An Asymptotic Analysis of Generative, Discriminative, and Pseudolikelihood Estimators

Percy Liang PLIANG@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics, University of California, Berkeley, CA, USA

Abstract

Statistical and computational concerns have motivated parameter estimators based on various forms of likelihood, e.g., joint, conditional, and pseudolikelihood. In this paper, we present a unified framework for studying these estimators, which allows us to compare their relative (statistical) efficiencies. Our asymptotic analysis suggests that modeling more of the data tends to reduce variance, but at the cost of being more sensitive to model misspecification. We present experiments validating our analysis.

Note: this version (updated Oct. 16, 2010) fixes some errors in the original ICML paper, strengthens some results, and provides more discussion.

1. Introduction

Probabilistic models play a prominent role in domains such as natural language processing, bioinformatics, and computer vision, where they provide methods for jointly reasoning about many interdependent variables. For prediction tasks, one generally models a conditional distribution over outputs given an input. There can be reasons, however, for pursuing alternatives to conditional modeling. First, we might be able to leverage additional statistical strength present in the input by using generative methods rather than discriminative ones. Second, the exact inference required for a full conditional likelihood could be intractable; in this case, one might turn to computationally more efficient alternatives such as pseudolikelihood (Besag, 1975).

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

The generative-discriminative distinction has received much attention in machine learning. The standing intuition is that while discriminative methods achieve lower asymptotic error, generative methods might be better when training data are limited. This intuition is supported by the theoretical comparison of Naive Bayes and logistic regression (Ng & Jordan, 2002) and the recent empirical success of hybrid methods (McCallum et al., 2006; Lasserre et al., 2006).

Computational concerns have also spurred the development of alternatives to the full likelihood; these methods can be seen as optimizing an alternate objective or performing approximate inference during optimization. Examples include pseudolikelihood (Besag, 1975), composite likelihood (Lindsay, 1988), tree-reweighted belief propagation (Wainwright et al., 2003), piecewise training (Sutton & McCallum, 2005), agreement-based learning (Liang et al., 2008), and many others (Varin, 2008).

We can think of all these schemes as simply different estimators operating in a single model family. In this work, we analyze the statistical properties of a class of convex composite likelihood estimators for exponential families, which contains the generative, discriminative, and pseudolikelihood estimators as special cases.

The main focus of our analysis is on prediction error. Standard tools from learning theory based on uniform convergence typically only provide upper bounds on this quantity. Moreover, they generally express estimation error in terms of the overall complexity of the model family.[1] In our case, since all estimators operate in the same model family, these tools are inadequate for comparing different estimators.

Instead, we turn to asymptotic analysis, a mainstay

[1] There are more advanced techniques such as local Rademacher complexities, which focus on the relevant regions of the model family, but these typically only apply to empirical risk minimization.

of theoretical statistics. There is much relevant statistical work on the estimators that we treat; note in particular that Lindsay (1988) used asymptotic arguments to show that composite likelihoods are generally less efficient than the joint likelihood. The majority of these results are, however, focused on parameter estimation. In the current paper, our focus is on prediction, and we also consider model misspecification.

We draw two main conclusions from our analysis: First, when the model is well-specified, conditioning on fewer variables increases statistical efficiency; this to some extent accounts for the better generalization enjoyed by generative estimators and the worse performance of pseudolikelihood estimators. Second, model misspecification can severely increase both the approximation and estimation errors of generative estimators. We confirm our theoretical results by comparing our three estimators on a toy example to verify the asymptotics and on a Markov model for part-of-speech tagging.

2. Exponential Family Estimators

In structured prediction tasks, we are interested in learning a mapping from an input space $\mathcal{X}$ to an output space $\mathcal{Y}$. Probabilistic modeling is a common platform for solving such tasks, allowing for the natural handling of missing data and the incorporation of latent variables.

In this paper, we focus on regular exponential families, which define distributions over an outcome space $\mathcal{Z}$ as follows:

$$p_\theta(z) \stackrel{\text{def}}{=} \exp\{\phi(z)^\top \theta - A(\theta)\} \text{ for } z \in \mathcal{Z}, \tag{1}$$

where $\phi(z) \in \mathbb{R}^d$ is a vector of sufficient statistics (features), $\theta \in \mathbb{R}^d$ is a vector of parameters, and $A(\theta) \stackrel{\text{def}}{=} \log \int e^{\phi(z)^\top \theta} \nu(dz)$ is the log-partition function. In our case, the outcomes are input-output pairs: $z = (x, y)$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

Exponential families include a wide range of popular models used in machine learning. For example, for a conditional random field (CRF) (Lafferty et al., 2001) defined on a graph $G = (V, E)$, we have an output variable for each node ($y = \{y_i\}_{i \in V}$), and the features are $\phi(x, y) = \sum_{i \in V} \phi_{\text{node}}(y_i, \mathbf{x}, i) + \sum_{(i,j) \in E} \phi_{\text{edge}}(y_i, y_j)$.

