Journal of Machine Learning Research 11 (2010) 2597-2633 Submitted 3/10; Revised 8/10; Published 10/10

Stochastic Composite Likelihood

Joshua V. Dillon JVDILLON@GATECH.EDU
Guy Lebanon LEBANON@CC.GATECH.EDU
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA

Editor: Fernando Pereira

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. Each of the estimators resolve the computation-accuracy trade-off differently, and taken together they span a continuous spectrum of computation-accuracy trade-off resolutions. We prove the consistency of the estimators, provide formulas for their asymptotic variance, statistical robustness, and computational complexity. We discuss experimental results in the context of Boltzmann machines and conditional random fields. The theoretical and experimental studies demonstrate the effectiveness of the estimators when the computational resources are insufficient. They also demonstrate that in some cases reduced computational complexity is associated with robustness thereby increasing statistical accuracy.

Keywords: Markov random fields, composite likelihood, maximum likelihood estimation

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

$$D = (X^{(1)}, \ldots, X^{(n)}), \quad X^{(i)} \in \mathbb{R}^m, \tag{1}$$

and is sampled iid from a parametric distribution p_{θ_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 log-likelihood function

$$\ell_n(\theta; D) = \sum_{i=1}^{n} \log p_\theta(X^{(i)}), \tag{2}$$
$$\hat{\theta}_n^{\text{ml}} = \arg\max_{\theta \in \Theta} \ell_n(\theta; D).$$

The use of the MLE is motivated by its consistency,[1] that is, $\hat{\theta}_n^{\text{ml}} \to \theta_0$ 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 θ_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 θ_0 and variance matrix $(nI(\theta_0))^{-1}$. More formally, we have the

1. The consistency $\hat{\theta}_n^{\text{ml}} \to \theta_0$ with probability 1 is sometimes called strong consistency in order to differentiate it from the weaker notion of consistency in probability $P(|\hat{\theta}_n^{\text{ml}} - \theta_0| < \varepsilon) \to 0$.

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

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

with ∇f representing the $r \times 1$ gradient vector of $f(\theta)$ with respect to θ. 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 log-likelihood (2) and its derivatives is impossible due to computational considerations. For instance this is the situation in many high dimensional exponential family distributions, including Markov random fields whose graphical structure contains cycles. 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.

In contrast to asymptotic efficiency, we view consistency as a less negotiable property and prefer to avoid inconsistent estimators if at all possible. This common viewpoint in statistics is somewhat at odds with recent advances in the machine learning literature promoting non-consistent estimators, for example using variational techniques (Jordan et al., 1999). Nevertheless, we feel that there is a consensus regarding the benefits of having consistent estimators over non-consistent ones.

In this paper, we propose a family of estimators, for use in situations where the computation of the MLE is intractable. In contrast to many 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. We also demonstrate that in some cases reduced computational complexity may in fact act as a regularizer, increasing robustness and therefore accomplishing both reduced computation and increased accuracy. This "win-win" situation conflicts with the conventional wisdom stating that moving from the MLE to pseudo likelihood and other related estimators result in a computational win but a statistical loss. Nevertheless we show that this occurs in some practical situations.

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. We include experiments on both simulated and real world data for several models including the Boltzmann machine, conditional random fields, and the Boltzmann linear chain model. Appendix B outlines a road-map of figures and the corresponding exposition.

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 log-likelihood and its gradient (Casella and Robert, 2004). Variational methods are techniques for conducting inference and learning based on tractable bounds (Jordan et al., 1999). A similar approach would be to conduct maximum likelihood estimation for a simpler model that is tractable.

Despite the substantial work on MCMC and variational methods, there are little practical results concerning the convergence and approximation rate of the resulting parameter estimators. Variational techniques are sometimes inconsistent and it is hard to analyze their asymptotic statistical behavior. In the case of MCMC, a number of asymptotic results exist (Casella and Robert, 2004), but since MCMC plays a role inside each gradient descent or EM iteration it is hard to analyze the asymptotic behavior of the resulting parameter estimates. An advantage of our framework is that we are able to directly characterize the asymptotic behavior of the estimator and relate it to the amount of computational savings.

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) and 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 include the experimental study of texture generation by Zhu and Liu (2002), the work of Xing et al. (2003) which focused on inference rather than parameter estimation, and the examination of generalization performance of small- and large- scale learning systems by Bottou and Bousquet (2008). The work of Liang and Jordan (2008) is also interesting in that the authors employ composite likelihood m-estimators and asymptotic arguments to compare the risk of discriminative and generative models. However, our work differs in theme and technique—we explore the tradeoff between computation and accuracy by way of a fundamentally different estimator.

Composite likelihood techniques, and consequently our work, can be thought of as local contrastive objectives (i.e., pseudo likelihood, contrastive divergence). Vickrey et al. (2010) present a non-local alternative in which the objective is not restricted to using the training label, but rather any assignment.

3. Stochastic Composite Likelihood

In many cases, the absence of a closed form expression for the normalization term prevents the computation of the log-likelihood (2) and its derivatives thereby severely limiting the use of the MLE. A popular example is Markov random fields, wherein the computation of the normalization term is often intractable (see Section 6 for more details). In this paper we propose alternative estimators based on the maximization of a stochastic variation of the composite likelihood.

We denote multiple samples using superscripts and individual dimensions using subscripts. Thus $X_j^{(r)}$ refers to the j-dimension of the r sample. Following standard convention we refer to random variables (RV) using uppercase letters and their corresponding values using lowercase letters. We also use the standard notations for extracting a subset of the dimensions of a random

variable

$$X_S \stackrel{\text{def}}{=} \{X_i : i \in S\}, \qquad X_{-j} \stackrel{\text{def}}{=} \{X_i : i \neq j\}. \tag{4}$$

We start by reviewing the pseudo log-likelihood function (Besag, 1974) associated with the data D (1),

$$p\ell_n(\theta; D) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sum_{j=1}^{m} \log p_\theta(X_j^{(i)} | X_{-j}^{(i)}). \tag{5}$$

The maximum pseudo likelihood estimator (MPLE) $\hat{\theta}_n^{\text{mpl}}$ is consistent, that is, $\hat{\theta}_n^{\text{mpl}} \to \theta_0$ with probability 1, but possesses considerably higher asymptotic variance than 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

$$p_\theta(X_j | X_{-j}) = p_\theta(X_j | \{X_k : k \neq j\}) = \frac{p_\theta(X)}{\sum_{x_j} p_\theta(X_1, \ldots, X_{j-1}, X_j = x_j, X_{j+1}, \ldots, X_m)}.$$

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 $A,B \subset \{1,\ldots,m\}$ satisfying $A \neq \emptyset = A \cap B$. The likelihood object associated with an m-pair (A,B) and X is $S_\theta(A,B) \stackrel{\text{def}}{=} \log p_\theta(X_A | X_B)$ where X_S is defined in* (4)*. The composite log-likelihood function (Lindsay, 1988) is a collection of likelihood objects defined by a finite sequence of m-pairs $(A_1,B_1),\ldots,(A_k,B_k)$*

$$c\ell_n(\theta; D) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sum_{j=1}^{k} \log p_\theta(X_{A_j}^{(i)} | X_{B_j}^{(i)}). \tag{6}$$

There is a certain lack of flexibility associated with the composite likelihood framework as each likelihood object is either selected or not for the entire sample $X^{(1)},\ldots,X^{(n)}$. There is no allowance for some objects to be selected more frequently than others. For example, available computational resources may allow the computation of the log-likelihood for 20% of the samples, and the pseudo likelihood for the remaining 80%. In the case of composite likelihood if we select the full likelihood component (or the pseudo likelihood or any other likelihood object) then this component is applied to all samples indiscriminately.

In SCL, different likelihood objects $S_\theta(A_j,B_j)$ may be selected for different samples with the possibility of some likelihood objects being selected for only a small fraction of the data samples.

The selection may be non-coordinated, in which case each component is selected or not independently of the other components. Or it may be coordinated in which case the selection of one component depends on the selection of the other ones. For example, we may wish to avoid selecting a pseudo likelihood component for a certain sample $X^{(i)}$ if the full likelihood component was already selected for it.

Another important advantage of stochastic selection is that the discrete parameterization of (6) defined by the sequence $(A_1,B_1),\ldots,(A_k,B_k)$ is less convenient for theoretical analysis. Each component is either selected or not, turning the problem of optimally selecting components into a hard combinatorial problem. The stochastic composite likelihood, which is defined below, enjoys continuous parameterization leading to more convenient optimization techniques and convergence analysis.

Definition 2 *Consider a finite sequence of m-pairs $(A_1,B_1),\ldots,(A_k,B_k)$, a data set $D=(X^{(1)},\ldots,X^{(n)})$, $\beta\in\mathbb{R}_+^k$, and n iid, length k, binary random vectors $Z^{(1)},\ldots,Z^{(n)}\overset{\text{i.i.d.}}{\sim}P(Z)$ with $\lambda_j\overset{\text{def}}{=}E(Z_j)>0$. The stochastic composite log-likelihood (SCL) is*

$$sc\ell_n(\theta;D,Z)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}m_\theta(X^{(i)},Z^{(i)}),\quad \textit{where}$$

$$m_\theta(X,Z)\overset{\text{def}}{=}\sum_{j=1}^{k}\beta_j Z_j\log p_\theta(X_{A_j}|X_{B_j}), \tag{7}$$

where, for brevity, we typically omit D,Z in favor of $sc\ell_n(\theta)$.

In other words, the SCL is a stochastic extension of (6) 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 either selected or not, depending on the values of the binary random variables $Z_1^{(i)},\ldots,Z_k^{(i)}$ and weighted by the constants $\beta_1,\ldots,\beta_k$. Note that $Z_j^{(i)}$ may in general depend on $Z_r^{(i)}$ but not on $Z_r^{(l)}$ or on $X^{(i)}$.

When we focus on examining different models for $P(Z)$ we sometimes parameterize it, for example by λ, that is, $P_\lambda(Z)$. This reuse of λ (it is also used in Definition 2) is a notational abuse. We accept it, however, as in most of the cases that we consider $\lambda_1,\ldots,\lambda_k$ from Definition 2 either form the parameter vector for $P(Z)$ or are part of it. Often we refer to a particular λ as a "policy" in order to emphasize its role as a "knob" in selecting particular *m*-pairs.

Some illustrative examples follow.

Independence. Factorizing $P_\lambda(Z_1,\ldots,Z_k)=\prod_j P_{\lambda_j}(Z_j)$ leads to $Z_j^{(i)}\sim\text{Ber}(\lambda_j)$ with complete independence among the indicator variables. For each sample $X^{(i)}$, each likelihood object $S(A_j,B_j)$ is selected or not independently with probability λ_j.

Multinomial. A multinomial model $Z\sim\text{Mult}(1,\lambda)$ implies that for each sample $Z^{(i)}$ a multivariate Bernoulli experiment is conducted with precisely one likelihood object being selected depending on the selection probabilities $\lambda_1,\ldots,\lambda_k$.

Product of Multinomials. A product of multinomials is formed by a partition of the dimensions to l disjoint subsets $\{1,\ldots,m\}=C_1\cup\cdots C_l$ where $Z_{C_i}\sim\text{Mult}(1,(\lambda_j:j\in C_i))$, that is,

$$P(Z)=\prod_{i=1}^{c}P_i\left(\{Z_j:j\in C_i\}\right),\quad \text{where } P_i \text{ is } \text{Mult}(1,(\lambda_j:j\in C_l)).$$

Loglinear Models. The distribution $P(Z)$ follows a hierarchical loglinear model (Bishop et al., 1975). This case subsumes the other cases above.

In analogy to the MLE and the MPLE, the maximum SCL estimator (MSCLE) $\hat{\theta}_n^{\text{msl}}$ estimates θ_0 by maximizing the SCL function. In contrast to the log-likelihood and pseudo log-likelihood functions, the SCL function and its maximizer are random variables that depend on the indicator variables $Z^{(1)},\ldots,Z^{(n)}$ in addition to the data D. As such, its behavior should be summarized by examining the limit $n\to\infty$. Doing so eliminates the dependency on particular realizations of $Z^{(1)},\ldots,Z^{(n)}$ in favor of the the expected frequencies $\lambda_j = \mathsf{E}_{P(Z)}Z_j$ which are non-random constants.

The statistical accuracy and computational complexity of the SCL estimator are continuous functions of the parameters (β,λ) (component weights and selection probabilities respectively) which vary continuously throughout their domain $(\lambda,\beta)\in\Lambda\times\mathbb{R}_+^k$. Choosing appropriate values of (λ,β) retrieves the special cases of MLE, MPLE, maximum composite likelihood with each selection being associated with a distinct statistical accuracy and computational complexity. The SCL framework allows selections of many more values of (λ,β) realizing a wide continuous spectrum of estimators, each resolving the accuracy-complexity tradeoff differently.

We include below a demonstration of the SCL framework in a simple low dimensional case. In the following sections we discuss in detail the statistical behavior of the MSCLE and its computational complexity. We conclude the paper with several experimental studies.

3.1 Boltzmann Machine Example

Before proceeding we illustrate the SCL framework using a simple example involving a Boltzmann machine (Hinton and Sejnowski, 1983). Section 8.1 describes the specifics of this model. We consider in detail three SCL policies: full likelihood (FL), pseudo likelihood (PL), and a stochastic combination of first and second order pseudo likelihood with the first order components $p(X_i|X_{-i})$ selected with probability λ and the second order components $p(X_i,X_j|X_{\{i,j\}^c})$ selected with probability $1-\lambda$.

Denoting the number of (binary) graph vertices, or nodes, by m, and the number of examples by n, the computational complexity of the FL function measured in FLOP[2] counts is $O\big(\binom{m}{2}(2^m+n)\big)$ (for the log-likelihood) and $O\big(\binom{m}{2}^2 2^m + n\binom{m}{2}\big)$ for the log-likelihood gradient.[3] The exponential growth in m prevents such computations for large graphs.

