Message Scheduling Methods for Belief Propagation

Christian Knoll[1(✉)], Michael Rath[1], Sebastian Tschiatschek[2], and Franz Pernkopf[1]

[1] Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria
christian.knoll@tugraz.at
[2] Learning and Adaptive Systems Group, Department of Computer Science, ETH Zurich, Zürich, Switzerland
sebastian.tschiatschek@inf.ethz.ch

Abstract. Approximate inference in large and densely connected graphical models is a challenging but highly relevant problem. Belief propagation, as a method for performing approximate inference in loopy graphs, has shown empirical success in many applications. However, convergence of belief propagation can only be guaranteed for simple graphs. Whether belief propagation converges depends strongly on the applied message update scheme, and specialized schemes can be highly beneficial. Yet, residual belief propagation is the only established method utilizing this fact to improve convergence properties. In experiments, we observe that residual belief propagation fails to converge if local oscillations occur and the same sequence of messages is repeatedly updated. To overcome this issue, we propose two novel message update schemes. In the first scheme we add noise to oscillating messages. In the second scheme we apply weight decay to gradually reduce the influence of these messages and consequently enforce convergence. Furthermore, in contrast to previous work, we consider the correctness of the obtained marginals and observe significant performance improvements when applying the proposed message update schemes to various Ising models with binary random variables.

Keywords: Residual belief propagation · Asynchronous message scheduling · Convergence analysis

1 Introduction

Probabilistic reasoning for complex distributions arises in many practical problems including computer vision, medical diagnosis systems, and speech processing [9]. These complex distributions are often modeled as probabilistic graphical models (PGMs). PGMs representing the joint distribution over many random

F. Pernkopf—This work was supported by the Austrian Science Fund (FWF) under the project number P25244-N15.

A. Appice et al. (Eds.): ECML PKDD 2015, Part II, LNAI 9285, pp. 295–310, 2015.
DOI: 10.1007/978-3-319-23525-7_18

variables (RVs) of practical problems are often complex and include many loops. Thus performing exact inference is increasingly intricate, in fact exact inference is intractable for general PGMs [1]. Message passing, a powerful method to approximate the marginal distribution, was first introduced to the field of machine learning as Belief Propagation (BP) by Pearl [19]. It is a parallel update scheme where messages are recursively exchanged between RVs until the marginal probabilities converge.

The conjecture that asynchronously updating the messages leads to better convergence performance of BP is widely accepted [2,5,22]. Thus, there was a recent interest in improving the performance of BP by applying dynamic message scheduling. One efficient way for scheduling is residual belief propagation (RBP) [2], where only the message that changes the most is sent. RBP has a provable convergence rate that is at least as good as the convergence rate of BP, while still providing good marginals. The quality of the obtained marginals in [2,24] is comparable to existing methods. Nonetheless, a detailed analysis of the quality of the marginals in comparison to the exact marginals is missing to the best of our knowledge. Dynamic message scheduling increases the number of graphs where BP converges. Yet, on graphs with many loops the occurrence of message oscillation is observed. In this case, a small set of messages is repeatedly selected for update and periodically takes the same values.

Inspired by this observation we introduce and investigate two different methods for dynamic message scheduling. The first method directly improves upon RBP if message oscillations occur. *Noise injection* belief propagation (NIBP) detects message oscillations of RBP. Adding random noise to the message that is propagated prevents these oscillations and improves convergence of BP. The second method is based on the assumption that messages repeatedly taking the same values do not contribute to convergence of the overall PGM. A sequence of oscillating messages does obviously not change the constraints in favor of convergence. We apply weight decay to the residual and consequently, support non oscillating messages to be updated. This way we avoid message oscillations before they even occur. *Weight decay* belief propagation (WDBP) solely changes the scheduling by the damping, whereas directly applying a damping term to the beliefs can also improve the convergence properties [21].

Our proposed methods are evaluated on different realizations of Ising grid graphs. Graphs of such structure have a rich history in statistical physics [8,15], and these models are appealing, as phase transitions can be analytically determined. Phase transitions separate convergent from divergent behavior and can be related to PGMs and the behavior of BP. It is shown in [25,26] how the fixed point solutions of BP correspond to the local minima of the Gibbs free energy.

On difficult Ising graphs we compare the performance of the proposed methods to RBP and asynchronous belief propagation (ABP). The convergence behavior is usually analyzed in terms of the number of times BP converges (i.e converged runs) and the speed of convergence (i.e. convergence rate). In addition, we compare the approximated marginals to the exact marginals, which are obtained by the junction tree algorithm [12].

Our two main findings are: (i) we show empirically that NIBP significantly increases the number of times convergence is achieved and (ii) WDBP accomplishes a quality of marginals superior to the remaining methods, while maintaining good convergence properties.

The rest of this paper is structured as follows. In Section 2 we give a short background on probabilistic graphical models and belief propagation. We introduce our proposed approach to message scheduling in Section 3 and relate it to existing methods. Our experimental results are presented and discussed in Section 4. Related work is deferred to Section 5 for the sake of reading flow. Section 6 summarizes the paper and provides some final conclusions.

2 Preliminaries

In this section we briefly introduce PGMs and the BP algorithm. Some applications and a detailed treatment of PGMs can be found in [11,20]. Let X be a binary random variable (RV) taking values $x \in \mathbb{S} = \{-1, 1\}$. We consider the finite set of random variables $\mathbf{X} = \{X_1, \ldots, X_N\}$.