From the density $p_\theta(z)$, we can compute event probabilities as follows:

$$p_\theta(z \in s) = \exp\{A(\theta; s) - A(\theta)\}, \tag{2}$$

where $A(\theta; s) = \log \int e^{\phi(z)^\top \theta} \mathbb{I}[z \in s] \nu(dz)$ is a conditional log-partition function.

2.1. Composite Likelihood Estimators

In this paper, we consider a class of *composite likelihood* estimators (Lindsay, 1988), which is incidentally equivalent to the multi-conditional learning framework of McCallum et al. (2006). A *composite likelihood* consists of a weighted sum of *component* likelihoods, each of which is the probability of one subset of variables conditioned on another. In this work, we only consider the case where the first set is all the variables.

We adopt the following more fundamental way of specifying the components: Each component r is defined by a *partitioning* of the outcome space $\mathcal{Z}$. We represent a partitioning by an associated *equivalence function* that maps each $z \in \mathcal{Z}$ to its partition:

Definition 1 (Equivalence function)**.** *An* equivalence function r *is a measurable map from* $\mathcal{Z}$ *to measurable subsets of* $\mathcal{Z}$ *such that for each* $z \in \mathcal{Z}$ *and* $z' \in r(z)$, $r(z) = r(z')$.

The component likelihood associated with r takes the following form:

$$p_\theta(z \mid z \in r(z)) = \exp\{\phi(z)^\top \theta - A(\theta; r(z))\}. \tag{3}$$

By maximizing this quantity, we are intuitively taking probability mass away from some neighborhood $r(z)$ of z and putting it on z.

Without loss of generality, assume the component weights sum to 1, so we can think of taking an expectation over a random component R drawn from some fixed distribution P^{r}. We then define the *criterion function*:

$$m_\theta(z) \stackrel{\text{def}}{=} \mathbb{E}_{R \sim P^{\text{r}}} \log p_\theta(z \mid z \in R(z)). \tag{4}$$

Given data points $Z^{(1)}, \ldots, Z^{(n)}$ drawn i.i.d. from some true distribution p^* (not necessarily in the exponential family), the maximum composite likelihood estimator is defined by averaging the criterion function over these data points:

$$\hat{\theta} = \operatorname*{argmax}_\theta \hat{\mathbb{E}} m_\theta(Z), \tag{5}$$

where $\hat{\mathbb{E}} m_\theta(Z) = \frac{1}{n} \sum_{i=1}^n m_\theta(Z^{(i)})$.

We can now place the three estimators of interest in our framework:

Generative: We have one component $r_{\text{g}}(x, y) = \mathcal{X} \times \mathcal{Y}$, which has one partition—the whole outcome space.

Fully discriminative: We have one component $r_{\text{d}}(x, y) = x \times \mathcal{Y}$. The outcomes in each partition have the same value of x, but different y.

$v^{\otimes} = vv^{\top}$
Parameter estimates
Z $= (X, Y)$ data point
p^* true distribution of the data
m_θ criterion function (defines the estimator)
$\mathcal{R}$ risk (expected log-loss)
$\hat{\theta}$ $= \operatorname{argmax}_\theta \hat{\mathbb{E}} m_\theta(Z)$ [empirical parameter estimate]
θ° $= \operatorname{argmax}_\theta \mathbb{E} m_\theta(Z)$ [limiting parameter vector]
Random variables for asymptotic analysis
R $\sim P^{\mathrm{r}}$ [choose composite likelihood component]
S $= R(Z)$ [neighborhood]
$r_{\mathrm{d}}(x, y) = x \times \mathcal{Y}$ [fully discriminative component]
$S_{\mathrm{d}} = r_{\mathrm{d}}(Z)$ [discriminative neighborhood]
ϕ $= \phi(Z)$, $Z \sim p^*$ [sample from true distribution]
$\phi^{\mathrm{m}} = \phi(Z^{\mathrm{m}})$, $Z^{\mathrm{m}} \sim p_{\theta^\circ}(\cdot \mid \cdot \in R(Z))$ [for estimation]
ϕ^{e} $= \phi(Z^{\mathrm{e}})$, $Z^{\mathrm{e}} \sim p_{\theta^\circ}(\cdot \mid \cdot \in r_{\mathrm{d}}(Z))$ [for prediction]

Table 1. Notation used in the paper. See Figure 1 for a visualization of the random variables.

