# Statistical and Computational Tradeoffs in Stochastic Composite Likelihood

**Joshua Dillon**
College of Computing
Georgia Institute of Technology
jvdillon@gatech.edu

**Guy Lebanon**
College of Computing
Georgia Institute of Technology
lebanon@cc.gatech.edu

September 21, 2009

**Abstract**

Maximum likelihood estimators are often of limited practical use due to the intensive computation they require. We propose a family of alternative estimators that maximize a stochastic variation of the composite likelihood function. We prove the consistency of the estimators, provide formulas for their asymptotic variance and computational complexity, and discuss experimental results in the context of Boltzmann machines and conditional random fields. The theoretical and experimental studies demonstrate the effectiveness of the estimators in achieving a predefined balance between computational complexity and statistical accuracy.

## 1 Introduction

Maximum likelihood estimation is by far the most popular point estimation technique in machine learning and statistics. Assuming that the data consists of $n$, $m$-dimensional vectors

$$D = \{X^{(1)}, \dots, X^{(n)}\} \subset \mathbb{R}^m, \tag{1}$$

and is sampled iid from a parametric distribution $p_{\theta_0}$ with $\theta_0 \in \Theta \subset \mathbb{R}^r$, a maximum likelihood estimator (mle) $\hat{\theta}_n^{\text{ml}}$ is a maximizer of the loglikelihood function

$$\ell_n(\theta\,; D) = \sum_{i=1}^{n} \log p_\theta(X^{(i)}). \tag{2}$$

The use of the mle is motivated by its consistency, i.e. $\hat{\theta}_n^{\text{ml}} \to \theta_0$ as $n \to \infty$ with probability 1 (Ferguson, 1996). The consistency property ensures that as the number $n$ of samples grows, the estimator will converge to the true parameter $\theta_0$ governing the data generation process.

An even stronger motivation for the use of the mle is that it has an asymptotically normal distribution with mean vector $\theta_0$ and variance matrix $(nI(\theta_0))^{-1}$. More formally, we have the following convergence in distribution as $n \to \infty$ (Ferguson, 1996)

$$\sqrt{n}\,(\hat{\theta}_n^{\text{ml}} - \theta_0) \rightsquigarrow N(0, I^{-1}(\theta_0)), \tag{3}$$

where $I(\theta)$ is the $r \times r$ Fisher information matrix

$$I(\theta) = \mathsf{E}_{p_\theta}\{\nabla \log p_\theta(X)(\nabla \log p_\theta(X))^\top\} \tag{4}$$

with $\nabla f$ represents the $r \times 1$ gradient vector of $f(\theta)$ with respect to $\theta$. The convergence (3) is especially striking since according to the Cramer-Rao lower bound, the asymptotic variance $(nI(\theta_0))^{-1}$ of the mle is the smallest possible variance for any estimator. Since it achieves the lowest possible asymptotic variance, the mle (and other estimators which share this property) is said to be asymptotically efficient.

The consistency and asymptotic efficiency of the mle motivate its use in many circumstances. Unfortunately, in some situations the maximization or even evaluation of the loglikelihood (2) and its derivatives is impossible due to computational considerations. This has lead to the proposal of alternative estimators under the premise that a loss of asymptotic efficiency is acceptable–in return for reduced computational complexity. Consistency however, is typically viewed as less negotiable and inconsistent estimators should be avoided if at all possible.

In this paper, we propose a family of estimators, for use in situations where the computation of the mle is intractable. In contrast to previously proposed approximate estimators, our estimators are statistically consistent and admit a precise quantification of both computational complexity and statistical accuracy through their asymptotic variance. Due to the continuous parameterization of the estimator family, we obtain an effective framework for achieving a predefined problem-specific balance between computational tractability and statistical accuracy. For the sake of concreteness, we focus on the case of estimating the parameters associated with Markov random fields. In this case, we provide a detailed discussion of the accuracy complexity tradeoff and experimental results for the Boltzmann machine and conditional random fields.

## 2 Related Work

There is a large body of work dedicated to tractable learning techniques. Two popular categories are Markov chain Monte Carlo (MCMC) and variational methods. MCMC is a general purpose technique for approximating expectations and can be used to approximate the normalization term and other intractable portions of the loglikelihood and its gradient (Casella and Robert, 2004). Variational methods are techniques for conducting inference and learning based on tractable bounds. Despite the substantial work on MCMC and variational methods, there are few results that are general enough to be practical while preserving clear results concerning convergence and approximation rate.

Our work draws on the composite likelihood method for parameter estimation proposed by Lindsay (1988) which in turn generalized the pseudo likelihood of Besag (1974). A selection of more recent studies on pseudo and composite likelihood are (Arnold and Strauss, 1991, Liang and Yu, 2003, Varin and Vidoni, 2005, Sutton and McCallum, 2007, Hjort and Varin, 2008). Most of the recent studies in this area examine the behavior of the pseudo or composite likelihood in a particular modeling situation. We believe that the present paper is the first to systematically examine statistical and computational tradeoffs in a general quantitative framework. Possible exceptions are (Zhu and Liu, 2002) which is an experimental study on texture generation, (Xing et al., 2003) which is focused on inference rather than parameter estimation, and (Liang and Jordan, 2008) which compares discriminative and generative risks.