The k-order PL function offers a practical alternative to FL (1-order PL corresponds to the traditional pseudo likelihood and 2-order PL its analog, $p(X_{\{i,j\}}|X_{\{i,j\}^c})$). The complexity of computing the corresponding SCL function is $O\big(\binom{m}{2}(\binom{m}{k}2^k+n)\big)$ for the objective function and $O\big(\binom{m}{2}^2\binom{m}{k}2^k + n\binom{m}{2}\big)$ for its gradient. The slower complexity growth of the k-order PL (polynomial in m instead of exponential) is offset by its reduced statistical accuracy, which we measure using the normalized asymptotic variance

$$\text{eff}(\hat{\theta}_n) = \frac{\det(\text{Asymp Var}(\hat{\theta}_n))}{\det(\text{Asymp Var}(\hat{\theta}_n^{\text{ml}}))} \tag{8}$$

2. FLOP is the number of FLoating point OPerations.

3. With memoization the complexity of the gradient can be reduced to $O\big(\binom{m}{2}2^m + n\binom{m}{2}\big)$ (at the cost of exponential 2^m storage). Note that this is only a polynomial improvement to an exponential complexity hence we lose no insight by making naive assumptions.

which is bounded from below by 1 (due to Cramer Rao lower bound) and its deviation from 1 reflects its inefficiency relative to the MLE.

The MLE thus achieves the best accuracy but it is computationally intractable. The first order and second order PL have higher asymptotic variance but are easier to compute. The SCL framework enables adding many more estimators filling in the gaps between ML, 1-order PL, 2-order PL, etc.

We illustrate three SCL functions in the context of a simple Boltzmann machine (five binary nodes, fourteen samples $X^{(1)},\ldots,X^{(14)}$, $\theta^{\text{true}} = (-1,-1,-1,-1,-1,1,1,1,1,1)$) in Figure 1. The top box refers to the full likelihood policy, that is, maximum likelihood. For each of the fourteen samples, the FL component is computed and their aggregation forms the SCL function which in this case equals the log-likelihood. The selection of the FL component for each sample is illustrated using a diamond box. The values under the boxes reflect the FLOP counts needed to compute the components and the total complexity associated with computing the entire SCL or log-likelihood is listed on the right. As mentioned above, the normalized asymptotic variance (8) is 1.

The pseudo likelihood function (5) is illustrated in the second box where each row correspond to one of the five PL components. As each of the five PL component is selected for each of the samples we have diamond boxes covering the entire 5×14 array. The shade of the diamond boxes reflects the complexity required to compute them enabling an easy comparison to the FL components in the top of the figure (note how the FL boxes are much darker than the PL boxes). The numbers at the bottom of each column reflect the FLOP marginal count for each of the fourteen samples and the numbers to the right of the rows reflect the FLOP marginal count for each of the PL components. In this case the FLOP count is less than half the FLOP count of the FL in top box (this reduction in complexity obtained by replacing FL with PL will increase dramatically for graphs with more than 5 nodes) but the asymptotic variance is 83% higher.[4]

The third SCL policy reflects a stochastic combination of first and second order pseudo likelihood components. Each first order component is selected with probability λ and each second order component is selected with probability $1-\lambda$. The result is a collection of 5 1-order PL components and 10 2-order components with only some of them selected for each of the fourteen samples. Again diamond boxes correspond to selected components which are shaded according to their FLOP complexity. The per-component FLOP marginals and per example FLOP marginals are listed as the bottom row and right-most column. The total complexity is somewhere between FL and PL and the asymptotic variance is reduced from the PL's 183% to 148%.

Additional insight may be gained at this point by considering Figure 3 which plots several SCL estimators as points in the plane whose *x* and *y* coordinates correspond to normalized asymptotic variance and computational complexity respectively. We turn at this point to considering the statistical properties of the SCL estimators.

4. Consistency and Asymptotic Variance of $\hat{\theta}_n^{\text{msl}}$

A nice property of the SCL framework is enabling mathematical characterization of the statistical properties of the estimator $\hat{\theta}_n^{\text{msl}}$. In this section we examine the conditions for consistency of the MSCLE and its asymptotic distribution and in the next section we consider robustness. The propositions below constitute novel generalizations of some well-known results in classical statistics. Proofs may be found in Appendix A. For simplicity, we assume that X is discrete and $p_\theta(x) > 0$.

4. The asymptotic variance of SCL functions is computed using formulas derived in the next section.

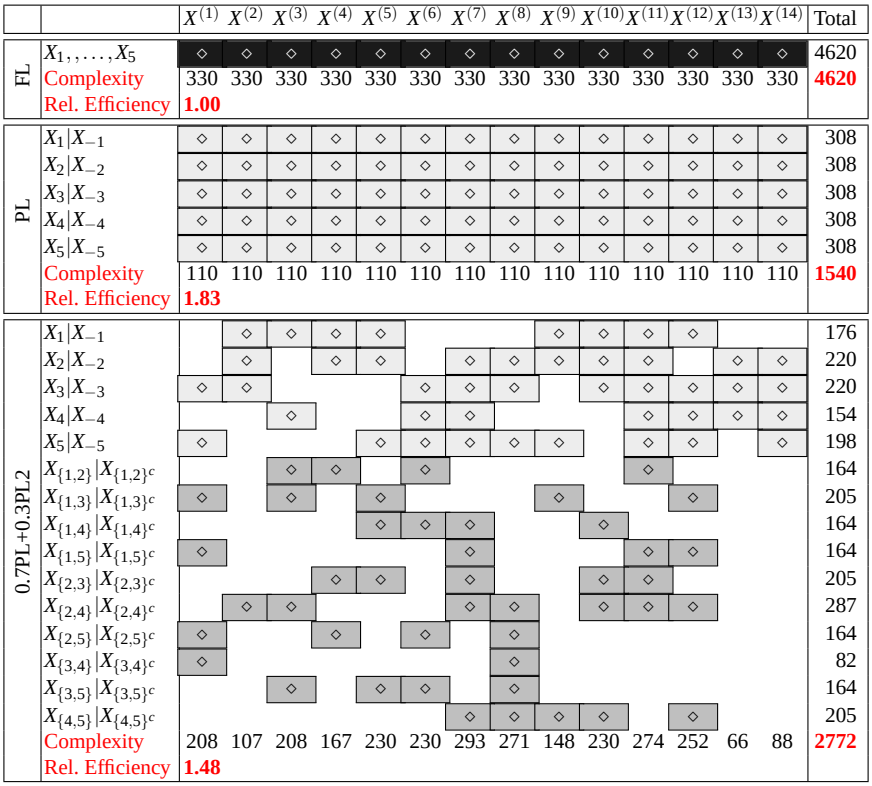

Figure 1: Sample runs of three different SCL policies for 14 examples $X^{(1)},\ldots,X^{(14)}$ drawn from a 5 binary node Boltzmann machine ($\theta^{\text{true}} = (-1,-1,-1,-1,-1,1,1,1,1,1)$). The policies are full likelihood (FL, top), pseudo likelihood (PL, middle), and a stochastic combination of first and second order pseudo likelihood with the first order components selected with probability 0.7 and the second order components with probability 0.3 (bottom).

The sample runs for the policies are illustrated by placing a diamond box in table entries corresponding to selected likelihood objects (rows corresponding to likelihood objects and columns to $X^{(1)},\ldots,X^{(14)}$). The FLOP counts of each likelihood object determines the shade of the diamond boxes while the total FLOP counts per example and per likelihood objects are displayed as table marginals (bottom row and right column for each policy). We also display the total FLOP count and the normalized asymptotic variance (8).

Even in the simple case of 5 nodes, FL is the most complex policy with PL requiring a third of the FL computation. 0.7PL+0.3PL2 is somewhere in between. The situation is reversed for the estimation accuracy-FL achieves the lowest possible normalized asymptotic variance of 1, PL is almost twice that, and 0.7PL+0.3PL2 somewhere in the middle. The SCL framework spans the accuracy-complexity spectrum. Choosing the right λ value obtains an estimator that suits available computational resources and required accuracy.

Definition 3 *A sequence of m-pairs $(A_1,B_1),\ldots,(A_k,B_k)$ is m-pair identifiable, or simply identifiable, of p_θ 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 *Let $\Theta \subset \mathbb{R}^r$ be an open set, $p_\theta(x) > 0$ and continuous and smooth in θ, and $(A_1,B_1),\ldots,(A_k,B_k)$ be a sequence of m-pairs for which $\{(A_j,B_j) : \forall j$ such that $\lambda_j > 0\}$ ensures identifiability. Then the sequence of SCL maximizers is strongly consistent, that is,*

$$P\left(\lim_{n\to\infty} \hat{\theta}_n = \theta_0\right) = 1.$$

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, that is,

$$\mathcal{S} = \{(A_1,B_1),\ldots,(A_m,B_m)\} \quad \text{where} \quad A_i = \{i\}, B_i = \{1,\ldots,m\} \setminus A_i \tag{9}$$

ensures identifiability and consequently the consistency of the MSCLE estimator. Furthermore, every selection of *m*-pairs that subsumes $\mathcal{S}$ in (9) similarly guarantees identifiability and consistency.

The proposition below establishes the asymptotic normality of the MSCLE $\hat{\theta}_n$. The asymptotic variance enables the comparison of SCL functions with different parameterizations (λ,β).

Proposition 2 *Making the assumptions of Proposition 1 as well as convexity of $\Theta \subset \mathbb{R}^r$ we have the following convergence in distribution*

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

where

$$\Upsilon^{-1} = \sum_{j=1}^{k} \beta_j\lambda_j Var_{\theta_0}(\nabla S_{\theta_0}(A_j,B_j)),$$

$$\Sigma = Var_{\theta_0}\left(\sum_{j=1}^{k} \beta_j\lambda_j \nabla S_{\theta_0}(A_j,B_j)\right).$$

The notation $\mathsf{Var}_{\theta_0}(Y)$ represents the covariance matrix of the random vector Y under p_{θ_0} while the notations $\xrightarrow{p}, \rightsquigarrow$ in the proof below denote convergences in probability and in distribution (Ferguson, 1996). ∇ represents the gradient vector with respect to θ.

When θ is a vector the asymptotic variance is a matrix. To facilitate comparison between different estimators we follow the convention of using the determinant, and in some cases the trace, to measure the statistical accuracy. See Chapter 4 of Serfling (1980) for some heuristic arguments for doing so. Figures 1,2,3 provide the asymptotic variance for some SCL estimators and describe how it can be used to gain insight into which estimator to use.

The fact that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a Gaussian with zero mean (for the MLE and similarly for SCL estimators as we show above) implies that the estimator's asymptotic behavior, up to $n^{-1/2}$ order, is determined exclusively by the asymptotic variance. That means that the estimator is essentially unbiased up to that order. Higher order statistical analysis (obtained using

Taylor series with more terms) show that the bias decays in the faster rate of n^{-1} (Cox and Snell, 1968). We thus follow the statistical convention of conducting first order asymptotic analysis and concentrate on the estimator's asymptotic variance.

The statistical accuracy of the SCL estimator depends on β (weight parameters) and λ (selection parameter). It is thus desirable to use the results in this section in determining what values of β, λ to use. Directly using the asymptotic variance is not possible in practice as it depends on the unknown quantity θ_0. However, it is possible to estimate the asymptotic variance using the training data. We describe this in Section 7.

5. Robustness of $\hat{\theta}_n^{\mathbf{msl}}$

We observed in our experiments (see Section 8) that the SCL estimator sometimes performed better on a held-out test set than did the maximum likelihood estimator. This phenomenon seems to be in contradiction to the fact that the asymptotic variance of the MLE is lower than that of the SCL maximizer. This is explained by the fact that in some cases the true model generating the data does not lie within the parametric family $\{p_\theta : \theta \in \Theta\}$ under consideration. For example, many graphical models (HMM, CRF, LDA, etc.) make conditional independence assumptions that are often violated in practice. In such cases the SCL estimator acts as a regularizer achieving better test set performance than the non-regularized MLE. We provide below a theoretical account of this phenomenon using the language of *m*-estimators and statistical robustness. Our notation follows the one in van der Vaart (1998).

We assume that the model generating the data is outside the model family $P(X) \notin \{p_\theta : \theta \in \Theta\}$ and we extend the notation of $m_\theta(X,Z)$ in (7) with,

$$\begin{aligned}
\psi_\theta(X,Z) &\stackrel{\text{def}}{=} \nabla m_\theta(X,Z), \\
\dot{\psi}_\theta(X,Z) &\stackrel{\text{def}}{=} \nabla^2 m_\theta(X,Z), \text{ and} \\
\Psi_n(\theta) &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X^{(i)}, Z^{(i)}),
\end{aligned}$$

noting that $\dot{\psi}_\theta(X,Z)$ is a matrix of second order derivatives.

Proposition 3 below generalizes the consistency result by asserting that $\hat{\theta}_n \to \theta_0$ where θ_0 is the point on $\{p_\theta : \theta \in \Theta\}$ that is closest to the true model P, as defined by

$$\theta_0 = \arg\max_{\theta\in\Theta} M(\theta) \quad \text{where} \quad M(\theta) \stackrel{\text{def}}{=} -\sum_{j=1}^{k} \beta_j \lambda_j D(P(X_{A_j}|X_{B_j}) || p_\theta(X_{A_j}|X_{B_j})),$$

or equivalently, θ_0 satisfies

$$\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\psi_{\theta_0}(X,Z) = 0.$$

When the SCL function reverts to the log-likelihood function, θ_0 becomes the KL projection of the true model P onto the parametric family $\{p_\theta : \theta \in \Theta\}$.

Proposition 3 *Assuming the conditions in Proposition 1 as well as* $\sup_{\theta:\|\theta-\theta_0\|\geq\varepsilon} M(\theta) < M(\theta_0)$ *for all* $\varepsilon > 0$ *we have* $\hat{\theta}_n^{msl} \to \theta_0$ *as* $n \to \infty$ *with probability 1.*

The added condition maintains that θ_0 is a well separated maximum point of M. In other words it asserts that only values close to θ_0 may yield a value of M that is close to the maximum $M(\theta_0)$. This condition is satisfied in the case of most exponential family models.

Proposition 4 *Assuming the conditions of Proposition 2 as well as* $E_{P(X)}E_{P(Z)}\|\psi_{\theta_0}(X,Z)\|^2 < \infty$, $E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0}(X)$ *exists and is non-singular,* $|\ddot{\Psi}_{ij}| = |\partial^2\psi_\theta(x)/\partial\theta_i\theta_j| < g(x)$ *for all* i, j *and* θ *in a neighborhood of* θ_0 *for some integrable* g*, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X^{(i)},Z^{(i)}) + o_P(1) \tag{10}$$

or equivalently

$$\hat{\theta}_n = \theta_0 - (E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0})^{-1}\frac{1}{n}\sum_{i=1}^{n}\psi_{\theta_0}(X^{(i)},Z^{(i)}) + o_P\left(\frac{1}{\sqrt{n}}\right). \tag{11}$$

Above, $f_n = o_P(g_n)$ means f_n/g_n converges to 0 with probability 1.

