Evidence Estimation for Bayesian Partially Observed MRFs

Yutian Chen
Department of Computer Science
University of California, Irvine
Irvine, CA 92697
yutianc@ics.uci.edu

Max Welling
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
m.welling@uva.nl

Abstract

Bayesian estimation in Markov random fields is very hard due to the intractability of the partition function. The introduction of hidden units makes the situation even worse due to the presence of potentially very many modes in the posterior distribution. For the first time we propose a comprehensive procedure to address one of the Bayesian estimation problems, approximating the evidence of partially observed MRFs based on the Laplace approximation. We also introduce a number of approximate MCMC-based methods for comparison but find that the Laplace approximation significantly outperforms these.

1 Introduction

Learning the parameters of a fully observed Markov random field is hard when the graphical representation has high treewidth. The reason is the intractability of the partition function. On the bright side, since the log-likelihood surface is concave we do not have to deal with the issue of local modes. As we add hidden units however, the likelihood surface may develop many modes and learning becomes considerably harder. Despite these difficulties a number of successful approaches have been proposed and analyzed (e.g. contrastive divergence (CD) (Hinton, 2002), persistent CD (Tieleman, 2008), MCMC-MLE (Geyer and Thompson, 1992)).

Bayesian estimation is hard even in the absence of partition functions, but also here powerful approximation methods have been developed (Neal, 1993; Carlin and Chib, 1995; Attias, 2000; Beal and Ghahramani, 2003; MacKay, 1998; Minka, 2001). The problem in the presence of a partition function is "doubly intractable" (in the language of Murray et al. (2006)). For instance, even running a Metropolis-Hastings MCMC algorithm would require the computation of the partition function for both the current parameters and the proposed parameters at every iteration. Yet, in the absence of hidden units the concavity of the likelihood extenuates the situation and indeed successful approximations have been proposed in the literature. For instance, Murray and Ghahramani (2004); Fan and Xing (2006) use Langevin dynamics with approximate gradients, Welling and Parise (2006); Parise and Welling (2006) use the Laplace approximation combined with belief propagation, Qi et al. (2005) use expectation propagation and Møller et al. (2006); Murray et al. (2006) use a nifty MCMC method for problems where perfect samples can be drawn. However, when we add hidden units the situation changes for the worse and many (possibly exponentially many) local modes may appear in the posterior distribution.

Thus, a Bayesian treatment of partially observed Markov random fields sits at the confluence of three sources of intractability: 1) Bayesian posterior estimation, 2) computation of the partition function and 3) multi-modality due to the presence of hidden units (perhaps one could say this class of problems is "triply intractable"). To the best of our knowledge there is no previous work on Bayesian estimation for partially observed MRFs.

In this paper we take the first attempt on this difficult task and propose a Laplace approximation to address the problem of evidence estimation for partially observed MRFs. The intrinsic intractabilities prevent us from providing a simple panacea, and we combine a series of techniques to achieve an accurate estimate. The Laplace approximation takes care of intractability 1. Intractability 2 is dealt with by annealed importance sampling (AIS) (Neal, 2001). Unfortunately, intractability 3 often sticks up its ugly head in the sense that it prevents us from finding the MAP state. Worse yet, modes may overlap creating plateaus breaking the Gaussian assumption in the Laplace method. Also, due to symmetries in the models many equivalent modes may

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

exist resulting in an under-estimate of the evidence. We identify these problems and introduce a number of effective corrections for them. We emphasize that, as our experiments show, each of these corrections is necessary and unavoidable in order to solve these issues.

We compare our proposed method with AIS on restricted Boltzmann machines (RBMs) for which the partition function remains tractable. We also propose a number of algorithms that adapt standard MCMC-based approaches in Bayesian estimation by replacing the required posterior samples with samples obtained from the approximate Langevin method of Murray and Ghahramani (2004). In all experiments the proposed Laplace method has outperformed all competitors by a significant margin.

2 Bayesian Model Selection

A Markov random field (MRF) model with visible variables $\mathbf{x}$ and hidden variables $\mathbf{z}$ can be represented as a log-linear model,

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[\boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}, \mathbf{z})\right] \quad (1)$$

where $\mathbf{f}(\mathbf{x}, \mathbf{z})$ is a vector of features for the state $(\mathbf{x}, \mathbf{z})$, $\boldsymbol{\lambda}$ specifies the associated parameters for each feature and $Z(\boldsymbol{\lambda})$ is the normalization constant, known as the partition function. In addition, we assume that the parameters are random variables subject to a prior distribution $p(\boldsymbol{\lambda})$.

An important quantity in Bayesian model selection is the log-marginal likelihood or evidence defined as,

$$\begin{aligned} \log p(D) &= \log \int_{\boldsymbol{\lambda}} d\boldsymbol{\lambda} p(D|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \\ &= \log \int_{\boldsymbol{\lambda}} d\boldsymbol{\lambda} \prod_n \frac{Z(\mathbf{x}_n, \boldsymbol{\lambda})}{Z(\boldsymbol{\lambda})} p(\boldsymbol{\lambda}) \end{aligned} \quad (2)$$

where $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is a set of N observations and $Z(\mathbf{x}_n, \boldsymbol{\lambda}) = \sum_{\mathbf{z}_n} \exp\left[\boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}_n, \mathbf{z}_n)\right]$. The merit of using evidence for model selection has been studied intensively (see review (Kadane and Lazar, 2004)). Although the evidence for an MRF is more difficult to compute than for a Bayesian network because the partition function in Eqn. 1 is usually intractable, a proper approximate approach still retains its advantage over other commonly used frequentist methods such as cross validation (CV).

