Supplementary Material for "Tensor Belief Propagation"

Andrew Wrigley [1] **Wee Sun Lee** [2] **Nan Ye** [3]

Appendix (Supplementary Material)

Proof of Consistency

To prove consistency of the algorithm, we introduce a few lemmas.

Lemma 1. Let $X_{1,n}, \ldots, X_{m,n}$ be random variables such that $X_{i,n} \xrightarrow{p} \mu_i$ as $n \to \infty$. Let $X_{i,n} \in [0, M]$ and $\mu_i \in [0, M]$ for $i = 1, \ldots, m$. Then $\sum_i X_{i,n} \xrightarrow{p} \sum_i \mu_i$ and $\prod_i X_{i,n} \xrightarrow{p} \prod_i \mu_i$.

Proof. Since $X_{i,n} \xrightarrow{p} \mu_i$, for any $\epsilon > 0$, and any $\delta > 0$, there exists an N_i such that for $n > N_i$, $P(|X_{i,n} - \mu_i| > \frac{\epsilon}{m}) \leq \frac{\delta}{m}$. Let $N = \max\{N_1, \ldots, N_m\}$, and assume $n > N$. Then for every i, $P(|X_{i,n} - \mu_i| > \frac{\epsilon}{m}) \leq \frac{\delta}{m}$. By the union bound, with probability at least $1 - \delta$, for every i, $|X_{i,n} - \mu_i| \leq \epsilon$, which implies $|\sum_i X_{i,n} - \sum_i \mu_i| \leq \epsilon$. Hence when $n > N$, $p(|\sum_i X_{i,n} - \sum_i \mu_i| > \epsilon) \leq \delta$. It follows that $\sum_i X_{i,n} \xrightarrow{p} \sum_i \mu_i$.

To show that $\prod_i X_{i,n} \xrightarrow{p} \prod_i \mu_i$, it suffices to show this for $m = 2$. The proof then follows by mathematical induction.

For any $\epsilon > 0$ and $\delta > 0$, there exists N_1 and N_2 such that for any $n > \max\{N_1, N_2\}$, we have $P(|X_{1,n} - \mu_1| > \frac{\epsilon}{3M}) < \frac{\delta}{2}$ and $P(|X_{2,n} - \mu_2| > \frac{\epsilon}{3M}) < \frac{\delta}{2}$. By the union bound, with probability at least $1 - \delta$, $|X_{1,n} - \mu_1| \leq \frac{\epsilon}{3M}$ and $|X_{2,n} - \mu_1| \leq \frac{\epsilon}{3M}$. Let $s_1, s_2 \in \{-1, 1\}$ and assume $\frac{\epsilon}{3M} \leq M$ (the result in the case $\frac{\epsilon}{3M} > M$ holds trivially). Then, with probability at least $1 - \delta$,

$$
\begin{aligned}
&|X_{1,n}X_{2,n} - \mu_1\mu_2| \\
&\leq \max_{s_1,s_2} \left| \left(\mu_1 + \frac{s_1\epsilon}{3M}\right)\left(\mu_2 + \frac{s_2\epsilon}{3M}\right) - \mu_1\mu_2 \right| \\
&= \max_{s_1,s_2} \left| \frac{\mu_1 s_2 \epsilon}{3M} + \frac{\mu_2 s_1 \epsilon}{3M} + \frac{s_1 s_2 \epsilon^2}{9M^2} \right| \\
&\leq \epsilon,
\end{aligned}
$$

where we have used $\mu_1, \mu_2 \leq M$ and $\frac{\epsilon}{3M} \leq M$. □

In the following, given two sequences of random variables $\{X_n\}$ and $\{Y_n\}$, we shall use $X_n \xrightarrow{p} Y_n$ to denote $X_n - Y_n \xrightarrow{p} 0$.

Lemma 2. For any random vectors X_n, Y_n, Y, if $X_n \xrightarrow{p} Y_n$ and $Y_n \xrightarrow{p} Y$ as $n \to \infty$, then $X_n \xrightarrow{p} Y$ as $n \to \infty$.