An undirected graphical model (UGM) consists of an undirected graph $G = (\mathbf{X}, \mathbf{E})$, where $\mathbf{X} = \{X_1, \ldots, X_N\}$ represents the nodes and $\mathbf{E}$ the edges. Two nodes X_i and X_j, $i \neq j$ can be connected by an undirected edge $e_{i,j} \in \mathbf{E}$ that specifies an interaction between these two nodes. Note that we use the same terminology for the nodes as for the RVs since there is a one-to-one relationship. The set of neighbors of X_i is defined by $\Gamma(X_i) = \{X_j \in \mathbf{X} \backslash X_i : e_{i,j} \in \mathbf{E}\}$. We use an UGM to model the joint distribution

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_{(i,j)\,:\, e_{i,j} \in \mathbf{E}} \Phi_{X_i, X_j}(x_i, x_j) \prod_{i=1}^{N} \Phi_{X_i}(x_i), \tag{1}$$

where the first product runs over all edges, and where Φ_{X_i, X_j} are the pairwise potentials and Φ_{X_i} is the local potential.

Our formulation of BP is similar to the one introduced in [25]. For a detailed introduction to the concept of BP we refer the reader to [19,29]. The messages are updated according to the following rule:

$$\mu_{i,j}^{n+1}(x_j) = \sum_{x_i \in \mathbb{S}} \Phi_{X_i, X_j}(x_i, x_j) \Phi_{X_i}(x_i) \prod_{X_k \in (\Gamma(X_i) \backslash \{X_j\})} \mu_{k,i}^{n}(x_i), \tag{2}$$

where $\mu_{i,j}^{n}(x_j)$ is the message from X_i to X_j of state x_j at iteration n.[1] Loosely speaking this means that X_i collects all messages from its neighbors $\Gamma(X_i)$ except for X_j. This product is then multiplied with the pairwise and local potentials $\Phi_{X_i, X_j}(x_i, x_j)$ and $\Phi_{X_i}(x_i)$. Finally the sum over all states of X_i is sent to X_j.

[1] Note that without loss of generality we will drop the superscript n where no ambiguities occur.

The marginals (or beliefs) $P(X_i = x_i)$ are obtained from all incoming messages according to

$$P(X_i = x_i) = \frac{1}{Z} \Phi_{X_i}(x_i) \prod_{X_k \in \Gamma(X_i)} \mu_{k,i}(x_i), \tag{3}$$

where $Z \in \mathbb{R}^+$ is the normalization constant ensuring that $\sum_{x_i \in \mathbb{S}} P(X_i = x_i) = 1$. When the specific realization is not relevant we use the shorthand notation $P(X_i)$ instead.

There is a rich history of statistical physicists studying the interaction in Ising models. The Edwards-Anderson model or Ising spin glass is an elegant abstraction that allows both, ferromagnetic and antiferromagnetic Ising models [14, p. 44]. Following the terminology of the Edwards-Anderson model we define the potentials of the model, such that we have a coupling $J_{i,j} \in \mathbb{R}$ and a local field $\theta_i \in \mathbb{R}$. Let the potentials be $\Phi_{X_i}(x_i) = \exp(\theta_i x_i)$ and $\Phi_{X_i,X_j}(x_i, x_j) = \exp(J_{i,j} x_i x_j)$. Plugging these potentials into (1), the Ising spin glass model defines the joint probability

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{(i,j)\,:\, e_{i,j} \in \mathbf{E}} J_{i,j} x_i x_j + \sum_{i=1}^{N} \theta_i x_i \right), \tag{4}$$

where the sum over $(i, j)\colon e_{i,j} \in \mathbf{E}$ runs over all edges of G and the second sum runs over all nodes. Spin glasses in this form offer a powerful generalization of the Ising model that allow for frustration.[2] Such models have been used to relate the convergence problem to the occurrence of phase transitions [4]. One can consequently derive a sharp bound for the parameter set $(J_{i,j}, \theta_i)$ and relate it to the convergence of loopy BP [18,25,26].

When analyzing the graph convergence over time, it is remarkable that certain subgraphs are almost converged after few iterations, while other regions are less stable. More formally we can introduce two subgraphs such that $G = G_c \cup G_{\bar{c}}$. We define the almost converged subgraph as $G_c = (\mathbf{X_c}, \mathbf{E_c})$, i.e. for all $(X_i, X_j) : e_{i,j} \in \mathbf{E_c}$ we have $\mu_{i,j}^{n+1}(x_j) \approx \mu_{i,j}^{n}(x_j)$. The second subgraph $G_{\bar{c}} = (\mathbf{X_{\bar{c}}}, \mathbf{E_{\bar{c}}})$ is less stable, i.e. $\mu_{i,j}^{n+1}(x_j) \not\approx \mu_{i,j}^{n}(x_j)$. Note that $G_{\bar{c}}$ may even include frustrated cycles such that convergence can never be reached.

3 Scheduling

For a given graph $G = (\mathbf{X}, \mathbf{E})$ we can define any message passing algorithm by basic operations on the alphabet of messages (cf. [14, p. 316]). The algorithm is converged if two successive messages show approximately the same value, i.e. $\mu_{i,j}^{n+1}(x_j) \approx \mu_{i,j}^{n}(x_j)$. At that point, updating the messages does not change their values, therefore we can also speak of a fixed point solution.

Note that in the original implementation of BP all messages are synchronously updated, i.e. to compute $\mu_{i,j}^{n+1}$ *all* messages at iteration n are used.

[2] Frustrated cycles have an overall parametrization, such that it is impossible to simultaneously satisfy all local constraints, i.e. convergence can never be achieved.