## 3 Stochastic Composite Likelihood

In many cases, the absence of a closed form expression for the normalization term prevents the computation of the loglikelihood (2) and its derivatives thereby severely limiting the use of the mle. A popular example are Markov random fields, wherein the computation of the normalization term is often intractable (see Section 5 for more details). In this paper we propose alternative estimators based on the maximization of a stochastic variation of the composite likelihood.

We start by defining the pseudo loglikelihood function (Besag, 1974) associated with the data $D$ of (1),

$$p\ell_n(\theta\,; D) = \sum_{i=1}^{n}\sum_{j=1}^{m} \log p_\theta(X_j^{(i)} | \{X_k^{(i)} : k \neq j\}). \tag{5}$$

The maximum pseudo likelihood estimator (mple) $\hat{\theta}_n^{\mathrm{mpl}}$ is consistent, but possesses considerably higher asymptotic variance than that of the mle's $(nI(\theta_0))^{-1}$. Its main advantage is that it does not require the computation of the normalization term as it cancels out in the probability ratio defining conditional distributions

$$\begin{aligned} p_\theta(X_j|\{X_k : k \neq j\}) = p_\theta(X)/ \\ \textstyle\sum_{X_j'} p_\theta(X_1, \ldots, X_{j-1}, X_j', X_{j+1}, \ldots, X_m). \end{aligned} \tag{6}$$

The mle and mple represent two different ways of resolving the tradeoff between asymptotic variance and computational complexity. The mle has low asymptotic variance but high computational complexity while the mple has higher asymptotic variance but low computational complexity. It is desirable to obtain additional estimators realizing alternative resolutions of the accuracy complexity tradeoff. To this end we define the stochastic composite likelihood whose maximization provides a family of consistent estimators with statistical accuracy and computational complexity spanning the entire accuracy-complexity spectrum.

Stochastic composite likelihood generalizes the likelihood and pseudo likelihood functions by constructing an objective function that is a stochastic sum of likelihood objects. We start by defining the notion of $m$-pairs and likelihood objects and then proceed to stochastic composite likelihood.

**Definition 1.** An $m$-pair $(A, B)$ is a pair of sets $A, B \subset \{1, \ldots, n\}$ satisfying $A \neq \emptyset = A \cap B$. The likelihood object associated with an $m$-pair $(A, B)$ and $X$ is $S_\theta(A, B) = \log p_\theta(X_A|X_B)$ where $X_S \overset{\text{def}}{=} \{X_j : j \in S\}$. We similarly define likelihood objects with respect to a dataset $D = \{X^{(1)}, \ldots, X^{(n)}\}$ as

$$S_\theta(n, A, B) = \sum_{i=1}^{n} \log p_\theta(X_A^{(i)}|X_B^{(i)}).$$

The Lindsay (1988) composite loglikelihood function, is a collection of likelihood objects defined by a finite sequence of $m$-pairs $(A_1, B_1), \ldots, (A_k, B_k)$

$$\begin{aligned} c\ell_n(\theta\,; D) &= \sum_{j=1}^{k} S_\theta(n, A_j, B_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \log p_\theta(X_{A_j}^{(i)}|X_{B_j}^{(i)}). \end{aligned} \tag{7}$$

There exists a certain lack of flexibility associated with the composite likelihood framework. Since each likelihood object $S_\theta(n, A, B)$ is either selected or not, there is no allowance for some objects to be selected more frequently than others. Allowing stochastic, rather than deterministic, selection of likelihood objects provides a higher degree of flexibility and a richer parametric family of estimators. Furthermore, the discrete parameterization of (7) defined by the sequence $(A_1, B_1), \ldots, (A_k, B_k)$ is less convenient for theoretical analysis than the continuous parameterization underlying the stochastic composite likelihood.

**Definition 2.** The stochastic composite loglikelihood (scl) associated with a finite sequence of $m$-pairs $(A_1, B_1), \ldots, (A_k, B_k)$ is

$$sc\ell_n(\theta\,; D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \beta_j Z_{ij} \log p_\theta(X_{A_j}^{(i)}|X_{B_j}^{(i)}). \tag{8}$$

where $\beta_j > 0$ and $Z_{ij} \sim \text{Ber}(\lambda_j)$ are independent binary Bernoulli rv with parameters $\lambda_j \in [0, 1]$.

In other words, the scl is a stochastic version of (7) where for each sample $X^{(i)}, i = 1, \ldots, n$, the likelihood objects $S(A_1, B_1), \ldots, S(A_k, B_k)$ are selected independently with probabilities $\lambda_1, \ldots, \lambda_k$. The positive weights $\beta_j$ provide additional flexibility by emphasizing different components more than others.