We illustrate this point in an example of learning the number of hidden variables in RBMs. To obtain a model with ground truth, we randomly generate 50 instances of RBMs of 10 visible and 5 hidden units with a Gaussian distributed prior. For each model, i.i.d. samples are drawn and candidate models with $1 \sim 10$ hidden units are then compared. A typical scenario is shown in Figure 1. Cross validation cannot find the true model, which is concordant with the

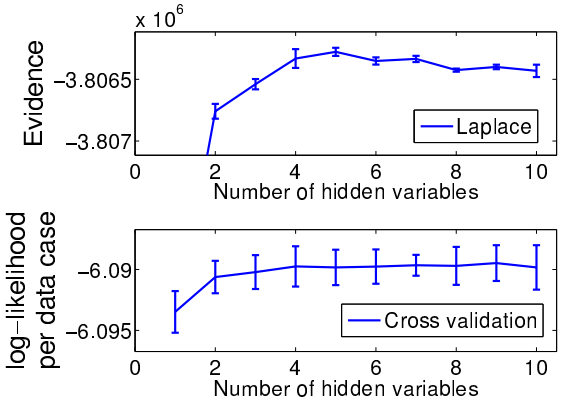


Figure 1: A typical example with 625K samples. Top: mean and standard deviation of estimated evidence by our Laplace method. Bottom: log-likelihood per data case in the validation set for CV.

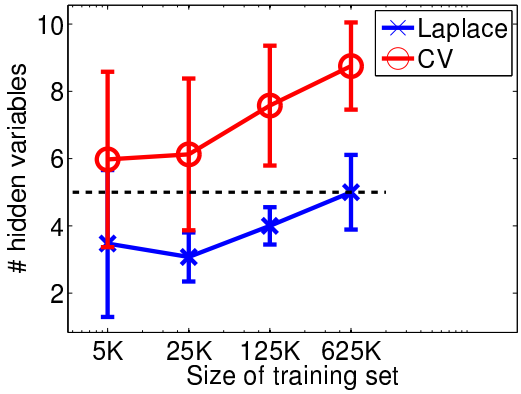


Figure 2: Mean and standard deviation of the estimated number of hidden units over 50 models against the size of the training set by the Laplace method and 10-fold cross validation.

fact that CV is not consistent in model selection (Yang, 2007). In Figure 2, our Bayesian method prefers simple structures when the training set is small and approaches the true model with more data. In contrast, the 10-fold CV tends to overestimate that number.

With this example in mind, we will be focused on how to estimate the evidence accurately using several approximate approaches in this paper.

3 Approaches to Bayesian Model Selection

3.1 Laplace Approximation

An MRF with hidden variables may have an exponentially large number of modes in the posterior distribution which makes integration over the entire parameter space NP hard. However, assuming that the maximum a posteriori (MAP) estimate $\boldsymbol{\lambda}^{\text{MP}}$ is unique (except for equivalent modes due to unidentifiability issues in the hidden variables), as the size

of D increases, the mass will concentrate around $\boldsymbol{\lambda}^{\text{MP}}$. It is therefore reasonable to approximate the marginal likelihood by integration on the single mode. The Laplace method approximates the posterior as a Gaussian distribution by Taylor expanding both the log-likelihood and the log-prior up to the second order around $\boldsymbol{\lambda}^{\text{MP}}$:

$$\log p(D|\boldsymbol{\lambda}) \approx \log p(D|\boldsymbol{\lambda}^{\text{MP}}) + \kappa^T \Delta\boldsymbol{\lambda} - \frac{1}{2}\Delta\boldsymbol{\lambda}^T (NC - \sum_{n=1}^{N} C_{\mathbf{x}_n})\Delta\boldsymbol{\lambda} \tag{3}$$

$$\log p(\boldsymbol{\lambda}) \approx \log p(\boldsymbol{\lambda}^{\text{MP}}) + \mathbf{g}^T \Delta\boldsymbol{\lambda} - \frac{1}{2}\Delta\boldsymbol{\lambda}^T \Lambda^{-1} \Delta\boldsymbol{\lambda} \tag{4}$$

where $\Delta\boldsymbol{\lambda} = \boldsymbol{\lambda} - \boldsymbol{\lambda}^{\text{MP}}$, $\kappa = N(\mathbf{E}_{\mathbf{x}\sim D}[\mathbf{E}_{p(\mathbf{z}|\mathbf{x},\boldsymbol{\lambda}^{\text{MP}})}[\mathbf{f}]] - \mathbf{E}_{p(\mathbf{x},\mathbf{z}|\boldsymbol{\lambda}^{\text{MP}})}[\mathbf{f}])$, $C = \text{Cov}(\mathbf{f})_{p(\mathbf{x},\mathbf{z}|\boldsymbol{\lambda}^{\text{MP}})}$ is the covariance of the features over the joint distribution of $\mathbf{x}$ and $\mathbf{z}$ and $C_{\mathbf{x}_n} = \text{Cov}(\mathbf{f})_{p(\mathbf{z}|\mathbf{x}_n,\boldsymbol{\lambda}^{\text{MP}})}$ is the conditional covariance matrix given $\mathbf{x}_n$. For the prior distribution, g and Λ^{-1} are the first and second derivative of $\log p(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}^{\text{MP}}$. Combining Eqn 3 and 4, we get the Laplace approximation for the log-marginal likelihood,

$$\log p(D) \approx \sum_{n=1}^{N} \log Z(\mathbf{x}_n, \boldsymbol{\lambda}^{\text{MP}}) - N \log Z(\boldsymbol{\lambda}^{\text{MP}}) + \log p(\boldsymbol{\lambda}^{\text{MP}}) - \frac{F}{2}\log(N) + \frac{F}{2}\log(2\pi) + \frac{1}{2}\log\det(N\Sigma) \tag{5}$$

where F is the number of features and $\Sigma^{-1} = NC - \sum_{n=1}^{N} C_{\mathbf{x}_n} + \Lambda^{-1}$ is the Hessian matrix of the posterior distribution at $\boldsymbol{\lambda}^{\text{MP}}$. Notice that the first order terms cancel at the MAP value.