Substituting the synchronous update rule by a sequential update rule, we obtain a flexibility in developing variants of BP. Exploiting this flexibility and changing the update schedule significantly influences the performance in practice, as reported in [2,14]. We are essentially interested in the advantages of different update schedules, therefore we solely consider sequential (or asynchronous) scheduling for the remainder of the work.

All variants of BP are compared to the performance of asynchronous belief propagation (ABP). ABP is based on a rudimentary sequential update rule, where all messages are considered equally important. Messages are selected according to round robin scheduling, i.e. according to a fixed order. Although no assumptions are made on a smart choice of the order, it can be observed that this simple message scheduling concept improves the convergence behavior [10,22].

We propose two modifications to BP to improve convergence properties. Either we change the calculation of the *message values* directly (NIBP), or we utilize alternative *message scheduling* (WDBP). In the following we describe these modifications in detail. Experimental results demonstrating the effectiveness of the proposed modifications can be found in Section 4.

3.1 Residual Belief Propagation

Residual belief propagation (RBP) utilizes a priority measure for each message and introduces dynamic scheduling [2]. The underlying assumption is that any message passed along an edge $e_{i,j} \in \mathbf{E_c}$ in the already converged subgraph does not contribute to the overall convergence. Thus focusing on the subgraph that has not converged $G_{\bar{c}} = (\mathbf{X_{\bar{c}}}, \mathbf{E_{\bar{c}}})$ is beneficial for convergence of the overall graph. As $G_{\bar{c}}$ is not converged, messages along edges $\bar{e}_{i,j} \in \mathbf{E}_{\bar{c}}$ vary considerably in every step.

This leads to the update rule of RBP, where the residual $r^n_{i,j} = |\mu^{n+1}_{i,j}(x_j) - \mu^n_{i,j}(x_j)|$ measures the distance between two messages.[3] The indices that maximize the residual

$$(k,l) = \underset{(i,j)}{\operatorname{argmax}}\, r^n_{i,j} \tag{5}$$

identify the message to update next, i.e.

$$\tilde{\mu}^{n+1}_{k,l}(x_l) = \mu^{n+1}_{k,l}(x_l). \tag{6}$$

Compared to ABP the number of graphs where RBP converges increases significantly [2]. Still, RBP computes all residuals although only the message with the most significant residual is sent. To further increase the convergence rate, the authors of [24] bound and approximate the message values for the estimation of the residual.

[3] Ultimately one would be interested in the distance to the fixed point, if it exists, $\lim_{n\to\infty} \mu^n_{i,j}(x_j)$. However, since $\lim_{n\to\infty} \mu^n_{i,j}(x_j)$ is not known, the time variation of the messages offers a valid surrogate (cf. [2]).

3.2 Noise Injection Belief Propagation

Investigating graphs with random Ising factors, where RBP fails to converge, we observe that a large part of the PGM is almost converged. We observe local frustrated cycles in $G_{\bar{c}}$, where the same message values are passed around repeatedly along the edges $\bar{e}_{i,j} \in \mathbf{E}_{\bar{c}}$. Noise injection belief propagation (NIBP) compares the current message $\mu^n_{i,j}$ to the last $L \in \mathbb{Z}^+$ messages for duplicate values. If older messages are in an δ-neighborhood, i.e. $|\mu^n_{i,j} - \mu^{n-l}_{i,j}| < \delta$ for any $l \in \{1, 2, \ldots, L\}$, although these messages are not converged, i.e. $\mu^{n+1}_{i,j} \not\approx \mu^n_{i,j}$, we conclude that the message values oscillate. If no oscillations are detected NIBP does not change the scheduling of RBP. Therefore, NIBP always converges if RBP does. If, however, message values oscillate Gaussian noise $\mathcal{N}(0, \sigma^2)$ is added to the message $\mu^{n+1}_{k,l}$ that is selected according to (5). The new update rule is then given as

$$\tilde{\mu}^{n+1}_{k,l}(x_l) = \mu^{n+1}_{k,l}(x_l) + \mathcal{N}(0, \sigma^2), \tag{7}$$

where X_k and X_l are the nodes that maximize the residual in (5) and $\mathcal{N}(0, \sigma^2)$ is the normal distribution with zero mean and standard deviation σ.

Loosely speaking we aim to introduce a relevant change to the system by injecting noise to the message selected for update $\mu^{n+1}_{k,l}$. Adding noise to the most influential part of the PGM, we assume that this minor change of one message propagates through the whole graph and leads to a stable fixed point. Pseudocode of the implementation can be found in Appendix A.

3.3 Weight Decay Belief Propagation

As mentioned above RBP fails to converge if message values oscillate. Obviously, repeatedly sending around the same messages along the same path does not contribute to achieving convergence. Weight decay belief propagation (WDBP) penalizes this behavior by damping the residual of messages along $\bar{e}_{i,j}$. Consequently, WDBP increases the relevance of G_c and further refines the parametrization of this subgraph. In doing so, messages $\mu_{i,j}$ between both subgraphs, where $X_i \in \mathbf{X_c}$ and $X_j \in \mathbf{X_{\bar{c}}}$ are re-evaluated, such that convergence can be achieved on the overall graph G.