Pseudolikelihood discriminative: Assume $y = \{y_i\}_{i \in V}$. For each $i \in V$, we have a component $r_i(x, y) = \{(x', y') : x' = x, y' \in \mathcal{Y}, y'_j = y_j \text{ for } j \neq i\}$. P^{r} is uniform over these components.

2.2. Prediction and Evaluation

Given a parameter estimate $\hat{\theta}$, we make predictions based on $p_{\hat{\theta}}(y \mid x)$. In this paper, we evaluate our model according to log-loss; the risk is the expected log-loss:

$$\mathcal{R}(\theta) = \mathbb{E}_{(X,Y) \sim p^*}[-\log p_\theta(Y \mid X)]. \tag{6}$$

The quality of an estimator is determined by the gap between the risk of the estimate $\mathcal{R}(\hat{\theta})$ and the Bayes risk $\mathcal{R}^* = H(Y \mid X)$. It will be useful to relate these two via the risk of $\theta^\circ = \operatorname{argmax}_\theta \mathbb{E}_{Z \sim p^*} m_\theta(Z)$, which leads to the following standard decomposition:

$$\underbrace{\mathcal{R}(\hat{\theta}) - \mathcal{R}^*}_{\text{total error}} = \underbrace{(\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\circ))}_{\text{estimation error}} + \underbrace{(\mathcal{R}(\theta^\circ) - \mathcal{R}^*)}_{\text{approx. error}}. \tag{7}$$

The estimation error is due to having only finite data; the approximation error is due to the intrinsic suboptimality of the estimator.[2]

3. Asymptotic Analysis

We first compute the asymptotic estimation errors of composite likelihood estimators in general (Sections 3.1 and 3.2). Then we use these results to compare the estimators of interest (Sections 3.3 and 3.4).

[2]Note that θ° is not necessarily the minimum risk parameter vector in the model family.

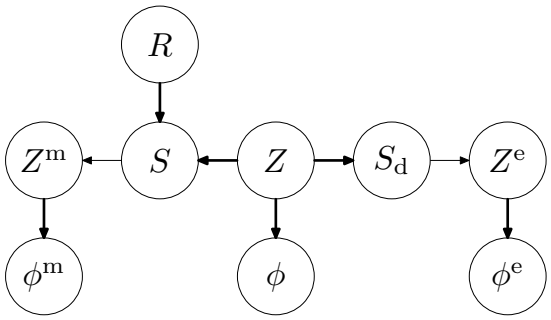


Figure 1. Graphical model showing the dependencies between the random variables in our analysis (these variables are defined formally in Table 1). Thick edges entering a node denote that the node is a deterministic function of its parents. Note the independence assumptions: conditioned on the neighborhood $S = R(Z)$ ($S_{\mathrm{d}} = r_{\mathrm{d}}(Z)$), ϕ is independent of ϕ^{m} (ϕ^{e}). This independence will be useful for decomposing variances in our analysis.

In this paper, we assume that our exponential family is identifiable.[3] Also assume that our estimators converge ($\hat{\theta} \xrightarrow{P} \theta^\circ$) and are consistent when the model is well-specified (if $p^* = p_{\theta^*}$, then $\theta^\circ = \theta^*$). Note, however, that in general we do not assume that our model is well-specified.

