we also perform a comparison of BP (point-wise and Bayesian) with SGD and another algorithm able to perform Bayesian predictions, Expectation Backpropagation (Soudry et al., 2014).

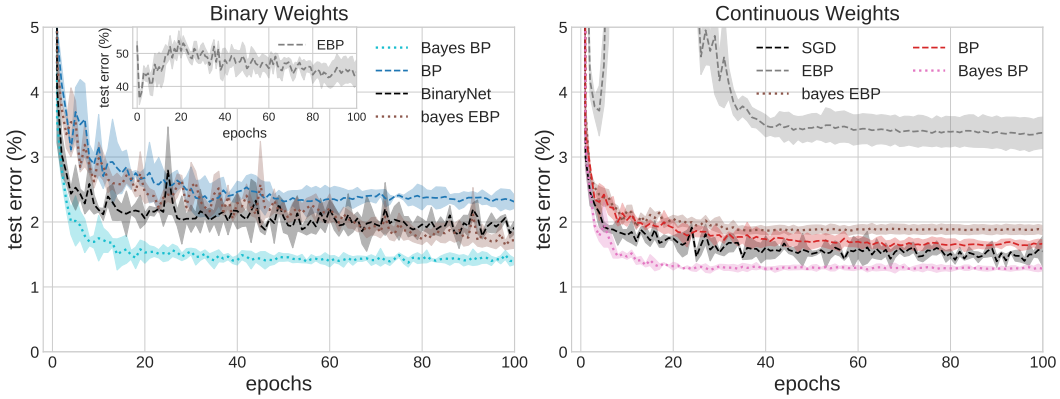

Figure 2: (Left) Test error curves for Bayesian and point-wise predictions for a MLP with 2 hidden layers of 101 units on the 2-class MNIST dataset. We report the results for (Left) binary and (Right) continuous weights. In both cases, we compare SGD, BP (point-wise and Bayesian) and EBP (point-wise and Bayesian). See Appendix B.3 for details.

4.5 Continual learning

Given the high local entropy (i.e. the flatness) of the solutions found by the BP-based algorithms (see Appendix B.5), we perform additional tests in a classic setting, continual learning, where the possibility of locally rearranging the solutions while keeping low training error can be an advantage. When a deep network is trained sequentially on different tasks, it tends to forget exponentially fast previously seen tasks while learning new ones (McCloskey & Cohen, 1989; Robins, 1995; Fusi et al., 2005). Recent work (Feng & Tu, 2021) has shown that searching for a flat region in the loss landscape can indeed help to prevent catastrophic forgetting. Several heuristics have been proposed to mitigate the problem (Kirkpatrick et al., 2017; Aljundi et al., 2018; Zenke et al., 2017; Laborieux et al., 2021) but all require specialized adjustments to the loss or the dynamics .

Here we show instead that our message passing scheme is naturally prone to learn multiple tasks sequentially, mitigating the characteristic memory issues of the gradient-based schemes without the need for explicit modifications. As a prototypical experiment, we sequentially trained a multi-layer neural network on 6 different versions of the MNIST dataset, where the pixels of the images have been randomly permuted (Goodfellow et al., 2013), giving a fixed budget of 40 epochs on each task. We present the results for a two hidden layer neural network with 2001 units on each layer (see Appendix B.3 for details). As can be seen in Fig. 3, at the end of the training the BP algorithm is able to reach good generalization performances on all the tasks. We compared the BP performance with BinaryNet, which already performs better than SGD with continuous weights (see the discussion in Laborieux et al. (2021)). While our BP implementation is not competitive with ad-hoc techniques specifically designed for this problem, it beats non-specialized heuristics. Moreover, we believe that specialized approaches like the one of Laborieux et al. (2021) can be adapted to message passing as well.

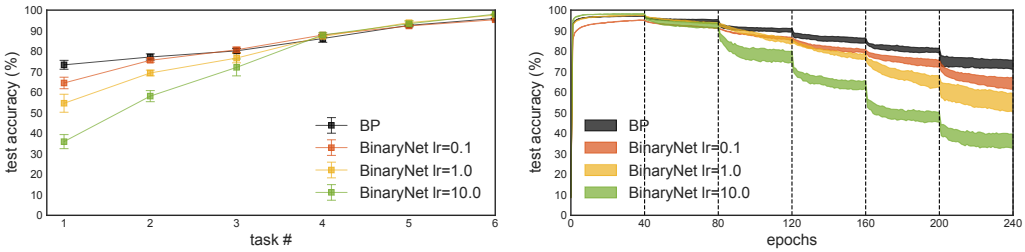

Figure 3: Performance of BP and BinaryNet on the permuted MNIST task (see text) for a two hidden layer network with 2001 units on each layer and binary weights and activations. The model is trained sequentially on 6 different versions of the MNIST dataset (the tasks), where the pixels have been permuted. (Left) Test accuracy on each task after the network has been trained on all the tasks. (Right) Test accuracy on the first task as a function of the number of epochs. Points are averages over 5 independent runs, shaded areas are errors on the mean.

5 Discussion and conclusions

While successful in many fields, message passing algorithms, have notoriously struggled to scale to deep neural networks training problems. Here we have developed a class of fBP-based message passing algorithms and used them within an update scheme, Posterior-as-Prior (PasP), that makes it possible to train deep and wide multilayer perceptrons by message passing.

We performed experiments with binary and continuous weights, and binary activations. Future work should try to include different activations, biases, batch-normalization, and convolutional layers as well. Another interesting direction is the algorithmic computation of the (local) entropy of the model from the messages.

Further theoretical work is needed for a more complete understanding of the robustness of our methods. Recent developments in message passing algorithms (Rangan et al., 2019) and related theoretical analysis (Goldt et al., 2020) could provide fruitful inspirations. While our algorithms can be used for approximate Bayesian inference, exact posterior calculation is still out of reach for message passing approaches and much technical work is needed in that direction.

Acknowledgments

We thank anonymous for useful discussions on the numerical implementation of the BP equations.

References

Michael Abbott, Dilum Aluthge, N3N5, Simeon Schaub, Carlo Lucibello, Chris Elrod, and Johnny Chen. Tullio.jl julia package, 2021. URL `https://github.com/mcabbott/Tullio.jl`.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

Arash Ardakani, Carlo Condo, and Warren J. Gross. Sparsely-connected neural networks: Towards efficient VLSI implementation of deep neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=r1fYuytex.

Carlo Baldassi, Alfredo Braunstein, Nicolas Brunel, and Riccardo Zecchina. Efficient supervised learning in networks with binary synapses. *Proceedings of the National Academy of Sciences*, 104(26):11079–11084, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0700324104. URL https://www.pnas.org/content/104/26/11079.

Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Phys. Rev. Lett.*, 115:128101, Sep 2015. doi: 10.1103/PhysRevLett.115.128101. URL https://link.aps.org/doi/10.1103/PhysRevLett.115.128101.

Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 113(48):E7655–E7662, 2016a. ISSN 0027-8424. doi: 10.1073/pnas.1608103113. URL https://www.pnas.org/content/113/48/E7655.

Carlo Baldassi, Federica Gerace, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Learning may need only a few bits of synaptic precision. *Phys. Rev. E*, 93:052313, May 2016b. doi: 10.1103/PhysRevE.93.052313. URL https://link.aps.org/doi/10.1103/PhysRevE.93.052313.

Carlo Baldassi, Fabrizio Pittorino, and Riccardo Zecchina. Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences*, 117(1): 161–170, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1908636117. URL https://www.pnas.org/content/117/1/161.

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. ISSN 0027-8424. doi: 10.1073/pnas.1802705116. URL https://www.pnas.org/content/116/12/5451.

Hans Bethe. Statistical theory of superlattices. *Proc. R. Soc. A*, 150:552, 1935.

Alfredo Braunstein and Riccardo Zecchina. Learning by message passing in networks of discrete synapses. *Phys. Rev. Lett.*, 96:030201, Jan 2006. doi: 10.1103/PhysRevLett.96.030201. URL https://link.aps.org/doi/10.1103/PhysRevLett.96.030201.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=B1YfAfcgl.

James Diffenderfer and Bhavya Kailkhura. Multi-prize lottery ticket hypothesis: Finding accurate binary neural networks by pruning a randomly weighted network. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=U_mat0b9iv.

Yu Feng and Yuhai Tu. The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

Alyson K Fletcher, Sundeep Rangan, and Philip Schniter. Inference in deep networks in high dimensions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1884–1888. IEEE, 2018.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ig-VyQc-MLK.

Stefano Fusi, Patrick J Drew, and Larry F Abbott. Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611, 2005.

Marylou Gabrié. Mean-field inference methods for neural networks. *Journal of Physics A: Mathematical and Theoretical*, 53(22):223002, 2020.

Robert Gallager. Low-density parity-check codes. *IRE Transactions on information theory*, 8(1): 21–28, 1962.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/be3087e74e9100d4bc4c6268cdbe8456-Paper.pdf.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10 (4):041044, 2020.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1510.00149.

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1861–1869. JMLR.org, 2015.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborová. Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265, 2016. doi: 10.1109/TIT.2016.2556702.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Jonathan Kuck, Shuvam Chakraborty, Hao Tang, Rachel Luo, Jiaming Song, Ashish Sabharwal, and Stefano Ermon. Belief propagation neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 667–678. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/07217414eb3fbe24d4e5b6cafb91ca18-Paper.pdf.

Axel Laborieux, Maxence Ernoult, Tifenn Hirtzlin, and Damien Querlioz. Synaptic metaplasticity in binarized neural networks. *Nature Communications*, 12(1):2549, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22768-y. URL https://doi.org/10.1038/s41467-021-22768-y.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

Antoine Maillard, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising. *arXiv preprint arXiv:2110.08775*, 2021.

Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Multi-layer generalized linear estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2098–2102, 2017. doi: 10.1109/ISIT.2017.8006899.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Marc Mézard. Mean-field message-passing equations in the hopfield model and its generalizations. *Physical Review E*, 95(2):022117, 2017.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Thomas P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, Inc., USA, 2009. ISBN 019857083X.

Jason T Parker, Philip Schniter, and Volkan Cevher. Bilinear generalized approximate message passing—part i: Derivation. *IEEE Transactions on Signal Processing*, 62(22):5839–5853, 2014.

Judea Pearl. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science . . . , 1982.

R. Peierls. On ising's model of ferromagnetism. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(3):477–481, 1936. doi: 10.1017/S0305004100019174.