In particular, we damp the residual of all messages of a node X_i based on the number of times a message has already been scheduled. More formally we rewrite (5), such that the indices of the selected message $\tilde{\mu}^{n+1}_{k,l}$ are given to

$$(k, l) = \underset{(i,j)}{\operatorname{argmax}} \frac{r^n_{i,j}}{\sum_{m=1}^{n} \mathbf{1}_{\mu^m_{i,j}}}, \tag{8}$$

where the indicator function $\mathbf{1}_{\mu^m_{i,j}} = 1$ if and only if $\mu^m_{i,j} = \tilde{\mu}^m_{i,j}$. Hence, the residual is divided by a factor corresponding to how often a certain message was selected. A detailed implementation is presented in Appendix A.

4 Experiments

In this section we evaluate the proposed methods and compare them to ABP and RBP. We evaluate all different types of scheduling with respect to the following measures: first the number of configurations where the algorithm converges will be considered, secondly we consider the rate of convergence, and finally we evaluate the quality of the marginals. To evaluate the marginals we obtain the approximate marginal distributions $\tilde{P}(X_i)$ and compare them to the exact marginal distributions $P(X_i)$, obtained by applying the junction tree algorithm [12,16]. Although the junction tree algorithm is intractable in general, the considered PGMs are simple enough to make exact inference computationally feasible. We quantify the quality of the marginals by computing the mean squared error (MSE) over all marginals. Note that the potential functions are identical for all compared methods.

Statistical physic provides exact statements regarding the performance of BP on Ising spin glasses, therefore such models are commonly used for evaluation of BP variants. In this work we perform message passing on Ising spin glasses of varying size with uniform and random coefficients.

For NIBP, the parameters of the additive Gaussian noise were optimized for different initialization and are zero mean and $\sigma = 0.25$. Simulations were either stopped after k_{max} iterations or if all messages converged, i.e. $\max_{i,j} |\mu_{i,j}^{n+1}(x_j) - \mu_{i,j}^{n}(x_j)| < \epsilon$ for all $i, j : i \neq j$, where $\epsilon = 10^{-3}$. Experiments on Ising grids with uniform parameters were stopped after $k_{max} = 4 \cdot 10^5$ iterations, whereas the experiments on Ising grids with random factors were stopped after $k_{max} = 2.5 \cdot 10^5$ iterations.

4.1 Fully Connected Graph with Uniform Parameters

We consider a fully connected Ising spin glass with $|\mathbf{X}| = 4$ binary spins, and uniform coupling $J_{i,j}$ and field θ_i among the four vertices. In the case of uniform parameters we introduce the shorthand notation (J, θ). Using such a model allows to compare our results to similar numerical experiments performed on this type of graphs in [18,25,26]. Figure 1 shows the complete graph for $|\mathbf{X}| = 4$.

Applying BP to this graph one can benefit from the rich history of statistical physics literature to discuss the effect of different messages schedules. For a fully connected Ising spin glass with uniform parameters the Gibbs measure is unique and the solution of BP is exactly equal to the one obtained by optimizing the Bethe approximation [26]. That is, there are certain regions in the 2-dimensional parameter-space (J, θ) where BP is guaranteed to converge. Nonetheless, there is a phase transition in the parameter space where BP does not converge. If $J \geq 0$ the model is known to be ferromagnetic and in fact reduces to the standard Ising model. The antiferromagnetic behavior is observed for $J < 0$, respectively [14].

In Figure 2a we show convergence of ABP and the transition to configurations (J, θ) where messages oscillate. The color encodes the logarithm of the number of iterations until convergence. We observe that reducing J increases

the difficulty of finding an equilibrium state. This, however, is intuitive since, the more negative J is, the more one node X_i tries to push its neighbors $\Gamma(X_i)$ into the opposite state.

Looking at Figure 2b we observe how RBP pushes the transition boundary and increases the set of coefficients where convergence is achieved. Finally Figure 2c and 2d show the performance of NIBP and WDBP respectively. Notably, both methods further increase the region of convergence. It can be seen that these boundaries are heavily blurred. For specific parameter configurations our proposed methods result in equilibrium state after many runs, where established methods fail to converge.

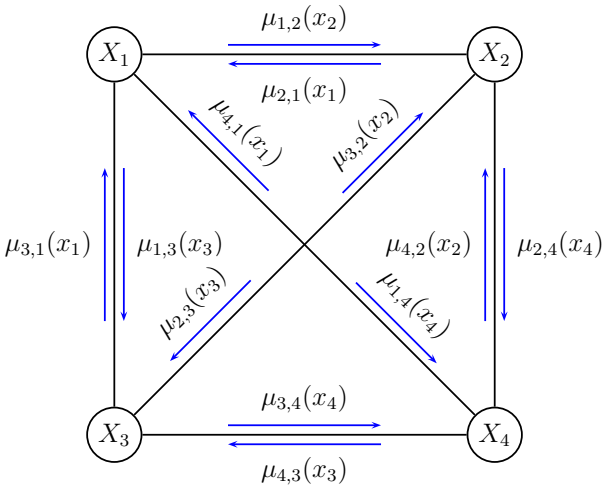
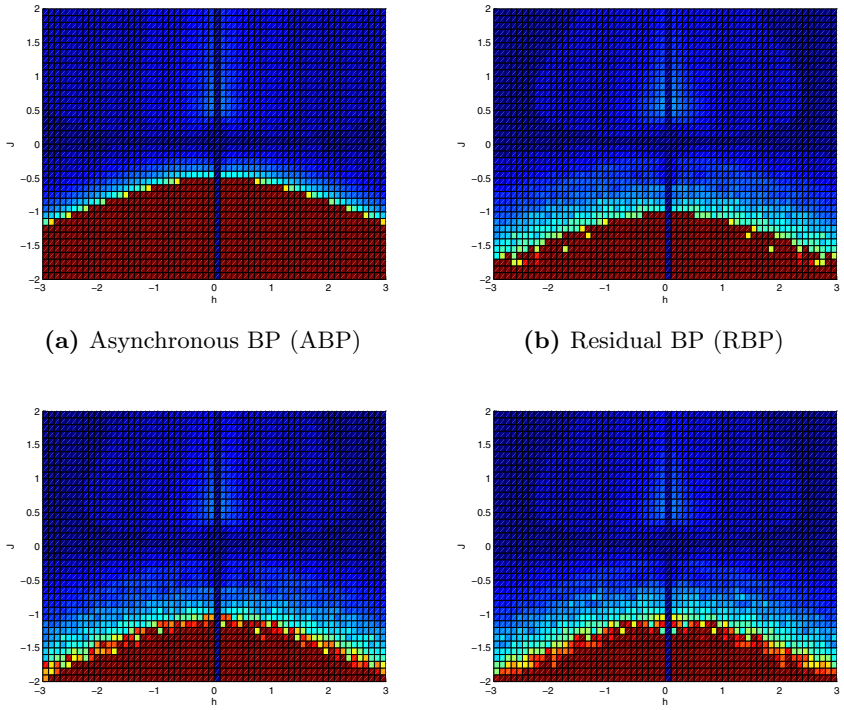