Corollary 1 *Assuming the conditions specified in Proposition 4 we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, (E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0})^{-1}(E_{P(X)}E_{P(Z)}\psi_{\theta_0}\psi_{\theta_0}^{\top})(E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0})^{-1}). \tag{12}$$

Equation (11) means that asymptotically, $\hat{\theta}_n$ behaves as θ_0 plus the average of iid RVs. As mentioned in van der Vaart (1998) this fact may be used to obtain a convenient expression for the asymptotic influence function, which measures the effect of adding a new observation to an existing large data set. Neglecting the remainder in (10) we have

$$\begin{aligned}
I(x,z) &\stackrel{\text{def}}{=} \hat{\theta}_n(X^{(1)},\ldots,X^{(n-1)},x,Z^{(1)},\ldots,Z^{(n-1)},z) - \hat{\theta}_{n-1}(X^{(1)},\ldots,X^{(n-1)},Z^{(1)},\ldots,Z^{(n-1)}) \\
&\approx -(\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_{\theta_0})^{-1}\left(\frac{1}{n}\sum_{i=1}^{n-1}\psi_{\theta_0}(X^{(i)},Z^{(i)}) + \frac{1}{n}\psi_{\theta_0}(w,z) - \frac{1}{n-1}\sum_{i=1}^{n-1}\psi_{\theta_0}(X^{(i)},Z^{(i)})\right) \\
&= -(\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_{\theta_0})^{-1}\frac{1}{n}\psi_{\theta_0}(w,z) + (\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_{\theta_0})^{-1}\frac{1}{n(n-1)}\sum_{i=1}^{n-1}\psi_{\theta_0}(X^{(i)},Z^{(i)}) \\
&= -\frac{1}{n}(\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_{\theta_0})^{-1}\psi_{\theta_0}(w,z) + o_P\left(\frac{1}{n}\right).
\end{aligned} \tag{13}$$

Corollary 1 and Equation (13) measure the statistical behavior of the estimator when the true distribution is outside the model family. In these cases it is possible that a computationally efficient SCL maximizer will result in higher statistical accuracy as well. This "win-win" situation of improving in both accuracy and complexity over the MLE is confirmed by our experiments in Section 8.

We finally note that the above analysis is not limited to misspecified models. For example, the influence function may be used to detect the robustness of $\hat{\theta}_n$ to outliers or rare events (it is desirable to be robust to such occurrences even if the model is not misspecified).

6. 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 $G = (V,E)$, $V = \{1,\ldots,m\}$ with a clique set $\mathcal{C}$ is given by the following exponential family model

$$p_\theta(x) = \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C(x_C) - \log Z(\theta)\right),$$
$$Z(\theta) = \sum_x \exp\left(\sum_{C\in\mathcal{C}} \theta_c f_C(x_C)\right). \tag{14}$$

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 (14) are given by (note the cancellation of $Z(\theta)$)

$$p_\theta(x_A|x_B) = \frac{\sum_{x'_{(A\cup B)^c}} \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C((x_A,x_B,x'_{(A\cup B)^c})_C)\right)}{\sum_{x'_{(A\cup B)^c}}\sum_{x''_A} \exp\left(\sum_{C\in\mathcal{C}} \theta_C f_C((x''_A,x_B,x'_{(A\cup B)^c})_C)\right)}. \tag{15}$$

whose computation is substantially faster. Specifically, The computation of (15) 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 $k \leq m!/(l!(m-l)!)$ 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.

7. Automatic Selection of β

As Proposition 2 indicates, the weight vector β and selection probabilities λ 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 λ independently of β. Conceptually, we are interested in resolving the accuracy-complexity tradeoff jointly for both β,λ before estimating θ by maximizing the SCL function. However, since the computational complexity depends only on λ we propose the following simplified problem: Select λ based on available computational resources, and then given λ, select the β (and θ) that will achieve optimal statistical accuracy.

Selecting β that minimizes the asymptotic variance is somewhat ambiguous as $\Upsilon\Sigma\Upsilon$ in Proposition 2 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,[5] and minimize

$$J(\beta) = \log\det(\Upsilon\Sigma\Upsilon) = \log\det\Sigma + 2\log\det\Upsilon. \tag{16}$$

5. See Chapter 4 of Serfling (1980) for a heuristic discussion motivating this measure.

A major complication with selecting β based on the optimization of (16) is that it depends on the true parameter value θ_0 which is not known at training time. This may be resolved, however, by noting that (16) is composed of covariance matrices under θ_0 which may be estimated using empirical covariances over the training set. To facilitate fast computation of the optimal β we also propose to replace the determinant in (16) with the product of the diagonal elements. Such an approximation is motivated by Hadamard's inequality (which states that for symmetric matrices $\det(M) \leq \prod_i M_{ii}$) and by Geršgorin's circle theorem (see below). This approximation works well in practice as we observe in the experiments section. We also note that the procedure described below involves only simple statistics that may be computed on the fly and does not contribute significant additional computation (nor do they require significant memory).

More specifically, we denote $K^{(ij)} = \mathsf{Cov}_{\theta_0}(\nabla S_{\theta_0}(A_i, B_i), \nabla S_{\theta_0}(A_j, B_j))$ with entries $K^{(ij)}_{st}$, and approximate the log det terms in (16) using

$$\log\det\Upsilon = -\log\det\sum_{j=1}^{k}\beta_j\lambda_j K^{(jj)} \approx -\sum_{l=1}^{r}\log\sum_{j=1}^{k}\beta_j\lambda_j K^{(jj)}_{ll}$$

$$\log\det\Sigma = \log\det\mathsf{Var}_{\theta_0}\left(\sum_{j=1}^{k}\beta_j\lambda_j\nabla S_{\theta_0}(A_j, B_j)\right) = \log\det\sum_{i=1}^{k}\sum_{j=1}^{k}\beta_i\lambda_i\beta_j\lambda_j K^{(ij)}$$

$$\approx \sum_{l=1}^{r}\log\sum_{i=1}^{k}\sum_{j=1}^{k}\beta_i\lambda_i\beta_j\lambda_j K^{(ij)}_{ll}.$$

We denote (assuming A is a $n \times n$ matrix) for $i \in \{1,\ldots,n\}$, $R_i(A) = \sum_{j\neq i}\left|A_{ij}\right|$ and let $D(A_{ii}, R_i(A))$ (D_i where unambiguous) be the closed disc centered at A_{ii} with radius $R_i(A)$. Such a disc is called a Geršgorin disc. The result below states that for matrices that are close to diagonal, the eigenvalues are close to the diagonal elements making our approximation accurate.

Theorem 1 (Geršgorin's circle theorem, for example, Horn and Johnson, 1990) *Every eigenvalue of A lies within at least one of the Geršgorin discs $D(A_{ii}, R_i(A))$. Furthermore, if the union of k discs is disjoint from the union of the remaining $n-k$ discs, then the former union contains exactly k and the latter $n-k$ eigenvalues of A.*

Algorithm 1 solves for θ, β jointly using alternating optimization. The second optimization problem $J(\beta;\cdot)$ is done using the approximation above and may be computed with minimal additional computation. The components of this objective are typically freely available when *sc*ℓ is minimized with Newton-type methods. In practice we found that such an approach leads to a selection of β that is close to optimal, despite loose convergence criteria for the minimization of the *sc*ℓ objective (see Sec. 8.3 and Figures 14, 20 for results).

8. Experiments

We demonstrate the asymptotic properties of $\hat{\theta}^{\mathrm{msl}}_n$ and explore the complexity-accuracy tradeoff for three different models-Boltzmann machine, linear Boltzmann MRF and conditional random fields. In terms of data sets, we consider synthetic data as well as data sets from sentiment prediction and text chunking domains.

In Appendix B we list all figures by subject. The basic road-map is to explore SCL for a theoretical Boltzmann machine and then to explore two data sets using both generative and discriminative models. We also demonstrate the effectiveness of the β heuristic for these experiments.

Algorithm 1 Calculate $\hat{\theta}^{msl}$

Require: $\{X_i\}_{i\in I}$ and λ, $\beta^{(0)}$
1: $t \leftarrow 1$
2: **while** $t <$ MAXITS **do**
3: $\quad\theta^{(t)} \leftarrow \arg\min sc\ell(\theta; \{X_i\}_{i\in I}, \lambda, \beta^{(t-1)})$
4: $\quad$**if** converged **then**
5: $\quad\quad$**return** θ
6: $\quad$**end if**
7: $\quad\beta^{(t)} \leftarrow \arg\min J(\beta; \{K^{(ij)}\}_{(i,j)\in J}, \lambda, \theta)$
8: $\quad t \leftarrow t+1$
9: **end while**
10: **return false**

8.1 Toy Example: Boltzmann Machines

We illustrate the improvement in asymptotic variance of the MSCLE associated with adding higher order Boltzmann machine likelihood components with increasingly higher probability. The Boltzmann machine can be parameterized as,

$$p_\theta(x) = \exp\left(\sum_{i<j}\theta_{ij}x_ix_j - \log\psi(\theta)\right), \quad x\in\{0,1\}^m.$$

To be able to accurately compute the asymptotic variance we use $m=5$ with θ 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.

Figure 2 displays the asymptotic variance, relative to the minimal variance of the MLE, for the cases of full likelihood (FL), pseudo likelihood ($|A_j|=1$) PL1, stochastic combination of pseudo likelihood and 2nd order pseudo likelihood ($|A_j|=2$) components λPL2$+(1-\lambda)$PL1, stochastic combination of 2nd order pseudo likelihood and 3rd order pseudo likelihood ($|A_j|=3$) components λPL3$+(1-\lambda)$PL2, and stochastic combination of 3rd order pseudo likelihood and 4th order pseudo likelihood ($|A_j|=4$) components λPL4$+(1-\lambda)$PL3.

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 λ 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 3 displays the asymptotic accuracy and complexity for different SCL policies for $m=9$ binary valued vertices of a Boltzmann machine. We explore three polices in which we denote pseudo likelihood components of size, or order, k. These policies include: $\lambda_1\beta_1$PL1$+\lambda_2(1-\beta_1)$PL2, $\lambda_1\beta_1$PL1$+\lambda_2(1-\beta_1)$PL3, $\lambda_1\beta_1$PL2$+\lambda_2(1-\beta_1)$PL3 (for multiple values of $\lambda_1,\lambda_2,\beta_1$). By taking different linear combinations of various sized pseudo likelihood components, we span a continuous spectrum of accuracy-complexity resolutions. The lower part of the diagram is the boundary of the achievable region (the optimal but unachievable place is the bottom left corner). SCL policies that

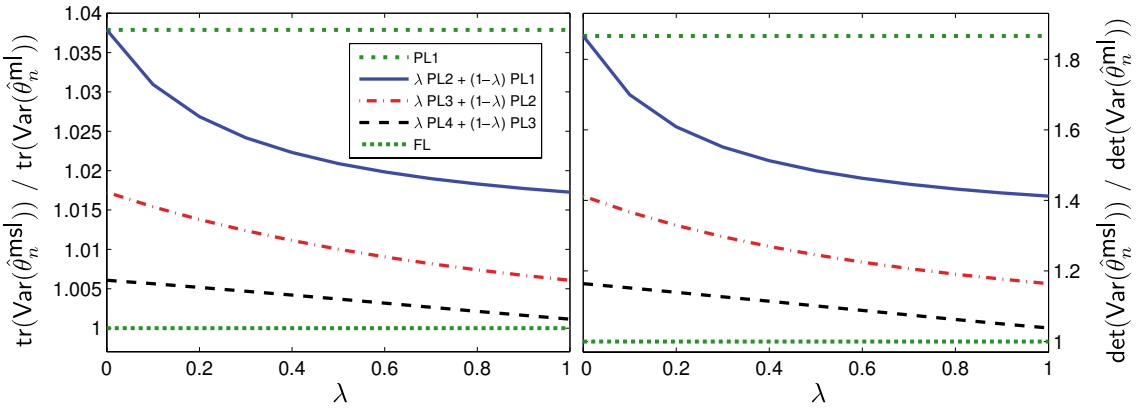

Figure 2: Asymptotic variance matrix, as measured by trace (left) and determinant (right), as a function of the selection probabilities for different stochastic versions of the SCL function.

lie to the right and top of that boundary may be improved by selecting a policy below and to the left of it.

8.2 Local Sentiment Prediction

Our first real world data set experiment involves local sentiment prediction using a conditional MRF model. The data set 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. As described in Mao and Lebanon (2007) (where more information may be found) we considered the task of predicting the local sentiment flow within these documents using regularized conditional random fields (CRFs) (see Figure 4 for a graphical diagram of the model in the case of four sentences).

As is common practice, we curtail overfitting through a L_2 regularizer, $\exp\{-(2n\sigma^2)^{-1}||\theta||_2^2\}$, which is strong when σ^2 is small and weak when σ^2 is large. We consider σ^2 a hyper-parameter and select it through cross-validation, unless noted otherwise.