Fabrizio Pittorino, Carlo Lucibello, Christoph Feinauer, Gabriele Perugini, Carlo Baldassi, Elizaveta Demyanenko, and Riccardo Zecchina. Entropic gradient descent algorithms and wide flat minima. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xjXg0bnoDmS.

Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019.

Rajesh P. N. Rao. *Neural models of Bayesian belief propagation.*, pp. 239–267. Bayesian brain: Probabilistic approaches to neural coding. MIT Press, Cambridge, MA, US, 2007. ISBN 026204238X (Hardcover); 978-0-262-04238-3 (Hardcover).

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2): 123–146, 1995.

Victor Garcia Satorras and Max Welling. Neural enhanced belief propagation on factor graphs. In *International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2021.

Daniel Soudry, Itay Hubara, and Ron Meir. Expectation backpropagation: Parameter-free training of multilayer neural networks with continuous or discrete weights. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/076a0c97d09cf1a0ec3e19c7f2529f2b-Paper.pdf.

George Stamatescu, Federica Gerace, Carlo Lucibello, Ian Fuss, and Langford B. White. Critical initialisation in continuous approximations of binary neural networks. 2020. URL https://openreview.net/forum?id=rylmoxrFDH.

Yi-Lin Sung, Varun Nair, and Colin Raffel. Training neural networks with fixed sparse masks, 2021.

Frederick Tung and Greg Mori. Clip-q: Deep network compression learning by in-parallel pruning-quantization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7873–7882, 2018. doi: 10.1109/CVPR.2018.00821.

Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E Turner, Jose Miguel Hernandez-Lobato, and Alexander L Gaunt. Deterministic variational inference for robust bayesian neural networks. *arXiv preprint arXiv:1810.03958*, 2018.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. *Understanding Belief Propagation and Its Generalizations*, pp. 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003. ISBN 1558608117.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

Qiuyun Zou, Haochuan Zhang, and Hongwen Yang. Multi-layer bilinear generalized approximate message passing. *IEEE Transactions on Signal Processing*, 69:4529–4543, 2021. doi: 10.1109/TSP.2021.3100305.

Appendices

CONTENTS

A BP-based message passing algorithms **14**

A.1 Preliminary considerations 14

A.2 Derivation of the BP equations 15

A.3 BP equations 18

A.4 BPI equations 19

A.5 MF equations 20

A.6 Derivation of the AMP equations 20

A.7 AMP equations 22

A.8 Activation Functions 23

A.9 The ArgMax layer 23

B Experimental details **25**

B.1 Hyper-parameters of the BP-based scheme 25

B.2 Damping scheme for the message passing 25

B.3 Architectures 26

B.4 Varying the dataset 26

B.5 Local energy 27

B.6 SGD implementation (BinaryNet) 27

B.7 Unit polarization and overlaps 28

B.8 Computational performance: varying batch-size 29

A BP-BASED MESSAGE PASSING ALGORITHMS

A.1 PRELIMINARY CONSIDERATIONS

Given a mini-batch $\mathcal{B} = \{(\boldsymbol{x}_n, y_n)\}_n$, the factor graph defined by Eqs. (1, 2, 18) is explicitly written as:

$$P(\mathcal{W}, \boldsymbol{x}^{1:L} \,|\, \mathcal{B}, \theta) \propto \prod_{\ell=0}^{L} \prod_{k,n} P^{\ell+1}\left(x_{kn}^{\ell+1} \,\middle|\, \sum_i W_{ki}^{\ell} x_{in}^{\ell} \right) \prod_{k,i,\ell} q_\theta(W_{ki}^{\ell}), \tag{18}$$

where $\boldsymbol{x}_n^0 = \boldsymbol{x}_n,\ \boldsymbol{x}_n^{L+1} = y_n$. The derivation of the BP equations for this model is straightforward albeit lengthy and involved. It is obtained following the steps presented in multiple papers, books, and reviews, see for instance (Mézard & Montanari, 2009; Zdeborová & Krzakala, 2016; Mézard, 2017), although it has not been attempted before in deep neural networks. It should be noted that a (common) approximation that we take here with respect to the standard BP scheme, is that messages are assumed to be Gaussian distributed and therefore parameterized by their mean and variance. This goes by the name of relaxed belied propagation (rBP), just referred to as BP throughout the paper.

We derive the BP equations in A.2 and present them all together in A.3.

From BP, we derive other 3 message passing algorithms useful for the deep network training setting, all of which are well known to the literature: BP-Inspired (BPI) message passing A.4, mean-field (MF) A.4, and approximate message passing (AMP) A.7. The AMP derivation is the more involved

and given in A.6. In all these cases, message updates can be divided in a forward pass and a backward pass, as also done in Fletcher et al. (2018) in a multi-layer inference setting. The BP algorithm is compactly reported in Algorithm 1.

In our notation, ℓ denotes the layer index, τ the BP iteration index, k an output neuron index, i an input neuron index, and n a sample index.

We report below, for convenience, some of the considerations also present in the main text.

Meaning of messages. All the messages involved in the message passing equations can be understood in terms of cavity marginals or full marginals (as mentioned in the introduction BP is also known as the Cavity Method, see (Mézard & Montanari, 2009)). Of particular relevance are the quantities m^ℓ_{ki} and σ^ℓ_{ki}, denoting the mean and variance of the weights W^ℓ_{ki}. The quantities $\hat{x}^\ell_{in}$ and Δ^ℓ_{in} instead denote mean and variance of the i-th neuron's activation in layer ℓ for a given input $\boldsymbol{x}_n$.

Scalar free energies. All message passing schemes can be expressed using the following scalar functions, corresponding to single neuron and single weight effective free-energies respectively:

$$\varphi^\ell(B, A, \omega, V) = \log \int \mathrm{d}x\, \mathrm{d}z\; e^{-\frac{1}{2}Ax^2+Bx}\, P^\ell\left(x|z\right) e^{-\frac{(\omega-z)^2}{2V}} \tag{19}$$

$$\psi(H, G, \theta) = \log \int \mathrm{d}w\; e^{-\frac{1}{2}G^2w^2+Hw}\, q_\theta(w) \tag{20}$$

These free energies will naturally arise in the derivation of the BP equations in Appendix A.2.

Notice that for common deterministic activations such as ReLU and sign, the function φ has analytic and smooth expressions that we give in Appendix A.8. Same goes for ψ when $q_\theta(w)$ is Gaussian (continuous weights) or a mixture of atoms (discrete weights). At the last layer we impose $P^{L+1}(y|z) = \mathbb{I}(y = \mathrm{sign}(z))$ in binary classification tasks and $P^{L+1}(y|\boldsymbol{z}) = \mathbb{I}(y = \arg\max(\boldsymbol{z}))$ in multi-class classification (see A.9). While in our experiments we use hard constraints for the final output and deterministic activation functions, therefore solving a constraint satisfaction problem, it would be interesting to also consider soft constraints and introduce a temperature, but this is beyond the scope of our work.

Binary weights. In our experiments we use ± 1 weights in each layer. Therefore each marginal can be parameterized by a single number and our prior/posterior takes the form

$$q_\theta(W^\ell_{ki}) \propto e^{\theta^\ell_{ki} W^\ell_{ki}} \tag{21}$$

The effective free energy function Eq. 20 becomes

$$\psi(H, G, \theta^\ell_{ki}) = \log 2\cosh(H + \theta^\ell_{ki}) \tag{22}$$

and the messages G can be dropped from the message passing.

Start and end of message passing. At the beginning of a new PasP iteration t, we reset the messages to zero and run message passing for $\tau_{\max}$ iterations. We then compute the new prior $q_{\theta^{t+1}}(\mathcal{W})$ from the posterior given by the message passing iterations.

A.2 Derivation of the BP equations

In order to derive the BP equations, we start with the following portion of the factor graph reported in Eq. 18 in the main text, describing the contribution of a single data example in the inner loop of the PasP updates:

$$\prod_{\ell=0}^{L} \prod_{k} P^{\ell+1}\left(x^{\ell+1}_{kn} \,\middle|\, \sum_i W^\ell_{ki} x^\ell_{in}\right) \quad \text{where } \boldsymbol{x}^0_n = \boldsymbol{x}_n,\ \boldsymbol{x}^{L+1}_n = y_n. \tag{23}$$

where we recall that the quantity x^ℓ_{kn} corresponds to the activation of neuron k in layer ℓ in correspondence of the input example n.

Let us start by analyzing the single factor:

$$P^{\ell+1}\left(x_{kn}^{\ell+1} \,\middle|\, \sum_i W_{ki}^{\ell} x_{in}^{\ell}\right) \tag{24}$$

We refer to messages that travel from input to output in the factor graph as *upgoing* or *upwards* messages, while to the ones that travel from output to input as *downgoing* or *backwards* messages.

Factor-to-variable-W messages The factor-to-variable-W messages read:

$$\hat{\nu}_{kn\to ki}^{\ell+1}(W_{ki}^{\ell}) \propto \int \prod_{i'\neq i} d\nu_{ki'\to n}^{\ell}(W_{ki'}^{\ell}) \prod_{i'} d\nu_{i'n\to k}^{\ell}(x_{i'n}^{\ell})\, d\nu_{\downarrow}(x_{kn}^{\ell+1})\, P^{\ell+1}\left(x_{kn}^{\ell+1} \,\middle|\, \sum_{i'} W_{ki'}^{\ell} x_{i'n}^{\ell}\right) \tag{25}$$

where $\nu_{\downarrow}$ denotes the messages travelling downwards (from output to input) in the factor graph.