Despite the approximation induced by the Taylor expansion, finding $\boldsymbol{\lambda}^{\text{MP}}$, estimating the first and second term of Eqn 5, and computing the covariance between features remain intractable. A second level of approximation given by the Bethe free energy has achieved good performance for a fully observed MRF (Parise and Welling, 2006; Welling and Parise, 2006), but suffers from a significant bias in estimating the MAP value $\boldsymbol{\lambda}^{\text{MP}}$ and potential non-positive definiteness in the Hessian matrix in the case of hidden units.

3.1.1 Our Proposed Laplace Procedure

Recent developments in training and sampling MRFs provide alternatives for improved accuracy. We combine these techniques and propose a comprehensive procedure outlined in algorithm 1. This procedure applies to partially observed MRFs if the following two conditions hold: 1) $\boldsymbol{\lambda}^{\text{MP}}$ is locally unique, 2) hidden units can be marginalized out efficiently given the observed variables. Condition 1 is for the Laplace approximation to hold, and condition 2 is for the convenience of computing the conditional covariance matrix $C_{\mathbf{x}}$ and $Z(\mathbf{x}, \boldsymbol{\lambda})$. The latter condition can be satisfied, e.g., when hidden units are structured as a chain or tree conditioned on $\mathbf{x}$.

Algorithm 1 Laplace Method for Partially Observed MRFs

1: Run persistent contrastive divergence to find a MAP estimate $\boldsymbol{\lambda}^{\text{MP}}$.
2: Run MCMC on $p(\mathbf{x}|\boldsymbol{\lambda}^{\text{MP}})$ to estimate the covariance matrix C.
3: Fine-tune $\boldsymbol{\lambda}^{\text{MP}}$ for a positive definite Hessian Σ^{-1} (section 3.1.2).
4: Compute $\log\det(\Sigma^{-1})$ and correct overlapping modes (section 3.1.3).
5: Run annealed importance sampling to estimate the partition function $Z(\boldsymbol{\lambda}^{\text{MP}})$.
6: Plug $\log\det(\Sigma)$ and $Z(\boldsymbol{\lambda}^{\text{MP}})$ into Eqn. 5 and count equivalent modes (section 3.1.4).

We train an MRF with persistent contrastive divergence (PCD) (Tieleman, 2008) (step 1), and estimate C with MCMC (step 2) in order to eliminate the bias in the Bethe free energy approximation in Parise and Welling (2006); Welling and Parise (2006). Annealed importance sampling (AIS) (Neal, 2001) in step 5 is able to estimate the partition function accurately at the cost of a slow annealing schedule. Fortunately, we only need to run AIS once at $Z(\boldsymbol{\lambda}^{\text{MP}})$ and it is therefore a feasible solution. The problem of multimodality in the posterior distribution are further addressed in step 3, 4 and 6.

Although the Laplace method has a time complexity of $\mathcal{O}(F^3)$ and space complexity of $\mathcal{O}(F^2)$ in computing the log-determinant of the Hessian matrix, Σ^{-1}, it is quite amenable for a model with thousands of parameters. Therefore, our proposed algorithm will work generally on any moderate-sized machine learning and statistical problems. In fact, the actual time spent in decomposing Σ^{-1} is negligible compared to searching for the MAP estimate $\boldsymbol{\lambda}^{\text{MP}}$ in our experiments with over 1000 parameters.

Another fact worth noticing is that all the additional approximation and corrections introduced in this section are aimed to improve the performance of the Laplace approximation under the *finite* data setting. Our algorithm approaches the pure Laplace approximation when the training data size as well as the sample size of the Monte Carlo method in step 2 approaches infinity. Furthermore, the Laplace approximation provides a consistent estimate to the marginal likelihood given we find the exact global MAP $\boldsymbol{\lambda}^{\text{MP}}$. While to satisfy the last condition is still intractable, it is indeed due to the multi-modal nature of our problem, and like any other approach we could at the best propose an approximate solution.

We now discuss the details of some proposed adaptations in Algorithm 1 in the following subsections.

3.1.2 Fine-tuning

Although at a local maximum of the posterior distribution $p(\boldsymbol{\lambda}|D)$, the Hessian matrix, Σ^{-1}, must (theoretically) be positive definite, two sources of noise in step $1 \sim 2$ may break the positive definiteness of the estimated matrix: (1) $\boldsymbol{\lambda}^{\text{MP}}$ is estimated by a stochastic optimization algorithm, (2) C is estimated through Monte Carlo methods. To address this issue, we further fine-tune the result by optimizing an approximate objective function based on the samples $\{x^{(t)}\}_{t=1}^{T}$ drawn in step 2.

The logarithm of the unnormalized posterior distribution can be written as

$$\begin{aligned}
&\log p(D|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \\
&= \sum_{n=1}^{N} \log Z(\mathbf{x}_n, \boldsymbol{\lambda}) - N \log Z(\boldsymbol{\lambda}^{\text{MP}}) \\
&- N \log \mathbf{E}_{p(\mathbf{x}|\boldsymbol{\lambda}^{\text{MP}})} \frac{Z(\mathbf{x}, \boldsymbol{\lambda})}{Z(\mathbf{x}, \boldsymbol{\lambda}^{\text{MP}})} + \log p(\boldsymbol{\lambda}) \\
&\approx \sum_{n=1}^{N} \log Z(\mathbf{x}_n, \boldsymbol{\lambda}) - N \log \frac{1}{T} \sum_{t=1}^{T} \frac{Z(\mathbf{x}^{(t)}, \boldsymbol{\lambda})}{Z(\mathbf{x}^{(t)}, \boldsymbol{\lambda}^{\text{MP}})} \\
&+ \log p(\boldsymbol{\lambda}) + \text{const.} \triangleq -\ell(\boldsymbol{\lambda}) + \text{const.} \qquad (6)
\end{aligned}$$

