Learning with Blocks: Composite Likelihood and Contrastive Divergence

Arthur U. Asuncion[1], **Qiang Liu**[1], **Alexander T. Ihler, Padhraic Smyth**
{asuncion,qliu1,ihler,smyth}@ics.uci.edu
Department of Computer Science
University of California, Irvine

Abstract

Composite likelihood methods provide a wide spectrum of computationally efficient techniques for statistical tasks such as parameter estimation and model selection. In this paper, we present a formal connection between the optimization of composite likelihoods and the well-known contrastive divergence algorithm. In particular, we show that composite likelihoods can be stochastically optimized by performing a variant of contrastive divergence with random-scan blocked Gibbs sampling. By using higher-order composite likelihoods, our proposed learning framework makes it possible to trade off computation time for increased accuracy. Furthermore, one can choose composite likelihood blocks that match the model's dependence structure, making the optimization of higher-order composite likelihoods computationally efficient. We empirically analyze the performance of blocked contrastive divergence on various models, including visible Boltzmann machines, conditional random fields, and exponential random graph models, and we demonstrate that using higher-order blocks improves both the accuracy of parameter estimates and the rate of convergence.

1 INTRODUCTION

Learning the parameters of graphical models through maximum likelihood estimation (MLE) is hard in many cases due to the general intractability of computing the partition function and its gradient. While it is possible to leverage the independence structure of a model to make these calculations manageable, many models, such as highly-connected Markov random fields, fall outside this tractable category, motivating the need for approximate parameter estimation techniques.

A general way to perform approximate maximum likelihood estimation is to use MCMC sampling to avoid the explicit calculation of the partition function or its gradient (Geyer, 1991; Snijders, 2002). With enough samples, MCMC-MLE becomes as accurate as MLE; however, the MCMC chains may take a long time to reach equilibrium and generate a sufficient number of samples. An alternative approach is to approximate the objective function itself by using pseudolikelihood. Maximum pseudolikelihood estimation (MPLE) is computationally fast; however, the estimates it produces may be inaccurate or unreliable (Besag, 1974; Geyer, 1991).

Composite likelihoods are higher-order generalizations of pseudolikelihood that unlock a potentially wide spectrum of estimation techniques that lie between MLE and MPLE. While composite likelihood methods have garnered significant interest among statisticians (e.g. Lindsay, 1988; Varin and Vidoni, 2005; Parner, 2001; Fearnhead and Donnelly, 2002), these techniques have yet to be fully explored by the machine learning community; a few exceptions include work by Liang and Jordan (2008) and Dillon and Lebanon (2009). Asymptotic analysis shows that maximum composite likelihood estimation (MCLE) is statistically more efficient than the maximum pseudolikelihood estimation (MPLE) (Liang and Jordan, 2008); however the computational cost also increases when higher-order relationships are used. Nonetheless it is possible to trade off computation time for increased accuracy by switching to higher-order composite likelihoods.

Meanwhile, contrastive divergence (CD) is a popular machine learning algorithm that has been used to learn Markov random fields and deep belief networks (Hinton, 2000). The CD algorithm iterates repeatedly between obtaining samples from the current model, used to calculate a gradient estimate, and optimizing the model parameters given that gradient. Interestingly, CD also provides a spectrum of estimation algorithms between MLE and MPLE. It

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

[1] Both authors contributed equally.

has been shown that CD based on one step of a random-scan Gibbs sampler (which we refer to as "CD-1" in this paper[1]) for models with only visible units is equivalent to a stochastic MPLE optimization (Hyvärinen, 2006). Furthermore, CD using exact samples from the current model – obtainable after sufficiently many steps of MCMC sampling ("CD-∞") – is equivalent to a stochastic MLE optimization. CD-n, which performs n sampling steps, is an algorithmic variant which, despite having less theoretical backing, allows one to operate between MLE and MPLE.

In this paper, we introduce *blocked contrastive divergence (BCD)* as another principled way to explore the middle ground between MLE and MPLE. Specifically, we show that CD based on a random-scan blocked Gibbs sampler is stochastically equivalent to maximum composite likelihood estimation. Not only does this connection provide an efficient algorithmic engine for optimizing composite likelihoods, but it also bridges two different fields and allows for the cross-fertilization of techniques. For instance, statisticians working on composite likelihood may benefit from techniques inspired by contrastive divergence, such as fast-weights persistent CD (FPCD) (Tieleman and Hinton, 2009). Likewise, from a CD point of view, this connection may provide tools to better understand theoretical properties of CD and generate improved versions of CD.

In the next section, we review the concepts of pseudolikelihood and composite likelihood. Then we present the theoretical connection between maximum composite likelihood estimation and blocked contrastive divergence. We empirically show that our blocked contrastive divergence algorithm performs well on models such as fully visible Boltzmann machines, conditional random fields, and exponential random graph models. Finally, we review related work and conclude with directions for future work.