We denote the means and variances of the incoming messages respectively with $m_{ki\to n}^{\ell}$, $\hat{x}_{in\to k}^{\ell}$ and $\sigma_{ki\to n}^{\ell}$, $\Delta_{in\to k}^{\ell}$:

$$m_{ki\to n}^{\ell} = \int d\nu_{ki\to n}^{\ell}(W_{ki}^{\ell})\, W_{ki}^{\ell} \tag{26}$$

$$\sigma_{ki\to n}^{\ell} = \int d\nu_{ki\to n}^{\ell}(W_{ki}^{\ell})\, \left(W_{ki}^{\ell} - m_{ki\to n}^{\ell}\right)^2 \tag{27}$$

$$\hat{x}_{in\to k}^{\ell} = \int d\nu_{in\to k}^{\ell}(x_{in}^{\ell})\, x_{in}^{\ell} \tag{28}$$

$$\Delta_{in\to k}^{\ell} = \int d\nu_{in\to k}^{\ell}(x_{in}^{\ell})\, \left(x_{in}^{\ell} - \hat{x}_{in\to k}^{\ell}\right)^2 \tag{29}$$

We now use the central limit theorem to observe that with respect to the incoming messages distributions - assuming independence of these messages - in the large input limit the preactivation is a Gaussian random variable:

$$\sum_{i'\neq i} W_{ki'}^{\ell} x_{i'n}^{\ell} \sim \mathcal{N}(\omega_{kn\to i}^{\ell}, V_{kn\to i}^{\ell}) \tag{30}$$

where:

$$\omega_{kn\to i}^{\ell} = \mathbb{E}_{\nu}\left[\sum_{i'\neq i} W_{ki'}^{\ell} x_{i'n}^{\ell}\right] = \sum_{i'\neq i} m_{ki'\to n}^{\ell}\, \hat{x}_{i'n\to k}^{\ell} \tag{31}$$

$$\begin{aligned} V_{kn\to i}^{\ell} &= Var_{\nu}\left[\sum_{i'\neq i} W_{ki'}^{\ell} x_{i'n}^{\ell}\right] \\ &= \sum_{i'\neq i}\left(\sigma_{ki'\to n}^{\ell}\, \Delta_{i'n\to k}^{\ell} + \left(m_{ki'\to n}^{\ell}\right)^2 \Delta_{i'n\to k}^{\ell} + \sigma_{ki'\to n}^{\ell}\, \left(\hat{x}_{i'n\to k}^{\ell}\right)^2\right) \end{aligned} \tag{32}$$

Therefore we can rewrite the outgoing messages as:

$$\hat{\nu}_{kn\to i}^{\ell+1}(W_{ki}^{\ell}) \propto \int dz\, d\nu_{in\to k}^{\ell}(x_{in}^{\ell})\, d\nu_{\downarrow}(x_{kn}^{\ell+1})\, e^{-\frac{\left(z-\omega_{kn\to i}-W_{ki}^{\ell}x_{in}^{\ell}\right)^2}{2V_{kn\to i}}}\, P^{\ell+1}\left(x_{kn}^{\ell+1} \,\middle|\, z\right) \tag{33}$$

We now assume $W_{ki}^{\ell} x_{in}^{\ell}$ to be small compared to the other terms. With a second order Taylor expansion we obtain:

$$\hat{\nu}^{\ell}_{kn\to i}(W^{\ell}_{ki}) \propto \int dz\ d\nu_{\downarrow}(x^{\ell+1}_{kn})\, e^{-\frac{(z-\omega_{kn\to i})^2}{2V_{kn\to i}}}\, P^{\ell+1}\left(x^{\ell+1}_{kn}\,\middle|\, z\right) \times \left(1+\frac{z-\omega_{kn\to i}}{V_{kn\to i}}\hat{x}^{\ell}_{in\to k}W^{\ell}_{ki}+\frac{(z-\omega_{kn\to i})^2-V_{kn\to i}}{2V_{kn\to i}}\left(\Delta+\left(\hat{x}^{\ell}_{in\to k}\right)^2\right)\left(W^{\ell}_{ki}\right)^2\right) \tag{34}$$

Introducing now the function:

$$\varphi^{\ell}(B,A,\omega,V) = \log\int \mathrm{d}x\,\mathrm{d}z\; e^{-\frac{1}{2}Ax^2+Bx}\, P^{\ell}\left(x|z\right) e^{-\frac{(\omega-z)^2}{2V}} \tag{35}$$

and defining:

$$g^{\ell}_{kn\to i} = \partial_{\omega}\varphi^{\ell+1}(B^{\ell+1},A^{\ell+1},\omega^{\ell}_{kn\to i},V^{\ell}_{kn\to i}) \tag{36}$$

$$\Gamma^{\ell}_{kn\to i} = -\partial^2_{\omega}\varphi^{\ell+1}(B^{\ell+1},A^{\ell+1},\omega^{\ell}_{kn\to i},V^{\ell}_{kn\to i}) \tag{37}$$

the expansion for the log-message reads:

$$\log\hat{\nu}^{\ell}_{kn\to i}(W^{\ell}_{ki}) \approx const + \hat{x}^{\ell}_{in\to k}\, g^{\ell}_{kn\to i}W^{\ell}_{ki} - \frac{1}{2}\left(\left(\Delta^{\ell}_{in\to k}+\left(\hat{x}^{\ell}_{in\to k}\right)^2\right)\Gamma^{\ell}_{kn\to i}-\Delta^{\ell}_{in\to k}\left(g^{\ell}_{kn\to i}\right)^2\right)\left(W^{\ell}_{ki}\right)^2 \tag{38}$$

Factor-to-variable-x messages The derivation of these messages is analogous to the factor-to-variable-W ones in Eq. 25 just reported. The final result for the log-message is:

$$\log\hat{\nu}^{\ell}_{kn\to i}(x^{\ell}_{in}) \approx const + m^{\ell}_{ki\to n}\, g^{\ell}_{kn\to i}x^{\ell}_{in} - \frac{1}{2}\left(\left(\sigma^{\ell}_{ki\to n}+\left(m^{\ell}_{ki\to n}\right)^2\right)\Gamma^{\ell}_{kn\to i}-\sigma^{\ell}_{ki\to n}\left(g^{\ell}_{kn\to i}\right)^2\right)\left(x^{\ell}_{in}\right)^2 \tag{39}$$

Variable-W-to-output-factor messages The message from variable W^{ℓ}_{ki} to the output factor kn reads:

$$\nu^{\ell}_{ki\to n}(W^{\ell}_{ki}) \propto P^{\ell}_{\theta_{ki}}(W^{\ell}_{ki})e^{\sum_{n'\neq n}\log\hat{\nu}^{\ell}_{kn'\to i}(W^{\ell}_{ki})} \approx P^{\ell}_{\theta_{ki}}(W^{\ell}_{ki})e^{H^{\ell}_{ki\to n}W^{\ell}_{ki}-\frac{1}{2}G^{\ell}_{ki\to n}\left(W^{\ell}_{ki}\right)^2} \tag{40}$$

where we have defined:

$$H^{\ell}_{ki\to n} = \sum_{n'\neq n}\hat{x}^{\ell}_{in'\to k}\, g^{\ell}_{kn'\to i} \tag{41}$$

$$G^{\ell}_{ki\to n} = \sum_{n'\neq n}\left(\left(\Delta^{\ell}_{in'\to k}+\left(\hat{x}^{\ell}_{in'\to k}\right)^2\right)\Gamma^{\ell}_{kn'\to i}-\Delta^{\ell}_{in'\to k}\left(g^{\ell}_{kn'\to i}\right)^2\right) \tag{42}$$

Introducing now the effective free energy:

$$\psi(H,G,\theta) = \log\int \mathrm{d}W\;\; P^{\ell}_{\theta}\left(W\right) e^{HW-\frac{1}{2}GW^2} \tag{43}$$

we can express the first two cumulants of the message $\nu^{\ell}_{ki\to n}(W^{\ell}_{ki})$ as:

$$m^{\ell}_{ki\to n} = \partial_H\psi(H^{\ell}_{ki\to n},G^{\ell}_{ki\to n},\theta_{ki}) \tag{44}$$

$$\sigma^{\ell}_{ki\to n} = \partial^2_H\psi(H^{\ell}_{ki\to n},G^{\ell}_{ki\to n},\theta_{ki}) \tag{45}$$

Variable-x-to-input-factor messages We can write the downgoing message as:

$$\begin{aligned}\nu_{\downarrow}(x_{in}^{\ell}) &\propto e^{\sum_k \log \hat{\nu}_{kn\to i}^{\ell}(x_{in}^{\ell})} \\ &\approx e^{B_{in}^{\ell} x - \frac{1}{2} A_{in}^{\ell} x^2}\end{aligned} \tag{46}$$

where:

$$B_{in}^{\ell} = \sum_n m_{ki\to n}^{\ell}\, g_{kn\to i}^{\ell} \tag{47}$$

$$A_{in}^{\ell} = \sum_n \left(\left(\sigma_{ki\to n}^{\ell} + \left(m_{ki\to n}^{\ell}\right)^2 \right) \Gamma_{kn\to i}^{\ell} - \sigma_{ki\to n}^{\ell} \left(g_{kn\to i}^{\ell+1}\right)^2 \right) \tag{48}$$

Variable-x-to-output-factor messages By defining the following cavity quantities:

$$B_{in\to k}^{\ell} = B_{in\to k}^{\ell} - m_{ki\to n}^{\ell}\, g_{kn\to i}^{\ell} \tag{49}$$

$$A_{in\to k}^{\ell} = A_{in\to k}^{\ell} - \left(\left(\sigma_{ki\to n}^{\ell} + \left(m_{ki\to n}^{\ell}\right)^2 \right) \Gamma_{kn\to i}^{\ell} - \sigma_{ki\to n}^{\ell} \left(g_{kn\to i}^{\ell}\right)^2 \right) \tag{50}$$

and the following non-cavity ones:

$$\omega_{kn}^{\ell} = \sum_i m_{ki\to n}^{\ell}\, \hat{x}_{in\to k}^{\ell} \tag{51}$$

$$V_{kn}^{\ell} = \sum_i \left(\sigma_{ki\to n}^{\ell}\, \Delta_{in\to k}^{\ell} + \left(m_{ki\to n}^{\ell}\right)^2 \Delta_{in\to k}^{\ell} + \sigma_{ki\to n}^{\ell} \left(\hat{x}_{i'n\to k}^{\ell}\right)^2 \right) \tag{52}$$

we can express the first 2 cumulants of the upgoing messages as:

$$\hat{x}_{in\to k}^{\ell} = \partial_B \varphi^{\ell}(B_{in\to k}^{\ell}, A_{in\to k}^{\ell}, \omega_{in}^{\ell-1}, V_{in}^{\ell-1}) \tag{53}$$

$$\Delta_{in\to k}^{\ell} = \partial_B^2 \varphi^{\ell}(B_{in\to k}^{\ell}, A_{in\to k}^{\ell}, \omega_{in}^{\ell-1}, V_{in}^{\ell-1}) \tag{54}$$

Wrapping it up Additional but straightforward considerations are required for the final input and output layers ($\ell = 0$ and $\ell = L + 1$ respectively), since they do not receive messages from below and above respectively. In the end, thanks to independence assumptions and the central limit theorem that we used throughout the derivations, we arrive to a closed set of equations involving the means and the variances (or otherwise the corresponding natural parameters) of the messages. Within the same approximation assumption, we also replace the cavity quantities corresponding to variances with the non-cavity counterparts. Dividing the update equations in a *forward* and *backward* pass, and ordering them using time indexes in such a way that we have an efficient flow of information, we obtain the set of BP equations presented in the main text Eqs. (6-17) and in the Appendix Eqs. (59-70).