$\ell(\boldsymbol{\lambda})$ is in fact an importance sampling estimator to $-\log p(D|\boldsymbol{\lambda})$, up to a constant, with a proposal distribution $p(\mathbf{x}|\boldsymbol{\lambda}^{\text{MP}})$ and weights $w^{(t)} = \frac{\tilde{w}^{(t)}}{\sum_{t=1}^{T} \tilde{w}^{(t)}}, \tilde{w}^{(t)} = \frac{Z(\mathbf{x}^{(t)}, \boldsymbol{\lambda})}{Z(\mathbf{x}^{(t)}, \boldsymbol{\lambda}^{\text{MP}})}$, which resembles the MCMC-MLE algorithm (Geyer and Thompson, 1992). $\ell(\boldsymbol{\lambda})$ does not involve the partition function and can be optimized by any traditional second order algorithm. When the optimization converges, the Hessian matrix

$$\frac{\partial^2 \ell(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2} = N \sum_{t=1}^{T} w^{(t)} C_{\mathbf{x}^{(t)}}(\boldsymbol{\lambda}) - \sum_{n=1}^{N} C_{\mathbf{x}_n}(\boldsymbol{\lambda}) + \Lambda^{-1} \qquad (7)$$

is guaranteed to be a positive definite importance estimator to Σ^{-1}. Unlike MCMC-MLE which can suffer from degeneration of the importance weights if the posterior distribution varies significantly, we start the optimization from a point very close to a local optimum and thus weight degeneration is rarely observed in our experiments.

3.1.3 Overlapping Modes

While the fine-tuning step solves most nonpositive definite problems, there are still some cases where the optimization does not converge or $\log\det(\Sigma^{-1})$ is sensitive to the value of $\boldsymbol{\lambda}^{\text{MP}}$. Further study shows that small or negative eigenvalues of Σ^{-1} often occur with overlapping modes along the corresponding eigen-directions as shown in Figure 3. In that case, the integration under the Laplace approximation (Gaussian function) either tends to overestimate the variance or completely fails.

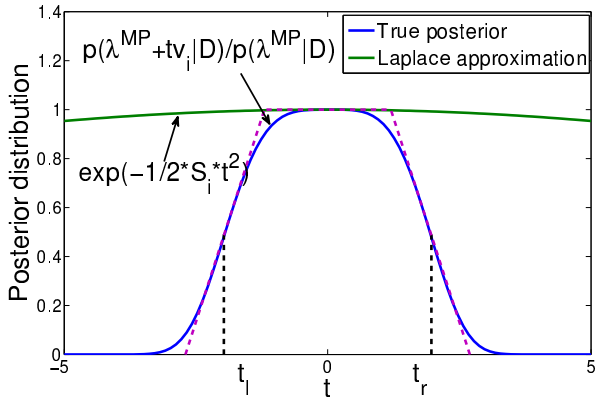


Figure 3: Posterior distribution vs Laplace approximation at $\boldsymbol{\lambda}^{\text{MP}}$ along eigenvector $\mathbf{v}_i$ with a small eigenvalue S_i. When multiple modes overlap, the curve could be close to flat or even convex and the integral under the approximation curve (green) could be much larger than that under the true posterior curve (blue).

To address this problem, we decompose the integration of the Gaussian function $\mathcal{N}(\boldsymbol{\lambda}; \boldsymbol{\lambda}^{\text{MP}}, \Sigma)$ as a product of integrals over all eigenvectors $\{\mathbf{v}_i\}_{i=1}^{F}$. For a pathological direction whose eigenvalue S_i is below a threshold S_{thresh}, we approximate $p(\boldsymbol{\lambda}|D)$ by a trapezoid instead (magenta curve in Figure 3). The integration along $\mathbf{v}_i$ is equal to a value $I = t_r - t_l$ with t_r and t_l the two bimedians. We use the estimator $\ell(\boldsymbol{\lambda})$ (Eqn. 6) to search for the points t_l and t_r, where the posterior density drops by a half along direction $\mathbf{v}_i$. We start at $t = 0$ and iteratively search in direction $\mathbf{v}_i$ for t_r ($-\mathbf{v}_i$ for t_l) using a doubling step size. When either $e^{-\ell(\boldsymbol{\lambda}^{\text{MP}} + t\mathbf{v}_i)} < 0.5$ or the effective sample size, $1/\sum_t (w^{(t)})^2$, drops below $0.9T$ we stop. The restriction of the effective sample size is to prevent degeneration of the importance sampling estimator. Finally, we perform a binary search in the interval $(0, t)$ ($(t, 0)$ for t_l) to find the location where either $e^{-\ell(\boldsymbol{\lambda}^{\text{MP}} + t\mathbf{v}_i)} = 0.5$ or $1/\sum_t (w^{(t)})^2 = 0.9T$. The trapezoid approximation increases accuracy and reduces variance of the estimated $\log\det(\Sigma^{-1})$. Even if we would have been better off using a Gaussian instead of a trapezoid we would only have introduced an error of no more than $|\log(2\sqrt{2\log(2)}/\sqrt{2\pi})| \approx 0.0625$.

