Does Better Inference mean Better Learning?

Andrew E. Gelfand, Rina Dechter & Alexander Ihler
Department of Computer Science
University of California, Irvine
{agelfand,dechter,ihler}@ics.uci.edu

Abstract

Maximum Likelihood learning of graphical models is not possible in problems where inference is intractable. In such settings it is common to use approximate inference (e.g. Loopy BP) and maximize the so-called "surrogate" likelihood objective. We examine the effect of using different approximate inference methods and, therefore, different surrogate likelihoods, on the accuracy of parameter estimation. In particular, we consider methods that utilize a control parameter to trade computation for accuracy. We demonstrate empirically that cheaper, but worse quality approximate inference methods should be used in the small data setting as they exhibit smaller variance and are more robust to model mis-specification.

1 Introduction

Graphical models offer a convenient and compact way to represent the joint distribution of many random variables. In a graphical model, each random variable is associated with a vertex and the joint distribution factorizes as a collection of potential functions defined over the cliques (complete subgraphs) of the graph. Each potential function is a positive function defined over the subset of variables within each clique. In the context of learning, these potentials are parameterized functions and given some data one seeks to find a setting of the parameters that optimize some criterion.

In this paper we focus on Markov Random Fields (MRFs), a statistical model used in many areas of computer science. An MRF is a probability distribution over a collection of discrete random variables, $\boldsymbol{y} = \{y_1, ..., y_M\}$, that can be written as:

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \exp\left(\sum\nolimits_{\alpha} \theta_\alpha(\boldsymbol{y}_\alpha) - \log Z(\boldsymbol{\theta})\right) = \exp\left(\boldsymbol{\theta}\cdot \boldsymbol{s}(\boldsymbol{y}) - \log Z(\boldsymbol{\theta})\right) \tag{1}$$

where $\boldsymbol{y}_\alpha$ is a subset of $\boldsymbol{y}$, $\theta_\alpha(\boldsymbol{y}_\alpha)$ is a real-valued potential function and $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}} \exp\left(\sum_\alpha \theta_\alpha(\boldsymbol{y}_\alpha)\right)$ is the partition function. MRFs can also be written in exponential family form, by identifying the vector of sufficient statistics as $\boldsymbol{s}(\boldsymbol{y}) = \{\delta(\boldsymbol{Y}_\alpha = \boldsymbol{y}_\alpha) \mid \forall\alpha, \boldsymbol{y}_\alpha\}$ and letting $\theta_\alpha(\boldsymbol{y}_\alpha)$ denote the component of the parameter vector $\boldsymbol{\theta}$ corresponding to the indicator $\delta(\boldsymbol{Y}_\alpha = \boldsymbol{y}_\alpha)$.

Our goal is to learn the parameters $\boldsymbol{\theta}$ given a data set of samples. We focus on finding a setting of parameters that maximize the likelihood of the data under the model. Such a setting cannot be found in practice because the likelihood and its gradients cannot be computed efficiently. The likelihood and its gradients can be approximated, however, by using an approximate inference method, such as loopy belief propagation (BP). In such cases, we can interpret the approximate inference method as exactly optimizing a 'surrogate' to the true likelihood function [1, 2].

In this paper, we examine the effect of using different approximate inference methods and, therefore, different surrogate likelihood functions, on the accuracy of parameter estimation. In particular, we consider approximate inference methods that utilize a control parameter - the *ibound* - to trade computation for accuracy [3, 4]. In such methods, a smaller ibound requires less memory and time, but typically provides a worse approximation and a worse surrogate to the true likelihood function.

The key findings of this empirical study are the following:

- Smaller *ibound* approximate inference methods should be used in the small data setting. Such methods have greater bias than larger *ibounds* methods, but exhibit smaller variance.

- Smaller *ibound* methods are more robust to model mis-specification in the small data setting.