2 LIKELIHOOD, PSEUDOLIKELIHOOD, AND COMPOSITE LIKELIHOOD

We briefly review the concepts of likelihood, pseudolikelihood, and composite likelihood. In this paper, we consider models which can be expressed in exponential family form,

$$p(x|\theta) = \exp(\theta^T u(x))/Z(\theta), \tag{1}$$

where $u(x)$ is a vector of the sufficient statistics and $Z(\theta)$ is the partition function which normalizes the distribution.

Suppose we have N independent observations from the model, $X = \{x^1, x^2, \ldots, x^N\}$, and the goal is to find the parameter θ which generated the observed data. A standard technique is to perform maximum likelihood estimation (MLE), which maximizes the loglikelihood,

$$\mathcal{L}(\theta|X) = \sum_{i=1}^{N} \log p(x^i|\theta). \tag{2}$$

The ML estimator, $\hat{\theta}^{ml} \equiv \arg\max_\theta \mathcal{L}(\theta|X)$ has several well-known properties such as asymptotic consistency and normality (Lehmann and Casella, 1998), so that

$$\sqrt{N}(\hat{\theta}^{ml} - \theta_*) \to \mathcal{N}(0, \mathcal{I}(\theta_*)^{-1}), \tag{3}$$

where θ_* is the true parameter and $\mathcal{I}(\theta)$ is the Fisher information. The asymptotic variance $\mathcal{I}(\theta_*)^{-1}/N$ is a measure of the estimator's statistical efficiency, and it is well-known that MLE is optimal in this regard since it achieves the Cramer-Rao lower bound — no other consistent estimator has lower asymptotic variance than the ML estimator.

While MLE has nice theoretical properties, its practical application is hindered by the general intractability of computing the partition function and its gradient. To see this, we substitute Eq. (1) into Eq. (2),

$$\mathcal{L}(\theta|X) \propto \frac{1}{N} \sum_{i}^{N} \theta^T u(x^i) - \log Z(\theta), \tag{4}$$

and take its gradient, which is a standard calculation,

$$\frac{\partial \mathcal{L}(\theta|X)}{\partial \theta} = \langle u(x) \rangle_0 - \langle u(x) \rangle_\theta. \tag{5}$$

where $\langle \cdot \rangle_0$ denotes the expected value over the empirical data distribution $p_0(x) = \frac{1}{N}\sum_i \delta(x - x^i)$, and $\langle \cdot \rangle_\theta$ denotes the expected value over the model distribution $p(x|\theta)$, i.e.,

$$\langle u(x) \rangle_\theta = \sum_x u(x) \exp(\theta^T u(x))/Z(\theta).$$

To exactly calculate this gradient, we would need to sum over a number of configurations which is exponential in the total number of variables.

Maximum pseudolikelihood estimation (MPLE) (Besag, 1974) avoids this computational issue entirely by optimizing a different objective function: the pseudolikelihood,

$$\mathcal{PL}(\theta|X) = \sum_{i=1}^{N} \sum_{j=1}^{M} \log p(x_j^i | x_{\neg j}^i, \theta), \tag{6}$$

where M is the number of variables in the model and $x_{\neg j}$ denotes the set of variables excluding variable j. The optimization of $\mathcal{PL}(\theta|X)$ does not require computing a difficult partition function, as it requires summing over only each x_i individually.

Like MLE, MPLE is also asymptotically consistent (Lindsay, 1988); however, in general it has larger asymptotic variance than MLE, and the MPLE is unable to achieve

[1] In the literature, "CD-1" usually refers to CD with one Gibbs sweep through all the variables. In this paper, "CD-1" refers to one Gibbs update step on a randomly selected variable.

the Cramer-Rao bound (Lindsay, 1988). Furthermore, empirical studies suggest that MPLE is less efficient in practice and may even fail for models such as Ising models (Geyer, 1991), conditional random fields (Kumar and Hebert, 2004), and exponential random graph models (van Duijn et al., 2009). Since pseudolikelihood overestimates local dependencies, problems may arise when the dependence among variables is very strong.

The concept of composite likelihood was introduced as an extension of pseudolikelihood to fill the gap between MLE and MPLE (Lindsay, 1988). The composite likelihood is defined as

$$\mathcal{CL}(\theta|X) = \sum_{i=1}^{N}\sum_{c=1}^{C} \log p(x^i_{A_c}|x^i_{B_c}, \theta), \qquad (7)$$

where the outer sum is over N observations, the inner sum is over C different likelihood terms, and A_c and B_c are subsets of the index set $I = \{1, 2, \ldots, M\}$ such that $A_c \neq \emptyset$ and $A_c \cap B_c = \emptyset$.