Our asymptotic analysis is driven by Taylor expansions, so we need to compute a few derivatives. The derivatives of the log-partition function are moments of the sufficient statistics (a standard result, see Wainwright and Jordan (2003), for example):

$$\dot{A}(\theta; s) = \mathbb{E}_{Z \sim p_\theta(\cdot \mid \cdot \in s)}(\phi(Z)) \tag{8}$$
$$\ddot{A}(\theta; s) = \operatorname{var}_{Z \sim p_\theta(\cdot \mid \cdot \in s)}(\phi(Z)). \tag{9}$$

From these moments, we can obtain the derivatives of m_{θ° and $\mathcal{R}$ (to simplify notation, we express these in terms of random variables whose distributions are defined in Table 1):

$$\dot{m}_{\theta^\circ} = \phi - \mathbb{E}(\phi^{\mathrm{m}} \mid Z) \tag{10}$$
$$\ddot{m}_{\theta^\circ} = -\mathbb{E}[\operatorname{var}(\phi^{\mathrm{m}} \mid S) \mid Z] \tag{11}$$
$$\dot{\mathcal{R}}(\theta^\circ) = \mathbb{E}(\phi^{\mathrm{e}} - \phi) \tag{12}$$
$$\ddot{\mathcal{R}}(\theta^\circ) = \mathbb{E} \operatorname{var}(\phi^{\mathrm{e}} \mid S). \tag{13}$$

3.1. Asymptotics of the Parameters

We first analyze how fast $\hat{\theta}$ converges to θ° by computing the asymptotic distribution of $\hat{\theta} - \theta^\circ$. In Section 3.2

[3]In the non-identifiable case, the analysis becomes more cluttered, but the results are essentially the same, since predictions depend on only the distributions induced by the parameters. See the longer version of this paper for an in-depth discussion.

we use this result to get the asymptotic distribution of the estimation error $\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\circ)$.

The following standard lemma will prove to be very useful in our analysis:

Lemma 1. *For random vectors X, Y, we have* $\operatorname{var}(X) = \mathbb{E}[\operatorname{var}(X \mid Y)] + \operatorname{var}[\mathbb{E}(X \mid Y)]$.

The important implication of this lemma is that conditioning on another variable Y reduces the variance of X. This lemma already hints at how conditioning on more variables can lead to poorer estimators: conditioning reduces the variance of the data, which can make it harder to learn about the parameters.

Another standard result we will use is the multivariate Cauchy-Schwarz inequality:

Lemma 2. *For random vectors X, Y with $\mathbb{E}(Y^{\otimes}) \succ 0$, we have $\mathbb{E}(X^{\otimes}) \succeq \mathbb{E}[XY^\top]\mathbb{E}(Y^{\otimes})^{-1}\mathbb{E}[YX^\top]$.*

The following theorem gives us the asymptotic variance of a general composite likelihood estimator:

Theorem 1 (Asymptotic distribution of the parameters)**.** *Assume $\hat{\theta} \xrightarrow{P} \theta^\circ$. Then*

$$\sqrt{n}(\hat{\theta} - \theta^\circ) \to \mathcal{N}(0, \Sigma). \tag{14}$$

The asymptotic variance is

$$\Sigma = \Gamma^{-1} + \Gamma^{-1}(C_{\mathrm{m}} - C_{\mathrm{c}})\Gamma^{-1}, \tag{15}$$

where

$$\Gamma = \mathbb{E}\operatorname{var}(\phi^{\mathrm{m}} \mid S) \tag{16}$$

is the sensitivity,

$$C_{\mathrm{c}} = \mathbb{E}\operatorname{var}[\mathbb{E}(\phi^{\mathrm{m}} \mid S) \mid Z] \tag{17}$$

is the component correction, and

$$C_{\mathrm{m}} = \mathbb{E}[\operatorname{var}(\phi \mid S) - \operatorname{var}(\phi^{\mathrm{m}} \mid S)] \tag{18}$$
$$+ \mathbb{E}[\mathbb{E}(\phi \mid S) - \mathbb{E}(\phi^{\mathrm{m}} \mid S)]^{\otimes} \tag{19}$$

is the misspecification correction.

Before we prove the theorem, let us use the decomposition in (15) to make several qualitative judgments. First, the *sensitivity* $\Gamma = \mathbb{E}\operatorname{var}(\phi^{\mathrm{m}} \mid S)$ is the expected amount of variation in the features given the (random) neighborhood $S = R(Z)$ of our data point Z. The larger the sensitivity, the more the data can tell us about the parameters, and thus the lower the asymptotic variance will be.

The *component correction* C_{c} intuitively measures the amount of variation in feature expectations $\mathbb{E}(\phi^{\mathrm{m}} \mid S)$ across different components R. C_{c} is zero for the generative and fully discriminative estimators, which have one component, but positive for the pseudolikelihood discriminative estimator, which has more than one component. The negative contribution of C_{c} to the asymptotic variance suggests that having many diverse components could in general be helpful. However, this diversity is somewhat at odds with having larger neighborhoods, which tends to increase the sensitivity Γ but reduce the component correction C_{c}.