2 Maximum Likelihood Estimation

Given a data set $\mathcal{D}_N = \{\boldsymbol{y}^{(n)}\}_{n=1}^{N}$ our goal is to find an estimate of $\boldsymbol{\theta}$. We focus on the maximum likelihood estimate (MLE), which for data set $\mathcal{D}_N$ is defined to be:

$$\boldsymbol{\theta}_N^{ML} = \arg\max_{\boldsymbol{\theta}} \ell_N(\boldsymbol{\theta}) \tag{2}$$

$$\ell_N(\boldsymbol{\theta}) = \frac{1}{N}\sum\nolimits_n \log p(\boldsymbol{y}^{(n)}; \boldsymbol{\theta}) = \bar{\boldsymbol{\mu}}_N \cdot \boldsymbol{\theta} - \log Z(\boldsymbol{\theta}) \tag{3}$$

where $\bar{\boldsymbol{\mu}}_N$ is a vector of empirical marginals with components computed from $\mathcal{D}_N$ as: $\bar{\mu}_N(\boldsymbol{Y}_\alpha = \boldsymbol{y}_\alpha) = \frac{1}{N}\sum_n \delta\left(\boldsymbol{Y}_\alpha^{(n)} = \boldsymbol{y}_\alpha\right)$ The likelihood function $\ell_N(\boldsymbol{\theta})$ is a concave function of $\boldsymbol{\theta}$. Thus, its optima can be found by standard numerical optimization methods if we can evaluate $\ell_N(\boldsymbol{\theta})$ and its gradients. Evaluating $\ell_N(\boldsymbol{\theta})$ requires computing $\log Z(\boldsymbol{\theta})$ and its derivatives are:

$$\frac{\partial \ell_N(\boldsymbol{\theta})}{\partial \theta_\alpha(\boldsymbol{y}_\alpha)} = \bar{\mu}_N(\boldsymbol{y}_\alpha) - \mu_{\boldsymbol{\theta}}(\boldsymbol{y}_\alpha) \tag{4}$$

where $\mu_\theta(\boldsymbol{y}_\alpha) = p(\boldsymbol{y}_\alpha; \boldsymbol{\theta})$ is a marginal probability in the model $p(\boldsymbol{y}; \boldsymbol{\theta})$. In general, computing $\log Z(\boldsymbol{\theta})$ and the marginals $\boldsymbol{\mu}_\theta$ are intractable. In the next section, we explore two different approximations of the likelihood that arise from approximating $\log Z(\boldsymbol{\theta})$.

2.1 Surrogate Likelihood Estimation

Variational inference methods replace the log partition function in Equation 3 with a tractable approximation. This gives rise to the following 'surrogate' likelihood function [2, 5]:

$$\tilde{\ell}_N(\boldsymbol{\theta}) = \bar{\boldsymbol{\mu}}_N \cdot \boldsymbol{\theta} - \log \tilde{Z}(\boldsymbol{\theta}) \tag{5}$$

In this paper, we consider approximations that utilize a control parameter - the *iBound* - to trade computation for accuracy. Let $\log \tilde{Z}(\boldsymbol{\theta}, i)$ denote an approximation to $\log Z(\boldsymbol{\theta})$ with an iBound of size i. We briefly review two approximate inference methods that utilize the iBound.

Weighted Mini-Bucket Elimination (WMB) [3] is an approximate inference method based on the Bucket Elimination (BE) algorithm [6]. BE is an exact inference algorithm that organizes its computations along a particular variable elimination order. A *bucket* is associated with each variable to be eliminated and each bucket is assigned a collection of (potential) functions defined over subsets of variables. Buckets are processed sequentially along the elimination order by combining the set of assigned functions and marginalizing out the bucket variable from the resultant combined function.

Mini-Bucket Elimination (MBE) [6] is an approximate method that partitions the set of functions assigned to a bucket so that at most i variables appear in any combined function. WMB extends MBE by using a weighted elimination (summation) operator. WMB provides the following upper bound:

$$\log Z(\boldsymbol{\theta}) \leq \log \tilde{Z}^{WMB}(\boldsymbol{\theta}, i)$$