Fig. 1. 2x2 Ising Spin Glass.

4.2 Ising Grids with Random Factors

From the experiments in Figure 2 we can hardly make any concrete statements regarding the convergence behavior. Hence, to further investigate the influence of WDBP and NIBP we consider Ising grids with many loops and random parameters. These graphs are often used for evaluation of the performance of BP, since BP is prone to diverge on those graphs. We consider grid graphs of size $N = |\mathbf{X}| = K \times K$ with binary spins and randomly initialized parameters $(J_{i,j}, \theta_i)$. Depending on the grid size K these parameters are uniformly sampled such that both $(J_{i,j}, \theta_i) \in [-\frac{K}{2}, \frac{K}{2}]$. Thus, besides increasing the size of the graph, the difficulty is implicitly increased as well.

The larger the values of the coupling and the local field are, the harder the resulting constraints for convergence are. Thus, although there is no structural change of the grid, inference becomes easier by reducing the range of the parameters. According to [24] the parameters have to provide sufficient difficulty to be of interest for analyzing convergence properties.



(a) Asynchronous BP (ABP)

(b) Residual BP (RBP)

(c) Noise Injection BP (NIBP)

(d) Weight Decay BP (WDBP)

Fig. 2. Convergence of various BP variants for a fully connected binary Ising spin glass with uniform parameters (J, θ). The color encodes the logarithm of the number of iterations until convergence; blue corresponds to convergence and red means that the method did not converge after $4 \cdot 10^5$ iterations. 2a shows the phase transition of ABP. Note how the RBP variants in 2b-2d increase the region of convergence.

The proportion of converged runs for different schedules is shown in Figure 3. We can see that ABP finds a fixed point in less scenarios than any other of the proposed variants, demonstrating the advantage of dynamic message scheduling. Looking at the overall performance we observe that NIBP converges most often throughout all experiments. The more complex the network, the better NIBP performs compared to other variants. WDBP outperforms RBP on all experiments and shows the best performance on the 7×7 graph, although the harder the problem, the slower it converges. Specifically, WDBP has a lower convergence rate than RBP. This observation is expected as damping the residual reduces the propagation of relevant messages even for relatively easy configurations.

It shall be noted that applying WDBP requires changing the residual to (8), where damping the residual implies some computational overhead. This overhead can be reduced partially with approximation of the messages according to [24].