Proof. For any $\epsilon > 0$ and $\delta > 0$, there exists an N such that for any $n > N$, we have $P(|X_n - Y_n| > \frac{\epsilon}{2}) < \frac{\delta}{2}$ and $P(|Y_n - Y| > \frac{\epsilon}{2}) < \frac{\delta}{2}$. Using the union bound, with probability at least $1 - \delta$, we have $|X_n - Y_n| \leq \frac{\epsilon}{2}$ and $|Y_n - Y| \leq \frac{\epsilon}{2}$. Hence with probability at least $1 - \delta$, $|X_n - Y| \leq \epsilon$. Thus $X_n \xrightarrow{p} Y$ as $n \to \infty$. □

Lemma 3. Let X_n is a random variable, and $Y_n = \frac{1}{n}\sum_{i=1}^{n} Y_{n,i}$ where $Y_{n,i}$ are i.i.d. random variables in $[0, M]$ for some constant M. Let the expectation of Y_n be X_n. Then $Y_n \xrightarrow{p} X_n$ as $n \to \infty$.

Proof. When $X_n = x$, we have $P(|Y_n - x| \leq \epsilon \mid X_n = x) \leq 1 - 2e^{-2n\epsilon^2/M^2}$ according to Hoeffding's inequality. Since this holds for any x, we have $P(|Y_n - X_n| \leq \epsilon) \leq 1 - 2e^{-2n\epsilon^2/M^2}$. It follows that $Y_n \xrightarrow{p} X_n$. □

Proof. (Proof of consistency) It suffices to show that at the beginning of each iteration, all the estimated messages are consistent.

Initially, none of the messages have been estimated, and it is vacuously true that all messages that have been estimated so far are consistent.

For the inductive case, it suffices to show that the message estimated at each iteration is consistent. Specifically, let $\tilde{m}^{(K)}_{t\to s}(\mathbf{x}_s)$ be the estimate of the true message $m_{t\to s}(\mathbf{x}_s)$ using K samples, we show that $\tilde{m}^{(K)}_{t\to s}(\mathbf{x}_s) \xrightarrow{p} m_{t\to s}(\mathbf{x}_s)$ as $K \to \infty$.

By the inductive assumption, $\tilde{m}^{(K)}_{u\to t}(\mathbf{x}_t)$ is consistent for each $u \in N(t) \setminus \{s\}$, where $N(t)$ is the set of neighbours for t. To simplify notation, we denote the true messages in $\{m_{u\to t}(\mathbf{x}_t) : u \in N(t) \setminus \{s\}\}$ by $m_1(\mathbf{x}_t), \ldots, m_l(\mathbf{x}_t)$ and denote their estimates by $\tilde{m}^{(K)}_1(\mathbf{x}_t), \ldots, \tilde{m}^{(K)}_l(\mathbf{x}_t)$.

[1]Australian National University, Canberra, Australia. [2]National University of Singapore, Singapore. [3]Queensland University of Technology, Brisbane, Australia. Correspondence to: Andrew Wrigley <andrew.wrigley@anu.edu.au>, Wee Sun Lee <leews@comp.nus.edu.sg>, Nan Ye <n.ye@qut.edu.au>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Let $\tilde{\Phi}_t^{(K)}(\mathbf{x}_t)$ be the estimate of the initial clique potential at the node. Each multiplication of factors to form the initial clique potential is done by sampling. Lemma 3 shows that each multiplication converges to its expected value. The expected value is in turn the product of two numbers, one of which may be a previously computed random variable. Lemma 1 shows that the product converges to the true value, and Lemma 2 chains the two process together to show that the estimate of the estimate for the initial clique potential converges.

Let $\tilde{v}_1^{(K)}(\mathbf{x}_t)$ be the estimate of $\tilde{\Phi}_t^{(K)}(\mathbf{x}_t)\tilde{m}_1^{(K)}(\mathbf{x}_t)$ obtained in the algorithm, and $\tilde{v}_j^{(K)}(\mathbf{x}_t)$ be the estimate of $\tilde{v}_{j-1}^{(K)}(\mathbf{x}_t)\tilde{m}_j^{(K)}(\mathbf{x}_t)$ for $2 \leq j \leq l$. Then we have $\tilde{m}_{t\to s}^{(K)}(\mathbf{x}_s) = \sum_{\mathbf{x}_t \setminus \mathbf{x}_s} \tilde{v}_l^{(K)}(\mathbf{x}_t)$.