3.1.4 Counting Equivalent Modes

It is well known that in models with hidden variables certain symmetries exist. Take restricted Boltzmann machines for example. The probability distribution is:

$$p(\mathbf{x}, \mathbf{z}|W, \alpha, \beta) = \frac{1}{Z} \exp\left(\sum_{i=1}^{K} z_i \left(W_i^T \mathbf{x} + \beta_i \right) + \alpha^T \mathbf{x} \right) \qquad (8)$$

where W is the interaction weights and α, β are biases. Exchanging the indices of two hidden variables will not change the probability of any observed state but will lead to two distinct modes in the posterior distribution. To com-

pensate for this effect we would simply need to add these $K!$ symmetries in the evidence. Unfortunately, life is not so easy because sometimes the modes corresponding to these symmetries are located so close together that the proposed Laplace approximation including the correction described in the previous section has already counted them in the total volume. Specifically, if there are R groups of hidden variables with each K_r hidden units which have very similar parameter values then exchanging the hidden units within these groups does *not* lead to different modes. Hence the total number of modes will be $\frac{K!}{K_1! \cdots K_R!}$.

To find the overlapping modes we leverage the Hessian matrix Σ^{-1}. Call two hidden units z_i, z_j equivalent and place them in the same group if $\exp\left(-\frac{1}{2}\frac{(\boldsymbol{\lambda}^{\text{MP}}-\boldsymbol{\lambda}_{i,j})^T}{2}\Sigma^{-1}\frac{\boldsymbol{\lambda}^{\text{MP}}-\boldsymbol{\lambda}_{i,j}}{2}\right) > \frac{1}{2}$ where $\boldsymbol{\lambda}_{i,j}$ is the parameter vector obtained by swapping all parameters associated with z_i and z_j.

3.1.5 Application to ML-BIC

The same techniques introduced for the Laplace approximation can also be applied to evaluate the Bayesian information criterion (BIC) at the maximum likelihood estimate (MLE). We compute $\text{Score}_{\text{ML}-\text{BIC}}$ by only retaining the first four terms in the log-marginal likelihood (5) with $\boldsymbol{\lambda}^{\text{MP}}$ replaced by the MLE $\boldsymbol{\lambda}^{\text{ML}}$.

3.2 MCMC Based Algorithms

MCMC approaches have been widely adopted for Bayesian model selection (Neal, 1993; Carlin and Chib, 1995). It is usually assumed that 1) one is able to run an MCMC sampling method to draw samples from the posterior distribution $p(\boldsymbol{\lambda}|D)$ and 2) it is tractable to compute the likelihood functions $p(D|\boldsymbol{\lambda})$. However, this is not the case for MRFs due to the intractability of $Z(\boldsymbol{\lambda})$. For assumption 1, several approximate MCMC methods have been introduced for Bayesian MRFs among which "brief" Langevin dynamics with gradients computed via contrastive divergence (Murray and Ghahramani, 2004; Fan and Xing, 2006). For assumption 2, since the partition function has to be evaluated for every sample drawn in a MCMC based method, AIS becomes impractical. Instead, we use belief propagation (BP) to approximate the likelihood function. Given samples from the posterior distribution, and a tool to compute the likelihood, we can now use any traditional sampling based algorithm for estimating the marginal likelihood of a Bayesian model.

In this paper, we propose and compare a number of algorithms that integrate brief Langevin sampling, belief propagation with one of the following methods: Harmonic mean ($\hat{p}_1$ in Newton and Raftery (1994), denoted as Harmonic-1), a modified version of Harmonic mean ($\hat{p}_4$ in Newton and Raftery (1994), Harmonic-4) to address stability issues, bridge sampling with Geometric mean (Meng and Wong, 1996) (Bridge-Geo), bridge sampling with optimal α (Meng and Wong, 1996) (Bridge-Opt), and deviance information criterion (DIC). The value of DIC can be informative in comparing different models but it is not a good approximation to the marginal likelihood.

3.3 Annealed Importance Sampling for Evidence

AIS is often treated as the gold standard for estimating the normalization constant of a distribution. In the posterior distribution $p(\boldsymbol{\lambda}|D) = \frac{1}{p(D)}p(D|\boldsymbol{\lambda})p(\boldsymbol{\lambda})$, the normalization term $p(D)$ can be estimated directly with AIS by running annealed Hybrid Monte Carlo (HMC) chains in parameter space, denoted as AIS-$p(D)$. This is another usage of AIS in this paper besides computing the partition function $Z(\boldsymbol{\lambda})$. However, this approach is impractical in general because evaluating $p(D|\boldsymbol{\lambda})$ in Eqn. 1 involves the intractable term, $Z(\boldsymbol{\lambda})$, and we cannot afford running another AIS to estimate $Z(\boldsymbol{\lambda})$ at every step of AIS-$p(D)$. Therefore, although our proposed Laplace approximation can handle large models, for the sake of evaluation we will study models at such a size that $Z(\boldsymbol{\lambda})$ can be computed exactly. This allows us to treat AIS-$p(D)$ as ground truth.

4 Experiments

We compare different approaches to estimating the evidence of a special class of partially observed MRFs, restricted Boltzmann machines, on the 20 Newsgroups dataset [1] with a vocabulary of 100 words. Hidden variables in an RBM are conditionally independent given the visible variables and thus it is tractable to compute $Z(\mathbf{x}, \boldsymbol{\lambda})$. We choose the Gaussian distributed prior with $\sigma_0 = 1$. Algorithms being compared include our Laplace method with (Laplace) and without (Laplace0) additional correction on overlapping and equivalent modes, penalized log-likelihood at $\boldsymbol{\lambda}^{\text{MP}}$ (MAP) which retains the first three terms in Eqn. 5, ML-BIC, Harmonic-1, Harmonic-4, Bridge-Geo, Bridge-Opt and DIC. Each method is run 100 times under every experiment setting. However, error bars are too small to be visible in the figures below. We first compute the evidence with a subset of the vocabulary, and then on the complete dataset. We also run 10-fold cross validation where we train models with persistent contrastive divergence and approximate the partition function in computing the test log-likelihood with annealed importance sampling, both of which have the same parameter setting as the Laplace method.