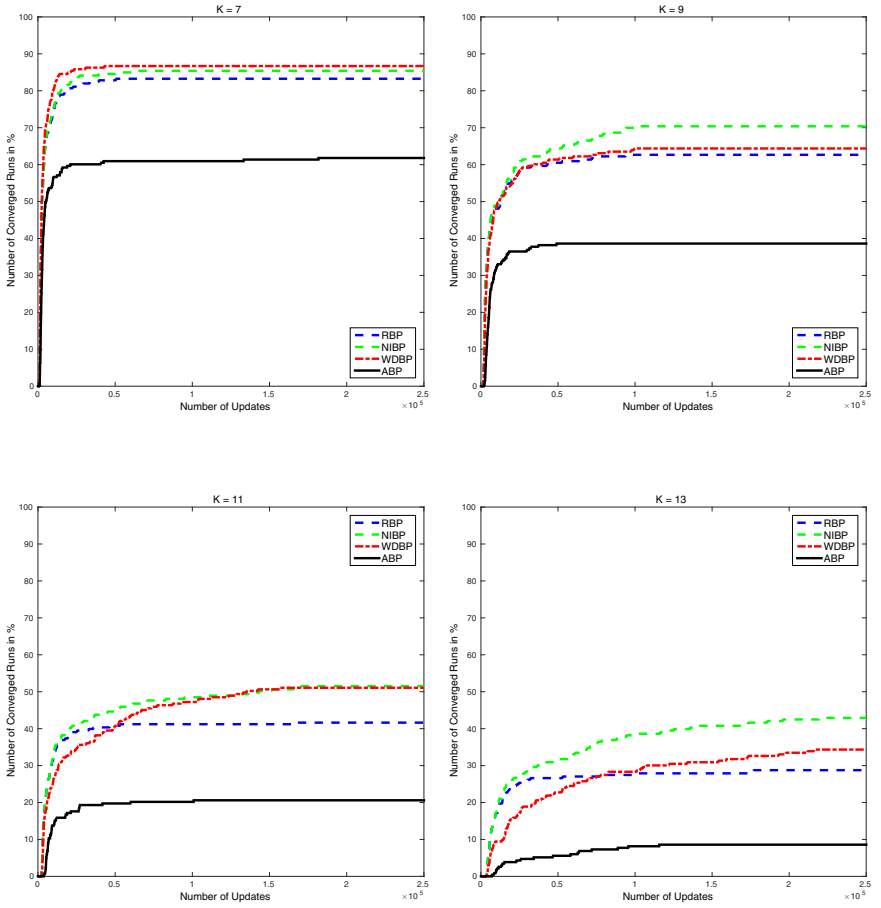


Fig. 3. Number of converged runs in percentage as a function of the number of message updates. All results were obtained by averaging over 233 random grid graphs. On graphs of the size $|\mathbf{X}| = K \times K$ we compare WDBP and NIBP to RBP and ABP.

All results were averaged over 233 runs with different random initialization of the pairs $(J_{i,j}, \theta_i)$ for $K \in \{7, 9, 11, 13\}$.

4.3 Quality of Marginals

Currently the influence of message scheduling was only evaluated in terms of the convergence rate and the number of graphs where BP converges. Here, we also evaluate the correctness of the approximated marginals $\tilde{P}(X_i)$, averaging the mean squared error (MSE) of all $N = K \times K = |\mathbf{X}|$ nodes, such that

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in \mathbb{S}} |\tilde{P}(X_i = a) - P(X_i = a)|^2, \tag{9}$$

where $P(X_i)$ are the exact marginals. Note that all RVs are binary and both, $\tilde{P}(X_i)$ and $P(X_i)$ are valid probability mass function, i.e. $\sum_{a \in \mathbb{S}} \tilde{P}(X_i) = 1$. Further applying symmetry properties it then follows that

$$MSE = \frac{2}{N} \sum_{i=1}^{N} |\tilde{P}(X_i = 1) - P(X_i = 1)|^2. \tag{10}$$

In Table 1 we present quantitative performance measures for all experiments. Solely considering the number of converged runs we can recapitulate the observation from Figure 3 that RBP converges in at least 20% of all experiments, where ABP did not. Both our proposed methods are able to further increase the convergence; throughout all experiments NIBP converges most often.

Still, in practice we are not only interested in the number of converged runs but also in the quality of the marginals. First we estimate the overall MSE based on (10) and average over all 233 randomly initialized graphs (MSE overall). Secondly, we average the MSE over all runs where the individual methods converged(MSE converged) – for ABP we estimate the MSE only for easy configurations, whereas the MSE for other variants includes harder configurations. Therefore, we finally estimate the MSE of all methods for those configurations where ABP converges to a fixed point(MSE ABP conv.).

It can be seen that ABP consistently shows the lowest MSE in terms of converged runs, i.e. averaging over all runs that converged with this method. This comes as no surprise as ABP converges only on graphs with relatively easy configurations. For these configurations we expect $\tilde{P}(X_i)$ to give a good approximation to the exact marginals $P(X_i)$. However, estimating the MSE of different methods for graphs where ABP is known to converge (MSE ABP conv.), we are surprised by the observations that the approximate marginals obtained by RBP or NIBP are consistently worse than the ones found by ABP. Still solely considering these easy graphs it is remarkable how well WDBP performs in terms of the MSE.

Note that by using an update rule based on RBP a lot of effort is put into locally complying with the constraints of $G_{\bar{c}}$ whereas ABP still puts a significant amount of resources into refining G_c. This clearly reduces the convergence rate but potentially boosts the correctness of the approximation.

We would expect the overall MSE, i.e. averaging over all 233 runs, is reduced using dynamic message scheduling. Despite ABP converges in less runs it still results in surprisingly good overall approximations of the marginals. In fact the obtained quality of the marginals is similar for ABP, RBP, and NIBP, supporting the empirical observations that ABP performs reasonable well on many graphs. Notably, it can also be seen that WDBP consistently reduces the overall MSE resulting in the best approximation of the marginals.

Looking at Table 1 we want to emphasize the superior overall performance of WDBP. The number of converged runs is significantly increased in comparison to ABP while a proper approximation accuracy is maintained.

Table 1. Performance of different BP schedules on Ising spin glasses of size $|\mathbf{X}| = K \times K$. The MSE is estimated between approximated $\tilde{P}(X_i)$ and exact $P(X_i)$ marginals. We average over all 233 runs (MSE overall), over runs where the individual methods converged (MSE converged), and over runs where ABP converged (MSE ABP conv.) We compare asynchronous BP (ABP), residual BP (RBP), noise injection BP (NIBP), and weight decay BP (WDBP).