Figure 5 shows the contour plots of train and test log-likelihood as a function of the SCL parameters: weight β and selection probability λ. 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. The figure demonstrates how the train log-likelihood 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 log-likelihood is also correlated with an increase in computational complexity as higher order likelihood components require more computation. Note however, that the test set behavior in the third and fourth rows shows an improvement in prediction accuracy associated with

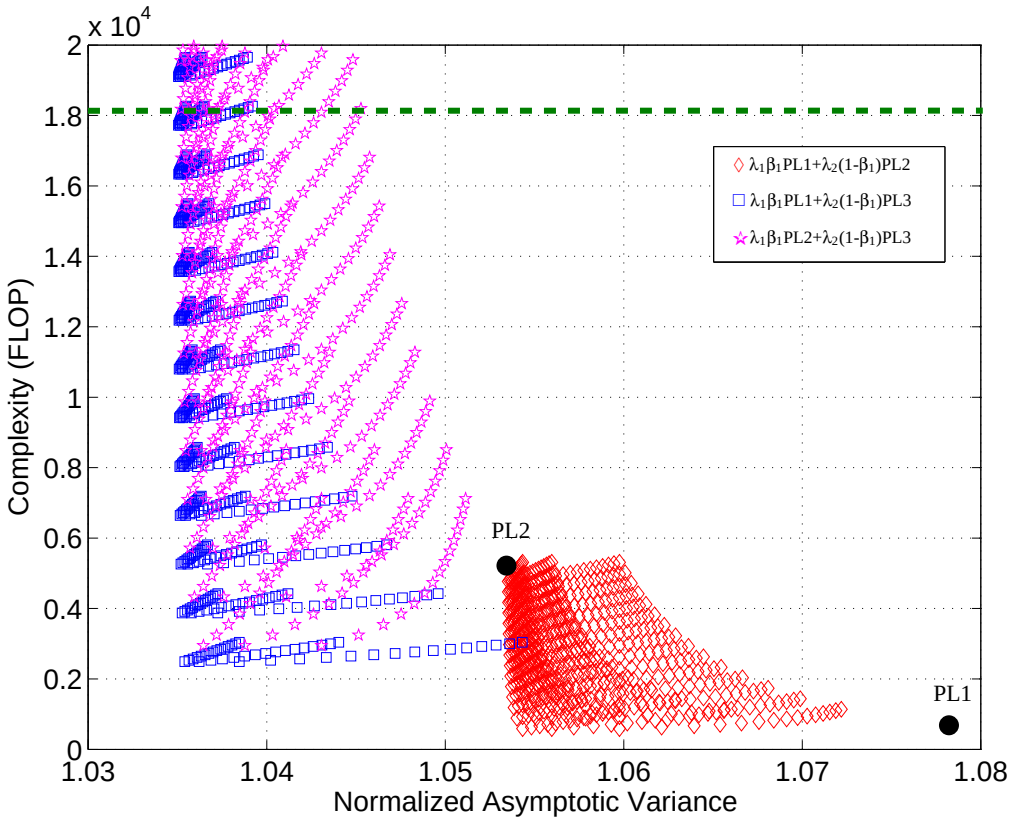

Figure 3: Computation-accuracy diagram for three SCL families:$\lambda_1\beta_1\text{PL1} + \lambda_2(1-\beta_1)\text{PL2}$, $\lambda_1\beta_1\text{PL1} + \lambda_2(1-\beta_1)\text{PL3}$, $\lambda_1\beta_1\text{PL2} + \lambda_2(1-\beta_1)\text{PL3}$ (for multiple values of $\lambda_1, \lambda_2, \beta_1$) for the Boltzmann machine with 9 binary nodes. The pure policies PL1 and PL2 are indicated by black circles and the computational complexity of the full likelihood indicated by a dashed line (corresponding normalized asymptotic variance is 1). The PL3 pure policy is beyond the scale of the diagram. As the graph size increases, the computational cost increases dramatically, in particular for the full likelihood policy and to a lesser extent for the pseudo likelihood policy.

decreasing the influence of full likelihood in favor of pseudo likelihood. The fact that this happens for (relatively) weak regularization, $\sigma^2 = 10$, and indicates that lower order pseudo likelihood has a regularization effect which improves prediction accuracy when the model is not regularized enough. We have encountered this phenomenon in other experiments as well and we will discuss it further in the following subsections.

Figure 6 displays the complexity and negative log-likelihoods (left:train, right:test) of different SCL estimators, sweeping through λ and β, 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

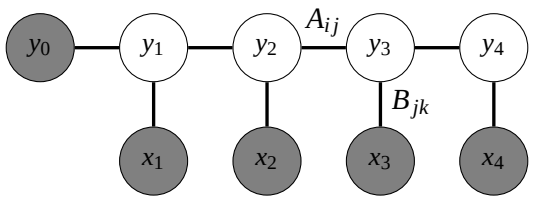

Figure 4: Graphical representation of a four token conditional random field (CRF). *A*, *B* are weight matrices and represent state-to-state transitions and state-to-observation outputs. Shading indicates the variable is conditioned upon while no shading indicates the variable is generated by the model.

achievable region with the exact position on that boundary depending on the required solution of the computation-accuracy tradeoff.

8.3 Text Chunking

This experiment consists of using sequential MRFs to divide sentences into "text chunks," that is, syntactically correlated sub-sequences, such as noun and verb phrases. Chunking is a crucial step towards full parsing. For example,[6] the sentence:

> He reckons the current account deficit will narrow to only # 1.8 billion in September.

could be divided as:

> [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September].

where NP, VP, and PP indicate noun phrase, verb phrase, and prepositional phrase.

We used the publicly available CoNLL-2000 shared task data set. It consists of labeled partitions of a subset of the Wall Street Journal (WSJ) corpus. Our training sets consisted of sampling 100 sentences without replacement from the the CoNLL-2000 training set (211,727 tokens from WSJ Sections 15-18). The test set was the same as the CoNLL-2000 testing partition (47,377 tokens from WSJ Section 20). Each of the possible 21,589 tokens, that is, words, numbers, punctuation, etc., are tagged by one of 11 chunk types and an O label indicating the token is not part of any chunk. Chunk labels are prepended with flags indicating that the token begins (B-) or is inside (I-) the phrase. Figure 7 lists all labels and respective frequencies. In addition to labeled tokens, the data set contains a part-of-speech (POS) column. These tags were automatically generated by the Brill tagger and must be incorporated into any model/feature set accordingly.

In the following, we explore this task using various SCL selection policies on two related, but fundamentally different sequential MRFs: Boltzmann chain MRFs and CRFs.

8.3.1 BOLTZMANN CHAIN MRF

Boltzmann chains are a generative MRF that are closely related to hidden Markov models (HMM). See MacKay (1996) for a discussion on the relationship between Boltzmann chain MRFs and

6. Taken from the CoNLL-2000 shared task, `http://www.cnts.ua.ac.be/conll2000/chunking/`.

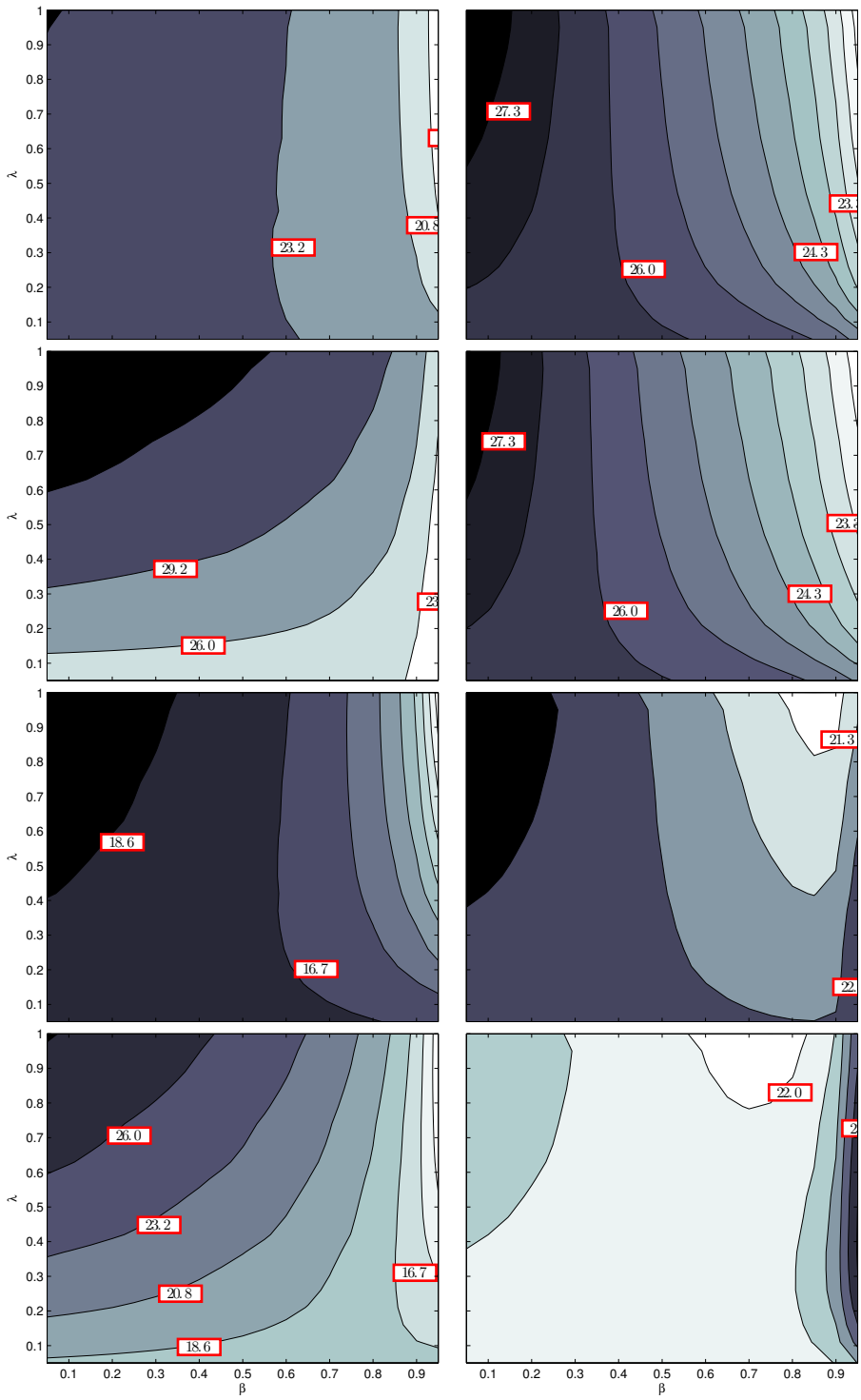

Figure 5: Train (left column) and test (right column) neg. log-likelihood contours for maximum SCL estimators for the CRF model. L_2 regularization, $\exp\{-(2n\sigma^2)^{-1}||\theta||_2^2\}$, parameters are $\sigma^2 = 1$ (rows 1,2) and $\sigma^2 = 10$ (rows 3,4). Rows 1,3 are stochastic mixtures of full (FL) and pseudo (PL1) log-likelihood components while rows 2,4 are PL1 and 2nd order pseudo (PL2). Note that weaker regularization resulted in higher accuracy at lower computation.

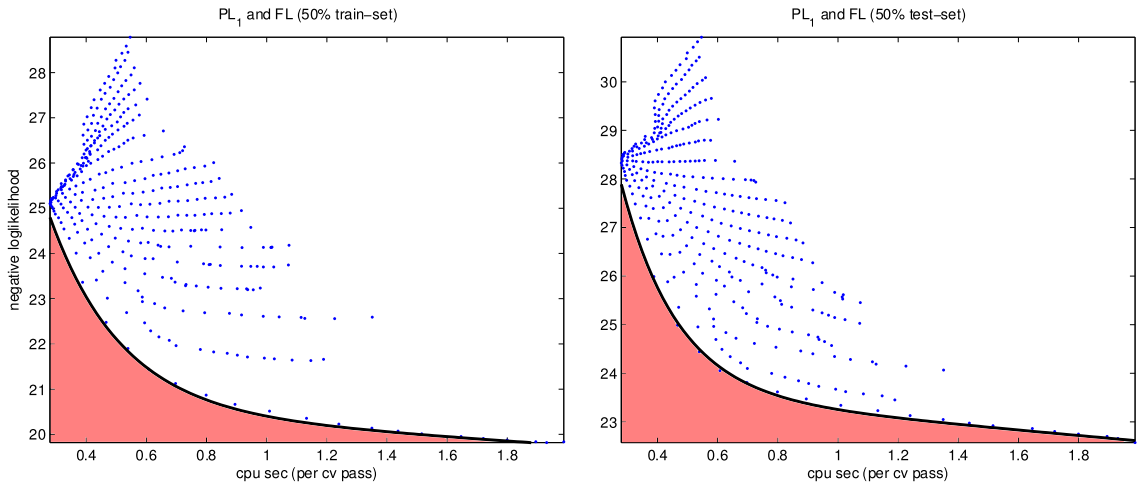

Figure 6: Scatter plot representing complexity and negative log-likelihood (left:train, right:test) of SCL functions for CRFs with L2 regularization parameter $\sigma^2 = 1/2$. The points represent different stochastic combinations of full and pseudo likelihood components. The shaded region represents impossible accuracy/complexity demands. Since the boundary of the obtainable region is empirical, the optimal beta always lies on this boundary. By varying λ, β we are able to smoothly span complexity (wall seconds) and accuracy.

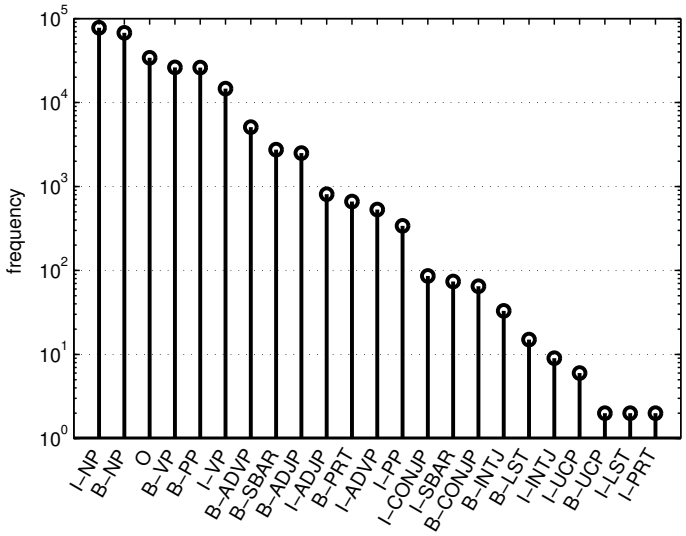

Figure 7: Label counts in CoNLL-2000 data set. Phrases such as noun (NP), verb (VP), and prepositional (PP) are demarcated by a "begin" tag (B-) and an "inside" tag (I-). Non-phrase entities are denoted as "other" (O).

HMMs. We consider SCL components of the form $p(X_2, Y_2|Y_1, Y_3)$, $p(X_2, X_3, Y_2, Y_3|Y_1, Y_4)$ which we refer to as first and second order pseudo likelihood (with higher order components generalizing in a straightforward manner).

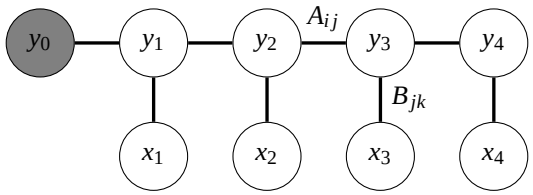

Figure 8: Graphical representation of a four token Boltzmann chain. A, B are weight matrices and represent preference in particular state-to-state transitions and state-to-feature emissions. Only the start state is conditioned upon while all others are generative.