We only consider RBMs with up to 9 hidden variables so that the partition function can be computed exactly in AIS-$p(D)$. We run one iteration of Hybrid Monte Carlo algorithm with 10 leapfrog steps at each temperature in a linear

[1] http://www.cs.nyu.edu/~roweis/data.html

annealing scheme, and the step size is chosen to keep the average acceptance rate around 90%. It takes up to 4.5 hours to run a single chain on the complete dataset, and we run 100 chains for an accurate estimate of the ground truth.

For the Laplace and ML-BIC methods, we trained RBMs with persistent CD with a $1/t$ annealing schedule for the step size. State samples, $\{x^{(t)}\}_{t=1}^{T}$, are drawn by Gibbs sampling. Running AIS with 100 chains to estimate the partition function takes less than half an hour on the large dataset. The fine-tuning is implemented by L-BFGS with an extra stopping criterion of 90% on the effective sample size to prevent importance weights degeneration (see section 3.1.3 for its definition). $S_{\text{thresh}} = 1/\sigma_0^2$ is used to detect overlapping modes.

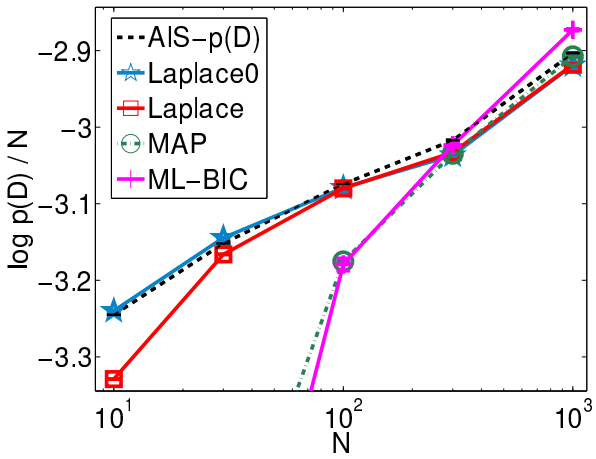
The performance of Langevin based methods strongly depends on the mixing rate of the sampling process. We use one step of Gibbs sampling at every iteration to minimize variance as suggested in Fan and Xing (2006) and choose the step size carefully so that a set of 10 Markov chains converge within about 2 hours. We assess the convergence by monitoring the "multivariate potential scale reduction factor" (MPSRF) with a threshold of 1.1. Running each chain takes up to 3 hours including drawing samples and computing the approximate likelihood with BP. For bridge sampling methods, we choose the prior distribution as $q_2(\boldsymbol{\lambda})$ in Meng and Wong (1996).

4.1 Newsgroups Dataset with 5 Words

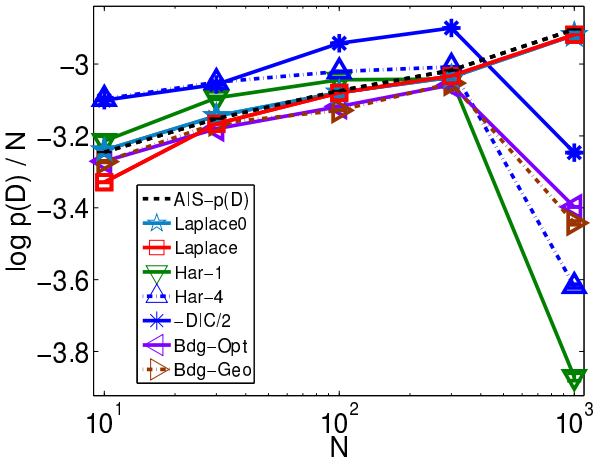
In the small dataset, we only use the 5 words with the highest frequencies in the 20 Newsgroups dataset. Subsets of training documents are randomly selected with the constraint that every word occurs in at least one document.

Figure 4 shows the evidence per document of an RBM with 3 hidden units as a function of the data size. We compare Laplace methods with MAP and ML-BIC in the upper panel which are all based on the evaluation around a single point of $\boldsymbol{\lambda}$, and with MCMC based methods in the bottom panel. Better methods should stay close to the black-dashed AIS-$p(D)$ curve.

Both Laplace0 and Laplace provide very accurate estimation of $p(D)$ for different sizes of the training set. ML-BIC and MAP over penalize due to the absence of the second order term. MCMC methods work well on small training sets, although they are still inferior to the Laplace methods. The sharp drop in scores of MCMC methods on $N = 1000$ results from a change in the step size of Langevin dynamics. As the Markov chain has difficulty converging on large datasets, we have to use a larger step size to achieve sufficient mixing within 2 hours which leads to large errors. We anticipate that Langevin based methods could also work reasonably well if we allowing a smaller step size and a longer running time. However, part of the error could also



(a) Comparison Laplace with ML-BIC, MAP



(b) Comparison Laplace with MCMC based methods

Figure 4: Log-marginal likelihood of a model with 5 visible variables and 3 hidden variables. The size of the training set N varies from 10 to 1000. Closer to AIS-$p(D)$ (black dashed) line means a better approximation. AIS-$p(D)$ almost completely overlaps with Laplace0

come from the bias introduced by the brief CD sampling as well as the Bethe free energy approximation.

Next, we study the posterior distribution approximated by brief Langevin sampling and the Laplace method, in comparison with samples drawn from a Hybrid Monte Carlo sampler where the partition function is computed exactly. Both sampling methods are initialized at $\boldsymbol{\lambda}^{\text{MP}}$, and use very small step sizes so that they do not escape from their modes. Typical histograms of parameter samples are shown in Figure 5. Both the Laplace approximation and Langevin dynamics fit the posterior distribution well without significant deviation.