Grid Size	Error Metric		ABP	RBP	NIBP	WDBP
$K = 7$	MSE	overall	0.0514	0.041	0.0382	**0.0330**
		converged	**0.0164**	0.0208	0.0218	0.0202
		ABP conv.	0.0164	0.0182	0.0150	**0.0130**
	Converged Runs		61.8%	83.26%	85.41%	**86.7%**
$K = 9$	MSE	overall	0.0706	0.0622	0.0538	**0.0486**
		converged	**0.0078**	0.0256	0.026	0.0230
		ABP conv.	**0.0078**	0.0190	0.0144	0.0112
	Converged Runs		38.63%	62.66%	**70.39%**	64.38%
$K = 11$	MSE	overall	0.0830	0.0914	0.0750	**0.0618**
		converged	**0.0106**	0.0340	0.0386	0.0258
		ABP conv.	**0.0106**	0.0262	0.0268	0.0152
	Converged Runs		20.6%	41.63%	**51.5%**	51.07%
$K = 13$	MSE	overall	0.1126	0.1274	0.1102	**0.0840**
		converged	**0.0286**	0.0642	0.0632	0.0314
		ABP conv.	0.0286	0.0746	0.0590	**0.0282**
	Converged Runs		8.58%	28.76%	**42.92%**	34.33%

5 Related Work

On trees and chains BP is guaranteed to converge, moreover BP obtains the optimal maximum a posterior assignment for PGMs with a single loop [27]. However, many graphs that represent a domain of the real world have an arbitrary structure, including loops. There is no general guarantee for BP to converge on such complex graphs [3,27]. Yet, it was shown empirically that BP can still give good results when applied to graphs with a complicated structure.

There are various approaches that aim to correct for the presence of loops such as loop correction [17] or the truncated loop series introduced in [6]. There are also many publications relating the fixed points of BP to extrema of approximate free energy functions from statistical physics [7,28]. It was shown in [28] how extrema of the Bethe free energy approximations correspond to the fixed points of BP. Using the generalization, the Kikuchi free energy function, generalized BP (GBP) was introduced in [28], which significantly improves the number of converged runs and the convergence rate compared to standard BP. Applying a concave- convex procedure to the Bethe and Kikuchi free energies the CCCP algorithm is introduced in [30] and results in slightly better results than those found by GBP. Convexified free energies [13] come with good convergence properties but still lack the empirical success. Linear programming relaxation can be

used to deal with frustrated cycles as well [23]. Long range correlations often lead to failure of BP [14] but can be handled through the cavity method [15].

6 Conclusion

In this paper, we introduced two novel methods for dynamic message scheduling. Refining the ideas of residual belief propagation (RBP), we further improve the number of converged runs on various difficult graphs.

The first method, noise injection belief propagation (NIBP) detects if RBP fails to find a fixed point, i.e. message values oscillate. Gaussian noise is then added to the selected message such that the overall configuration is modified to achieve convergence. Our assumption is that this noise injection propagates to other parts of the network and counteracts the oscillations. Still if RBP converges, NIBP is guaranteed to converge as well.

The second method, weight decay belief propagation (WDBP) obviates the need for oscillation detection. Each time a message is selected for an update, the importance of the message for potential future updates is reduced. Thus, WDBP implicitly reduces the priority of subgraphs that oscillate and forces the overall graph to a fixed point.

Both proposed methods are applied to various Ising grids and are evaluated in comparison to other sequential message passing algorithms. Our main evaluation is based on convergence properties and the correctness of the marginals. In all experiments both methods, NIBP and WDBP converge more often than RBP and asynchronous belief propagation (ABP).

NIBP achieves the highest convergence rate and number of converged runs. Still, considering the mean squared error of the marginals we notice that ABP leads to surprisingly good marginals. Applying RBP and NIBP to increase the number of converged runs comes with a sacrifice of the approximation accuracy of the marginals.

We further compare the MSE between the exact and the approximated marginals in different scenarios. This quality aspect has not been mentioned in previous work. Only considering easy graphs, where ABP converges, we are surprised by the observation that ABP consistently outperforms RBP or NIBP in terms of the quality of the approximated marginals. The quality of the marginals obtained by WDBP on these graphs is remarkable and superior to all compared methods.

By all means the above results highlight how the message passing scheduling influences the performance of belief propagation. Still, the convergence rate of both, NIBP and WDBP can potentially be further improved by using an estimation of the residual [24] instead of computing the messages for every step.

Acknowledgments. This work was supported by the Austrian Science Fund (FWF) under the project number P25244-N15.

Appendix A: Pseudocode

We present the pseudocode for NIBP and WDBP. Removing the if then else clause in line 8 to 11 of NIBP and substituting it with $\mu_u^{old} \leftarrow \mu_u^{new}$ reduces

Algorithm 1 to RBP. The maximum number of iterations is denoted by $k_{max} = 2.5 \cdot 10^5$ and $\epsilon = 10^{-3}$. NrOfMessages denotes the overall number of messages in the graph.

Algorithm 1. Noise Injection Belief Propagation (NIBP)

input : Graph $G = (\mathbf{X}, \mathbf{E})$
output: Converged messages μ^{old}

initialization
for $i \leftarrow 1$ **to** NrOfMessages **do**
  $\mu_i^{new} \leftarrow$ `ComputeUpdate`(μ_i^{old})
  $r_i \leftarrow |\mu_i^{old} - \mu_i^{new}|$
$k \leftarrow 1$
while $k < k_{max}$ **and** $\max |\mu^{old} - \mu^{new}| > \epsilon$ **do**
  $u \leftarrow \text{argmax}_i\, r$
  if `OscillationDetection`(μ_u^{old}, L) **then**
    $\mu_u^{old} \leftarrow \mu_u^{new} + \mathcal{N}(0, \sigma)$
  else
    $\mu_u^{old} \leftarrow \mu_u^{new}$
  for $j \leftarrow 1$ **to** NrOfMessages **do**
    $\mu_j^{new} \leftarrow$ `ComputeUpdate`(μ_j^{old})
    $r_j \leftarrow |\mu_j^{new} - \mu_j^{old}|$
  $k = k + 1$