The nature of the Boltzmann chain constrains our feature set to only encode the particular token present at each position, or time index. In doing so we avoid having to model additional dependencies across time steps and dramatically reduce computational complexity. Although SCL is precisely motivated by high treewidth graphs, we wish to include the full likelihood for demonstrative purposes—in practice, this is often not possible. Although POS tags are available we do not include them in these features since the dependence they share on neighboring tokens and other POS tags is unclear. For these reasons our time-sliced feature vector, x_i, has only a single-entry one and cardinality matching the size of the vocabulary (21,589 tokens).

As in Section 8.2, we control overfitting through a L_2 regularizer, $\exp\{-(2n\sigma^2)^{-1}||\theta||_2^2\}$, which is strong when σ^2 is small and weak when σ^2 is large. Here again we choose σ^2 via cross-validation unless otherwise noted. More often though, we show results for several representative σ^2 to demonstrate the roles of λ and β in $\hat{\theta}_n^{msl}$.

Figures 9 and 10 depict train and test negative log-likelihood, that is, perplexity, for the SCL estimator $\hat{\theta}_{100}^{msl}$ with a pseudo/full likelihood selection policy (PL1/FL). As is our convention, weight β and selection probability λ correspond to the higher order component, in this case full likelihood. The lower order pseudo likelihood component is always selected and has weight $1-\beta$. As expected the test set perplexity dominates the train-set perplexity. As was the situation in Sec. 8.2, we note that the lower order component serves to regularize the full likelihood, as evident by the abnormally large σ^2.

We next demonstrate the effect of using a 1st order/2nd order pseudo likelihood selection policy (PL1/PL2). Recall, our notion of pseudo likelihood never entails conditioning on x, although in principle it could. Figures 11 and 12 show how the policy responds to varying both λ and β. Figure 13 depicts the empirical tradeoff between accuracy and complexity. Figure 14 highlights the effectiveness of the β heuristic. See captions for additional comments.

8.3.2 CRFs

Conditional random fields are the discriminative counterpart of Boltzmann chains (cf. Figures 4 and 8). Since x is not jointly modeled with y, we are free to include features with non-independence across time steps without significantly increasing the computational complexity. Here our notion of

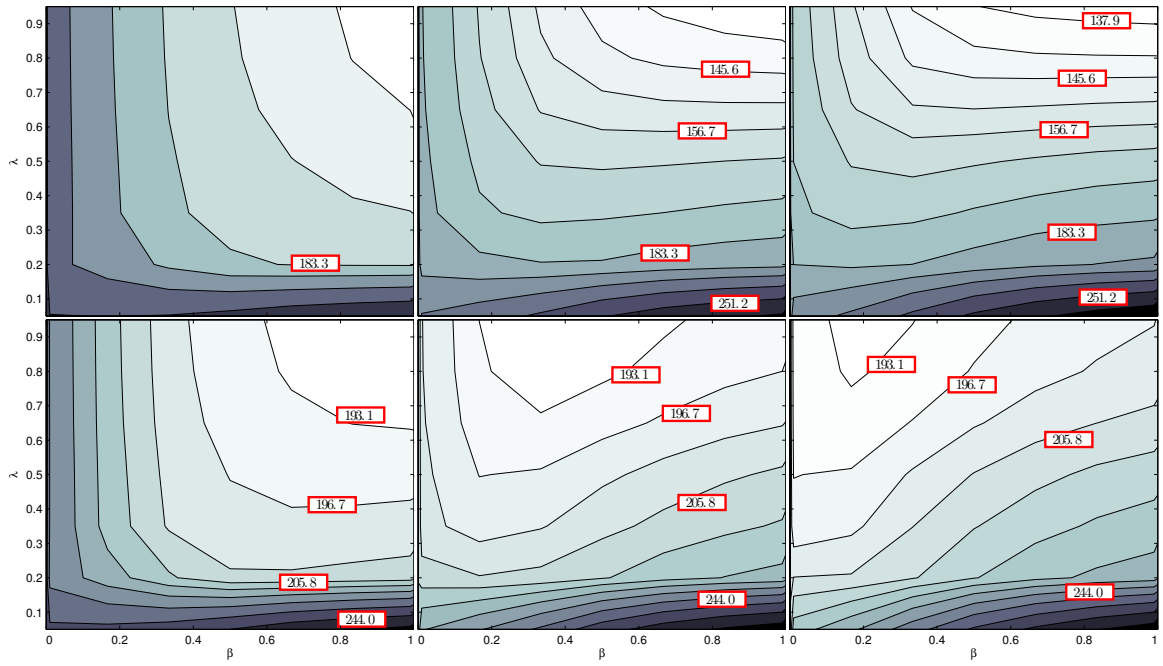

Figure 9: Train set (top) and test set (bottom) negative log-likelihood (perplexity) for the Boltzmann chain MRF with pseudo/full likelihood selection policy (PL1/FL). The x-axis, β, corresponds to relative weight placed on FL and and the y-axis, λ, corresponds to the probability of selecting FL. PL1 is selected with probability 1 and weight $1-\beta$. Contours and labels are fixed across columns. Results averaged over several cross-validation folds, that is, resampling both the train set and the PL1/FL policy. Columns from left to right correspond to weaker regularization, $\sigma^2 = \{500, 2500, 5000\}$. The best achievable test set perplexity is about 190.

Unsurprisingly the test set perplexity dominates the train set perplexity at each σ^2 (column). For a desired level of accuracy (contour) there exists a computationally favorable regularizer. Hence $\hat{\theta}_n^{msl}$ acts as both a regularizer and mechanism for controlling accuracy and complexity.

pseudo likelihood is more traditional, for example, $p(Y_2|Y_1, Y, 3, X_2)$ and $p(Y_2, Y_3|Y_1, Y, 4, X_2, X_3)$ are valid 1st and 2nd order pseudo likelihood components.

We employ a subset of the features outlined in Sha and Pereira (2003) which proved competitive for the CoNLL-2000 shared task. Our feature vector was based on seven feature categories, resulting in a total of 273,571 binary features (i.e., $\sum_i f_i(x_t) = 7$). The feature categories consisted of word unigrams, POS unigrams, word bigrams (forward and backward), and POS bigrams (forward and backward) as well as a stopword indicator (and its complement) as defined by Lewis et al. (2004). The set of possible feature/label pairs is much larger than our set—we use only those features supported by the CoNLL-2000 data set, that is, those which occur at least once. Thus we modeled 297,041 feature/label pairs and 847 transitions for a total of 297,888 parameters. As before, we use

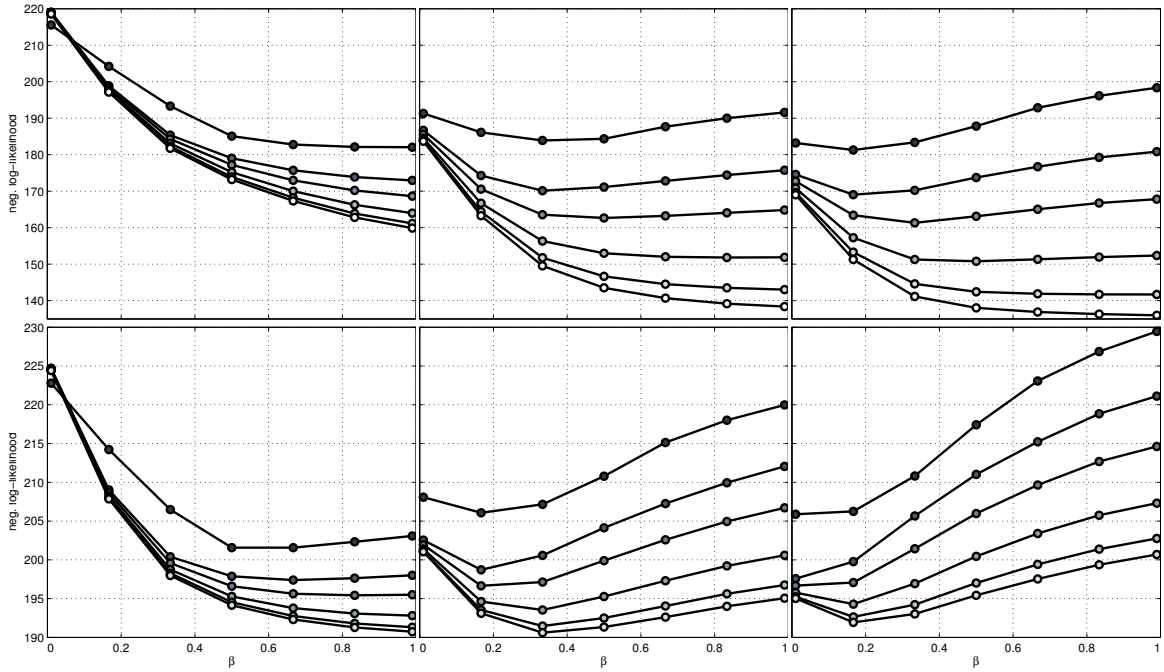

Figure 10: Train set and test set perplexities for the Boltzmann chain MRF with PL1/FL selection policy (see above layout description). The x-axis is again β and the y-axis perplexity. Lighter shading indicates FL is selected with increasing frequency. Note that as the regularizer is weakened the range in perplexity spanned by λ increases and the lower bound decreases. This indicates that the approximating power of $\hat{\theta}_n^{msl}$ increases when unencumbered by the regularizer and highlights its secondary role as a regularizer.

the L_2 regularizer, $\exp\{-(2\sigma^2)^{-1}||\theta||_2^2\}$, which is strong when σ^2 is small and weak when σ^2 is large.

We demonstrate learning on two selection policies: pseudo/full likelihood (Figures 15 and 16) and 1st/2nd order pseudo likelihood (Figures 17 and 18). In both selection polices we note a significant difference from the Boltzmann chain, β has less impact on both train and test perplexity. Intuitively, this seems reasonable as the component likelihood range and variance are constrained by the conditional nature of CRFs. Figure 19 demonstrates the empirical accuracy/complexity tradeoff and Figure 20 depicts the effectiveness of the β heuristic. See captions for further comments.

8.4 Complexity/Regularization Win-Win

It is interesting to contrast the test log-likelihood behavior in the case of mild and stronger L_2 regularization. In the case of weaker or no regularization, the test log-likelihood shows different behavior than the train log-likelihood. 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.

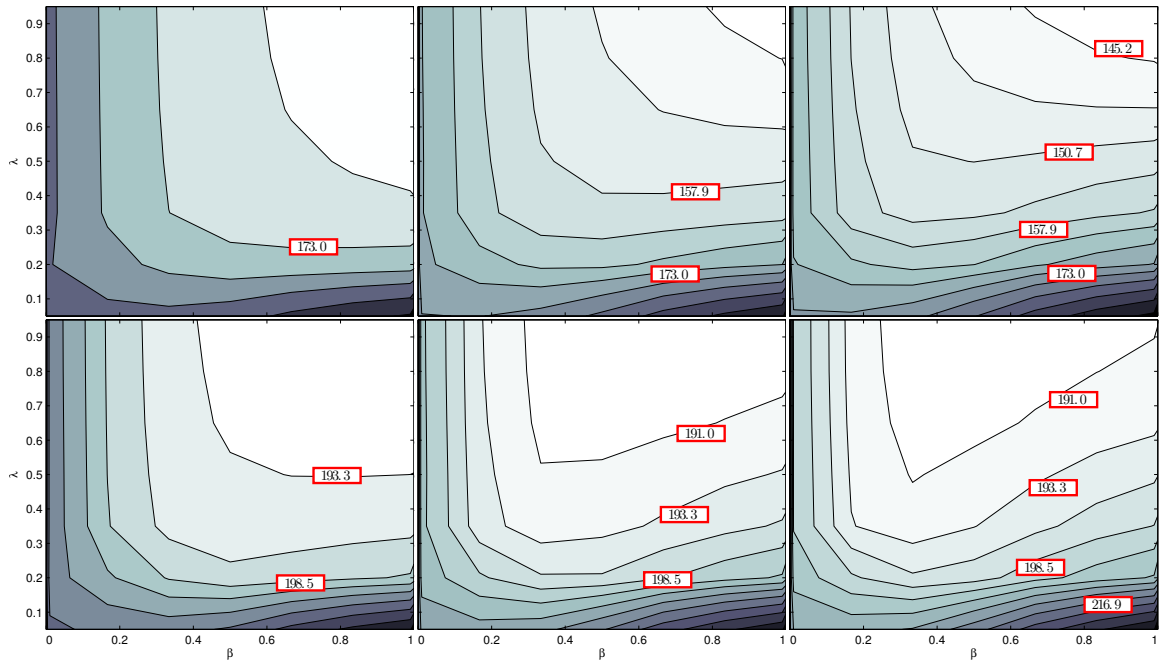

Figure 11: Train set (top) and test set (bottom) perplexity for the Boltzmann chain MRF with 1st/2nd order pseudo likelihood selection policy (PL1/PL2). The x-axis corresponds to PL2 weight and the y-axis the probability of its selection. PL1 is selected with probability 1 and weight $1-\beta$. Columns from left to right correspond to $\sigma^2 = \{5000, 10000, 15000\}$. See Figure 9 for more details. The best achievable test set perplexity is about 189.5.

In comparing these results to PL1/FL, we note that the test set contours exhibit less perplexity for larger areas. In particular, perplexity is lower at smaller λ values, meaning a computational saving over PL1/FL at a given level of accuracy.

In Figure 5 we note this phenomenon when comparing $\sigma^2 = 1$ to $\sigma^2 = 10$ across the selection policies PL1/FL and PL1/PL2. That is, the weaker regularization and more restrictive selection policy, that is, PL1/PL2, is able to achieve comparable test set perplexity.

For the text chunking experiments, we observe a striking win-win when using the Boltzmann chain MRF, Figures 9 and 11. Notice that as regularization is decreased (comparing from left to right), the contours are pulled closer to the x-axis. This means that we are achieving the same perplexity at reduced levels of computational complexity. The CRF however, only exhibits the win-win to a minor extent. We delve deeper into why this is might be the case in the following section.