The random variable $\tilde{v}_1^{(K)}(\mathbf{x}_t)$ is the average of K i.i.d. random variables with expected value $\tilde{\Phi}_t(\mathbf{x}_t)\tilde{m}_1^{(K)}(\mathbf{x}_t)$. By the construction of $\tilde{v}_1$ and the assumption that each rank-1 tensor value is in $[0, M]$, we have that each of the K random variables are in the range of $[0, M^2]$. It follows from Lemma 3 that $\tilde{v}_1^{(K)}(\mathbf{x}_t) \xrightarrow{p} \tilde{\Phi}_t(\mathbf{x}_t)\tilde{m}_1^{(K)}(\mathbf{x}_t)$ as $K \to \infty$. Since we also have $\tilde{m}_1^{(K)}(\mathbf{x}_t) \xrightarrow{p} m_1(\mathbf{x}_t)$ and $\tilde{\Phi}_t^{(K)}(\mathbf{x}_t) \xrightarrow{p} \Phi_t^{(K)}(\mathbf{x}_t)$, it follows from Lemma 1 and Lemma 2 that $\tilde{v}_1^{(K)}(\mathbf{x}_t) \xrightarrow{p} \Phi_t(\mathbf{x}_t)m_1(\mathbf{x}_t)$.

Using induction, we can similarly show that $\tilde{v}_l^{(K)}(\mathbf{x}_t) \xrightarrow{p} \Phi_t(\mathbf{x}_t)m_1(\mathbf{x}_t)\dots m_l(\mathbf{x}_t)$. Summing over $\mathbf{x}_t \setminus \mathbf{x}_s$ on both sides and applying Lemma 1 again, we have $\tilde{m}_{t\to s}^{(K)}(\mathbf{x}_s) \to m_{t\to s}(\mathbf{x}_s)$. Since this holds for any $\mathbf{x}_s$, the convergence holds for all $\mathbf{x}_s$. □

Decomposition for Ising models

In the case of Ising models, and any pairwise MRFs with Ising potentials of the form $\phi_{ij}(x_i, x_j) = \exp(w_{ij}x_ix_j)$, the 2×2 potential tables are in general rank-2 of the form

$$\begin{pmatrix} \theta & \frac{1}{\theta} \\ \frac{1}{\theta} & \theta \end{pmatrix}, \quad \theta = \exp(w_{ij}).$$

For tables of this particular form, we note there is a natural rank-2 decomposition that one can compute quickly by assuming terms in the decomposition are symmetric, by solving

$$\begin{pmatrix} \theta & \frac{1}{\theta} \\ \frac{1}{\theta} & \theta \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \otimes \begin{pmatrix} y \\ x \end{pmatrix} + \begin{pmatrix} y \\ x \end{pmatrix} \otimes \begin{pmatrix} x \\ y \end{pmatrix} \tag{1}$$

for $\theta \leq 1$, and

$$\begin{pmatrix} \theta & \frac{1}{\theta} \\ \frac{1}{\theta} & \theta \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \otimes \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} y \\ x \end{pmatrix} \otimes \begin{pmatrix} y \\ x \end{pmatrix} \tag{2}$$

for $\theta > 1$. Specifically, to solve (1) we solve

$$2xy = \theta$$
$$x^2 + y^2 = \frac{1}{\theta}$$

which yields

$$x = \frac{1}{2}\left(\sqrt{\theta + \frac{1}{\theta}} \pm \sqrt{\frac{1}{\theta} - \theta}\right),$$
$$y = \frac{1}{2}\left(\sqrt{\theta + \frac{1}{\theta}} \mp \sqrt{\frac{1}{\theta} - \theta}\right)$$

or

$$x = \frac{1}{2}\left(-\sqrt{\theta + \frac{1}{\theta}} \pm \sqrt{\frac{1}{\theta} - \theta}\right),$$
$$y = \frac{1}{2}\left(-\sqrt{\theta + \frac{1}{\theta}} \mp \sqrt{\frac{1}{\theta} - \theta}\right)$$

(2) is solved analogously. In each case, we use the first solution and weight each rank-1 term equally.