Algorithm 2. Weight Decay Belief Propagation (WDBP)

input : Graph $G = (\mathbf{X}, \mathbf{E})$
output: Converged messages μ^{old}

initialization
for $i \leftarrow 1$ **to** NrOfMessages **do**
  $\mu_i^{new} \leftarrow$ `ComputeUpdate`(μ_i^{old})
  $r_i \leftarrow |\mu_i^{old} - \mu_i^{new}|$
  NrUpdates (i) $\leftarrow 1$
$k \leftarrow 1$
while $k < k_{max}$ **and** $\max |\mu^{old} - \mu^{new}| > \epsilon$ **do**
  $u \leftarrow \text{argmax}_i\, r$
  $\mu_u^{old} \leftarrow \mu_u^{new}$
  NrUpdates (u) $\leftarrow$ NrUpdates (u) + 1
  for $j \leftarrow 1$ **to** NrOfMessages **do**
    $\mu_j^{new} \leftarrow$ `ComputeUpdate`(μ_j^{old})
    $r_j \leftarrow \frac{|\mu_j^{new} - \mu_j^{old}|}{\text{NrUpdates(j)}}$
  $k = k + 1$

References

1. Cooper, G.F.: The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence **42**(2), 393–405 (1990)
2. Elidan, G., McGraw, I., Koller, D.: Residual belief propagation: Informed scheduling for asynchronous message passing. In: Conference on Uncertainty in Artificial Intelligence (UAI) (2006)
3. Frey, B.J., MacKay, D.J.: A revolution: Belief propagation in graphs with cycles. In: Neural Information Processing Systems (NIPS), pp. 479–485 (1998)
4. Georgii, H.O.: Gibbs Measures and Phase Transitions, vol. 9 (2011)
5. Goldberger, J., Kfir, H.: Serial schedules for belief-propagation: analysis of convergence time. IEEE Transactions on Information Theory **54**(3), 1316–1319 (2008)
6. Gómez, V., Mooij, J.M., Kappen, H.J.: Truncating the loop series expansion for belief propagation. The Journal of Machine Learning Research (2007)
7. Heskes, T.: On the uniqueness of loopy belief propagation fixed points. Neural Computation **16**(11), 2379–2413 (2004)
8. Ising, E.: Beitrag zur Theorie des Ferromagnetismus. Zeitschrift für Physik A Hadrons and Nuclei **31**(1), 253–258 (1925)
9. Jordan, M.I.: Graphical models. Statistical Science, pp. 140–155 (2004)
10. Kfir, H., Kanter, I.: Parallel versus sequential updating for belief propagation decoding. Physica A: Statistical Mechanics and its Applications **330**(1)
11. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT press (2009)
12. Lauritzen, S.L., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society. Series B (Methodological), 157–224 (1988)
13. Meshi, O., Jaimovich, A., Globerson, A., Friedman, N.: Convexifying the bethe free energy. In: Conference on Uncertainty in Artificial Intelligence (UAI), pp. 402–410. AUAI Press (2009)
14. Mezard, M., Montanari, A.: Information, Physics, and Computation. Oxford University Press (2009)
15. Mézard, M., Parisi, G.: The Bethe lattice spin glass revisited. The European Physical Journal B-Condensed Matter and Complex Systems **20**(2), 217–233 (2001)
16. Mooij, J.M.: libdai: A free and open source c++ library for discrete approximate inference in graphical models. The Journal of Machine Learning Research **11** (2010)
17. Mooij, J.M., Kappen, H.J.: Loop corrections for approximate inference on factor graphs. Journal of Machine Learning Research **8**, 1113–1143 (2007)
18. Mooij, J.M., Kappen, H.J.: Sufficient conditions for convergence of the sum-product algorithm. IEEE Transactions on Information Theory **53**(12), 4422–4437 (2007)
19. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Representation and Reasoning Series. Morgan Kaufmann Publishers (1988)
20. Pernkopf, F., Peharz, R., Tschiatschek, S.: Introduction to Probabilistic Graphical Models (2014)
21. Pretti, M.: A message-passing algorithm with damping. Journal of Statistical Mechanics: Theory and Experiment **2005**(11), P11008 (2005)
22. Sharon, E., Litsyn, S., Goldberger, J.: Efficient serial message-passing schedules for LDPC decoding. IEEE Transactions on Information Theory **53**(11), 4076–4091 (2007)

23. Sontag, D., Choe, D.K., Li, Y.: Efficiently searching for frustrated cycles in MAP inference. In: Conference on Uncertainty in Artificial Intelligence (UAI) (2012)
24. Sutton, C.A., McCallum, A.: Improved dynamic schedules for belief propagation. In: Conference on Uncertainty in Artificial Intelligence (UAI) (2007)
25. Taga, N., Mase, S.: On the convergence of belief propagation algorithm for stochastic networks with loops. Citeseer (2004)
26. Tatikonda, S.C., Jordan, M.I.: Loopy belief propagation and Gibbs measures. In: Conference on Uncertainty in Artificial Intelligence (UAI) (2002)
27. Weiss, Y.: Correctness of local probability propagation in graphical models with loops. Neural Computation **12**(1), 1–41 (2000)
28. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Bethe free energy, Kikuchi approximations, and belief propagation algorithms. Neural Information Processing Systems (NIPS) **13** (2001)
29. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Understanding belief propagation and its generalizations. Exploring Artificial Intelligence in the New Millennium **8**, 239–269 (2003)
30. Yuille, A.L.: CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. Neural Computation **14**(7) (2002)