In analogy to the mle and the mple, the maximum scl estimator (mscle) $\hat{\theta}_n^{\text{msl}}$ estimates $\theta_0$ by maximizing the scl function. In contrast to the loglikelihood and pseudo loglikelihood functions, the scl function and its maximizer are random variables that depend on the indicator variables $Z_{ij}$ in addition to $D$. As such, its behavior should be summarized by examining its expectation or its behavior in the limit $n \to \infty$. Different selections of the continuous parameters $(\lambda, \beta) \in [0, 1]^k \times \mathbb{R}_+^k$ underlying the scl function result in different asymptotic variance and computational complexity. As a result the accuracy and complexity of $\hat{\theta}_n^{\text{msl}}$ become continuous functions over the parametric space $[0, 1]^k \times \mathbb{R}_+^k$ which include as special cases the mle, mple, and maximum quasi likelihood (Hjort and Varin, 2008) estimators. Different selections of $(\lambda, \beta) \in [0, 1]^k \times \mathbb{R}_+^k$ represent estimators $\hat{\theta}_n^{\text{msl}}$ achieving different resolutions of the accuracy-complexity tradeoff.

# 4 Statistical Properties of $\hat{\theta}_n^{\text{msl}}$

The statistical properties of the mscle depend on the selection probabilities and positive weights $(\lambda, \beta) \in [0, 1]^k \times \mathbb{R}_+^k$ while the computational properties depend only on $\lambda$. Under some mild conditions $\hat{\theta}_n^{\text{msl}}$ may be shown to be a consistent estimator whose asymptotic distribution is Gaussian with a certain variance matrix that is larger or equal to the optimal variance expressed by the inverse Fisher information. For simplicity, we assume that $X$ is discrete and $p_\theta(x) > 0$.

**Definition 3.** A sequence of $m$-pairs $(A_1, B_1), \ldots, (A_k, B_k)$ ensures identifiability of $p_\theta$ if the map $\{p_\theta(X_{A_j}|X_{B_j}) : j = 1, \ldots, k\} \mapsto p_\theta(X)$ is injective. In other words, there exists only a single collection of conditionals $\{p_\theta(X_{A_j}|X_{B_j}) : j = 1, \ldots, k\}$ that does not contradict the joint $p_\theta(X)$.

Proposition 1 below generalizes the Shannon-Kolmogorov information inequality.

**Proposition 1.** *Let $(A_1, B_1), \ldots, (A_k, B_k)$ be a sequence of $m$-pairs that ensures identifiability of $p_\theta, \theta \in \Theta$ and $\alpha_1, \ldots, \alpha_k$ positive constants. Then*

$$\sum_{j=1}^{k} \alpha_k \, D(p_\theta(X_{A_j}|X_{B_j}) \,||\, p_{\theta'}(X_{A_j}|X_{B_j})) \geq 0$$

*where equality holds iff $\theta = \theta'$.*

*Proof.* The inequality follows from applying Jensen's inequality for each conditional KL divergence

$$\begin{aligned} &-D(p_\theta(X_{A_j}|X_{B_j}) \,||\, p_{\theta'}(X_{A_j}|X_{B_j})) \\ &\quad = \mathsf{E}_{p_\theta} \log \frac{p_{\theta'}(X_{A_j}|X_{B_j})}{p_\theta(X_{A_j}|X_{B_j})} \leq \log E_{p_\theta} \frac{p_{\theta'}(X_{A_j}|X_{B_j})}{p_\theta(X_{A_j}|X_{B_j})} \\ &\quad = \log 1 = 0. \end{aligned}$$

For equality to hold we need each term to be 0 which follows only if $p_\theta(X_{A_j}|X_{B_j}) \equiv p_{\theta'}(X_{A_j}|X_{B_j})$ for all $j$ which, assuming identifiability, holds iff $\theta = \theta'$. □

**Proposition 2.** *Let $\lambda \in [0,1]^k$ and $(A_1, B_1), \ldots, (A_k, B_k)$ be a sequence of $m$-pairs for which $\{(A_j, B_j) : \forall j \text{ such that } \lambda_j > 0\}$ ensures identifiability. We also assume that $\Theta \subset \mathbb{R}^r$ is an open set and $p_\theta(x) > 0$ and is continuous and smooth in $\theta$. Then there exists a strongly consistent sequence of scl maximizers, i.e. $\hat{\theta}_n^{msl} \to \theta_0$ as $n \to \infty$ with probability 1.*

The proof technique below generalizes Wald's proof for the consistency of the mle.

*Proof.* The scl function, modified slightly by a linear combination with a term that is constant in $\theta$ is

$$\begin{aligned} sc\ell'(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \beta_j \Big( & Z_{ij} \log p_\theta(X_{A_j}^{(i)}|X_{B_j}^{(i)}) \\ & - \lambda_j \log p_{\theta_0}(X_{A_j}^{(i)}|X_{B_j}^{(i)}) \Big). \end{aligned}$$

By the strong law of large numbers, the above expression converges as $n \to \infty$ to its expectation

$$\mu(\theta) = -\sum_{j=1}^{k} \beta_j \lambda_j \, D(p_\theta(X_{A_j}|X_{B_j}) \,||\, p_{\theta_0}(X_{A_j}|X_{B_j})).$$

If we restrict ourselves to the compact set $S = \{\theta : c_1 \leq \|\theta - \theta_0\| \leq c_2\}$ then $|\log p_\theta(x)| < K(x) < \infty, \forall \theta \in S$. As a result, the conditions for the uniform strong law of large numbers (Ferguson, 1996) hold on $S$ leading to

$$P \left\{ \lim_{n \to \infty} \sup_{\theta \in S} |sc\ell'(\theta) - \mu(\theta)| = 0 \right\} = 1. \tag{9}$$