Figure 6 shows the log-marginal likelihood on a dataset of 100 documents and different numbers of hidden units K. The Laplace methods again provide the most accurate

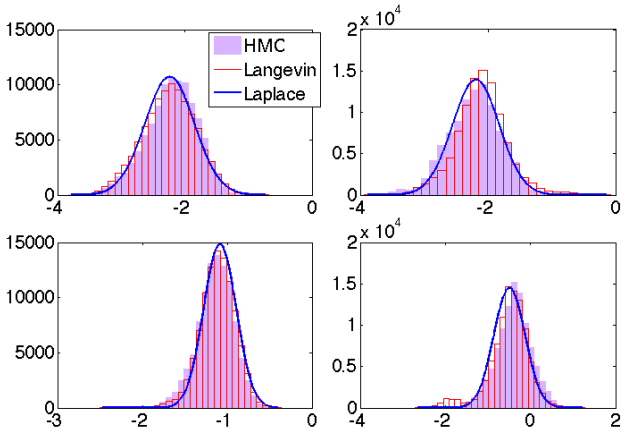


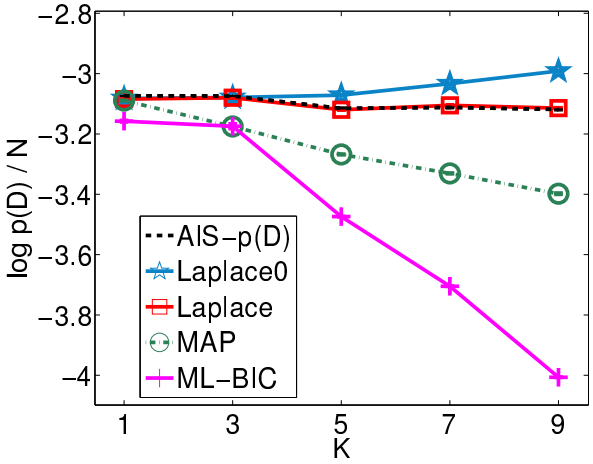
Figure 5: Histogram of samples from accurate Hybrid Monte Carlo (pink solid), Langevin with brief sampling (red) and Laplace approximation (blue)

scores. With increasing K, the chance of having overlapping modes increases rapidly and Laplace0 tends to give an over-estimate of the integral (see Figure 3) while Laplace with mode-corrections improves the robustness.

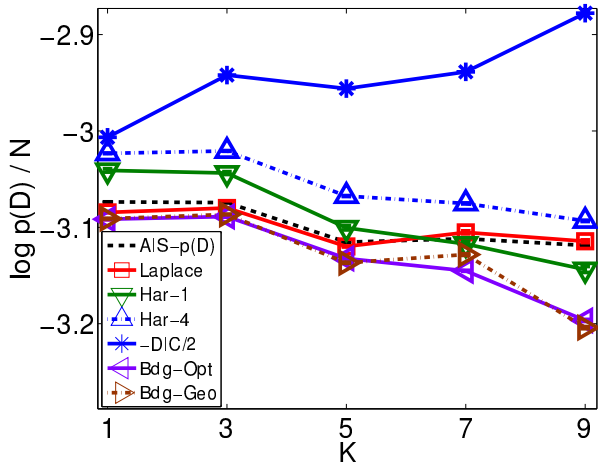
4.2 Newsgroups Dataset with the Full Vocabulary

On the full vocabulary, we show the comparison on a dataset of 100 documents in Figure 7. Evaluating the evidence on a larger set is straightforward for the Laplace method, but it becomes much more difficult for AIS-$p(D)$. For Langevin dynamics we collect 10k samples after the burn-in period with a subsampling interval of 500 iterations. For bridge sampling methods, we draw an additional 10k samples from the prior distribution (100k samples for Bdg-Opt/Geo*).

The results for all the methods are shown in Figure 7. K varies from 1 to 9. We also compute the scores with $K = 11$ and 13 for the Laplace method to illustrate its potential for larger models. MAP and ML-BIC suffer seriously from the increasing number of parameters, and those algorithms depending on only posterior samples show inferior performance to the Laplace method possibly due to the large error from the Langevin dynamics. Even worse are the bridge sampling methods which completely fail, because in the high dimensional parameter space ($\sim$ 1k parameters when $K = 9$) the likelihood of samples from the prior distribution has a variance that is too large. Increasing the number of samples drawn from the prior helps to some extent (Bdg-Opt/Geo*) and a better choice of q_2 might as well be considered. Although Laplace and Laplace0 fit the AIS-$p(D)$ curve very well, we do observe an increasing gap in the inset. This is mainly because the Laplace approximation cannot cover the increasing number of local modes. The score of Laplace is slightly higher than Laplace0 because it corrects for additional equivalent modes.



(a) Comparison Laplace with ML-BIC, MAP



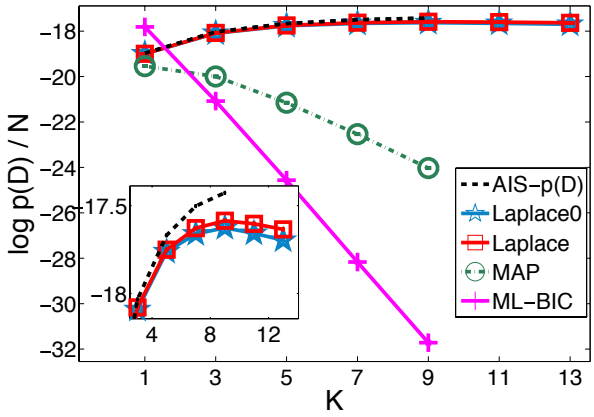
(b) Comparison Laplace with MCMC based methods

Figure 6: Log-marginal likelihood of a model with 5 visible variables and 100 training data items. The number of hidden variables K varies from 1 to 9.