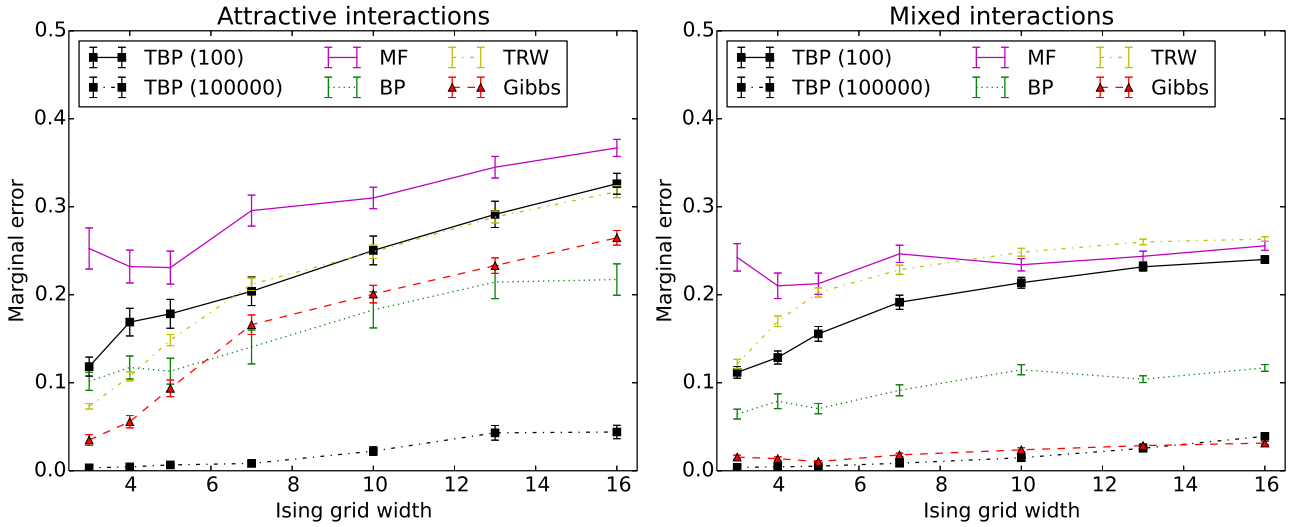
Parameters for BP, MF, TRW, Gibbs

The following parameters were used for the existing approximate inference algorithms within the libDAI package:

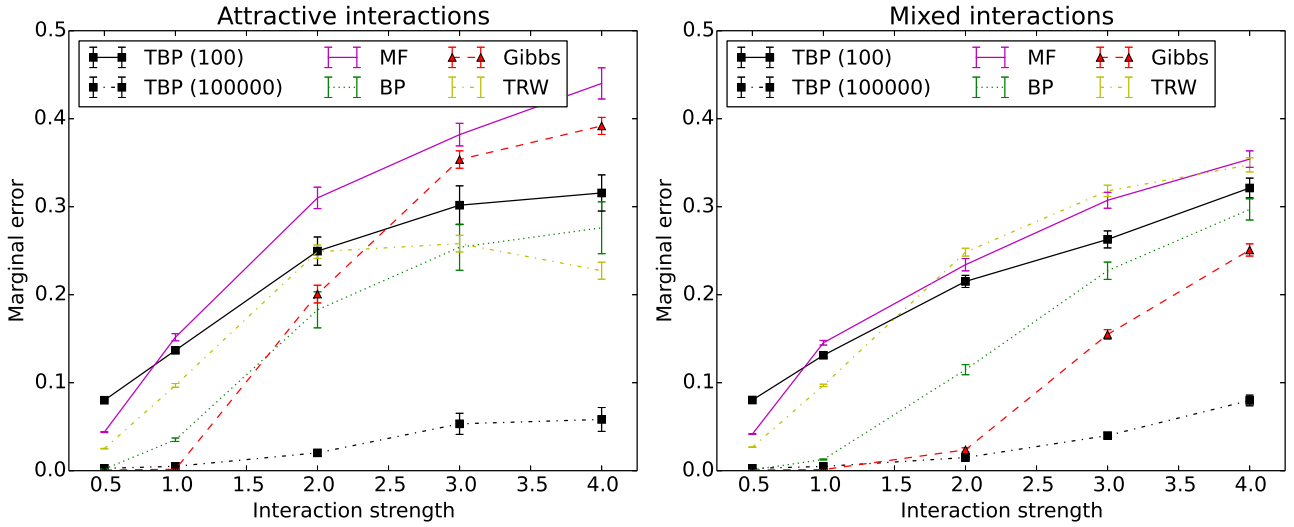
- Loopy BP: Update schedule sequential using a random sequence; maximum 10^4 iterations; tolerance for convergence 10^{-12}
- Mean-field: Maximum 10^4 iterations; tolerance for convergence 10^{-12}
- Tree-reweighted BP: Sequential updates using a random sequence; tree sample size of 10^4 used to set weights; tolerance for convergence 10^{-12}
- Gibbs: Burn-in 100 passes; restart chain with random initialisation every 1000 passes; record one sample per pass (pass = cycle once over all variables); running time limited as indicated in text.

Supplementary Results

Additional results on the Ising model with different grid sizes and different interaction strengths are shown here.