By Proposition 1, $\mu(\theta)$ is non-positive and is zero iff $\theta = \theta_0$. Since the function $\mu(\theta)$ is continuous it attains its negative supremum on the compact $S$: $\sup_{\theta \in S} \mu(\theta) < 0$. Combining this fact with (9) we have that there exists $N$ such that for all $n > N$ the scl maximizers on $S$ achieves strictly negative values of $sc\ell'(\theta)$ with probability 1. However, since $sc\ell'(\theta)$ can be made to achieve values arbitrarily close to zero under $\theta = \theta_0$, we have that $\hat{\theta}_n^{\mathrm{msl}} \notin S$ for $n > N$. Since $c_1, c_2$ were chosen arbitrarily $\hat{\theta}_n^{\mathrm{msl}} \to \theta_0$ with probability 1. □

The above proposition indicates that to guarantee consistency, the sequence of $m$-pairs needs to satisfy Definition 3. It can be shown that a selection equivalent to the pseudo likelihood function, i.e.,

$$A_i = \{i\}, B_i = \{1, \dots, m\} \setminus A_i, i = 1, \dots, k, \tag{10}$$

ensure identifiability and consequently the consistency of the mscle estimator. Furthermore, every selection of $m$-pairs that includes as a subset (10) similarly guarantees identifiability and consistency.

**Proposition 3.** *Making the assumptions of Proposition 2 as well as convexity of $\Theta \subset \mathbb{R}^r$ we have*

$$\sqrt{n}(\hat{\theta}_n^{msl} - \theta_0) \rightsquigarrow N\left(0, \Upsilon\Sigma\Upsilon\right) \tag{11}$$

*where* $\Upsilon^{-1} = \sum_{j=1}^k \beta_j \lambda_j \mathsf{Var}_{\theta_0}(V_j)$*,* $V_j = \nabla S_{\theta_0}(A_j, B_j)$*, and* $\Sigma = \mathsf{Var}_{\theta_0}(\sum_{j=1}^k \beta_j\lambda_j V_j)$*.*

The notation $\mathsf{Var}_{\theta_0}(Y)$ represents the covariance matrix of the random vector $Y$ under $p_{\theta_0}$ while the notations $\xrightarrow{\text{P}}$, $\rightsquigarrow$ in the proof below denote convergences in probability and in distribution (Ferguson, 1996).

*Proof.* By the mean value theorem and convexity of $\Theta$ there exists $\eta \in (0,1)$ for which $\theta' = \theta_0 + \eta(\hat{\theta}_n^{\text{msl}} - \theta_0)$ and

$$\nabla sc\ell_n(\hat{\theta}_n^{\text{msl}}) = \nabla sc\ell_n(\theta_0) + \nabla^2 sc\ell_n(\theta')(\hat{\theta}_n^{\text{msl}} - \theta_0)$$

where $\nabla f(\theta)$ and $\nabla^2 f(\theta)$ are the $r \times 1$ gradient vector and $r \times r$ matrix of second order derivatives of $f(\theta)$. Since $\hat{\theta}_n$ maximizes the scl, $\nabla sc\ell_n(\hat{\theta}_n^{\text{msl}}) = 0$ and

$$\sqrt{n}(\hat{\theta}_n^{\text{msl}} - \theta_0) = -\sqrt{n}(\nabla^2 sc\ell_n(\theta'))^{-1}\nabla sc\ell_n(\theta_0). \tag{12}$$

By Proposition 2 we have $\hat{\theta}_n^{\text{msl}} \xrightarrow{\text{P}} \theta_0$ which implies that $\theta' \xrightarrow{\text{P}} \theta_0$ as well. Furthermore, by the law of large numbers and the fact that if $W_n \xrightarrow{\text{P}} W$ then $g(W_n) \xrightarrow{\text{P}} g(W)$ for continuous $g$,

$$\begin{aligned}(\nabla^2 sc\ell_n(\theta'))^{-1} &\xrightarrow{\text{P}} (\nabla^2 sc\ell_n(\theta_0))^{-1} \\ &\xrightarrow{\text{P}} \left(\sum_{j=1}^k \beta_j\lambda_j \mathsf{E}_{\theta_0} \nabla^2 S_{\theta_0}(A_j, B_j)\right)^{-1} \\ &= -\left(\sum_{j=1}^k \beta_j\lambda_j \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j))\right)^{-1}.\end{aligned} \tag{13}$$

For the remaining term in (12) we have

$$\sqrt{n}\,\nabla sc\ell_n(\theta_0) = \sum_{j=1}^k \beta_j \sqrt{n}\,\frac{1}{n}\sum_{i=1}^n W_{ij}$$

where the random vectors $W_{ij} = Z_{ij}\nabla \log p_\theta(X_{A_j}^{(i)}|X_{B_j}^{(i)})$ have expectation 0 and variance matrix $\mathsf{Var}_{\theta_0}(W_{ij}) = \lambda_j \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j))$. By the central limit theorem

$$\sqrt{n}\,\frac{1}{n}\sum_{i=1}^n W_{ij} \rightsquigarrow N\left(0, \lambda_j \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j))\right).$$