Generalized Belief Propagation (GBP) [4, 7] can utilize the iBound parameter just like MBE to limit the size of the largest combined function. However, rather than providing a bound on the log partition function GBP provides an approximation resulting from the following free energy optimization:

$$\log Z(\boldsymbol{\theta}) \approx \log \tilde{Z}^{GBP}(\boldsymbol{\theta}, i) = \max_{\boldsymbol{\mu} \in \mathcal{M}_L} \boldsymbol{\mu} \cdot \boldsymbol{\theta} + \tilde{H}(\boldsymbol{\mu}, i) \tag{6}$$

where $\tilde{H}(\boldsymbol{\mu}, i)$ is an approximate entropy of the following form:

$$\tilde{H}(\boldsymbol{\mu}, i) = -\sum\nolimits_{R \in \mathcal{R}} \kappa_R \sum\nolimits_{\boldsymbol{y}_R} \mu(\boldsymbol{y}_R) \log \mu(\boldsymbol{y}_R) \tag{7}$$

and where $\mathcal{R}$ is a collection of regions on subsets of at most i variables ($|\boldsymbol{y}_R| \leq i$) and κ_R are over-counting numbers. Equation 6 optimizes the vector of pseudo-marginals, $\boldsymbol{\mu}$, subject to: $\boldsymbol{\mu} \in \mathcal{M}_L = \{\boldsymbol{\mu} \geq 0 | \sum_{\boldsymbol{y}_R \setminus y_{R'}} \mu(\boldsymbol{y}_R) = \mu(\boldsymbol{y}_{R'})\ \forall \boldsymbol{y}_{R'} \subset \boldsymbol{y}_R\ R, R' \in \mathcal{R},\ \sum_{\boldsymbol{y}_R} \mu(\boldsymbol{y}_R) = 1\ \forall \boldsymbol{y}_R\}$.

As noted in [8, 9], the GBP surrogate likelihood is a concave function of $\boldsymbol{\theta}$. However, it is non-smooth which means that utilizing numerical optimization methods (e.g. L-BFGS) to optimize it are unprincipled and cause great practical difficulties. As a result, in the experiments section we consider a form of GBP in [10] that ignores any terms in $\tilde{H}(\boldsymbol{\mu}, i)$ with negative counting numbers.

3 Surrogate Likelihood Error Decomposition

Assume our data samples $\boldsymbol{y}^{(n)}$ are drawn iid from $p(\boldsymbol{y};\boldsymbol{\theta}^\star)$, where $\boldsymbol{\theta}^\star \in \Theta$ is a member of some parametric family of distributions. Further, assume that our model is statistically identifiable so that if $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{y}_\alpha) = p(\boldsymbol{y}_\alpha;\boldsymbol{\theta}^\star)$ for all components $\alpha, \boldsymbol{y}_\alpha$ we have that $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$ [11].

Let $\boldsymbol{\theta}^{ML} = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\boldsymbol{\theta})\right]$ be the *idealistic* MLE. Note that because the model is assumed to be identifiable we have that $\boldsymbol{\theta}^{ML} = \boldsymbol{\theta}^\star$. By *idealistic*, we simply mean that this MLE is unattainable for the following three practical reasons. First, we are often interested in learning models for which the true class Θ is unknown. As a result, we restrict attention to an easy-to-specify family of models $\underline{\Theta} \subset \Theta$, such as pairwise MRFs. Let $\underline{\boldsymbol{\theta}}^\star = \arg\max_{\boldsymbol{\theta}\in\underline{\Theta}} \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\boldsymbol{\theta})\right]$ be the best parameter estimate in this restricted family of models. Second, note that $\mathbb{E}_{\boldsymbol{\theta}^\star}[\cdot]$ is the expectation over the true and unknown distribution $p(\boldsymbol{y};\boldsymbol{\theta}^\star)$. In practice, we don't have access to the marginals of $p(\boldsymbol{y};\boldsymbol{\theta}^\star)$ and instead use our samples to compute an empirical approximation to $\mathbb{E}_{\boldsymbol{\theta}^\star}$. Let $\underline{\boldsymbol{\theta}}_N = \arg\max_{\boldsymbol{\theta}\in\underline{\Theta}} \frac{1}{N}\sum_n \left[\log p(\boldsymbol{y}^{(n)};\boldsymbol{\theta})\right] = \arg\max_{\boldsymbol{\theta}\in\underline{\Theta}} \ell_N(\boldsymbol{\theta})$ be the empirical optimum in the restricted family of models. Finally, it is often infeasible to optimize the true likelihood $\ell_N(\boldsymbol{\theta})$, so we instead optimize a surrogate to it $\tilde{\ell}_N(\boldsymbol{\theta})$. Let $\tilde{\underline{\boldsymbol{\theta}}}_N = \arg\max_{\boldsymbol{\theta}\in\underline{\Theta}} \tilde{\ell}_N(\boldsymbol{\theta})$ be the empirical optimum found by our approximate inference method.