The estimator obtained by optimizing the composite likelihood is called the maximum composite likelihood estimator (MCLE). The definition of composite likelihood is quite general and contains several special cases. If $A_c = I$, $B_c = \emptyset$, and $C = 1$, Eq. (7) is the full loglikelihood. When $B_c = \emptyset$, Eq. (7) is sometimes referred to as a marginal composite likelihood, and when $B_c = \neg A_c$, where $\neg A_c = I/A_c$, it is referred to as a conditional composite likelihood. In this paper, we restrict our focus to conditional composite likelihoods. We also assume that all A_c in one composite likelihood have the same size, and we let their size $n = |A_c|$ be the order of the composite likelihood. In this case, if $A_c = \{c\}$ for $c = 1, 2, \ldots, M$, Eq. (7) is the pseudolikelihood.

Like MPLE, MCLE has been shown to be asymptotically consistent (Lindsay, 1988); furthermore, MCLE has an asymptotic variance that is smaller than MPLE but larger than MLE (Liang and Jordan, 2008; Dillon and Lebanon, 2009). It is widely believed that by increasing the size of A_c (and correspondingly decreasing the size of B_c), one can capture more dependency relations in the model and increase the accuracy of the estimates. One can also select specific A_c and B_c subsets to fit the characteristics of specific models. Various forms of composite likelihood have been successfully used in applications such as spatial statistics (Varin and Vidoni, 2005), survival analysis (Parner, 2001), and genetics (Fearnhead and Donnelly, 2002).

The drawback of using a composite likelihood as opposed to the pseudolikelihood is the increased computational cost of optimizing the product of higher-order likelihood terms. The computational cost is in general exponential in the size of the largest subset A_c. However, it is possible to choose the composite likelihood terms in such a way that they match the dependence structure of the actual model. For certain models with sparse dependence structure, we show that one may choose the set A_c such that the conditional probability is computationally tractable (see Figure 4), allowing the optimization of higher-order composite likelihoods to be computationally efficient. Later in the paper we demonstrate this method on conditional random fields.

3 BLOCKED CONTRASTIVE DIVERGENCE

Contrastive divergence (CD) is a widely used machine learning algorithm (Hinton, 2000) which has been successfully used to learn a variety of models, including restricted Boltzmann machines and Markov random fields (Carreira-Perpiñán and Hinton, 2005; Hinton and Salakhutdinov, 2006). We briefly review the CD framework and introduce a variant of CD based on composite likelihood.

To introduce contrastive divergence, recall that the gradient of the likelihood is defined as

$$\frac{\partial \mathcal{L}(\theta|X)}{\partial \theta} = \langle u(x)\rangle_0 - \langle u(x)\rangle_\theta. \qquad (8)$$

As mentioned earlier, calculating the second term of the gradient is usually not tractable due to the presence of the partition function. Nonetheless, given the ability to sample from the model distribution $p(x|\theta)$, one can perform a Monte Carlo approximation of the gradient. One way to obtain a sample from the model distribution is to perform Gibbs sampling across the individual variables x_j until the Markov chain converges to the equilibrium distribution. This approach is referred to as CD-∞ (Carreira-Perpiñán and Hinton, 2005), since the Markov chain is guaranteed to converge after an infinite number of Gibbs sampling steps.

Hinton (2002) observed that it can be effective to run just a few steps of the sampler (after initializing the sampler from the observed data distribution $p_0(x) = \frac{1}{N}\sum_i \delta(x - x^i)$), in order to obtain an approximate gradient. Running one step of the sampler is known as CD-1, while running n steps is known as CD-n. The intuition is that one obtains a rough estimate of the gradient even though the sampler has not reached the equilibrium distribution. Formally, contrastive divergence uses the modified gradient,

$$\frac{\partial \mathcal{CD}(\theta|X)}{\partial \theta} = \langle u(x)\rangle_0 - \langle u(x)\rangle_n, \qquad (9)$$

where $\langle\cdot\rangle_n$ represents the average over samples drawn from the n-th step of a Markov chain governed by $p(x|\theta)$, starting from the data distribution $p_0(x)$.

It has been shown that the gradient for CD-1 is the stochastic version of the gradient used in maximum pseudolikelihood estimation (Hyvärinen, 2006). Thus, CD-1 corresponds to MPLE, CD-∞ corresponds to MLE, and CD-n is an algorithm that is between MPLE and MLE.

We propose a new approach, *blocked contrastive divergence (BCD)*, which performs blocked Gibbs sampling in the MCMC part of CD. Just as CD-1 is equivalent to MPLE, we show that BCD is equivalent to a stochastic maximum composite likelihood estimation (MCLE). To see this, consider the gradient of the composite likelihood,

$$\frac{\partial \mathcal{CL}(\theta|X)}{\partial \theta} = \langle u(x) \rangle_0 - \langle u(x) \rangle_{\mathcal{CL}}, \tag{10}$$

where

$$\langle u(x) \rangle_{\mathcal{CL}} = \frac{1}{C} \sum_c \langle \langle u(x) \rangle_{p(x_{A_c}|x_{\neg A_c}, \theta)} \rangle_0,$$

where $\langle u(x) \rangle_{p(x_{A_c}|x_{\neg A_c}, \theta)}$ denotes the expectation of $u(x)$ with respect to the conditional distribution $p(x_{A_c}|x_{\neg A_c}, \theta)$ for fixed values of $x_{\neg A_c}$, and $\langle \cdot \rangle_0$ still denotes the average over data distribution $p_0(x)$. The derivation of Eq. (10) is presented in the Appendix.