The sum $\sqrt{n}\,\nabla sc\ell_n(\theta_0) = \sum_{j=1}^k \beta_j\sqrt{n}\,\frac{1}{n}\sum_{i=1}^n W_{ij}$ is asymptotically Gaussian as well with mean zero since it converges to a sum of Gaussian distributions with mean zero. Since in the general case the random variables $\sqrt{n}\,\frac{1}{n}\sum_{i=1}^n W_{ij}$, $j = 1, \dots, k$ are correlated, the asymptotic variance matrix of $\sqrt{n}\,\nabla sc\ell_n(\theta_0)$ needs to account for cross covariance terms leading to

$$\sqrt{n}\,\nabla sc\ell_n(\theta_0) \rightsquigarrow N\left(0, \mathsf{Var}_{\theta_0}\left(\sum_{j=1}^k \beta_j\lambda_j \nabla S_{\theta_0}(A_j, B_j)\right)\right). \tag{14}$$

We finish the proof by combining (12), (13) and (14) using Slutsky's theorem. □

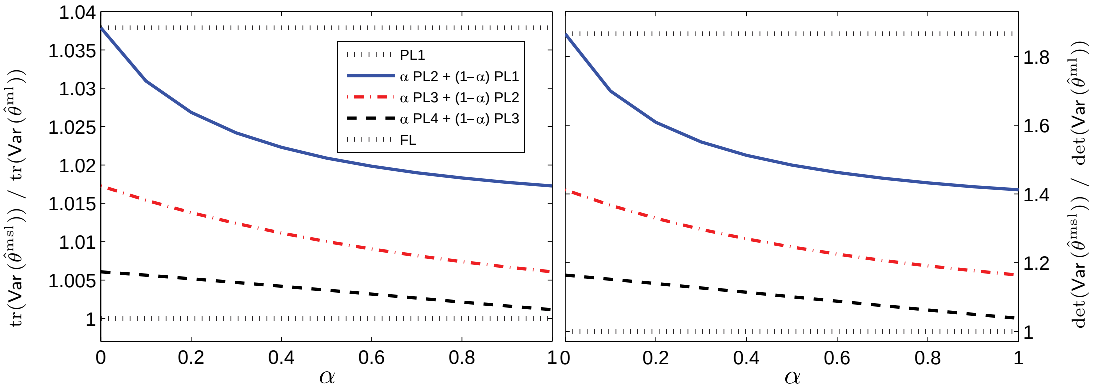

Figure 1: Asymptotic variance matrix, as measured by trace (left) and determinant (right), as a function of the selection probabilities for different stochastic versions of the scl function.

# 5 Stochastic Composite Likelihood for Markov Random Fields

Markov random fields (MRF) are some of the more popular statistical models for complex high dimensional data. Approaches based on pseudo likelihood and composite likelihood are naturally well-suited in this case due to the cancellation of the normalization term in the probability ratios defining conditional distributions. More specifically, a MRF with respect to a graph $G = (V, E)$, $V = \{1, \dots, m\}$ with a clique set $\mathcal{C}$ is given by the following exponential family model

$$P_\theta(x) = \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C(x_C) - \log Z(\theta)\right),$$
$$Z(\theta) = \sum_x \exp\left(\sum_{C\in\mathcal{C}} \theta_c f_C(x_C)\right). \tag{15}$$

The primary bottlenecks in obtaining the maximum likelihood are the computations $\log Z(\theta)$ and $\nabla \log Z(\theta)$. Their computational complexity is exponential in the graph's treewidth and for many cyclic graphs, such as the Ising model or the Boltzmann machine, it is exponential in $|V| = m$.

In contrast, the conditional distributions that form the composite likelihood of (15) are given by

$$P_\theta(x_A|x_B) = \frac{\cancel{Z(\theta)} \sum_{x'_{(A\cup B)^c}} \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C((x_A, x_B, x'_{(A\cup B)^c})_C)\right)}{\cancel{Z(\theta)} \sum_{x'_{(A\cup B)^c}} \sum_{x''_A} \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C((x''_A, x_B, x'_{(A\cup B)^c})_C)\right)}. \tag{16}$$

The computation of (16) depends on the size of the sets $A$ and $(A\cup B)^c$ and their intersections with the cliques in $\mathcal{C}$. In general, selecting small $|A_j|$ and $B_j = (A_j)^c$ leads to efficient computation of the composite likelihood and its gradient. For example, in the case of $|A_j| = l, |B_j| = m - l$ with $l \ll m$ we have that $k \le m!/(l!(m-l)!)$ and the complexity of computing the $c\ell(\theta)$ function and its gradient may be shown to require time that is at most exponential in $l$ and polynomial in $m$.

Computing the $sc\ell(\theta)$ function and its gradient depends on the Bernoulli parameters $\lambda \in [0,1]^k$ and the sequence of $m$-pairs $(A_1, B_1), \dots, (A_k, B_k)$. Selecting a sequence of $m$ pairs that includes all $A_i = \{i\}, B_i = \{1, \dots, m\} \setminus A_i$ pairs ensures consistency. Adding pairs $(A_j, B_j)$ with larger sets $|A_j|$ enables obtaining a specific complexity number within a wide spectrum of available complexities by choosing appropriate mixing parameters $\lambda$.