A.3 BP EQUATIONS

We report here the end result of the derivation in last section, the complete set of BP equations also presented in the main text Eqs. (6-17).

Initialization At $\tau = 0$:

$$B_{in\to k}^{\ell,0} = 0 \tag{55}$$

$$A_{in}^{\ell,0} = 0 \tag{56}$$

$$H_{ki\to n}^{\ell,0} = 0 \tag{57}$$

$$G_{ki}^{\ell,0} = 0 \tag{58}$$

Forward Pass At each $\tau = 1, \dots, \tau_{max}$, for $\ell = 0, \dots, L$:

$$\hat{x}^{\ell,\tau}_{in\to k} = \partial_B \varphi^\ell(B^{\ell,\tau-1}_{in\to k}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in}) \tag{59}$$

$$\Delta^{\ell,\tau}_{in} = \partial^2_B \varphi^\ell(B^{\ell,\tau-1}_{in}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in}) \tag{60}$$

$$m^{\ell,\tau}_{ki\to n} = \partial_H \psi(H^{\ell,\tau-1}_{ki\to n}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) \tag{61}$$

$$\sigma^{\ell,\tau}_{ki} = \partial^2_H \psi(H^{\ell,\tau-1}_{ki}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) \tag{62}$$

$$V^{\ell,\tau}_{kn} = \sum_i \left(\left(m^{\ell,\tau}_{ki}\right)^2 \Delta^{\ell,\tau}_{in} + \sigma^{\ell,\tau-1}_{ki} \left(\hat{x}^{\ell,\tau}_{i'n}\right)^2 + \sigma^{\ell,\tau-1}_{ki} \Delta^{\ell,\tau}_{in} \right) \tag{63}$$

$$\omega^{\ell,\tau}_{kn\to i} = \sum_{i'\neq i} m^{\ell,\tau}_{ki'\to n}\, \hat{x}^{\ell,\tau}_{i'n\to k} \tag{64}$$

Backward Pass For $\tau = 1, \dots, \tau_{max}$, for $\ell = L, \dots, 0$:

$$g^{\ell,\tau}_{kn\to i} = \partial_\omega \varphi^{\ell+1}(B^{\ell+1,\tau}_{kn}, A^{\ell+1,\tau}_{kn}, \omega^{\ell,\tau}_{kn\to i}, V^{\ell,\tau}_{kn}) \tag{65}$$

$$\Gamma^{\ell,\tau}_{kn} = -\partial^2_\omega \varphi^{\ell+1}(B^{\ell+1,\tau}_{kn}, A^{\ell+1,\tau}_{kn}, \omega^{\ell,\tau}_{kn}, V^{\ell,\tau}_{kn}) \tag{66}$$

$$A^{\ell,\tau}_{in} = \sum_k \left(\left(\left(m^{\ell,\tau}_{ki}\right)^2 + \sigma^{\ell,\tau}_{ki} \right) \Gamma^{\ell,\tau}_{kn} - \sigma^{\ell,\tau}_{ki} \left(g^{\ell,\tau}_{kn}\right)^2 \right) \tag{67}$$

$$B^{\ell,\tau}_{in\to k} = \sum_{k'\neq k} m^{\ell,\tau}_{k'i\to n}\, g^{\ell,\tau}_{k'n\to i} \tag{68}$$

$$G^{\ell,\tau}_{ki} = \sum_n \left(\left(\left(\hat{x}^{\ell,\tau}_{in}\right)^2 + \Delta^{\ell,\tau}_{in} \right) \Gamma^{\ell,\tau}_{kn} - \Delta^{\ell,\tau}_{in} \left(g^{\ell,\tau}_{kn}\right)^2 \right) \tag{69}$$

$$H^{\ell,\tau}_{ki\to n} = \sum_{n'\neq n} \hat{x}^{\ell,\tau}_{in'\to k}\, g^{\ell,\tau}_{kn'\to i} \tag{70}$$

A.4 BPI EQUATIONS

The BP-Inspired algorithm (BPI) is obtained as an approximation of BP replacing some cavity quantities with their non-cavity counterparts. What we obtain is a generalization of the single layer algorithm of Baldassi et al. (2007).

Forward pass.

$$\hat{x}^{\ell,\tau}_{in} = \partial_B \varphi^\ell \left(B^{\ell,\tau-1}_{in}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in}\right) \tag{71}$$

$$\Delta^{\ell,\tau}_{in} = \partial^2_B \varphi^\ell \left(B^{\ell,\tau-1}_{in}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in}\right) \tag{72}$$

$$m^{\ell,\tau}_{ki} = \partial_H \psi(H^{\ell,\tau-1}_{ki}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) \tag{73}$$

$$\sigma^{\ell,\tau}_{ki} = \partial^2_H \psi(H^{\ell,\tau-1}_{ki}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) \tag{74}$$

$$V^{\ell,\tau}_{kn} = \sum_i \left(\left(m^{\ell,\tau}_{ki}\right)^2 \Delta^{\ell,\tau}_{in} + \sigma^{\ell,\tau}_{ki} (\hat{x}^{\ell,\tau}_{in})^2 + \sigma^{\ell,\tau}_{ki} \Delta^{\ell,\tau}_{in} \right) \tag{75}$$

$$\omega^{\ell,\tau}_{kn} = \sum_i m^{\ell,\tau}_{ki}\, \hat{x}^{\ell,\tau}_{in} \tag{76}$$

Backward pass.

$$g_{kn\to i}^{\ell,\tau} = \partial_\omega \varphi^{\ell+1}\left(B_{kn}^{\ell+1,\tau}, A_{kn}^{\ell+1,\tau}, \omega_{kn}^{\ell,\tau} - m_{ki}^{\ell,\tau}\hat{x}_{ai}^{\ell,\tau}, V_{kn}^{\ell,\tau}\right) \tag{77}$$

$$\Gamma_{kn}^{\ell,\tau} = -\partial_\omega^2 \varphi^{\ell+1}\left(B_{kn}^{\ell+1,\tau}, A_{kn}^{\ell+1,\tau}, \omega_{kn}^{\ell,\tau}, V_{kn}^{\ell,\tau}\right) \tag{78}$$

$$A_{in}^{\ell,\tau} = \sum_k \left((m_{ki}^{\ell,\tau})^2 + \sigma_{ki}^{\ell,\tau}\right)\Gamma_{kn}^{\ell,\tau} - \sigma_{ki}^{\ell,\tau}\left(g_{kn}^{\ell,\tau}\right)^2 \tag{79}$$

$$B_{in}^{\ell,\tau} = \sum_k m_{ki}^{\ell,\tau} g_{kn\to i}^{\ell,\tau} \tag{80}$$

$$G_{ki}^{\ell,\tau} = \sum_n \left((\hat{x}_{in}^{\ell,\tau})^2 + \Delta_{in}^{\ell,\tau}\right)\Gamma_{kn}^{\ell,\tau} - \Delta_{in}^{\ell,\tau}\left(g_{kn}^{\ell,\tau}\right)^2 \tag{81}$$

$$H_{ki}^{\ell,\tau} = \sum_n \hat{x}_{in}^{\ell,\tau} g_{kn\to i}^{\ell,\tau} \tag{82}$$

A.5 MF EQUATIONS

The mean-field (MF) equations are obtained as a further simplification of BPI, using only non-cavity quantities. Although the simplification appears minimal at this point, we empirically observe a non-negligible discrepancy between the two algorithms in terms of generalization performance and computational time.

Forward pass.

$$\hat{x}_{in}^{\ell,\tau} = \partial_B \varphi^{\ell}\left(B_{in}^{\ell,\tau-1}, A_{in}^{\ell,\tau-1}, \omega_{in}^{\ell-1,\tau}, V_{in}^{\ell-1,\tau}\right) \tag{83}$$

$$\Delta_{in}^{\ell,\tau} = \partial_B^2 \varphi^{\ell}\left(B_{in}^{\ell,\tau-1}, A_{in}^{\ell,\tau-1}, \omega_{in}^{\ell-1,\tau}, V_{in}^{\ell-1,\tau}\right) \tag{84}$$

$$m_{ki}^{\ell,\tau} = \partial_H \psi(H_{ki}^{\ell,\tau-1}, G_{ki}^{\ell,\tau-1}, \theta_{ki}^{\ell}) \tag{85}$$

$$\sigma_{ki}^{\ell,\tau} = \partial_H^2 \psi(H_{ki}^{\ell,\tau-1}, G_{ki}^{\ell,\tau-1}, \theta_{ki}^{\ell}) \tag{86}$$

$$V_{kn}^{\ell,\tau} = \sum_i \left(\left(m_{ki}^{\ell,\tau}\right)^2 \Delta_{in}^{\ell,\tau} + \sigma_{ki}^{\ell,\tau}(\hat{x}_{in}^{\ell,\tau})^2 + \sigma_{ki}^{\ell,\tau}\Delta_{in}^{\ell,\tau}\right) \tag{87}$$

$$\omega_{kn}^{\ell,\tau} = \sum_i m_{ki}^{\ell,\tau}\hat{x}_{in}^{\ell,\tau} \tag{88}$$

Backward pass.

$$g_{kn}^{\ell,\tau} = \partial_\omega \varphi^{\ell+1}\left(B_{kn}^{\ell+1,\tau}, A_{kn}^{\ell+1,\tau}, \omega_{kn}^{\ell,\tau}, V_{kn}^{\ell,\tau}\right) \tag{89}$$

$$\Gamma_{kn}^{\ell,\tau} = -\partial_\omega^2 \varphi^{\ell+1}\left(B_{kn}^{\ell+1,\tau}, A_{kn}^{\ell+1,\tau}, \omega_{kn}^{\ell,\tau}, V_{kn}^{\ell,\tau}\right) \tag{90}$$

$$A_{in}^{\ell,\tau} = \sum_k \left((m_{ki}^{\ell,\tau})^2 + \sigma_{ki}^{\ell,\tau}\right)\Gamma_{kn}^{\ell,\tau} - \sigma_{ki}^{\ell,\tau}\left(g_{kn}^{\ell,\tau}\right)^2 \tag{91}$$

$$B_{in}^{\ell,\tau} = \sum_k m_{ki}^{\ell,\tau} g_{kn}^{\ell,\tau} \tag{92}$$

$$G_{ki}^{\ell,\tau} = \sum_n \left((\hat{x}_{in}^{\ell,\tau})^2 + \Delta_{in}^{\ell,\tau}\right)\Gamma_{kn}^{\ell,\tau} - \Delta_{in}^{\ell,\tau}\left(g_{kn}^{\ell,\tau}\right)^2 \tag{93}$$

$$H_{ki}^{\ell,\tau} = \sum_n \hat{x}_{in}^{\ell,\tau} g_{kn}^{\ell,\tau} \tag{94}$$