We can approximate $\langle u(x) \rangle_{\mathcal{CL}}$ with a Monte Carlo approximation. To this end, we define a *random-scan blocked Gibbs sampler* (RSBG) based on $\{A_c\}$ as a blocked Gibbs sampler which randomly selects one of the subsets A_c with equal probability $\frac{1}{C}$ and updates the elements of x_{A_c} jointly conditional on the other elements $x_{\neg A_c}$ using the full conditional probability $p(x_{A_c}|x_{\neg A_c}, \theta)$. From the definition of RSBG, one can see that $\langle u(x) \rangle_{\mathcal{CL}}$ is in fact the average of samples drawn from the first step of the RSBG sampler starting from data distribution $p_0(x)$. Thus, CD using this RSBG sampler is indeed a stochastic version of MCLE. With a sufficiently large number of samples, this Monte Carlo approximation approaches the gradient of the conditional composite likelihood with blocks $\{A_c\}$.

The tight connection between CD and composite likelihood implies that maximum composite likelihood estimation can be implemented by using a variant of CD based on the blocked Gibbs sampler. Since higher-order composite likelihoods generally have higher statistical efficiency, BCD is expected to have higher efficiency as the blocking size n increases. It is reasonable to expect contrastive divergence to behave better by using the blocked Gibbs sampler. This is confirmed by the experiments presented later in the paper.

This connection also promotes the transfer of ideas between CD and composite likelihood. One advantage of composite likelihood, mainly developed within the statistics community, is the existence of theoretical properties such as concavity and asymptotic consistency. Meanwhile, contrastive divergence has spawned many practical algorithmic improvements, such as CD-n and fast weights persistent CD (FPCD) (Tieleman and Hinton, 2009), which perform well but are lacking in theoretical justification. One can combine CD-n or PCD with our blocked sampler to improve performance, although it may no longer be clear which objective functions are being optimized. Furthermore, it is well known that composite likelihoods are concave and have unique maximal points (Lindsay, 1988). Thus, this guarantees from a theoretical viewpoint that CD based on one step of the RSBG sampler, as the stochastic version of MCLE, should also have unique solutions.

4 EXPERIMENTAL ANALYSIS

We apply blocked contrastive divergence to visible Boltzmann machines, conditional random fields, and exponential random graph models, and we show that blocking can improve both the accuracy of the estimates and the algorithm's rate of convergence. In the following sections, we use the notation "Bn-CDm" to denote the version of CD with m steps of the random-scan blocked Gibbs sampler with block size n. We are interested in the behavior of Bn-CD1, which is equivalent to nth-order composite likelihood, as well as B1-CDn (or CD-n), which is widely used in the contrastive divergence literature.

4.1 BOLTZMANN MACHINES

Boltzmann machines are widely-used models in computer science and statistical physics which are useful as testbeds for evaluating parameter estimation methods. In this set of experiments, we use a visible Boltzmann machine (VBM) consisting of 8 binary variables, with a matrix of pairwise potentials W and higher-order potentials weighted by θ:

$$p(x|W, \theta) = \frac{1}{Z} \exp(x^T W x + \theta_1 \prod_{i=1:4} x_i + \theta_2 \prod_{i=5:8} x_i + \\ \theta_3 \prod_{i=3:6} x_i + \theta_4 \prod_{i=1:8} x_i).$$

We include higher-order potentials in this model to highlight the fact that BCD can outperform standard CD in these regimes. For each algorithm, we use a gradient descent step size of 0.001. In the first experiment, we set W to be a matrix which exhibits a block structure (namely, 0.5 appearing in matrix positions [1,2],[3,4],[5,6],[7,8] and their transposed entries, and 0 elsewhere); we also arbitrarily set $\theta_1 = 0.8$, $\theta_2 = -0.5$, $\theta_3 = 4$, and $\theta_4 = 3$. Using these parameters, we sampled 2000 data cases to create the training set and 500 data cases to create a test set. Using the training set as the data, we ran our algorithms for thousands of iterations. We use randomly selected subsets of variables of size n as blocks in Bn-CD1.

Figure 1 shows the performance of BCD with various block sizes relative to B1-CDn when plotted over both iterations and time. In each iteration, Bn-CD1 calculates conditional probabilities over 2^n configurations while B1-CDn only calculates $2n$ probabilities; thus we are able to obtain an accurate estimate of the actual time it would take to run these algorithms. Performance is measured by the average loglikelihood of the test cases given the current model parameters. Since the model is small, we can calculate these loglikelihoods exactly. We see that BCD significantly outperforms standard CD, which is not surprising because the high-order potentials define strong dependency structures in the model, and the n steps of regular Gibbs sampling in B1-CDn would converge more slowly than one step of n-order blocked Gibbs sampling in Bn-CD1.