The total error in our estimate $\mathcal{E}$ can then be written as [12]:

$$\mathcal{E} = \mathbb{E}\left[E_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\boldsymbol{\theta}^\star)\right] - E_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\tilde{\underline{\boldsymbol{\theta}}}_N)\right]\right] = \mathcal{E}_{\text{Model}} + \mathcal{E}_{\text{Estimation}} + \mathcal{E}_{\text{Optimization}} \tag{8}$$

where the outer expectation is taken with respect to the random choice of data set and

- $\mathcal{E}_{\text{Model}} = \mathbb{E}\left[E_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\boldsymbol{\theta}^\star)\right] - E_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\underline{\boldsymbol{\theta}}^\star)\right]\right]$ is the model error that measures how well models in $\underline{\Theta}$ can model the optimal solution $\boldsymbol{\theta}^\star \in \Theta$.
- $\mathcal{E}_{\text{Estimation}} = \mathbb{E}\left[E_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\underline{\boldsymbol{\theta}}^\star)\right] - E_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\underline{\boldsymbol{\theta}}_N)\right]\right]$ is the estimation error that measures the error due to optimizing an empirical likelihood using N samples.
- $\mathcal{E}_{\text{Optimization}} = \mathbb{E}\left[E_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\underline{\boldsymbol{\theta}}_N)\right] - E_{\boldsymbol{\theta}^\star}\left[\log p(\boldsymbol{y};\tilde{\underline{\boldsymbol{\theta}}}_N)\right]\right]$ is the optimization error that measures the error introduced by approximate inference.

We utilize this decomposition in the experiments section to systematically study the optimization error introduced by different approximate inference methods and assess the robustness of the different inference methods to varying levels of model and estimation error.

4 Experimental Results

We conducted a variety of experiments to study how the choice of approximate inference method affects the accuracy of the learned parameters. For simplicity, we focused on pairwise MRFs on a $d \times d$ grid with the standard 4-neighbor connectivity. Each variable in the model is k-ary - i.e. $y_i \in \{1,..,K\}$. Each vertex in the grid has an associated unary potential, sampled as $\theta_i(y_i) \sim \mathcal{N}(0,\sigma_i^2)$, and each edge has a pairwise potential, sampled as $\theta_{ij}(y_i,y_j) \sim \mathcal{N}(0,\sigma_{ij}^2)$. To ensure that the model is identifiable, we force the pairwise potentials to be symmetric $\theta_{ij}(y_i,y_j) = \theta_{ij}(y_j,y_i)$ and set $\theta_i(y_i = 1) = 0$ and $\theta_{ij}(y_i = 1, y_j) = 0$ for $y_j \in \{1,..,K\}$[13]. A set of N samples is generated from each such identifiable model and from this set of samples we find a parameter estimate using different inference-based likelihood surrogates.

We begin with an experiment for which $\underline{\Theta} = \Theta$ so there is no model error. Figure 1 shows the estimation accuracy for the WMB method with iBound $\{2,4,8\}$ and for GBP with iBound $\{2,4\}$. Note that GBP with iBound 2 is a convexified loopy BP and for GBP with iBound 4 we choose the set of regions $\mathcal{R}$ to be comprised of the faces and all interior edges and vertices of the planar grid [14]. The reported errors are averaged across 8 random models generated with $\sigma_i = 0.2$ and $\sigma_{ij} = 0.4$. Notice in the left plot that the error of all approximate inference methods decreases as the sample size N increases. Interestingly, we see that higher iBound methods have greater error when N is small. For example, WMB with iBound 2 (WMB_2) lies below WMB_8 for $N < 1000$. The right-hand plot shows the optimization error for the different methods, computed by substituting the true marginals $\boldsymbol{\mu}_{\boldsymbol{\theta}^\star}$ for $\bar{\boldsymbol{\mu}}_N$ in Equation 5. Notice that the optimization error for the higher iBound methods is smaller than the lower iBound methods. This is expected behavior as higher iBound methods will, in general, produce more accurate marginals and log partition function estimates.