8.5 λ, σ^2 Interplay

Throughout these experiments we fixed σ^2 and either swept over (λ, β) or used the heuristic to evaluate $(\lambda, \beta(\lambda))$. Motivated by the sometimes weak win-win (cf. Section 8.4) we now consider

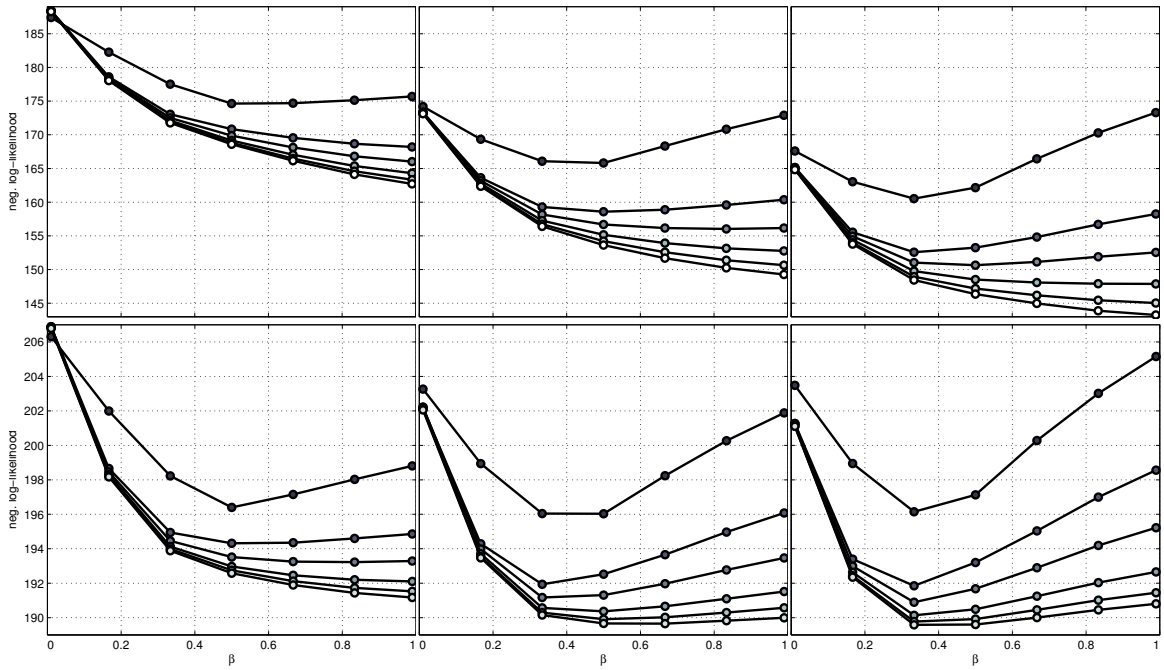

Figure 12: Train (top) and test (bottom) perplexities for the Boltzmann chain MRF with PL1/PL2 selection policy (x-axis:PL2 weight, y-axis:perplexity; see above and previous).

PL1/PL2 outperforms PL1/FL (Fig. 10) test perplexity despite PL1/FL including FL as a special case (i.e., $(\lambda,\beta)=(1,1)$). We speculate that the regularizer's indirect connection to the training samples precludes it from preventing certain types of overfitting. See Sec. 8.4 for more discussion.

how the optimal σ^2 changes as a function of λ. In Figure 21 we used the β heuristic to evaluate train and test perplexity over a (λ,σ^2) grid. We used CRFs and the text chunking task as outlined in Section 8.3.2.

For the PL1/FL policy, we observe that for small enough λ the optimal σ^2, that is, the σ^2 with smallest test perplexity, has considerable range. At some point there are enough samples of the higher-order component to stabilize the choice of regularizer, noting that it is still weaker than the optimal full likelihood regularizer. Conversely, the PL1/PL2 regularizer has an essentially constant optimal regularizer which is relatively much weaker.

As a result, we believe that the lack of win-win for the chunking CRF follows from two effects. In the case of the PL1/FL policy the contour plots are misleading since there is no single σ^2 that performs well across all $\lambda\in[0,1]$. For the PL1/PL2 there is simply little change in regularization necessary across λ.

9. Discussion

The proposed estimator family facilitates computationally efficient estimation in complex graphical models. In particular, different (β,λ) parameterizations of the stochastic composite likelihood en-

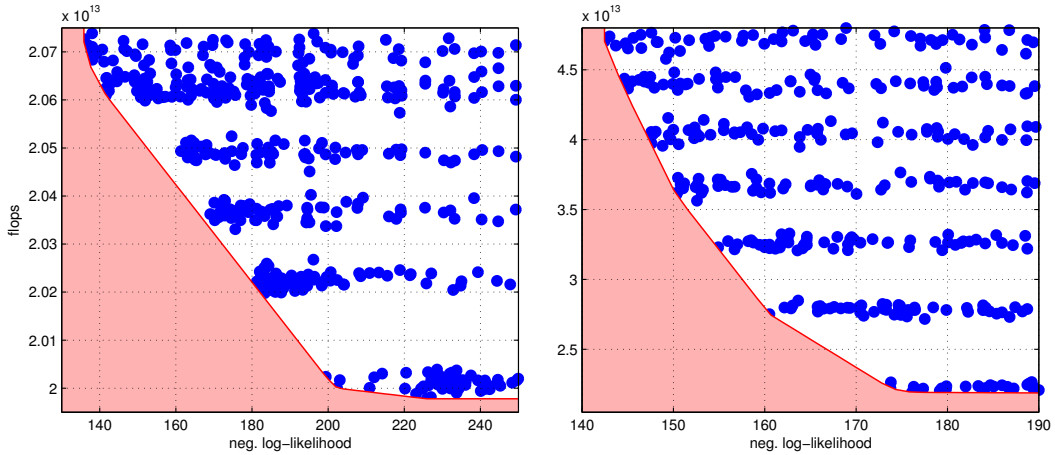

Figure 13: Accuracy and complexity tradeoff for the Boltzmann chain MRF with PL1/FL (left) and PL1/PL2 (right) selection policies. Each point represents the negative log-likelihood (perplexity) and the number of flops required to evaluate the composite likelihood and its gradient under a particular instantiation of the selection policy. The shaded region is the convex hull of the points and represents empirically unobtainable combinations of computational complexity and accuracy. Particularly interesting is the difference between policies and against the discriminative CRF, cf. Figure 19.

ables 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.

It is interesting to note that the SCL framework may be generalized to random m-estimators beyond likelihood objects. That is, instead of a fixed m-function we may consider a linear combination of stochastic objects (appearing or not with some probability). Such estimators go beyond traditional m-estimator but may be analyzed using techniques similar to the ones developed in this paper. Although not a random m-estimator, the work of Dillon et al. (2010) borrows SCL concepts to facilitate budgeted semi-supervised learning. This too would benefit from a random m-estimator interpretation and indeed many machine learning tasks may fit nicely into such a framework.

The SCL framework may be useful for a wide variety of intractable graphical models. Besides the examples presented here, it may be particularly suited for large scale models from statistical physics, exponential random graph models, and models from computational biology. A particularly nice feature is that the above computation may be trivially parallelized thus leading to effective computation on large clusters and cloud computing.

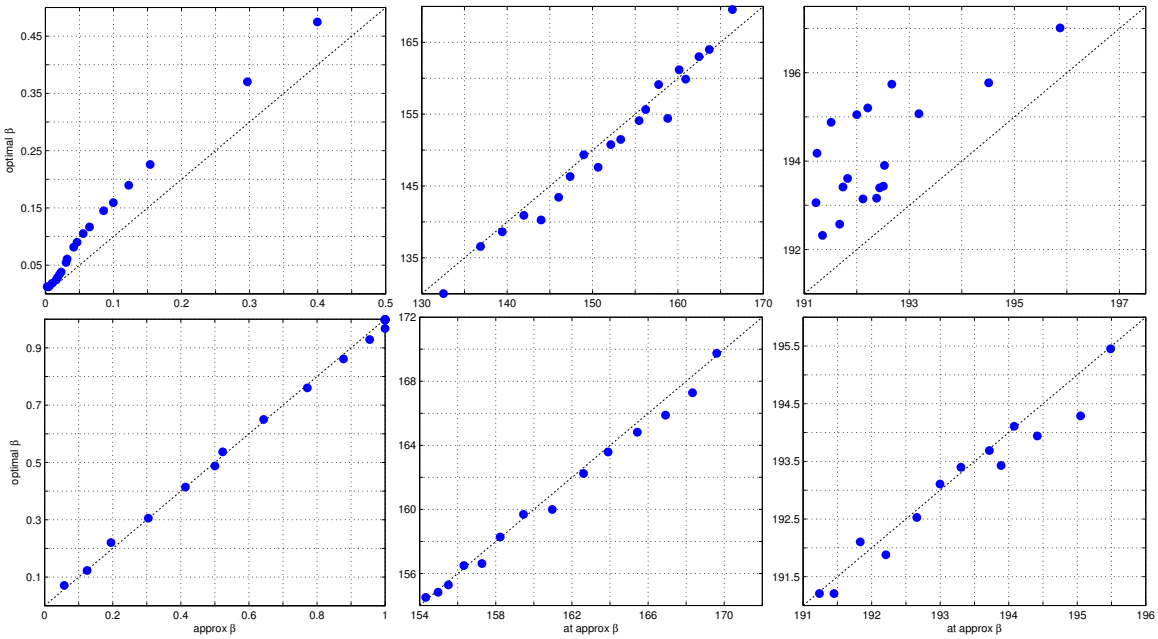

Figure 14: Demonstration of the effectiveness of the β heuristic, that is, using $\hat{\theta}^{msl}$ as a plug-in estimate for θ_0 to periodically re-estimate β during gradient descent. Results are for the Boltzmann chain with PL1/FL (top) and PL1/PL2 (bottom) selection policies. The x-axis is the value at the heuristically found β and the y-axis the value at the optimal β. The optimal β was found by evaluating over a β grid and choosing that with the smallest train set perplexity. The first column depicts the best performing β against the heuristic β. The second and third columns depict the training and testing perplexities (resp.) at the best performing β and heuristically found β. For all three columns, we assess the effectiveness of the heuristic by its nearness to the diagonal (dashed line).

For the PL1/PL2 policy the heuristic closely matched the optimal (all bottom row points are on diagonal). The heuristic out-performed the optimal on the test set and had slightly higher perplexity on the training set. It is a positive result, albeit somewhat surprising, and is attributable to either coarseness in the grid or improved generalization by accounting for variability in $\hat{\theta}^{msl}$.

Acknowledgments

The authors acknowledge useful discussion with P. Liang, M. Welling, P. Smyth, S.-C. Zhu, and C. Sutton. The presentation of the paper was substantially improved following useful feedback from Fernando Pereira and the anonymous referees.

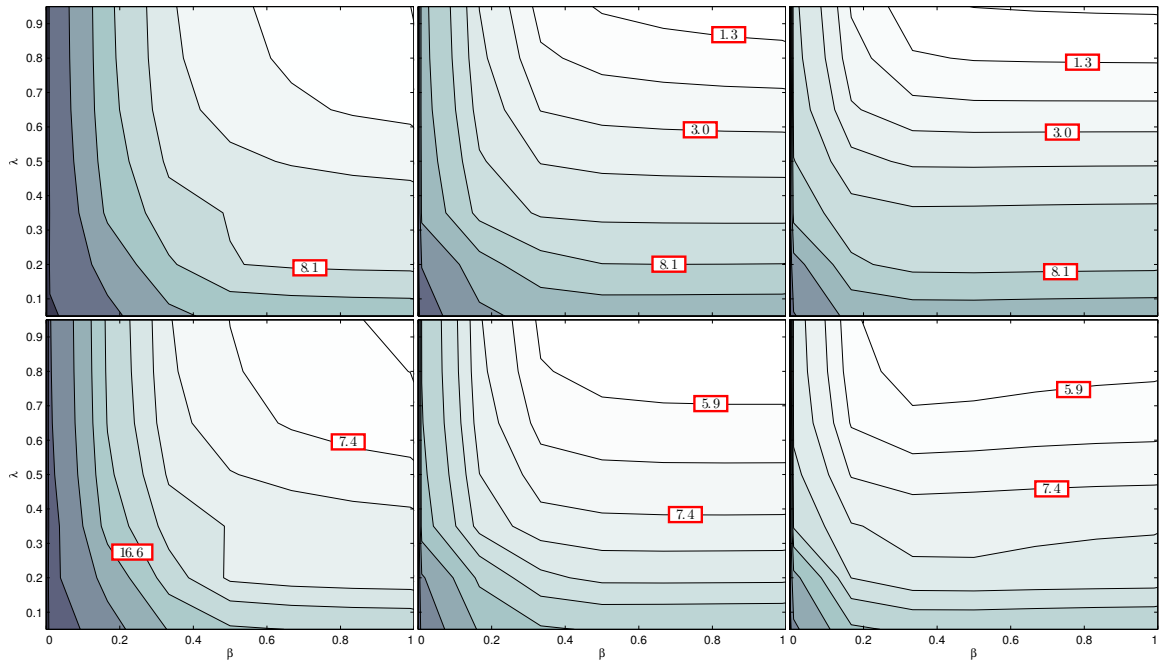

Figure 15: Train set (top) and test set (bottom) perplexity for the CRF with pseudo/full likelihood selection policy (PL1/FL). The x-axis corresponds to FL weight and the y-axis the probability of its selection. PL1 is selected with probability 1 and weight $1-\beta$. Columns from left to right correspond to $\sigma^2 = \{100, 1000, 5000\}$. See Figure 9 for more details. The best achievable test set perplexity is about 5.5.

Although we cannot directly compare CRFs to its generative counterpart, we observe some strikingly different trends. It is immediately clear that the CRF is less sensitive to the relative weighting of components than is the Boltzmann chain. This is partially attributable to a smaller range of the objective—the CRF is already conditional hence the per-component perplexity range is reduced.

Appendix A. Proofs

The proofs below generalize the classical consistency and asymptotic efficiency of the MLE (Ferguson, 1996) and the corresponding results for *m*-estimators (van der Vaart, 1998). They follow similar lines as the proofs in Ferguson (1996) and van der Vaart (1998), with the necessary modifications due to the stochasticity of the SCL function. We assume below that $p_\theta(X) > 0$ and that X is a discrete and finite RV.

The following lemma generalizes Shannon's inequality (Cover and Thomas, 2005) for the KL divergence. We will use it to prove consistency of the SCL estimator.