In Figure 2(a), we perform the same experiment 100 times but with a random set of parameters. The strength for W (which uses the same block structure as before) is chosen uniformly from [0,1] and each θ is chosen uniformly from [-2,2]. We restrict our focus to comparing B3-CD1 and B1-CD4, whose time complexities are the same per iteration. We show the test-likelihood achieved by B3-CD1 and B1-CD4 after 10,000 iterations, and we see that B3-CD1 performs as well or better than B1-CD4 over 100 different models. We also perform an experiment in Figure 2(b) where we let the strength of both W and θ vary from 0 to 1 (each point is a different experiment). We see that blocked contrastive divergence stays stable and obtains accurate parameter estimates even when the strength of the model potentials increases significantly. In contrast, standard CD starts to become unstable and inaccurate for models with higher potential strengths. This plot suggests that blocked contrastive divergence is especially useful when the variables in the model are strongly connected. Overall, these results indicate that BCD rarely performs worse than CD and in many cases performs significantly better.

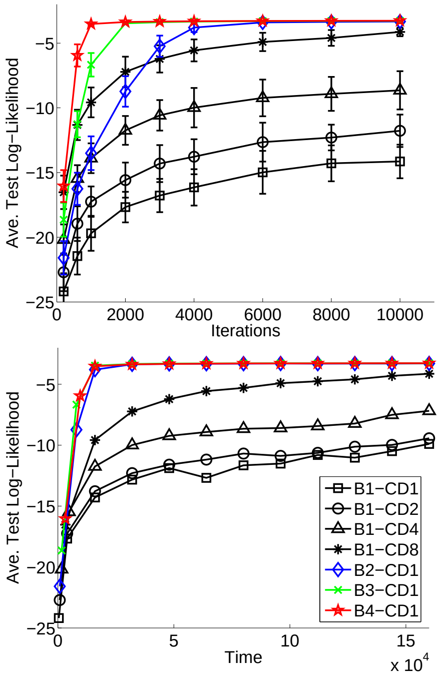


Figure 1: Performance of BCD when compared to standard CD-n on a visible Boltzmann machine. (a) TOP: Average test loglikelihood as a function of the number of CD iterations. (b) BOTTOM: Average test loglikelihood as a function of rescaled time (error bars suppressed for clarity). The same legend is shared between the two plots.

4.2 CONDITIONAL RANDOM FIELDS

We apply our methods on a conditional random field (CRF) for image segmentation (Kumar and Hebert, 2004; Vishwanathan et al., 2006). Let x be a binary image, where $x_j = \pm 1$ is the label of the j-th pixel. Let y be a noisy observation of x. In this CRF, the posterior distribution $p(x|y)$ is directly specified as an Ising model,

$$p(x|y,\theta) = \frac{1}{Z}\exp(\sum_j w^T h_j(y)x_j + \sum_{i\sim j} v^T h_{ij}(y)x_i x_j), \tag{11}$$

where $i \sim j$ means pixels i and j are adjacent in the image. Meanwhile, $h_j(y)$ and $h_{ij}(y)$ are the node features and edge features. In CRFs, we can incorporate prior knowledge by defining specific features using that knowledge. The features $h_j(y)$ and $h_{ij}(y)$ can be complicated functions which depend on the whole vector of y, and not just local values like y_i and y_j. In this setting, the features are defined as $h_j(y) = [1, y_j]$ and $h_{ij}(y) = [1, c_{ij}]$, where c_{ij} is calculated by running a Canny edge detector on the noisy image y. We take $c_{ij} = 1$ if there is no edge detected both on pixel i and j, and $c_{ij} = 0$ otherwise.

We analyze the performance of Bn-CD1 on this CRF. To decrease the blocked sampling computation time of Bn-CD1, which is $O(d^n)$ in worst case, where d is the domain size of x_i (for the binary image example, $d = 2$), we select blocks whose corresponding graphs have tree structures when performing RSBG – see Figure 4 for the possible sub-tree structures. We perform forward-backward sampling on those subtree structures to draw samples from the full conditional distribution $p(x_{A_c}|x_{\neg A_c})$. This procedure only requires a time complexity of $O(nd^2)$ rather than $O(d^n)$. Thus, we can directly compare Bn-CD1 with B1-CDn, which also has time complexity linear in n.

In our experiment, we use a chess board image with 32×32 pixels, and we generate 15 noisy images by adding Gaussian noise $\mathcal{N}(0,1)$ to the original image. We use 10 of the noisy images (a total of 10,240 pixels) for training and 5 images (a total of 5,120 pixels) for testing. The parameters $[w, v]$ are estimated by using Bn-CD1 and B1-CDn. The step size of gradient descent is 0.001 and the maximum number of iterations is 1000. Since exact inference is intractable, we use loopy belief propagation to estimate the MAP image on test data, and we calculate the average error rate of the MAP image with respect to the original image. Figure 3 suggests that the error rate decreases as the blocking size increases, and Bn-CD1 outperforms B1-CDn for the same value of n. Note that since the time complexities of Bn-CD1 and B1-CDn are approximately the same due to the use of tree-structured blocks, we obtain significant computational savings by using Bn-CD1 in this case.