# 6 Controlling Efficiency through $\beta$

As Proposition 3 indicates, the weight vector $\beta$ and selection probabilities $\lambda$ play an important role in the statistical accuracy of the estimator through its asymptotic variance. The computational complexity, on the other hand, is determined by $\lambda$ independently of $\beta$. Conceptually, we are interested in resolving the accuracy-complexity tradeoff jointly for both $\beta, \lambda$ before estimating $\theta$ by maximizing the scl function. We simplify this objective by choosing the selection probabilities $\lambda$ based on available computational resources and computing time. Since the computational complexity does not depend on $\beta$ we can then proceed to select the $\beta$ that maximizes the statistical accuracy of the estimator given the selection probabilities $\lambda$.

Selecting $\beta$ that minimizes the asymptotic variance is somewhat ambiguous as $\Upsilon\Sigma\Upsilon$ in Proposition 3 is an $r \times r$ positive semidefinite matrix. A common solution is to consider the determinant as a one dimensional measure of the size of the variance matrix, and minimize

$$J(\beta) = \log\det(\Upsilon\Sigma\Upsilon) = \log\det\Sigma + 2\log\det\Upsilon \tag{17}$$

There are two significant drawbacks associated with the optimization of (17). It depends on the true parameter value $\theta_0$ which is not known at training time. Additionally, introducing a secondary optimization problem into the iterative maximization of the scl function undermines the motivation of scl as a computationally efficient approximate estimation technique.

We propose to address both issues by constructing an estimator for the determinants $\log\det\Sigma$, $\log\det\Upsilon$ based on the empirical variance

$$\mathsf{Var}_{p_{\theta_0}}(g(X)) \approx \mathsf{E}_{\tilde{p}}(g(X) - \mathsf{E}_{\tilde{p}}(g(X)))^2$$

where $\tilde{p}(z) = \frac{1}{n}\sum_i \delta_{\{z=x^{(i)}\}}$ is the empirical distribution associated with the available training set. We note that the corresponding estimators of $\log\det\Sigma, \log\det\Upsilon$ can be computed without the knowledge of $\hat{\theta}_n^{\text{msl}}$. As a consequence, we can determine the optimal $\beta$ before solving the scl maximization problem.

Estimating $\log\det\Sigma, \log\det\Upsilon$ can be performed very quickly due to a decomposition similar to the inclusion-exclusion principle. However, we omit the details due to lack of space.

# 7 Experiments

We demonstrate the asymptotic properties of $\hat{\theta}_n^{\text{msl}}$ for the Boltzmann machine and explore the complexity-accuracy tradeoff associated with several stochastic versions of $sc\ell(\theta)$ for CRFs.

## 7.1 Boltzmann Machines

We illustrate the improvement in asymptotic variance of the mscle associated with adding higher order likelihood components with increasing probabilities in context of the Boltzmann machine $p_\theta(x) = \exp(\sum_{i<j}\theta_{ij}x_ix_j - \log\psi(\theta)), x \in \{0,1\}^m$. To be able to accurately compute the asymptotic variance we use $m = 5$ with $\theta$ being a $\binom{5}{2}$ dimensional vector with half the components $+1$ and half $-1$. Since the asymptotic variance of $\hat{\theta}_n^{\text{msl}}$ is a matrix we summarize its size using either its trace or determinant.

We plot in Figure 1 the asymptotic variance, relative to the minimal variance of the mle, for the cases of full likelihood (FL), pseudo likelihood ($|A_j| = 1$) $\text{PL}_1$, stochastic combination of pseudo likelihood and 2nd order pseudo likelihood ($|A_j| = 2$) components $\alpha\text{PL}_2 + (1-\alpha)\text{PL}_1$, stochastic combination of 2nd order pseudo likelihood and 3rd order pseudo likelihood ($|A_j| = 3$) components $\alpha\text{PL}_3 + (1-\alpha)\text{PL}_2$, and stochastic combination of 3rd order pseudo likelihood and 4th order pseudo likelihood ($|A_j| = 4$) components $\alpha\text{PL}_4 + (1-\alpha)\text{PL}_3$.

The graph demonstrates the computation-accuracy tradeoff as follows: (a) pseudo likelihood is the fastest but also the least accurate, (b) full likelihood is the slowest but the most accurate, (c) adding higher order components reduces the asymptotic variance but also requires more computation, (d) the variance reduces with the increase in the selection probability $\alpha$ of the higher order component, and (e) adding 4th order components brings the variance very close the lower limit and with each successive improvement becoming smaller and smaller according to a law of diminishing returns.