The *misspecification correction* C_{m} is zero when the model is well-specified (in this case, $\phi^{\mathrm{m}} \mid S \stackrel{d}{=} \phi \mid S$), but is in general nonzero under model misspecification. In this latter case, one incurs a nonzero approximation error (defined in (7)) as expected, but the more subtle point is that there is also a nonzero effect on estimation error.

Proof. The standard asymptotic normality result for M-estimators (Theorem 5.21 of van der Vaart (1998)), which includes composite likelihood estimators, gives us the asymptotic variance:

$$\Sigma = (\mathbb{E}\ddot{m}_{\theta^\circ})^{-1}(\mathbb{E}\dot{m}_{\theta^\circ}^{\otimes})(\mathbb{E}\ddot{m}_{\theta^\circ})^{-1}. \tag{20}$$

The remainder of the proof simply re-expresses Σ in terms of the quantities in our theorem. Subtracting and adding an intermediate ϕ^{m} to (10) and taking the outer product yields:

$$\mathbb{E}\dot{m}_{\theta^\circ}^{\otimes} = \mathbb{E}[(\phi - \phi^{\mathrm{m}}) + (\phi^{\mathrm{m}} - \mathbb{E}(\phi^{\mathrm{m}} \mid Z))]^{\otimes}.$$

We distribute this outer product over the sum of the two terms, yielding (i) $\mathbb{E}[(\phi - \phi^{\mathrm{m}})^{\otimes}]$, (ii) $\mathbb{E}[\operatorname{var}(\phi^{\mathrm{m}} \mid Z)]$, and (iii) the pair of cross terms, which is exactly $-2\mathbb{E}[\operatorname{var}(\phi^{\mathrm{m}} \mid Z)]$ (since conditioned on Z, ϕ is a constant).

To handle (i), we condition on S, rendering ϕ and ϕ^{m} conditionally independent (see Figure 1). Then we subtract and add various expectations and observe that all the cross terms are zero due to (conditional) independence:

$$\mathbb{E}[(\phi - \phi^{\mathrm{m}})^{\otimes}] \tag{21}$$
$$= \mathbb{E}\mathbb{E}\big[\big((\phi - \mathbb{E}(\phi \mid S)) - (\phi^{\mathrm{m}} - \mathbb{E}(\phi^{\mathrm{m}} \mid S)) + \tag{22}$$
$$(\mathbb{E}(\phi \mid S) - \mathbb{E}(\phi^{\mathrm{m}} \mid S))\big)^{\otimes} \mid S\big] \tag{23}$$
$$= \mathbb{E}\operatorname{var}(\phi \mid S) + \mathbb{E}\operatorname{var}(\phi^{\mathrm{m}} \mid S) + \tag{24}$$
$$\mathbb{E}(\mathbb{E}(\phi \mid S) - \mathbb{E}(\phi^{\mathrm{m}} \mid S))^{\otimes} \tag{25}$$
$$= C_{\mathrm{m}} + 2\mathbb{E}\operatorname{var}(\phi^{\mathrm{m}} \mid S). \tag{26}$$

Summarizing our progress so far, we have

$$\mathbb{E}\dot{m}_{\theta^\circ}^{\otimes} = C_{\mathrm{m}} + 2\mathbb{E}\operatorname{var}(\phi^{\mathrm{m}} \mid S) - \mathbb{E}\operatorname{var}(\phi^{\mathrm{m}} \mid Z).$$

We now apply Lemma 1 (with $X = \phi^{\mathrm{m}}$ and $Y = R$ conditioned on Z) to decompose the last term:

$$\mathbb{E}\operatorname{var}(\phi^{\mathrm{m}} \mid Z) = \mathbb{E}\operatorname{var}(\phi^{\mathrm{m}} \mid S) + \mathbb{E}\operatorname{var}[\mathbb{E}(\phi^{\mathrm{m}} \mid S) \mid Z].$$

Note that we write S in place of R, Z because ϕ^{m} only depends on R and Z through $S = R(Z)$. Putting these results together, and substituting the definitions of C_{c} and Γ, we get

$$\mathbb{E}\dot{m}_{\theta^\circ}^{\otimes} = C_{\mathrm{m}} + \Gamma - C_{\mathrm{c}}.$$