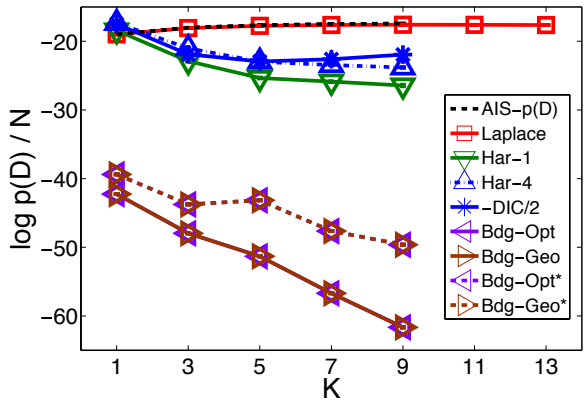
Lastly, although our main interest in this paper is to provide an accurate estimate to the marginal likelihood, we also follow the example in the introduction and show how the 10-fold cross validation performs on this experiment in Figure 8. Similar to Figure 1, while there is a slight tendency of decrease (increase) in the small (large) data set, the one standard deviation intervals overlap with each other by so much that CV provides little confidence on model selection in these experiments.

5 Conclusion

For the first time we have proposed and evaluated a method for what is perhaps the hardest class of Bayesian estimation problems: partially observed Markov random fields. The method we propose is based on the Laplace approximation, annealed importance sampling to estimate the par-

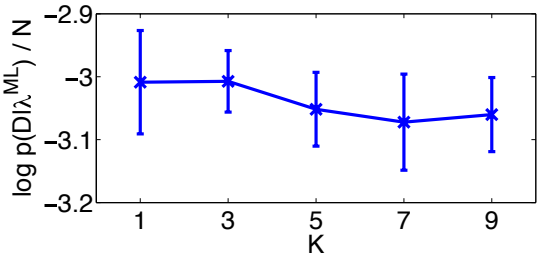


(a) Comparison Laplace with ML-BIC, MAP

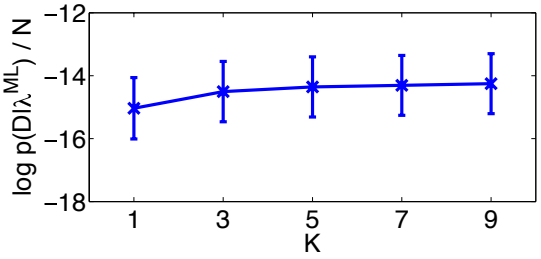


(b) Comparison Laplace with MCMC based methods

Figure 7: Log-marginal likelihood of a model on Newsgroups dataset with 100 words and 100 training documents. The number of hidden variables K varies from 1 to 9. Closer to AIS-$p(D)$ (black dashed) line means a better approximation.



(a) 5 words.



(b) Full vocabulary.

Figure 8: Average and standard deviation of the log-likelihood per data point using cross validation with 100 training data items. The number of hidden variables K varies from 1 to 9.

tition function and a series of correction terms to deal with the multi-modality of the posterior distribution. In all cases where we were able to estimate the ground truth, our method seems to work very well. In particular, the Laplace approximation seems superior to MCMC-based methods which draw approximate samples using Langevin dynamics combined with contrastive divergence.

While this represents a first comparative study we believe much work still needs to be done on more datasets of various sizes and difficulty and different MRF models. Also, while there are a lot of interesting problems in machine learning and statistics that require no more than $\mathcal{O}(1000)$ model parameters, it would be worthwhile to consider scaling up our algorithm further for large-scale problems. In that case the computational burden of decomposing and storing the precision matrix would become noticeable and we may want to exploit certain properties of the covariance/precision matrix such as sparsity (irrelevant parameters are weakly correlated) and parameter sharing/clustering (parameters for similar features tend to have similar values). We leave this for future investigation. But given the positive results obtained in this study we do believe that approximate Bayesian estimation for this class of models is feasible and can become a useful tool for model selection and model averaging for partially observed MRF models.

References

G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

T. Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the International Conference on Machine Learning*, volume 25, pages 1064–1071, 2008.

C.J. Geyer and E.A. Thompson. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54 (3):657–699, 1992.

R.M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, Computer Science, 1993.

B.P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society, Series B*, 57:473484, 1995.

H. Attias. A variational Bayesian framework for graphical models. In *NIPS*, volume 12, 2000.

M.J. Beal and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics*, pages 453–464. Oxford University Press, 2003.

D. MacKay. Choice of basis for Laplace approximation. *Machine Learning*, 33 (1):77–86, 1998.

T. Minka. Expectation propagation for approximate Bayesian inference. In *Proc. of the Conf. on Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

I. Murray, Z. Ghahramani, and D.J.C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Pittsburgh, PA, 2006.

I. Murray and Z. Ghahramani. Bayesian learning in undirected graphical models: approximate MCMC algorithms. In *Proceedings of the 14th Annual Conference on Uncertainty in AI*, pages 392–399, 2004.

G. Fan and E.P. Xing. Bayesian exponential family harmoniums. Technical report, 2006. CMU-MLD Technical Report 06-103.

M. Welling and S. Parise. Bayesian random fields: The Bethe-Laplace approximation. In *Proc. of the Conf. on Uncertainty in Artificial Intelligence*, pages 512–519, Pittsburgh, PA, 2006.

S. Parise and M. Welling. Bayesian model scoring in Markov random fields. In *Neural Information Processing Systems*, 2006.

Y. Qi, M. Szummer, and T.P. Minka. Bayesian conditional random fields. In *Artificial Intelligence and Statistics*, 2005.

J. Møller, A. Pettitt, K. Berthelsen, and R. Reeves. An efficient Markov chain Monte Carlo method for distributions with intractable normalisation constants. *Biometrica*, 93, 2006. to appear.

R.M. Neal. Annealed importance sampling. In *Statistics and Computing*, pages 125–139, 2001.

J.B. Kadane and N.A. Lazar. Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465):279–290, 2004.

Y. Yang. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6): 2450–2473, 2007.

M.A. Newton and A.E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48, 1994.

X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.