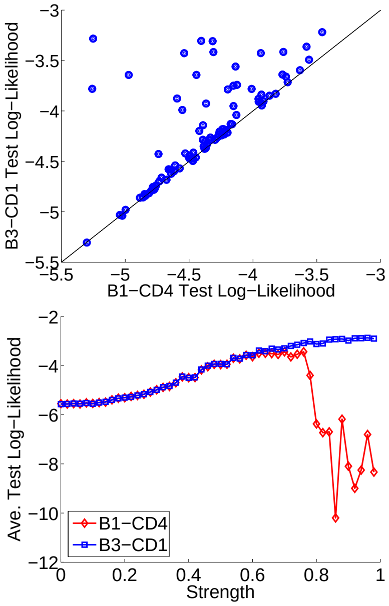


Figure 2: A comparison between B1-CD4 and B3-CD1 on VBMs. (a) TOP: Each dot is a separate model with random parameters. (b) BOTTOM: The performance of the algorithms, as the strength of the model's potentials increases.

4.3 EXPONENTIAL RANDOM GRAPH MODELS

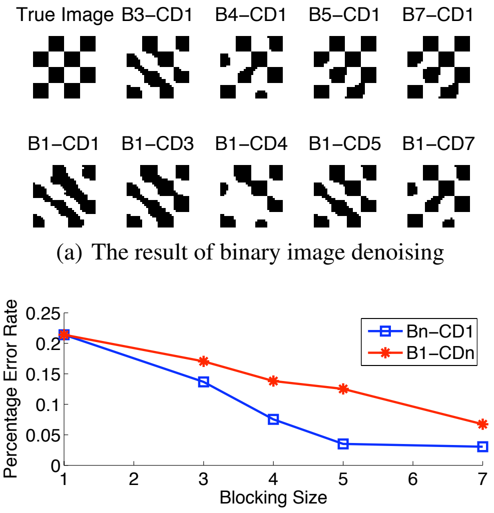
The exponential random graph model (ERGM) is a flexible family of models for complex networks which is widely used in social network analysis (see Robins et al. (2007) for an introduction to ERGMs).

Assume we have a graph x, and suppose x_{ij} denotes the edge between i-th node and j-th node, with $x_{ij} = 1$ indicating the existence of an edge between nodes i and j, and $x_{ij} = 0$ otherwise. The distribution over networks for an exponential random graph model is the following,

$$P(x|\theta) = \frac{1}{Z(\theta)} \exp(\theta^T u(x)) \qquad (12)$$

where $u(x) = [u_1(x), \ldots, u_k(x)]$ are the global features of the network, such as the number of edges, the number of k-stars, and the number of triangles.

An important problem in ERGM research is to estimate the parameter vector θ given the observed network data. MPLE and MCMC-MLE are two of the main methods that have been widely investigated. The use of MPLE on ERGMs was first suggested by Strauss and Ikeda (1990). Although MPLE is computationally fast, it is known that estimates from MPLE are inaccurate for certain networks (Robins et al., 2007). MCMC-MLE is a more accurate method for parameter estimation in ERGMs (Snijders, 2002; van Duijn et al., 2009). However, MCMC-MLE requires the Markov chain to converge and, hence, requires large amounts of computation, especially when the network size is large.



(a) The result of binary image denoising

(b) The percentage error rates

Figure 3: A comparison between B1-CDn and Bn-CD1 on a CRF. (a) TOP: A visual depiction of the results of binary image denoising. (b) BOTTOM: Percentage error rates of B1-CDn and Bn-CD1 with blocking size n. The error decreases as the blocking size increases, and Bn-CD1 outperforms B1-CDn. The blocks A_c in Bn-CD1 are randomly selected from the tree-structured subsets in Figure 4.

To trade off computation time for accuracy, we use Bn-CD1 to estimate the parameters of the ERGM. In our experiment, we used the Lazega social network data used in Snijders et al. (2006). In this network of 36 nodes, an edge denotes a tie between corporate lawyers in the network. Since each possible undirected edge is a different variable, our model contains 630 variables. The network statistics that we employ in our model are the number of edges, the number of 2-stars, and the number of triangles. Thus, there are three parameters to fit in our model.