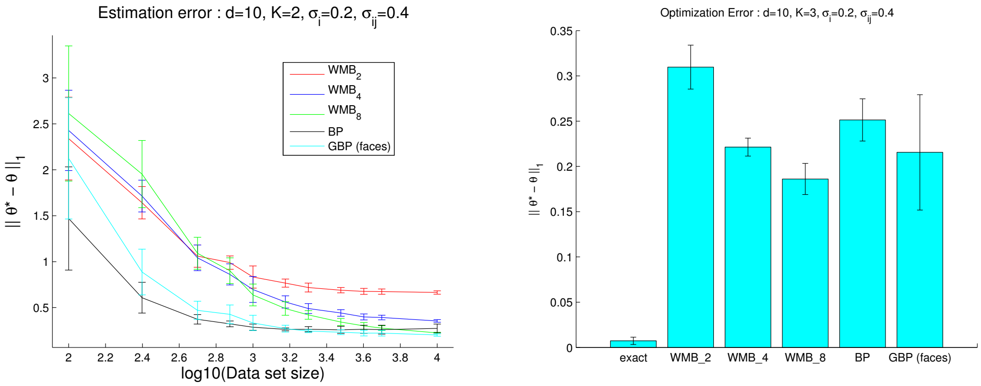

Figure 1: **Left**: Estimation Error, $|\boldsymbol{\theta}^{\star} - \boldsymbol{\theta}|_1$, for different approximate inference methods as a function of sample size N. **Right**: Optimization Error for the different inference methods.

4.1 Bias-Variance Comparison

The previous experiment suggested that lower iBound methods are on average more accurate than higher iBound methods in the small data setting. To better understand this behavior, we conducted a simple experiment to analyze the bias and variance of the different approximate inference-based estimators. In particular, we generated one random model $\boldsymbol{\theta}^{\star}$ with σ_i and σ_{ij} set as before. Then we generated 50 different data sets from $p(\boldsymbol{y}; \boldsymbol{\theta}^{\star})$ for each $N \in \{100, 250, 500, 1000, 2000, 3000, 4000, 5000\}$. Figure 2 shows the bias and variance for exact inference and WMB and GBP with iBounds of $\{2, 4\}$. Note that the estimator using WMB_2 has much smaller variance than WMB_4 and exact inference for small N. This helps to explain why the error for WMB_2 was below WMB_8 for $N < 1000$ in the previous experiment.

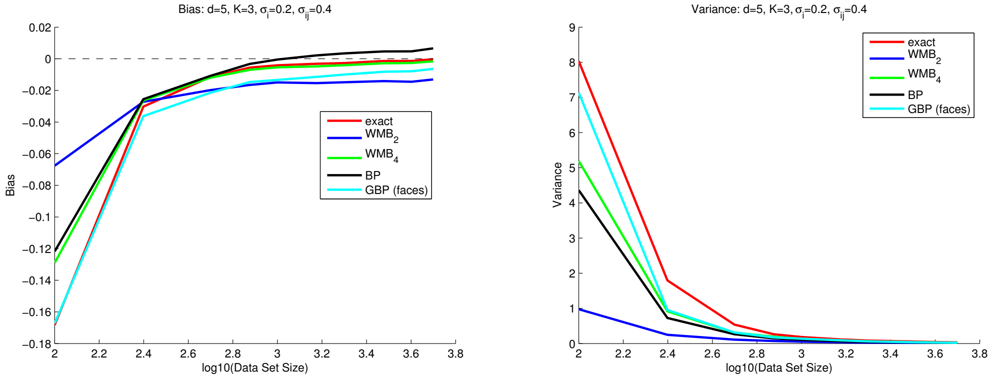

Figure 2: Bias (**Left**) and Variance (**Right**) for a 5×5 grid with $K = 3$.