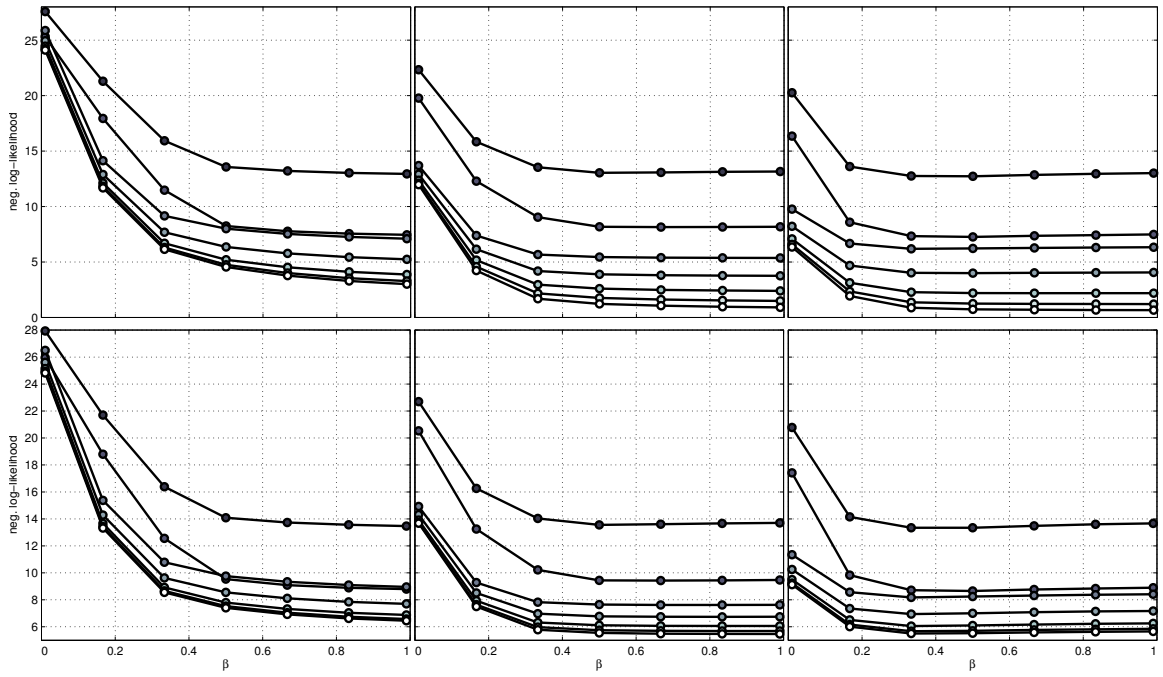

Figure 16: Train (top) and test (bottom) perplexities for a CRF with PL1/FL selection policy (x-axis:FL weight, y-axis:perplexity; see above and Fig. 10).

Perhaps more evidently here than above, we note that the significance of a particular β is less than that of the Boltzmann chain. However, for large enough σ^2, the optimal $\beta \neq 1$. This indicates the dual role of PL1 as a regularizer. Moreover, the left panel calls attention to the interplay between β, λ, and σ^2. See Sec. 8.5 for more discussion.

Lemma 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

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

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 $\theta = \theta'$. ■

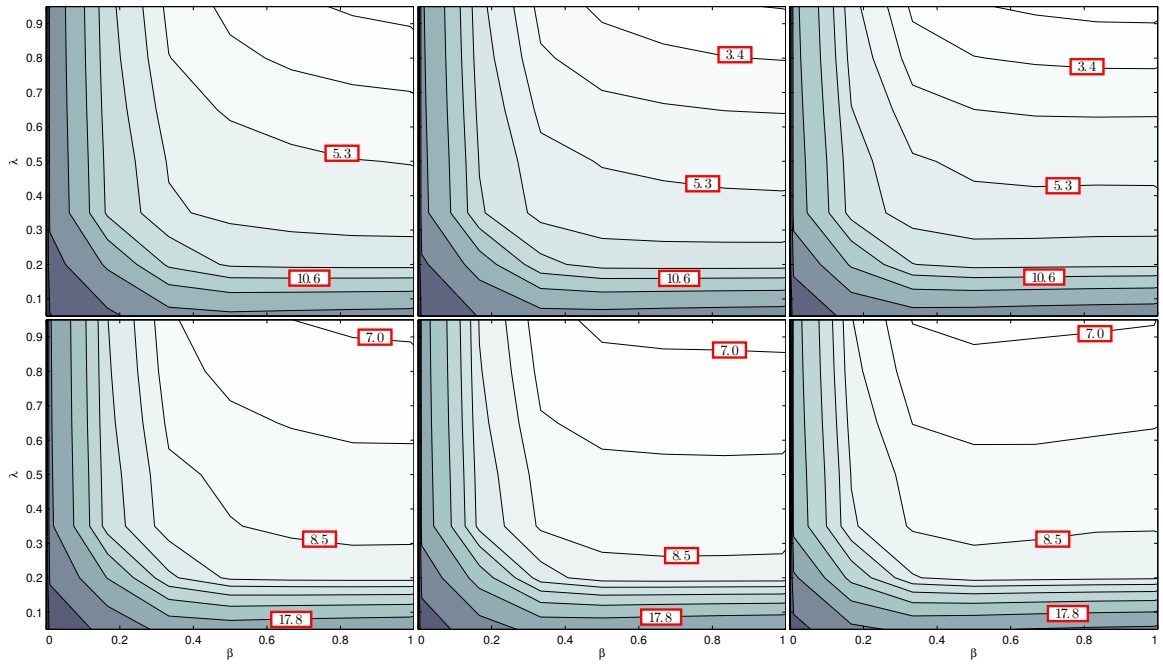

Figure 17: Train set (top) and test set (bottom) perplexity for a CRF with 1st/2nd order pseudo likelihood selection policy (PL1/PL2). The x-axis, β, represents the relative weight placed on PL2 and the y-axis, λ, the probability of selecting PL2. PL1 is selected with probability 1. Columns from left to right correspond to weaker regularization, $\sigma^2 = \{10000, 20000, 40000\}$. See Figure 15 for more details.

Proposition 5 *Let $\Theta \subset \mathbb{R}^r$ be an open set, $p_\theta(x) > 0$ and continuous and smooth in θ, and $(A_1, B_1), \ldots, (A_k, B_k)$ be a sequence of m-pairs for which $\{(A_j, B_j) : \forall j$ such that $\lambda_j > 0\}$ ensures identifiability. Then the sequence of SCL maximizers is strongly consistent, that is,*

$$P\left(\lim_{n\to\infty} \hat{\theta}_n = \theta_0\right) = 1.$$

Proof The SCL function, modified slightly by a linear combination with a term that is constant in θ is

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

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

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

If we restrict ourselves to the compact set $S = \{\theta : c_1 \le \|\theta - \theta_0\| \le c_2\}$ then

$$\sup_{\theta\in S}\sup_{Z}\left|\sum_{j=1}^{k} Z_j \beta_j \log p_\theta(X_{A_j}|X_{B_j}) - \lambda_j \beta_j \log p_{\theta_0}(X_{A_j}|X_{B_j})\right| < K(x) < \infty$$

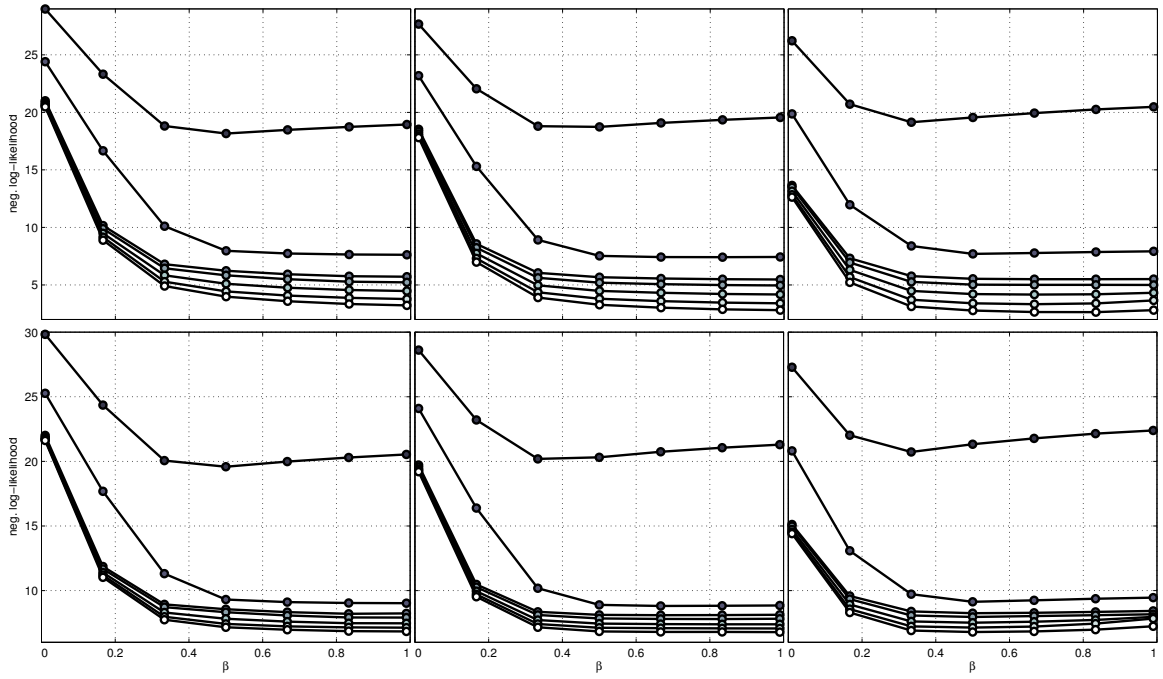

Figure 18: Train (top) and test (bottom) perplexities for a CRF with PL1/PL2 selection policy (x-axis:PL2 weight, y-axis:perplexity; see above and Fig. 10).

Although increasing λ only brings minor improvement to both the training and testing perplexities, it is worth noting that the test perplexity meets that of the PL1/FL. Still though, the overall lack of resolution here suggests that smaller values of λ would better span a range of perplexities and at reduced computational cost.

where $K(x)$ is a function satisfying $\mathsf{E}K(X) < \infty$. As a result, the conditions for the uniform strong law of large numbers (Ferguson, 1996) hold on S leading to

$$P\left\{\lim_{n\to\infty}\sup_{\theta\in S}|scl'(\theta)-\mu(\theta)|=0\right\}=1. \tag{17}$$

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 (17) 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 $\hat{\theta}_n^{\text{msl}}\notin S$ for $n>N$. Since c_1,c_2 were chosen arbitrarily $\hat{\theta}_n^{\text{msl}}\to\theta_0$ with probability 1. ■

Proposition 6 *Making the assumptions of Proposition 1 as well as convexity of $\Theta\subset\mathbb{R}^r$ we have the following convergence in distribution*

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

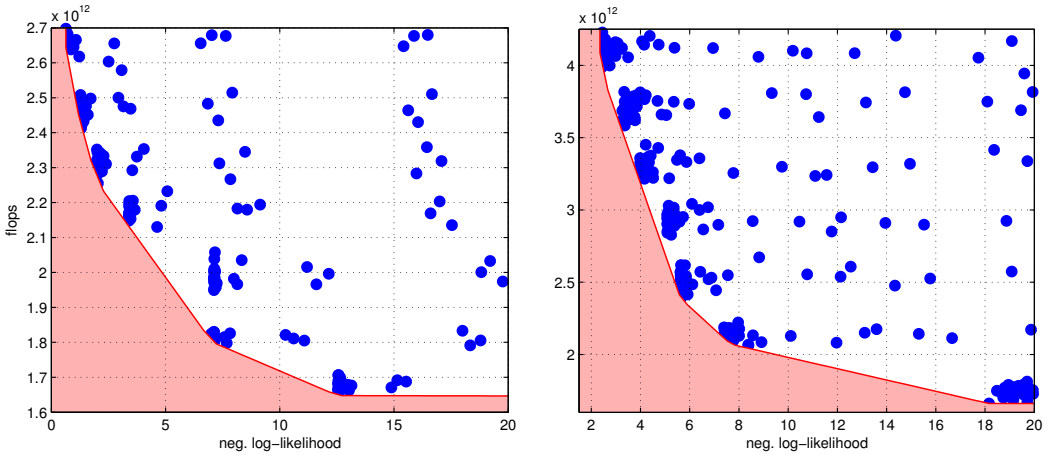

Figure 19: Accuracy and complexity tradeoff for the CRF with PL1/FL (left) and PL1/PL2 (right) selection policies. Each point represents the negative log-likelihood (perplexity) and the number of flops required to evaluate the composite likelihood and its gradient under a particular instance of the selection policy. The shaded region is the convex hull of the points and represents empirically unobtainable combinations of computational complexity and accuracy. σ^2. Particularly interesting is the difference between policies and against the generative Boltzmann chain, cf. Figure 13.

where

$$\Upsilon^{-1} = \sum_{j=1}^{k} \beta_j \lambda_j \mathit{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j))$$

$$\Sigma = \mathit{Var}_{\theta_0}\left(\sum_{j=1}^{k} \beta_j \lambda_j \nabla S_{\theta_0}(A_j, B_j)\right).$$

The notation $\mathsf{Var}_{\theta_0}(Y)$ represents the covariance matrix of the random vector Y under p_{θ_0} while the notations $\xrightarrow{\mathrm{p}}, \rightsquigarrow$ in the proof below denote convergences in probability and in distribution (Ferguson, 1996).