Since there is no ground truth for the parameters for this data set, we ran MCMC-MLE using the widely-used statnet R package (Handcock et al., 2008) and used those estimates as the ground truth. As is usually the case in social network analysis, we only have one data point (a single network of lawyers) to use when fitting the three parameters. Nevertheless, it is still possible to successfully perform blocked contrastive divergence. In Figure 4.3, we show the results of running B1-CD1, B2-CD1, B3-CD1, and B4-CD1 on the Lazega data for 20,000 iterations. We use an initial gradient descent step size of 0.001 which is gradually decreased. We plot the L1 error between the current parameter estimates of the algorithm and the parameter estimates obtained from MCMC-MLE, and we see that using BCD with higher-order blocks improves the accuracy of the estimates when compared to B1-CD1. As in the case of the visible Boltzmann machines, we choose random blocks of variables of the same order. Even using a block size of 2

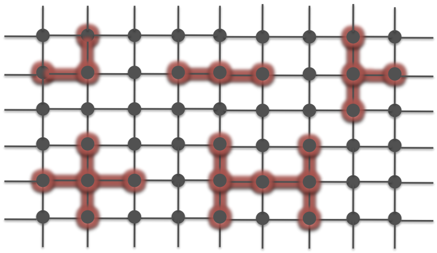


Figure 4: An illustration of tree-structured composite likelihood blocks for an Ising model. These configurations are used in Bn-CD1 in Figure 3: B3-CD1 randomly selects blocks of shape “L” or “–”, B4-CD1 uses blocks of shape “T”, B5-CD1 uses shape “+”, and B7-CD1 uses shape “H”. The locations and orientations of the blocks are randomly drawn at each iteration of Gibbs sampling.

shows a substantial amount of improvement over standard CD. We also see that a relatively small blocking size of 4, which does not require a lot of calculation in practice, allows the algorithm to estimate parameters which are close to the ones found by statnet’s MCMC-MLE routine.

We also performed experiments on simulated networks (which we sampled from ground truth parameters) and found that BCD also performs well in these cases.

5 RELATED WORK

The theory and application of composite likelihood techniques has garnered increasing interest among statisticians, e.g. Besag (1974), Lindsay (1988), Parner (2001), Fearnhead and Donnelly (2002), and Varin and Vidoni (2005). The machine learning community has also recently begun to explore composite likelihood techniques. Liang and Jordan (2008) compare discriminative and generative risks and provide an asymptotic analysis of a unified framework of techniques which includes composite likelihood. Dillon and Lebanon (2009) propose stochastic composite loglikelihood (SCL) as a method to decrease the computation time for optimizing the composite likelihood. In their approach, each block A_c is selected by a certain probability; they also examine statistical and computational tradeoffs for different composite likelihoods, using asymptotic variance as a measurement. While our random-scan blocked Gibbs sampling strategy is similar to the idea of SCL, we formally connect MCLE to contrastive divergence and thus obtain a practical algorithm for optimizing composite likelihoods. We also investigate the use of tree-structured composite likelihoods to increase computational efficiency.

Our work is also influenced by Hyvärinen (2006), who developed the connection between pseudolikelihood and contrastive divergence for visible Boltzmann machines. Another result by the same author (Hyvärinen, 2007) uncovers the connection between pseudolikelihood, contrastive divergence, and score matching. For instance, score matching is equivalent to a version of CD that performs Langevin updates instead of Gibbs sampling. Thus, there exists an entire tapestry of related estimation techniques based on the contrastive divergence framework.

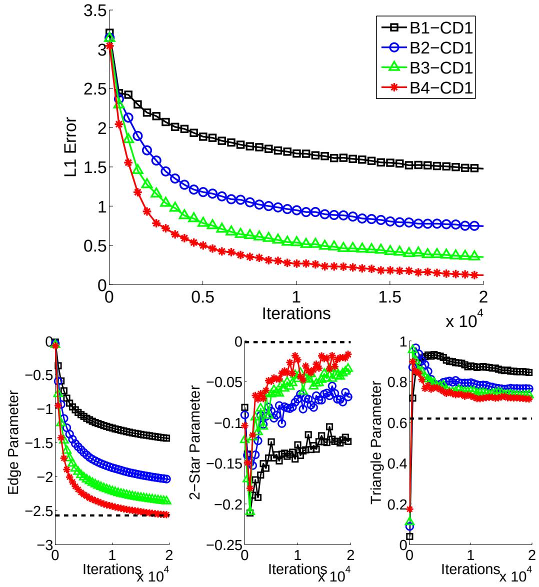


Figure 5: Performance of BCD with various block sizes on an ERGM. (a) TOP: L1 error of BCD as the algorithm progresses. (b) BOTTOM: Evolution of ERGM parameters. Dashed lines indicate values obtained by MCMC-MLE. Higher-order blocks lead to more accurate estimates.

We note that a trivial form of blocked Gibbs sampling is usually performed when applying contrastive divergence to models like restricted Boltzmann machines. In RBMs, all the variables in a hidden or visible layer are conditionally independent from one another given the other layer, and thus they can be easily sampled jointly. In contrast, our work applies blocked sampling to dependent sets of variables and highlights the benefits of including such nontrivial joint sampling steps in speeding MCMC convergence. Moreover, we have focused on models with only visible variables. One can extend our blocked sampling approach to models with hidden variables, although the connection to composite likelihoods is less clear.