A.6 DERIVATION OF THE AMP EQUATIONS

In order to obtain the AMP equations, we approximate cavity quantities with non-cavity ones in the BP equations Eqs. (59-70) using a first order expansion. We start with the mean activation:

$$
\begin{aligned}
\hat{x}^{\ell,\tau}_{in\to k} =& \partial_B \varphi^\ell(B^{\ell,\tau-1}_{in} - m^{\ell,\tau-1}_{ki\to n}\, g^{\ell,\tau-1}_{kn\to i}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in}) \\
\approx& \partial_B \varphi^\ell(B^{\ell,\tau-1}_{in}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in}) \\
&- m^{\ell,\tau-1}_{ki\to n}\, g^{\ell,\tau-1}_{kn\to i} \partial_B^2 \varphi^\ell(B^{\ell,\tau-1}_{in}, A^{\ell,\tau-1}_{in}, \omega^{\ell-1,\tau}_{in}, V^{\ell-1,\tau}_{in}) \\
\approx& \hat{x}^{\ell,\tau}_{in} - m^{\ell,\tau-1}_{ki} g^{\ell,\tau-1}_{kn} \Delta^{\ell,\tau}_{in}
\end{aligned} \tag{95}
$$

Analogously, for the weight's mean we have:

$$
\begin{aligned}
m^{\ell,\tau}_{ki\to n} &= \partial_H \psi(H^{\ell,\tau-1}_{ki} - \hat{x}^{\ell,\tau-1}_{in\to k}\, g^{\ell,\tau-1}_{kn\to i}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) \\
&\approx \partial_H \psi(H^{\ell,\tau-1}_{ki}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) - \hat{x}^{\ell,\tau-1}_{in\to k}\, g^{\ell,\tau-1}_{kn\to i} \partial_H^2 \psi(H^{\ell,\tau-1}_{ki}, G^{\ell,\tau-1}_{ki}, \theta^\ell_{ki}) \\
&\approx m^{\ell,\tau}_{ki} - \hat{x}^{\ell,\tau-1}_{in}\, g^{\ell,\tau-1}_{kn}\, \sigma^{\ell,\tau}_{ki}\,.
\end{aligned} \tag{96}
$$

This brings us to:

$$
\begin{aligned}
\omega^{\ell,\tau}_{kn} =& \sum_i m^{\ell,\tau}_{ki\to n}\, \hat{x}^{\ell,\tau}_{in\to k} \\
\approx& \sum_i m^{\ell,\tau}_{ki}\, \hat{x}^{\ell,\tau}_{in} - g^{\ell,\tau-1}_{kn} \sum_i \sigma^{\ell,\tau}_{ki}\, \hat{x}^{\ell,\tau}_{in}\, \hat{x}^{\ell,\tau-1}_{in} - g^{\ell,\tau-1}_{kn} \sum_i m^{\ell,\tau}_{ki} m^{\ell,\tau-1}_{ki} \Delta^{\ell,\tau}_{in} \\
&+ (g^{\ell,\tau-1}_{kn})^2 \sum_i \sigma^{\ell,\tau}_{ki} m^{\ell,\tau-1}_{ki} \hat{x}^{\ell,\tau-1}_{in} \Delta^{\ell,\tau}_{in}
\end{aligned} \tag{97}
$$

Let us now apply the same procedure to the other set of cavity messages:

$$
\begin{aligned}
g^{\ell,\tau}_{kn\to i} =& \partial_\omega \varphi^{\ell+1}(B^{\ell+1,\tau}_{kn}, A^{\ell+1,\tau}_{kn}, \omega^{\ell,\tau}_{kn} - m^{\ell,\tau}_{ki\to n}\, \hat{x}^{\ell,\tau}_{in\to k}, V^{\ell,\tau}_{kn}) \\
\approx& \partial_\omega \varphi^{\ell+1}(B^{\ell+1,\tau}_{kn}, A^{\ell+1,\tau}_{kn}, \omega^{\ell,\tau}_{kn}, V^{\ell,\tau}_{kn}) \\
&- m^{\ell,\tau}_{ki\to n}\, \hat{x}^{\ell,\tau}_{in\to k} \partial_\omega^2 \varphi^{\ell+1}(B^{\ell+1,\tau}_{kn}, A^{\ell+1,\tau}_{kn}, \omega^{\ell,\tau}_{kn}, V^{\ell,\tau}_{kn}) \\
\approx& g^{\ell,\tau}_{kn} + m^{\ell,\tau}_{ki}\, \hat{x}^{\ell,\tau}_{in} \Gamma^{\ell,\tau}_{kn}
\end{aligned} \tag{98}
$$

$$
\begin{aligned}
B^{\ell,\tau}_{in} =& \sum_k m^{\ell,\tau}_{ki\to n}\, g^{\ell,\tau}_{kn\to i} \\
\approx& \sum_k m^{\ell,\tau}_{ki}\, g^{\ell,\tau}_{kn} - \hat{x}_{in} \sum_k \left(g^{\ell,\tau}_{kn}\right)^2 \sigma^{\ell,\tau}_{ki} + \hat{x}^{\ell,\tau}_{in} \sum_k (m^{\ell,\tau}_{ki})^2 \Gamma^{\ell,\tau}_{kn} \\
&- (\hat{x}^{\ell,\tau}_{in})^2 \sum_k \sigma^{\ell,\tau}_{ki} m^{\ell,\tau}_{ki} g^{\ell,\tau}_{kn} \Gamma^{\ell,\tau}_{kn}
\end{aligned} \tag{99}
$$

$$
\begin{aligned}
H^{\ell,\tau}_{ki} =& \sum_n \hat{x}^{\ell,\tau}_{in\to k}\, g^{\ell,\tau}_{kn\to i} \\
\approx& \sum_n \hat{x}^{\ell,\tau}_{in}\, g^{\ell,\tau}_{kn} + m^{\ell,\tau}_{ki} \sum_n \left(\hat{x}^{\ell,\tau}_{in}\right)^2 \Gamma^{\ell,\tau}_{kn} - m^{\ell,\tau}_{ki} \sum_n (g^{\ell,\tau}_{kn})^2 \Delta^{\ell,\tau}_{in} \\
&- (m^{\ell,\tau}_{ki})^2 \sum_n g^{\ell,\tau}_{kn} \Gamma^{\ell,\tau}_{kn} \Delta^{\ell,\tau}_{in} \hat{x}^{\ell,\tau}_{in}
\end{aligned} \tag{100}
$$

We are now able to write down the full AMP equations, that we present in the next section.

A.7 AMP EQUATIONS

In summary, in the last section we derived the AMP algorithm as a closure of the BP messages passing over non-cavity quantities, relying on some statistical assumptions on messages and interactions. With respect to the MF message passing, we find some additional terms that go under the name of Onsager corrections. In-depth overviews of the AMP (also known as Thouless-Anderson-Palmer (TAP)) approach can be found in Refs. (Zdeborová & Krzakala, 2016; Mézard, 2017; Gabrié, 2020). The final form of the AMP equations for the multi-layer perceptron is given below.

Initialization At $\tau = 0$:

$$B_{in}^{\ell,0} = 0 \tag{101}$$
$$A_{in}^{\ell,0} = 0 \tag{102}$$
$$H_{ki}^{\ell,0} = 0 \text{ or some values} \tag{103}$$
$$G_{ki}^{\ell,0} = 0 \text{ or some values} \tag{104}$$
$$g_{kn}^{\ell,0} = 0 \tag{105}$$

Forward Pass At each $\tau = 1, \dots, \tau_{max}$, for $\ell = 0, \dots, L$:

$$\hat{x}_{in}^{\ell,\tau} = \partial_B \varphi^{\ell}(B_{in}^{\ell,\tau-1}, A_{in}^{\ell,\tau-1}, \omega_{in}^{\ell-1,\tau}, V_{in}^{\ell-1,\tau}) \tag{106}$$
$$\Delta_{in}^{\ell,\tau} = \partial_B^2 \varphi^{\ell}(B_{in}^{\ell,\tau-1}, A_{in}^{\ell,\tau-1}, \omega_{in}^{\ell-1,\tau}, V_{in}^{\ell-1,\tau}) \tag{107}$$
$$m_{ki}^{\ell,\tau} = \partial_H \psi(H_{ki}^{\ell,\tau-1}, G_{ki}^{\ell,\tau-1}, \theta_{ki}^{\ell}) \tag{108}$$
$$\sigma_{ki}^{\ell,\tau} = \partial_H^2 \psi(H_{ki}^{\ell,\tau-1}, G_{ki}^{\ell,\tau-1}, \theta_{ki}^{\ell}) \tag{109}$$
$$V_{kn}^{\ell,\tau} = \sum_i \left(\left(m_{ki}^{\ell,\tau} \right)^2 \Delta_{in}^{\ell,\tau} + \sigma_{ki}^{\ell,\tau} \left(\hat{x}_{i'n}^{\ell,\tau} \right)^2 + \sigma_{ki}^{\ell,\tau} \Delta_{in}^{\ell,\tau} \right) \tag{110}$$
$$\begin{aligned} \omega_{kn}^{\ell,\tau} = & \sum_i m_{ki}^{\ell,\tau} \hat{x}_{in}^{\ell,\tau} - g_{kn}^{\ell,\tau-1} \sum_i \sigma_{ki}^{\ell,\tau} \hat{x}_{in}^{\ell,\tau} \hat{x}_{in}^{\ell,\tau-1} - g_{kn}^{\ell,\tau-1} \sum_i m_{ki}^{\ell,\tau} m_{ki}^{\ell,\tau-1} \Delta_{in}^{\ell,\tau} \\ & + (g_{kn}^{\ell,\tau-1})^2 \sum_i \sigma_{ki}^{\ell,\tau} m_{ki}^{\ell,\tau-1} \hat{x}_{in}^{\ell,\tau-1} \Delta_{in}^{\ell,\tau} \end{aligned} \tag{111}$$