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

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

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, $\nabla sc\ell_n(\hat{\theta}_n^{\mathrm{msl}}) = 0$ and

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{msl}} - \theta_0) = -\sqrt{n}(\nabla^2 sc\ell_n(\theta'))^{-1} \nabla sc\ell_n(\theta_0). \tag{18}$$

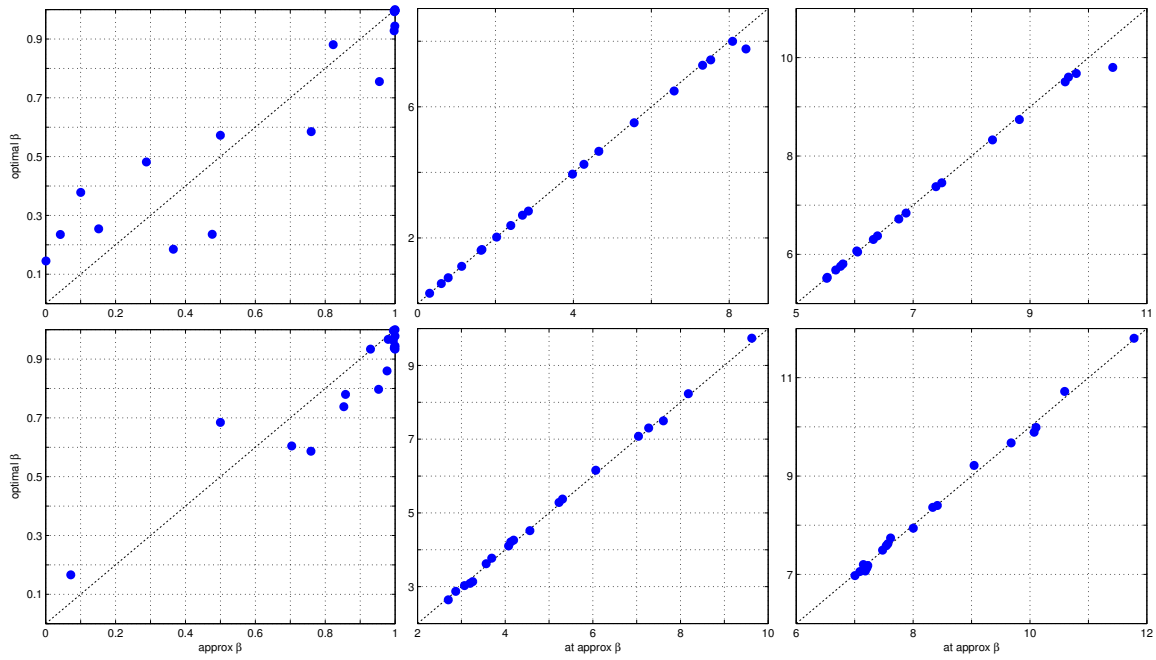

Figure 20: Demonstration of the effectiveness of the β heuristic. Results are for the CRF with PL1/FL (top) and PL1/PL2 (bottom) selection policies. The x-axis is the value at the heuristically found β and the y-axis the value at the optimal β. The first column depicts the best performing β against the heuristic β. The second and third columns depict the training and testing perplexities (resp.) at the best performing β and heuristically found β. For all three columns, we assess the effectiveness of the heuristic by its nearness to the diagonal (dashed line). See Fig. 14 for more details.

The optimal and heuristic β match train and test perplexities for both policies. The actual β value however does not seem to match as well as the Boltzmann chain. However, if we note the flatness of the β grid (cf. Fig. 16 and 18) this result is unsurprising and can be disregarded as an indication of the heuristic's performance.

By Proposition 1 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,

$$\begin{aligned}(\nabla^2 sc\ell_n(\theta'))^{-1} &\xrightarrow{\text{p}} (\nabla^2 sc\ell_n(\theta_0))^{-1} \\ &\xrightarrow{\text{p}} \left(\sum_{j=1}^{k} \beta_j \lambda_j \mathsf{E}_{\theta_0} \nabla^2 S_{\theta_0}(A_j, B_j)\right)^{-1} \\ &= -\left(\sum_{j=1}^{k} \beta_j \lambda_j \mathsf{Var}_{\theta_0}(\nabla S_{\theta_0}(A_j, B_j))\right)^{-1}.\end{aligned} \tag{19}$$

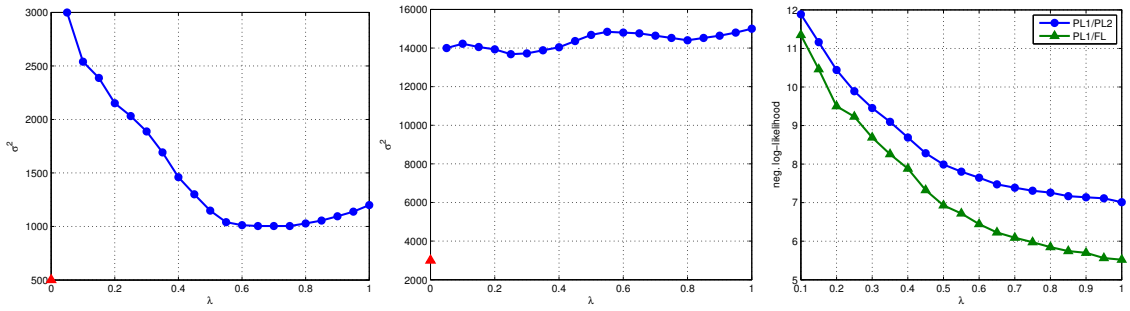

Figure 21: Optimal regularization parameter as a function of $(\lambda, \hat{\beta}(\lambda))$ for PL1/FL (left) and PL1/PL2 (center) CRF selection policies. In the left figure, PL1/FL, λ represents the probability of including FL into the objective. A few FL samples add uncertainty to the objective thus a weaker regularizer is preferable. As more FL samples are incorporated, this effect diminishes but still acts to regularize since the full likelihood (only) best regularization is $\sigma^2 = 500$ (red triangle). The center figure, PL1/PL2, exhibits only a minor change as λ (the probability of incorporating PL2) is increased. It is however, best served by a much weaker regularizer than PL2 alone (red triangle).

The right figure depicts the test-set perplexity as a function of λ using the optimal σ^2 (small λ values were clipped as their performance is quite poor). Note that the perplexity is lowest when both components are always selected ($\lambda = 1$) and that the PL1/FL policy outperforms the PL1/PL2 policy as expected.

For the remaining term in (18) 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 $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} W_{ij}$, $j = 1, \ldots, k$ are correlated, the asymptotic variance matrix of $\sqrt{n}\nabla sc\ell_n(\theta_0)$ needs to account for cross covariance terms leading to

$$\sqrt{n}\nabla sc\ell_n(\theta_0) \rightsquigarrow N\left(0, \mathsf{Var}_{\theta_0}\left(\sum_{j=1}^{k} \beta_j \lambda_j \nabla S_{\theta_0}(A_j, B_j)\right)\right). \tag{20}$$

We finish the proof by combining (18), (19) and (20) using Slutsky's theorem. ■

Recall our notation for the case that the true model $P \notin \{p_\theta : \theta \in \Theta\}$.

$$\psi_\theta(X,Z) \stackrel{\text{def}}{=} \nabla m_\theta(X,Z)$$
$$\dot{\psi}_\theta(X,Z) \stackrel{\text{def}}{=} \nabla^2 m_\theta(X,Z) \quad \text{(matrix of second order derivatives)}$$
$$\Psi_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \psi_\theta(X^{(i)},Z^{(i)}).$$

Proposition 7 *Assuming the conditions in Proposition 1 as well as* $\sup_{\theta:\|\theta-\theta_0\|\geq\varepsilon} M(\theta) < M(\theta_0)$ *for all* $\varepsilon > 0$ *we have* $\hat{\theta}_n^{msl} \to \theta_0$ *as* $n \to \infty$ *with probability 1.*

Proof We assert

$$P\left\{\lim_{n\to\infty}\sup_{\theta\in S}|scl'(\theta)-\mu(\theta)| = 0\right\} = 1. \tag{21}$$

on the compact set $S = \{\theta : c_1 \leq \|\theta-\theta_0\| \leq c_2\}$ as in the proof of Proposition 1. We proceed similarly along the lines of Proposition 1, with the necessary modification due to the fact that the true model is outside the parametric family.

Since the function $\mu(\theta)$ is continuous it attains its negative supremum on the compact S: $\sup_{\theta\in S}\mu(\theta) < \mu(\theta_0) \geq 0$. Combining this fact with (21) 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 $\mu(\theta_0)$ as $\hat{\theta}_n \to \theta_0$, we have that $\hat{\theta}_n^{\text{msl}} \notin S$ for $n > N$. Since c_1, c_2 were chosen arbitrarily $\hat{\theta}_n^{\text{msl}} \to \theta_0$ with probability 1. ■

Proposition 8 *Assuming the conditions of Proposition 2 as well as* $E_{P(X)}E_{P(Z)}\|\psi_{\theta_0}(X,Z)\|^2 < \infty$, $E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0}(X)$ *exists and is non-singular,* $|\ddot{\Psi}_{ij}| = |\partial^2\psi_\theta(x)/\partial\theta_i\theta_j| < g(x)$ *for all* i, j *and* θ *in a neighborhood of* θ_0 *for some integrable* g*, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\theta_0}(X^{(i)},Z^{(i)}) + o_P(1)$$

or equivalently

$$\hat{\theta}_n = \theta_0 - (E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0})^{-1}\frac{1}{n}\sum_{i=1}^n \psi_{\theta_0}(X^{(i)},Z^{(i)}) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof By Taylor's theorem there exists a random vector $\tilde{\theta}_n$ on the line segment between θ_0 and $\hat{\theta}_n$ for which

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0).$$

which we re-arrange as

$$\begin{aligned}\sqrt{n}\dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \sqrt{n}\frac{1}{2}(\hat{\theta}_n - \theta_0)^\top \ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) &= -\sqrt{n}\Psi_n(\hat{\theta}_n) \\ &= -\sqrt{n}\Psi_n(\theta_0) + o_P(1)\end{aligned} \tag{22}$$

where the second equality follows from the fact that $\hat{\theta}_n \xrightarrow{p} \theta_0$ and continuous functions preserves converges in probability.

Since $\dot{\Psi}_n(\theta_0)$ converges by the law of large numbers to $\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_\theta(X,Z)$ and $\ddot{\Psi}_n(\tilde{\theta}_n)$ converges to a matrix of bounded values in the neighborhood of θ_0 (for large n), the lhs of (22) is

$$\sqrt{n}\left(\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_\theta(X,Z) + o_P(1) + \frac{1}{2}(\hat{\theta}_n - \theta_0)O_P(1)\right)(\hat{\theta}_n - \theta_0)$$
$$= \sqrt{n}(\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_\theta(X,Z) + o_P(1))(\hat{\theta}_n - \theta_0)$$

since $\hat{\theta}_n - \theta_0 = o_P(1)$ and $o_P(1)O_P(1) = o_P(1)$ (the notation $O_P(1)$ denotes stochastically bounded and it applies to $\ddot{\Psi}_n(\tilde{\theta}_n)$ as described above). Putting it together we have

$$\sqrt{n}(\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_\theta(X,Z) + o_P(1))(\hat{\theta}_n - \theta_0) = -\sqrt{n}\Psi_n(\theta_0) + o_P(1).$$

Since the matrix $\mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\dot{\psi}_\theta(X,Z) + o_P(1)$ converges to a non-singular matrix, multiplying the equation above by its inverse finishes the proof. ■

Corollary 2 *Assuming the conditions specified in Proposition 4 we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, (E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0})^{-1}(E_{P(X)}E_{P(Z)}\psi_{\theta_0}\psi_{\theta_0}^\top)(E_{P(X)}E_{P(Z)}\dot{\psi}_{\theta_0})^{-1}).$$

Proof Equation (12) follows from (10) by noticing that due to the central limit theorem $\Psi_n(\theta_0)$ (as it is an average of n iid RVs with expectation 0)

$$\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n} \psi_{\theta_0}(X^{(i)}, Z^{(i)}) \rightsquigarrow N(0, \mathsf{E}_{P(X)}\mathsf{E}_{P(Z)}\psi_{\theta_0}\psi_{\theta_0}^\top).$$

Substituting this in the right hand side of (10) and accounting for the modified variance due to the matrix inverse results in (12). ■

Appendix B. List of Figures

The following lists figures by subject.

B.1 Case Study

- Boltzmann Machines
  - Figure 1 Tabular comparison of policies for computation and accuracy
  - Figure 2 Theoretical analysis of asymptotic variance for trace and determinant
  - Figure 3 Computation/accuracy tradeoff

B.2 Experimental Study

- Labeling sentiment with CRFs
  - Figure 4 CRF graphical model
  - Figure 5 PL1/FL & PL1/PL2 for different σ^2 as a function of $\beta \times \lambda$
  - Figure 6 Computation/accuracy tradeoff with empirical unachievable region

- Chunking CoNLL-2000
  Figure 7 CoNLL-2000 data set label counts

  ... generatively (Boltzmann Chains)
  Figure 8 Boltzmann Chain graphical model
  Figure 13 Computation/accuracy tradeoff with empirical unachievable region
  Figure 9 PL1/FL train & test results as a function of $\beta \times \lambda$
  Figure 10 PL1/FL train & test results as a function of β
  Figure 11 PL1/PL2 train & test results as a function of $\beta \times \lambda$
  Figure 12 PL1/PL2 train & test results as a function of β

  ... discriminatively (CRFs)
  Figure 4 CRF graphical model
  Figure 19 Computation/accuracy tradeoff with empirical unachievable region
  Figure 15 PL1/FL train & test results as a function of $\beta \times \lambda$
  Figure 16 PL1/FL train & test results as a function of β
  Figure 17 PL1/PL2 train & test results as a function of $\beta \times \lambda$
  Figure 18 PL1/PL2 train & test results as a function of β

- β heuristic
  Figure 14 for Boltzmann Chains
  Figure 20 for CRFs
  Figure 21 Optimal regularizing parameter as a function of λ

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). *Journal of the Royal Statistical Society B*, 36(2):192–236, 1974.

Y. Bishop, S. Fienberg, and P. Holland. *Discrete multivariate analysis: theory and practice*. MIT press, 1975.

Léon Bottou and Olivier Bousquet. Learning using large datasets. In *Mining Massive DataSets for Security*. IOS Press, 2008.

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

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, second edition, 2005.

D. R. Cox and E. J. Snell. A general definition of residuals (with discussion). *Journal of the Royal Statistical Society B*, 1968.

J. V. Dillon, K. Balasubramanian, and G. Lebanon. Asymptotic analysis of generative semi-supervised learning. In *Proc. of the International Conference on Machine Learning*, 2010.

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

G. Hinton and T. Sejnowski. Optimal perceptual inference. In *Proc. Computer Vision and Pattern Recognition*, 1983.

N. Hjort and C. Varin. ML, PL, and QL in markov chain models. *Scandinavian Journal of Statistics*, 35(1):64–82, 2008.

R. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

G. Liang and B. Yu. Maximum pseudo likelihood estimation in network tomography. *IEEE Trans. Signal Process*, 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.

D. J. C. MacKay. Equivalence of linear boltzmann chains and hidden markov models. *Neural Computation*, 8(1):178–181, 1996.

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.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley, 1980.

F. Sha and F. Pereira. Shallow parsing with conditional random fields. *Proceedings of HLT-NAACL*, pages 213–220, 2003.

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

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

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

D. Vickrey, C. Lin, and D. Koller. Non-local contrastive objectives. In *Proc. of the International Conference on Machine Learning*, 2010.

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

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