6 CONCLUSION

In this paper, we have studied the relationship between composite likelihood and blocked contrastive divergence. We have shown that one can improve the accuracy of parameter estimates using blocked contrastive divergence for various models, especially when there exist strong dependencies. We also proposed the use of efficient tree-structured composite likelihoods for models with sparse dependence structures. This approach decreases the time complexity of MCLE from being exponential in block size to being linear in block size, making the optimization of high-order composite likelihoods possible in practice.

There remain many open issues regarding the efficient optimization of composite likelihoods. Optimally choosing the sizes and shapes of the blocks to use in the composite likelihood is still an open area of research. Furthermore, combining blocked contrastive divergence with other computational tricks, such as persistent CD with fast weights, provides interesting directions for future work.

Acknowledgements

This work was supported in part by NSF under grant number IIS-0083489 (PS, AA) and an NSF Graduate Fellowship (AA), by ONR/MURI under grant number N00014-08-1-1015 (PS, AA, QL), by NIH-NIAMS under grant AR-44882 BIRT revision (AI, QL), and by Google (PS).

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36(2):192–236.

Carreira-Perpiñán, M. Á. and Hinton, G. (2005). On contrastive divergence learning. In *10th International Conference on AI and Statistics*, pages 33–40.

Dillon, J. and Lebanon, G. (2009). Statistical and computational tradeoffs in stochastic composite likelihood. In *12th International Conference on AI and Statistics*, pages 129–136.

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society. Series B*, 64(4):657–680.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163.

Handcock, M., Hunter, D., Butts, C., Goodreau, S., and Morris, M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1548.

Hinton, G. (2000). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504.

Hyvärinen, A. (2006). Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Computation*, 18(10):2283–2292.

Hyvärinen, A. (2007). Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, 18(5):1529–1531.

Kumar, S. and Hebert, M. (2004). Discriminative fields for modeling spatial dependencies in natural images. In *Advances in Neural Information Processing Systems 16*, pages 1531 – 1538, Cambridge, MA. MIT Press.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Verlag.

Liang, P. and Jordan, M. I. (2008). An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *25th International Conference on Machine Learning*, pages 584–591, New York. ACM.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1):22139.

Parner, E. (2001). A composite likelihood approach to multivariate survival data. *Scandinavian Journal of Statistics*, 28(2):295–302.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173 – 191.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, pages 99–153.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212.

Tieleman, T. and Hinton, G. (2009). Using fast weights to improve persistent contrastive divergence. In *26th International Conference on Machine Learning*, pages 1033–1040, New York. ACM.

van Duijn, M. A., Gile, K. J., and Handcock, M. S. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52 – 62.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528.

Vishwanathan, S., Schraudolph, N., Schmidt, M., and Murphy, P. (2006). Accelerated training of conditional random fields with stochastic gradient methods. In *23rd International Conference on Machine Learning*, pages 969–976, New York. ACM.

Appendix

We derive Eq. (10). Starting from the conditional CL (Eq. 7),

$$\begin{aligned}
\mathcal{CL}(\theta|X) &= \sum_{c=1}^{C}\sum_{i=1}^{N}\log p(x^i_{A_c}|x^i_{\neg A_c},\theta)\\
&= C\sum_{i=1}^{N}\log\frac{\exp(\theta^T\mu(x^i))}{Z(\theta)} - \sum_{c=1}^{C}\sum_{i=1}^{N}\log\sum_{x_{A_c}}\frac{\exp(\theta^T\mu(x_{A_c},x^i_{\neg A_c}))}{Z(\theta)}\\
&\propto \frac{1}{N}\sum_{i=1}^{N}\theta^T\mu(x^i) - \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N}\sum_{i=1}^{N}\log\sum_{x_{A_c}}\exp(\theta^T\mu(x_{A_c},x^i_{\neg A_c})),
\end{aligned}$$

where the second line uses the definition of conditional probability and Eq. 1, and the third line cancels $Z(\theta)$ and divides by CN. While the gradient of the first term is $\langle\mu(x)\rangle_0$, the gradient of the second term (the "negative" part) is below,

$$\begin{aligned}
\frac{\partial\mathcal{CL}^-}{\partial\theta} &= \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{x_{A_c}}\mu(x_{A_c},x^i_{\neg A_c})\exp(\theta^T\mu(x_{A_c},x^i_{\neg A_c}))}{\sum_{x_{A_c}}\exp(\theta^T\mu(x_{A_c},x^i_{\neg A_c}))}\\
&= \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N}\sum_{i=1}^{N}\sum_{x_{A_c}}\mu(x_{A_c},x^i_{\neg A_c})p(x_{A_c}|x^i_{\neg A_c},\theta)\\
&= \frac{1}{C}\sum_{c}^{C}\langle\langle u(x)\rangle_{p(x_{A_c}|x^i_{\neg A_c},\theta)}\rangle_0.
\end{aligned}$$

For a Monte Carlo approximation of this gradient, one can obtain samples by randomly choosing a block c and a data point i, and performing one blocked Gibbs step using $p(x_{A_c}|x^i_{\neg A_c},\theta)$.