Backward Pass

$$g_{kn}^{\ell,\tau} = \partial_\omega \varphi^{\ell+1}(B_{kn}^{\ell+1,\tau}, A_{kn}^{\ell+1,\tau}, \omega_{kn\to i}^{\ell,\tau}, V_{kn}^{\ell,\tau}) \tag{112}$$
$$\Gamma_{kn}^{\ell,\tau} = -\partial_\omega^2 \varphi^{\ell+1}(B_{kn}^{\ell+1,\tau}, A_{kn}^{\ell+1,\tau}, \omega_{kn}^{\ell,\tau}, V_{kn}^{\ell,\tau}) \tag{113}$$
$$A_{in}^{\ell,\tau} = \sum_k \left(\left(\left(m_{ki}^{\ell,\tau} \right)^2 + \sigma_{ki}^{\ell,\tau} \right) \Gamma_{kn}^{\ell,\tau} - \sigma_{ki}^{\ell,\tau} \left(g_{kn}^{\ell,\tau} \right)^2 \right) \tag{114}$$
$$\begin{aligned} B_{in}^{\ell,\tau} = & \sum_k m_{ki}^{\ell,\tau} g_{kn}^{\ell,\tau} - \hat{x}_{in} \sum_k \left(g_{kn}^{\ell,\tau} \right)^2 \sigma_{ki}^{\ell,\tau} + \hat{x}_{in}^{\ell,\tau} \sum_k (m_{ki}^{\ell,\tau})^2 \Gamma_{kn}^{\ell,\tau} \\ & - (\hat{x}_{in}^{\ell,\tau})^2 \sum_k \sigma_{ki}^{\ell,\tau} m_{ki}^{\ell,\tau} g_{kn}^{\ell,\tau} \Gamma_{kn}^{\ell,\tau} \end{aligned} \tag{115}$$
$$G_{ki}^{\ell,\tau} = \sum_n \left(\left(\left(\hat{x}_{in}^{\ell,\tau} \right)^2 + \Delta_{in}^{\ell,\tau} \right) \Gamma_{kn}^{\ell,\tau} - \Delta_{in}^{\ell,\tau} \left(g_{kn}^{\ell,\tau} \right)^2 \right) \tag{116}$$
$$\begin{aligned} H_{ki}^{\ell,\tau} = & \sum_n \hat{x}_{in}^{\ell,\tau} g_{kn}^{\ell,\tau} + m_{ki}^{\ell,\tau} \sum_n \left(\hat{x}_{in}^{\ell,\tau} \right)^2 \Gamma_{kn}^{\ell,\tau} - m_{ki}^{\ell,\tau} \sum_n (g_{kn}^{\ell,\tau})^2 \Delta_{in}^{\ell,\tau} \\ & - (m_{ki}^{\ell,\tau})^2 \sum_n g_{kn}^{\ell,\tau} \Gamma_{kn}^{\ell,\tau} \Delta_{in}^{\ell,\tau} \hat{x}_{in}^{\ell,\tau} \end{aligned} \tag{117}$$

A.8 ACTIVATION FUNCTIONS

A.8.1 SIGN

In most of our experiments we use sign activations in each layer. With this choice, the neuron's free energy 19 takes the form

$$\varphi(B, A, \omega, V) = \log\left(\frac{1}{2}\sum_{x\in\{-1,+1\}} e^{Bx}\,\mathcal{H}\left(-\frac{x\omega}{\sqrt{V}}\right)\right) + \frac{1}{2}\log(2\pi V), \tag{118}$$

where

$$\mathcal{H} = \frac{1}{2}\,\mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right). \tag{119}$$

Notice that for sign activations the messages A can be dropped.

A.8.2 RELU

For $ReLU(x) = \max(0, x)$ activations the free energy 19 becomes

$$\varphi(B, A, \omega, V) = \int \mathrm{d}x\mathrm{d}z\; e^{-\frac{1}{2}Ax^2+Bx}\,\delta(x - \max(0, z))\; e^{-\frac{(\omega-z)^2}{2V}} \tag{120}$$

$$= \log\left(H\left(\frac{\omega}{\sqrt{V}}\right) + \frac{\mathcal{N}(\omega; B/A, V + \frac{1}{A})}{A\mathcal{N}(B; 0, A)}\,H\left(-\frac{BV+\omega}{\sqrt{V+AV^2}}\right)\right) + \frac{1}{2}\log(2\pi V), \tag{121}$$

where

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{2\pi\Sigma}}\,e^{-\frac{(x-\mu)^2}{2\Sigma}}. \tag{122}$$

A.9 THE ARGMAX LAYER

In order to perform multi-class classification, we have to perform an argmax operation on the last layer of the neural network. Call z_k, for $k = 1, \ldots, K$, the Gaussian random variables output of the last layer of the network in correspondence of some input $\boldsymbol{x}$. Assuming the correct label is class k^*, the effective partition function Z_{k^*} corresponding to the output constraint reads:

$$Z_{k^*} = \int \prod_k dz_k\,\mathcal{N}(z_k; \omega_k, V_k) \prod_{k\neq k^*} \Theta(z_{k^*} - z_k) \tag{123}$$

$$= \int dz_{k^*}\,\mathcal{N}(z_{k^*}; \omega_{k^*}, V_{k^*}) \prod_{k\neq k^*} \mathcal{H}\left(-\frac{z_{k^*} - \omega_k}{\sqrt{V_k}}\right) \tag{124}$$

Here $\Theta(x)$ is the Heaviside indicator function and we used the definition of $\mathcal{H}$ from Eq. 119. The integral on the last line cannot be expressed analytically, therefore we have to resort to approximations.

A.9.1 APPROACH 1: JENSEN INEQUALITY

Using the Jensen inequality we obtain:

$$\phi_{k^*} = \log Z_{k^*} = \log \mathbb{E}_{z\sim\mathcal{N}(\omega_{k^*}, V_{k^*})} \prod_{k\neq k^*} \mathcal{H}\left(-\frac{z - \omega_k}{\sqrt{V_k}}\right) \tag{125}$$

$$\geq \sum_{k\neq k^*} \mathbb{E}_{z\sim\mathcal{N}(\omega_{k^*}, V_{k^*})} \log \mathcal{H}\left(-\frac{z - \omega_k}{\sqrt{V_k}}\right) \tag{126}$$

Reparameterizing the expectation we have:

$$\tilde{\phi}_{k^*} = \sum_{k\neq k^*} \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)} \log \mathcal{H}\left(-\frac{\omega_{k^*} + \epsilon\sqrt{V_{k^*}} - \omega_k}{\sqrt{V_k}}\right) \tag{127}$$

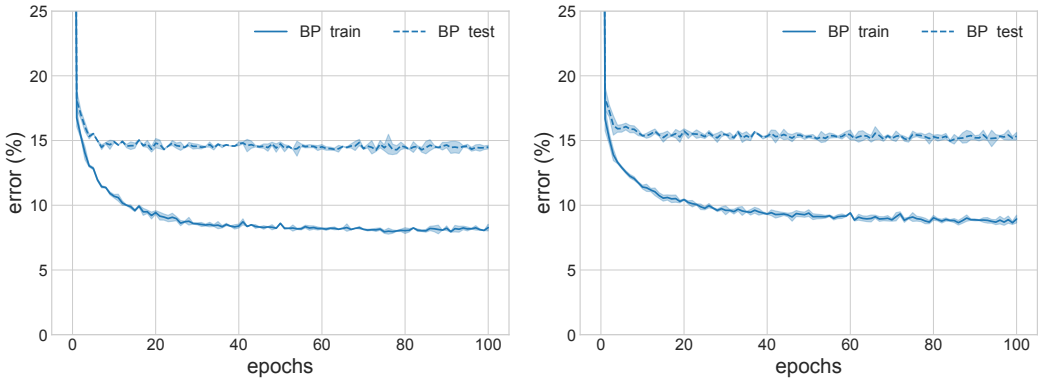

Figure 4: MLP with 2 hidden layers with 101 hidden units each, batch-size 128 on the Fashion-MNIST dataset. In the first two layers we use the BP equations, while in the last layer the ArgMax ones. (Left) ArgMax layer first version; (Right) ArgMax layer second version. Even if it is possible to reach similar accuracies with the two versions, we decide to use the first one as it is simpler to use.

The derivative $\partial_{\omega_k} \tilde{\phi}_{k^*}$ and $\partial^2_{\omega_k} \tilde{\phi}_{k^*}$ that we need can then be estimated by sampling (once) ϵ:

$$\partial_{\omega_k} \tilde{\phi}_{k^*} = \begin{cases} -\frac{1}{\sqrt{V_k}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \mathcal{K}\left(-\frac{\omega_{k^*} + \epsilon\sqrt{V_{k^*}} - \omega_k}{\sqrt{V_k}}\right) & k \neq k^* \\ \sum_{k' \neq k^*} \frac{1}{\sqrt{V_{k'}}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \mathcal{K}\left(-\frac{\omega_{k^*} + \epsilon\sqrt{V_{k^*}} - \omega_{k'}}{\sqrt{V_{k'}}}\right) & k = k^* \end{cases} \tag{128}$$

where we have defined:

$$\mathcal{K}(x) = \frac{\mathcal{N}(x)}{\mathcal{H}(x)} = \frac{\sqrt{2/\pi}}{\text{erfcx}(x/2)} \tag{129}$$

A.9.2 Approach 2: Jensen again

A further simplification is obtained by applying Jensen inequality again to 127 but in the opposite direction, therefore we renounce to having a bound and look only for an approximation. We have the new effective free energy:

$$\tilde{\phi}_{k^*} = \sum_{k \neq k^*} \log \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \mathcal{H}\left(-\frac{\omega_{k^*} + \epsilon\sqrt{V_{k^*}} - \omega_k}{\sqrt{V_k}}\right) \tag{130}$$

$$= \sum_{k \neq k^*} \log \mathcal{H}\left(-\frac{\omega_{k^*} - \omega_k}{\sqrt{V_k + V_{k^*}}}\right) \tag{131}$$

This gives, for $k \neq k^*$:

$$\partial_{\omega_k} \tilde{\phi}_{k^*} = \begin{cases} -\frac{1}{\sqrt{V_k + V_{k^*}}} \mathcal{K}\left(-\frac{\omega_{k^*} - \omega_k}{\sqrt{V_k + V_{k^*}}}\right) & k \neq k^* \\ \sum_{k' \neq k^*} \frac{1}{\sqrt{V_{k'} + V_{k^*}}} \mathcal{K}\left(-\frac{\omega_{k^*} - \omega_{k'}}{\sqrt{V_{k'} + V_{k^*}}}\right) & k = k^* \end{cases} \tag{132}$$

Notice that $\partial_{\omega_{k^*}} \tilde{\phi}_{k^*} = -\sum_{k \neq k^*} \partial_{\omega_k} \tilde{\phi}_{k^*}$. In last formulas we used the definition of $\mathcal{K}$ in Eq. 129.