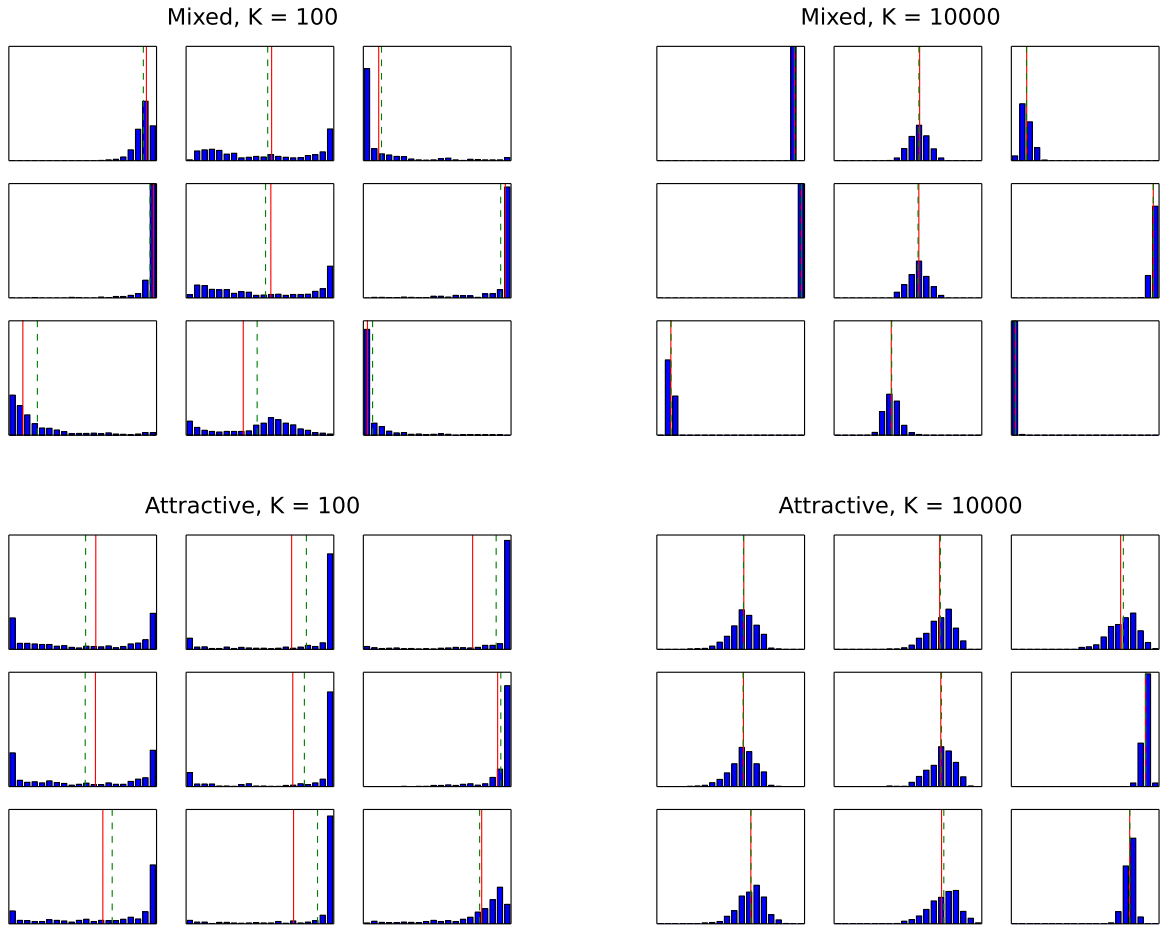
Ising models: Effect of model size N on marginal error. 100: sample size $K = 100$, 100000: sample size $K = 100000$. Gibbs running time matches the running time of TBP with $K = 100000$.



Ising models: Effect of interaction strength on performance of approximate inference algorithms. Gibbs sampling matches TBP (100000) runtime.

Distribution of estimated marginals

To give an indication of the variance of the estimates, we show histograms of the marginal estimates on the Ising models.



Estimated values of $P(X_i = 1)$ for 500 runs of tensor propagation for small versus large multiplication sample size K. Each histogram shows marginal estimates for a single node in the 10×10 Ising model grouped into 20 bins. Nodes shown are from the upper-left 3×3 corner of the grid. Solid red vertical lines indicate the true marginal and dashed green vertical lines show the mean of the 500 marginal estimates. The two mixed plots use the same Ising model instance, as do the two attractive plots.