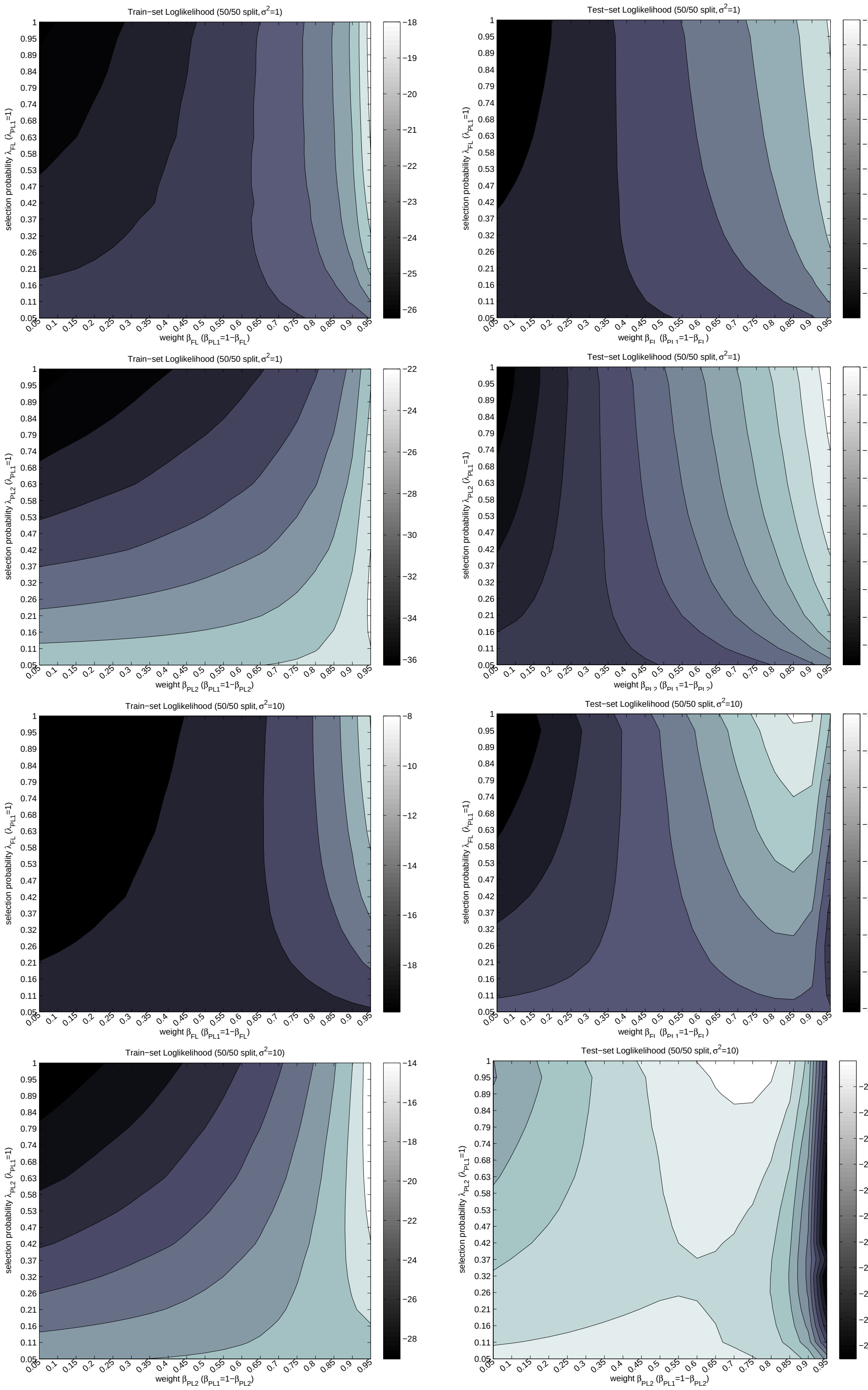

Figure 2: Train (left) and test (right) loglikelihood contours for maximum scl estimators for the CRF model. $L_2$ regularization parameters are $\sigma^2 = 1$ (rows 1,2) and $\sigma^2 = 10$ (rows 3,4). Rows 1,3 are stochastic mixtures of full (FL) and pseudo ($\text{PL}_1$) loglikelihood components while rows 2,4 are pseudo ($\text{PL}_1$) and 2nd order pseudo.

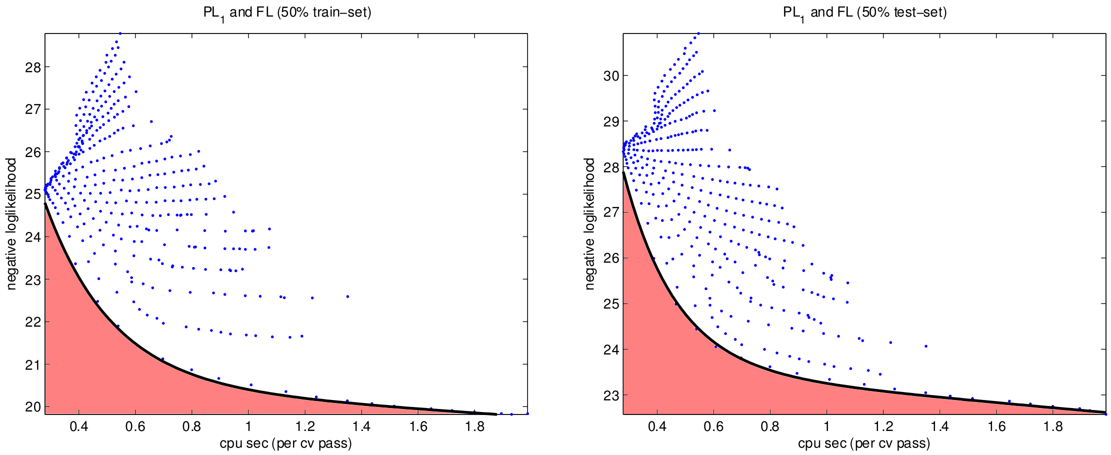

Figure 3: Scatter plot representing complexity and negative loglikelihood (left:train, right:test) of scl functions for CRFs with regularization parameter $\sigma^2 = 1/2$. The points represent different stochastic combinations of full and pseudo likelihood components. The shaded region represents impossible accuracy/complexity demands.