We show in Fig. 4 the negligible effect of using the two ArgMax versions when using BP on the layers before the last one (on which we apply ArgMax).

B EXPERIMENTAL DETAILS

B.1 HYPER-PARAMETERS OF THE BP-BASED SCHEME

We include here a complete list of the hyper-parameters present in the BP-based algorithms. Notice that, like in the SGD type of algorithms, many of them can be fixed or it is possible to find a prescription for their value that works in most cases. However, we expect future research to find even more effective values of the hyper-parameters, in the same way it has been done for SGD. These hyper-parameters are: the mini-batch size bs; the parameter ρ (that has to be tuned similarly to the learning rate in SGD); the damping parameter α (that performs a running smoothing on the BP fields along the dynamics by adding a fraction of the field at the previous iteration, see Eqs. (133, 134)); the initialization coefficient ϵ that we use to to sample the parameters of our prior distribution $q_\theta(\mathcal{W})$ according to $\theta_{ki}^{\ell,t=0} \sim \epsilon\mathcal{N}(0,1)$. Different choices of ϵ correspond to different initial distribution of the weights' magnetization $m_{ki}^{\ell} = \tanh(\theta_{ki}^{\ell})$, as is shown in Fig. 5); the number of internal steps of reinforcement $\tau_{\max}$ and the associated intensity of the internal reinforcement r. The performances of the BP-based algorithms are robust in a reasonable range of these hyper-parameters. A more principled choice of a good initialization condition could be made by adapting the technique from Stamatescu et al. (2020).

Notice that among these parameters, the BP dynamics at each layer is mostly sensitive to ρ and α, so that in general we consider them layer-dependent. See Sec. B.7 for details on the effect of these parameters on the learning dynamics and on layer polarization (i.e. how the BP dynamics tends to bias the weights towards a single point-wise configuration with high probability). Unless otherwise stated we fix some of the hyper-parameters, in particular: $bs = 128$ (results are consistent with other values of the batch-size, from $bs = 1$ up to $bs = 1024$ in our experiments), $\epsilon = 1.0$, $\tau_{\max} = 1$, $r = 0$.

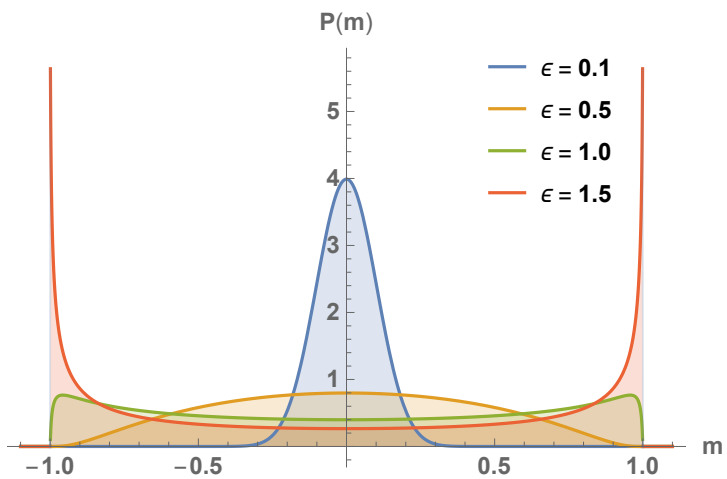

Figure 5: Initial distribution of the magnetizations varying the parameter ϵ. The initial distribution is more concentrated around ± 1 as ϵ increases (i.e. it is more bimodal and the initial configuration is more polarized).

B.2 DAMPING SCHEME FOR THE MESSAGE PASSING

We use a damping parameter $\alpha \in (0,1)$ to stabilize the training, changing the updated rule for the weights' means as follows

$$\tilde{m}_{ki}^{\ell,\tau} = \partial_H \psi(H_{ki}^{\ell,\tau-1}, G_{ki}^{\ell,\tau-1}, \theta_{ki}^{\ell}) \tag{133}$$

$$m_{ki}^{\ell,\tau} = \alpha\, m_{ki}^{\ell,\tau-1} + (1-\alpha)\, \tilde{m}_{ki}^{\ell,\tau} \tag{134}$$

B.3 ARCHITECTURES

In the experiments in which we vary the architecture (see Sec. 4.1), all simulations of the BP-based algorithms use a number of internal reinforcement iterations $\tau_{\max} = 1$. Learning is performed on the totality of the training dataset, the batch-size is $bs = 128$, the initialization coefficient is $\epsilon = 1.0$.

For all architectures and all BP approximations, we use $\alpha = 0.8$ for each layer, apart for the 501-501-501 MLP in which we use $\alpha = (0.1, 0.1, 0.1, 0.9)$. Concerning the parameter ρ, we use $\rho = 0.9$ on the last layer for all architectures and BP approximations. On the other layers we use: for the 101-101 and the 501-501 MLPs, $\rho = 1.0001$ for all BP approximations; for the 101-101-101 MLP, $\rho = 1.0$ for BP and AMP while $\rho = 1.001$ for MF; for the 501-501-501 MLP $\rho = 1.0001$ for all BP approximations. For the BinaryNet simulations, the learning rate is $lr = 10.0$ for all MLP architectures, giving the better performance among the learning rates we have tested, $lr = 100, 10, 1, 0.1, 0.001$.

We notice that while we need some tuning of the hyper-parameters to reach the performances of BinaryNet, it is possible to fix them across datasets and architectures (e.g. $\rho = 1$ and $\alpha = 0.8$ on each layer) without in general losing more than 20% (relative) of the generalization performances, demonstrating that the BP-based algorithms are effective for learning also with minimal hyper-parameter tuning.

The experiments on the Bayesian error are performed on a MLP with 2 hidden layers of 101 units on the MNIST dataset (binary classification). Learning is performed on the totality of the training dataset, the batch-size is $bs = 128$, the initialization coefficient is $\epsilon = 1.0$. In order to find the pointwise configurations we use $\alpha = 0.8$ on each layer and $\rho = (1.0001, 1.0001, 0.9)$, while to find the Bayesian ones we use $\alpha = 0.8$ on each layer and $\rho = (0.9999, 0.9999, 0.9)$ (these value prevent an excessive polarization of the network towards a particular pointwise configurations).

For the continual learning task (see Sec. 4.5) we fixed $\rho = 1$ and $\alpha = 0.8$ on each layer as we empirically observed that polarizing the last layer helps mitigating the forgetting while leaving the single-task performances almost unchanged.

B.4 VARYING THE DATASET

When varying the dataset (see Sec. 4.3), all simulation of the BP-based algorithms use a number of internal reinforcement iterations $\tau_{\max} = 1$. Learning is performed on the totality of the training dataset, the batch-size is $bs = 128$, the initialization coefficient is $\epsilon = 1.0$. For all datasets (MNIST (2 classes), FashionMNIST (2 classes), CIFAR-10 (2 classes), MNIST, FashionMNIST, CIFAR-10) and all algorithms (BP, AMP, MF) we use $\rho = (1.0001, 1.0001, 0.9)$ and $\alpha = 0.8$ for each layer. Using in the first layers values of $\rho = 1 + \epsilon$ with $\epsilon \geq 0$ and sufficiently small typically leads to good results.

For the BinaryNet simulations, the learning rate is $lr = 10.0$ (both for binary classification and multi-class classification), giving the better performance among the learning rates we have tested, $lr = 100, 10, 1, 0.1, 0.001$. In Tab. 2 we report the final train errors obtained on the different datasets.

Dataset	BinaryNet	BP	AMP	MF
MNIST (2 classes)	0.05 ± 0.05	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
FashionMNIST (2 classes)	0.3 ± 0.1	0.06 ± 0.01	0.06 ± 0.01	0.09 ± 0.01
CIFAR10 (2 classes)	1.2 ± 0.5	0.37 ± 0.01	0.4 ± 0.1	0.9 ± 0.2
MNIST	0.09 ± 0.01	0.12 ± 0.01	0.12 ± 0.01	0.03 ± 0.01
FashionMNIST	4.0 ± 0.5	3.4 ± 0.1	3.7 ± 0.1	2.5 ± 0.2
CIFAR10	13.0 ± 0.9	4.7 ± 0.1	4.7 ± 0.2	9.2 ± 0.5

Table 2: Train error (%) on Fashion-MNIST of a multilayer perceptron with 2 hidden layers of 501 units each for BinaryNet (baseline), BP, AMP and MF. All algorithms are trained with batch-size 128 and for 100 epochs. Mean and standard deviations are calculated over 5 random initializations.

B.5 LOCAL ENERGY

We adapt the notion of flatness used in (Jiang et al., 2020; Pittorino et al., 2021), that we call local energy, to configurations with binary weights. Given a weight configuration $\boldsymbol{w} \in \{\pm 1\}^N$, we define the local energy $\delta E_{\text{train}}(\boldsymbol{w}, p)$ as the average difference in training error $E_{\text{train}}(\boldsymbol{w})$ when perturbing $\boldsymbol{w}$ by flipping a random fraction p of its elements:

$$\delta E_{\text{train}}(\boldsymbol{w}, p) = \mathbb{E}_{\boldsymbol{z}}\, E_{\text{train}}(\boldsymbol{w} \odot \boldsymbol{z}) - E_{\text{train}}(\boldsymbol{w}), \tag{135}$$

where $\odot$ denotes the Hadamard (element-wise) product and the expectation is over i.i.d. entries for $\boldsymbol{z}$ equal to -1 with probability p and to $+1$ with probability $1-p$. We report the resulting local energy profiles (in a range $[0, p_{\max}]$) in Fig. 6 right panel for BP and BinaryNet. The relative error grows slowly when perturbing the trained configurations (notice the convexity of the curves). This shows that both BP-based and SGD-based algorithms find configurations that lie in relatively flat minima in the energy landscape. The same qualitative phenomenon holds for different architectures and datasets.