4.2 Robustness to Model Mis-specification

Last, we consider an experiment in which $\underline{\Theta} \subset \Theta$. In particular, we generate our data from a pairwise grid with 8-neighbor connectivity, but learn a model with only 4-neighbor connectivity. The pairwise potentials on the *new* diagonal edges are drawn from $\mathcal{N}(0, \sigma_{ik}^2)$ and we increase σ_{ik} from 0 to 1 to increase the level of model error. Figure 3 plots the estimation error of the different inference methods as a function of mis-specification, σ_{ik}, for $N = 100$ and $N = 10000$. Notice that WMB_2 is more robust to model error than any other method when $N = 100$. However, for $N = 10000$ the situation flips and it is least robust.

References

[1] J. Domke, “Learning graphical model parameters with approximate marginal inference,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2454–2467, 2013.

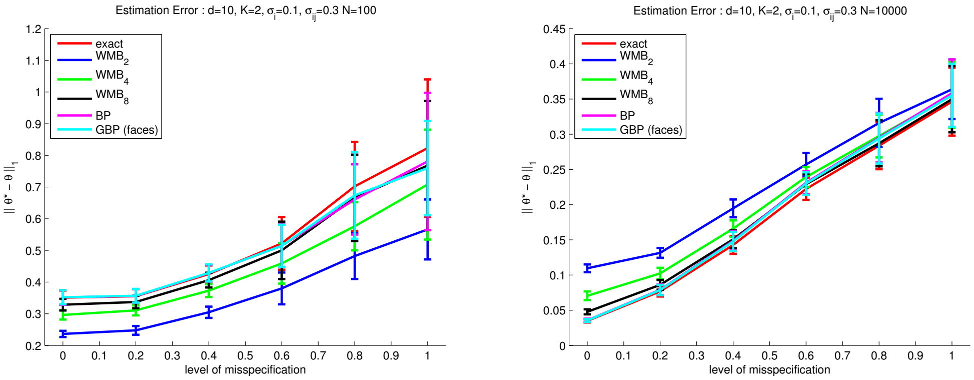

Figure 3: Estimation error for $N = 100$ (**Left**) and $N = 10000$ (**Right**) for a 10×10 binary grid.

[2] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1-2, pp. 1–305, Jan. 2008.

[3] Q. Liu and A. Ihler, "Bounding the partition function using hölder's inequality," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ser. ICML '11, L. Getoor and T. Scheffer, Eds. New York, NY, USA: ACM, June 2011, pp. 849–856.

[4] J. Yedidia, W. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," MERL, Tech. Rep., 2002, technical Report TR-2002-35.

[5] M. J. Wainwright, "Estimating the "wrong" graphical model: Benefits in the computation-limited setting," *J. Mach. Learn. Res.*, vol. 7, pp. 1829–1859, 2006.

[6] R. Dechter, "Bucket elimination: A unifying framework for reasoning," *Artificial Intelligence*, vol. 113, no. 12, pp. 41 – 85, 1999.

[7] J. Yedidia, W. Freeman, and Y. Weiss, "Generalized belief propagation," vol. 13, 2000.

[8] U. Heinemann and A. Globerson, "What cannot be learned with bethe approximations," in *UAI*, 2011.

[9] S. Nowozin, "Constructing composite likelihoods in general random fields," in *ICML Workshop on Inferning: Interactions between Inference and Learning*, 2013.

[10] T. Heskes, "Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies," *Journal of Machine Learning Research*, vol. 26, no. 153-190, 2006.

[11] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, no. 1, pp. 1–25, 2013.

[12] L. Bottou, "Stochastic gradient tricks," in *Neural Networks, Tricks of the Trade, Reloaded*, ser. Lecture Notes in Computer Science (LNCS 7700), G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Springer, 2012, pp. 430–445.

[13] X. Zhou and S. C. Schmidler, "Bayesian parameter estimation in ising and potts models: A comparative study with applications to protein modeling," Duke University, Tech. Rep., 2009.

[14] A. Gelfand and M. Welling, "Generalized belief propagation on tree robust structured region graphs," in *In 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, ser. UAI '12, 2012.