## 7.2 Conditional Random Fields

To demonstrate the complexity-accuracy tradeoff in a more realistic scenario we experimented with regularized maximum scl estimators for conditional random fields (CRF). We trained and tested the CRF models on local sentiment prediction data obtained in Mao and Lebanon (2007). The data consisted of 249 movie review documents having an average of 30.5 sentences each with an average of 12.3 words from a 12633 word vocabulary. Each sentence was manually labeled as one of five sentimental designations: very negative, negative, objective, positive, or very positive.

Figure 2 contains the contour plots of train and test loglikelihood as a function of the scl parameters: weight $\beta$ and selection probability $\lambda$. The likelihood components were mixtures of full and pseudo ($|A_j| = 1$) likelihood (rows 1,3) and pseudo and 2nd order pseudo ($|A_j| = 2$) likelihood (rows 2,4). $A_j$ identifies a set of labels corresponding to adjacent sentences over which the probabilistic query is evaluated. Results were averaged over 100 cross validation iterations with 50% train-test split. We used BFGS quasi-Newton method for maximizing the regularized scl functions. Figure 2 demonstrates how the train loglikelihood increases with increasing the weight and selection probability of full likelihood in rows 1,3 and of 2nd order pseudo likelihood in rows 2,4. This increase in train loglikelihood is also correlated with an increase in computational complexity as higher order likelihood components require more computation.

It is interesting to contrast the test loglikelihood behavior in the case of mild ($\sigma = 10$) and stronger ($\sigma = 1$) $L_2$ regularization. In the case of weaker or no regularization, the test loglikelihood shows different behavior than the train loglikelihood. Adding a lower order component such as pseudo likelihood acts as a regularizer that prevents overfitting. Thus, in cases that are prone to overfitting reducing higher order likelihood components improves both performance as well as complexity. This represents a win-win situation in contrast to the classical view where the mle has the lowest variance and adding lower order components reduces complexity but increases the variance.

Figure 3 displays the complexity and negative loglikelihoods (left:train, right:test) of different scl estimators, sweeping through $\lambda$ and $\beta$, as points in a two dimensional space. The shaded area near the origin is unachievable as no scl estimator can achieve high accuracy and low computation at the same time. The optimal location in this 2D plane is the curved boundary of the achievable region with the exact position on that boundary depending on the required solution of the computation-accuracy tradeoff. Note that a particular $\lambda$ indeed has a dominant $\beta$, however relative comparison of $\lambda$ is meaningless as its choice is a function of available computational resources and time.

# 8 Discussion

The proposed estimator family facilitates computationally efficient estimation in complex graphical models. In particular, different parameterizations of the stochastic likelihood enables the resolution of the complexity-accuracy tradeoff in a domain and problem specific manner. The framework is generally suited for Markov random fields, including conditional graphical models and is theoretically motivated. When the model is prone to overfit, stochastically mixing lower order components with higher order ones acts as a regularizer and results in a win-win situation of improving test-set accuracy and reducing computational complexity at the same time.

## Acknowledgements

The authors acknowledge useful discussion with P. Liang and S.-C. Zhu. The proof of Proposition 2 was clarified following useful feedback from the anonymous referees.

# References

B. Arnold and D. Strauss. Pseudolikelihood estimation: some examples. *Sankhya B*, 53:233–243, 1991.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J Roy Statist Soc B*, 36 (2):192–236, 1974.

R. Casella and C. Robert. *Monte Carlo Statistical Methods*. Springer Verlag, second edition, 2004.

T. S. Ferguson. *A Course in Large Sample Theory*. Chapman & Hall, 1996.

N. Hjort and C. Varin. ML, PL, and QL in markov chain models. *Scand J Stat*, 35(1):64–82, 2008.

G. Liang and B. Yu. Maximum pseudo likelihood estimation in network tomography. *IEEE T Signal Proces*, 51(8): 2043–2053, 2003.

P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proc. of the International Conference on Machine Learning*, 2008.

B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

Y. Mao and G. Lebanon. Isotonic conditional random fields and local sentiment flow. In *Advances in Neural Information Processing Systems 19*, pages 961–968, 2007.

C. Sutton and A. McCallum. Piecewise pseudolikelihood for efficient training of conditional random fields. In *Proc. of the International Conference on Machine Learning*, 2007.

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

E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence*, 2003.

S.-C. Zhu and X. Liu. Learning in Gibbsian fields: How accurate and how fast can it be? *IEEE T Pattern Anal*, 24 (7):1001–1006, 2002.