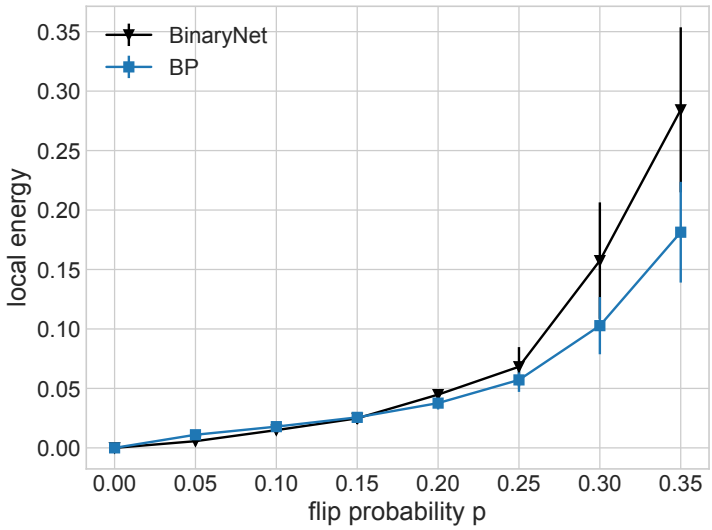

Figure 6: Local energy curve of the point-wise configuration found by the BP algorithm compared with BinaryNet on a MLP with 2 hidden layers of 101 units on the 2-class MNIST dataset.

B.6 SGD IMPLEMENTATION (BINARYNET)

We compare the BP-based algorithms with SGD training for neural networks with binary weights and activations as introduced in BinaryNet (Hubara et al., 2016). This procedure consists in keeping a continuous version of the parameters w which is updated with the SGD rule, with the gradient calculated on the binarized configuration $w_b = \text{sign}(w)$. At inference time the forward pass is calculated with the parameters w_b. The backward pass with binary activations is performed with the so called *straight-through estimator*.

Our implementation presents some differences with respect to the original proposal of the algorithm in (Hubara et al., 2016), in order to keep the comparison as fair as possible with the BP-based algorithms, in particular for what concerns the number of parameters. We do not use biases nor batch normalization layers, therefore in order to keep the pre-activations of each hidden layer normalized we rescale them by $\frac{1}{\sqrt{N}}$ where N is the size of the previous layer (or the input size in the case of the pre-activations afferent to the first hidden layer). The standard SGD update rule is applied (instead of Adam), and we use the binary cross-entropy loss. Clipping of the continuous configuration w in $[-1, 1]$ is applied. We use Xavier initialization (Glorot & Bengio, 2010) for the continuous weights.

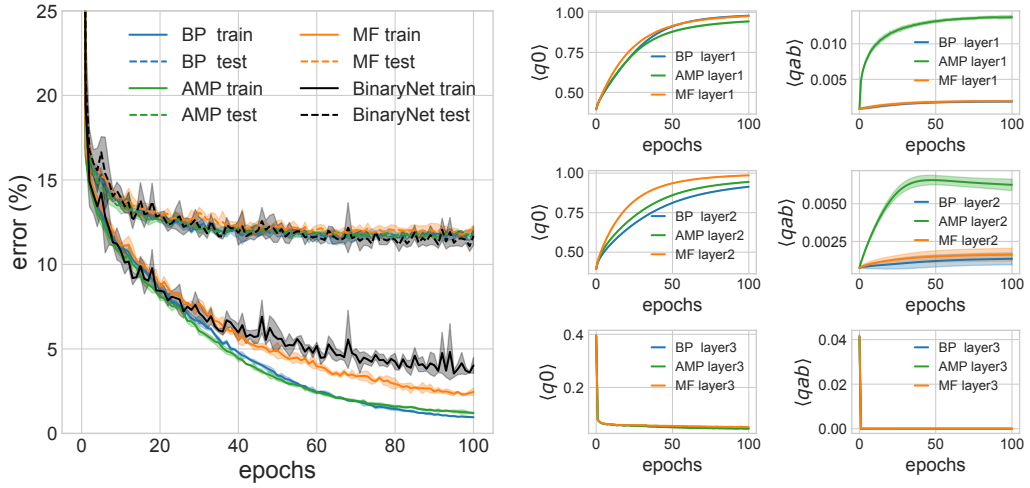

Figure 7: (Right panels) Polarizations $\langle q_0 \rangle$ and overlaps $\langle q_{ab} \rangle$ on each layer of a MLP with 2 hidden layers of 501 units on the Fashion-MNIST dataset (multi-class), the batch-size is $bs = 128$. (Right) Corresponding train and test error curves.

B.7 Unit polarization and overlaps

We define the self-overlap or polarization of a given unit a as $q_0^a = \frac{1}{N}\sum_i (w_i^a)^2$, where N is the number of parameters of the unit and $\{w_i^a\}_{i=1}^N$ its weights. It quantifies how much the unit is polarized towards a unique point-wise binary configuration ($q_0^a = 1$ corresponding to high confidence in a given configurations while $q_0^a = 0$ to low). The overlap between two units a and b (considered in the same layer) is $q_{ab} = \frac{1}{N}\sum w_i^a w_i^b$. The number of parameters N is the same for units belonging to the same fully connected layer. We denote by $\langle q_0 \rangle = \frac{1}{N_{out}}\sum_{a=1}^{N_{out}} q_0^a$ and $\langle q_{ab} \rangle = \frac{1}{N_{out}}\sum_{a<b}^{N_{out}} q_{ab}$ the mean polarization and mean overlap in a given layer (where N_{out} is the number of units in the layer).

The parameters ρ and α govern the dynamical evolution of the polarization of each layer during training. A value $\rho \gtrapprox 1$ has the effect to progressively increase the units polarization during training, while $\rho < 1$ disfavours it. The damping α which takes values in $[0, 1]$ has the effect to slow the dynamics by a smoothing process (the intensity of which depends on the value of α), generically favoring convergence. Given the nature of the updates in Algorithm 1, each layer presents its own dynamics given the values of ρ_ℓ and $\alpha\ell$ at layer ℓ, that in general can differ from each other.

We find that it is is beneficial to control the polarization layer-per-layer, see Fig. 7 for the corresponding typical behavior of the mean polarization and the mean overlaps during training. Empirically, we have found that (as we could expect) when training is successful the layers polarize progressively towards $q_0 = 1$, i.e. towards a precise point-wise solution, while the overlaps between units in each hidden layer are such that $q_{ab} \ll 1$ (indicating low symmetry between intra-layer units, as expected for a non-redundant solution). To this aim, in most cases $\alpha\ell$ can be the same for each layer, while tuning ρ_ℓ for each layer allows to find better generalization performances in some cases (but is not strictly necessary for learning).

In particular, it is possible to use the same value ρ_ℓ for each layer before the last one ($\ell < L$ where L is the number of layers in the network), while we have found that the last layer tends to polarize immediately during the dynamics (probably due to its proximity to the output constraints). Empirically, it is usually beneficial for learning that this layer does not or only slightly polarize, i.e. $\langle q_0 \rangle \ll 1$ (this can be achieved by imposing $\rho_L < 1$). Learning is anyway possible even when the last layer polarizes towards $\langle q_0 \rangle = 1$ along the dynamics, i.e. by choosing ρ_L sufficiently large.

As a simple general prescription in most experiments we can fix $\alpha = 0.8$ and $\rho_L = 0.9$, therefore leaving $\rho_{\ell<L}$ as the only hyper-parameter to be tuned, akin to the learning rate in SGD. Its value has to be very close to 1.0 (a value smaller than 1.0 tends to depolarize the layers, without focusing on a

particular point-wise binary configuration, while a value greater than 1.0 tends to lead to numerical instabilities and parameters' divergence).

B.8 COMPUTATIONAL PERFORMANCE: VARYING BATCH-SIZE

In order to compare the time performances of the BP-based algorithms with our implementation of BinaryNet, we report in Fig. 8 the time in seconds taken by a single epoch of each algorithm in function of the batch-size, on a MLP of 2 layers of 501 units on Fashion-MNIST. We test both algorithms on a NVIDIA GeForce RTX 2080 Ti GPU. Multi-class and binary classification present a very similar time scaling with the batch-size, in both cases comparable with BinaryNet. Let us also notice that BP-based algorithms are able to reach generalization performances comparable to BinaryNet for all the values of the batch-size reported in this section.

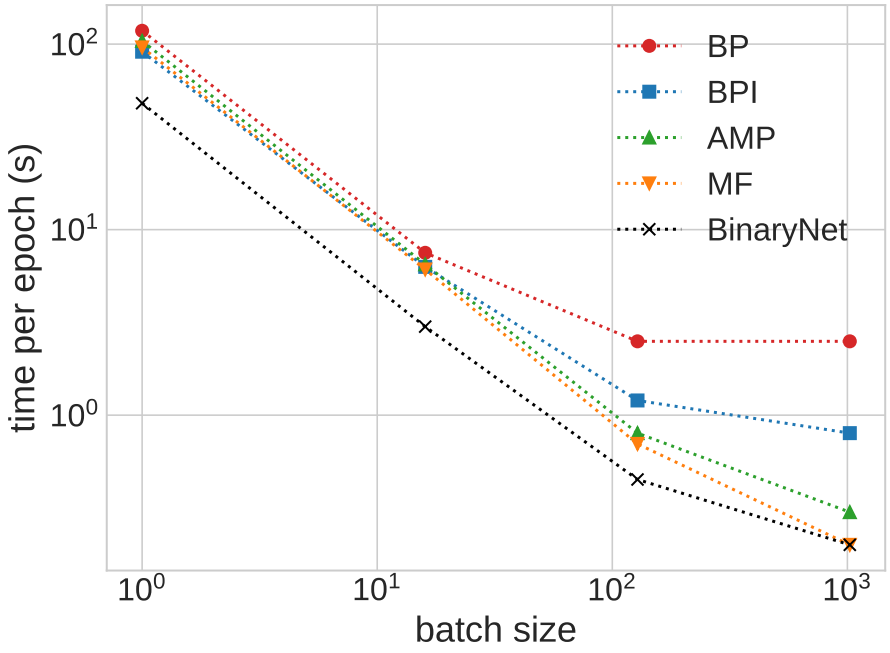

Figure 8: Algorithms time scaling with the batch-size on a MLP with 2 hidden layers of 501 hidden units each on the Fashion-MNIST dataset (multi-class classification). The reported time (in seconds) refers to one epoch for each algorithm.