From (11), we have $\mathbb{E}\ddot{m}_{\theta^\circ} = -\Gamma$. Some simple algebra yields the formula of the asymptotic variance (15). □

3.2. Asymptotics of the Risk

The following theorem turns Theorem 1 from a statement about the asymptotic distribution of the parameters into one about the risk:

Theorem 2 (Asymptotic distribution of the risk)**.** *Let Σ be the asymptotic variance as defined in (15). Denote $\dot{\mathcal{R}} \stackrel{\text{def}}{=} \dot{\mathcal{R}}(\theta^\circ)$ and $\ddot{\mathcal{R}} \stackrel{\text{def}}{=} \ddot{\mathcal{R}}(\theta^\circ)$. Then*

$$\sqrt{n}(\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\circ)) \xrightarrow{d} \mathcal{N}\left(0, \dot{\mathcal{R}}^\top \Sigma \dot{\mathcal{R}}\right). \tag{27}$$

Furthermore, if $\dot{\mathcal{R}} = 0$, then

$$n(\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\circ)) \xrightarrow{d} \frac{1}{2}\operatorname{tr} \mathcal{W}\left(\ddot{\mathcal{R}}^{\frac{1}{2}} \Sigma \ddot{\mathcal{R}}^{\frac{1}{2}}, 1\right), \tag{28}$$

where $\mathcal{W}(V, n)$ is the Wishart distribution with n degrees of freedom.

Proof. Perform a Taylor expansion of the risk function around θ°:

$$\mathcal{R}(\hat{\theta}) = \mathcal{R}(\theta^\circ) + \dot{\mathcal{R}}^\top(\hat{\theta} - \theta^\circ) + \frac{1}{2}(\hat{\theta} - \theta^\circ)^\top \ddot{\mathcal{R}}(\hat{\theta} - \theta^\circ) + o(||\hat{\theta} - \theta^\circ||^2). \tag{29}$$

We use a standard argument known as the delta method (van der Vaart, 1998). Multiplying (29) on both sides by $\sqrt{n}$, rearranging terms, and applying Slutsky's theorem, we get (27). However, when $\dot{\mathcal{R}} = 0$, the first-order term of the expansion (29) is zero, so we must consider the second-order term to get a non-degenerate distribution. Note that $\ddot{\mathcal{R}}$ is positive semidefinite. Multiplying (29) by n and rearranging yields the following:

$$n(\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\circ)) = \frac{1}{2}\operatorname{tr}\left([\ddot{\mathcal{R}}^{\frac{1}{2}}\sqrt{n}(\hat{\theta} - \theta^\circ)]^{\otimes}\right) + \cdots$$

Since $\ddot{\mathcal{R}}^{\frac{1}{2}}\sqrt{n}(\hat{\theta} - \theta^\circ) \xrightarrow{d} \mathcal{N}(0, \ddot{\mathcal{R}}^{\frac{1}{2}}\Sigma\ddot{\mathcal{R}}^{\frac{1}{2}})$, applying the continuous mapping theorem with the outer product function yields a Wishart as the limiting distribution. Thus, $n(\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\circ))$ is asymptotically equal in distribution to $\frac{1}{2}$ times the trace of a sample from that Wishart distribution. □

We can also understand (28) in the following way. Let $V = \ddot{\mathcal{R}}^{\frac{1}{2}}\Sigma\ddot{\mathcal{R}}^{\frac{1}{2}}$. Note that $\frac{1}{2}\operatorname{tr}\mathcal{W}(V, 1) \stackrel{d}{=} \frac{1}{2}\operatorname{tr}(V\mathcal{W}(I, 1))$, which is the distribution of a weighted sum of independent χ^2_1 variables, where the weights are determined by the diagonal elements of V. The mean of this distribution is $\frac{1}{2}\operatorname{tr}(V)$ and the variance is $\operatorname{tr}(V \bullet V)$, where $\bullet$ denotes elementwise product.

An important question is when we obtain the ordinary $O(n^{-\frac{1}{2}})$ convergence (27) versus the much better $O(n^{-1})$ convergence (28). A sufficient condition for $O(n^{-1})$ convergence is $\dot{\mathcal{R}}(\theta^\circ) = 0$. When the model is well-specified, this is true for any consistent estimator.

Even if the model is misspecified, the fully discriminative estimator still achieves the $O(n^{-1})$ rate. The reason is that whenever a training criterion m_θ is the same (up to constants) as the test criterion $\mathcal{R}(\cdot)$, $\dot{\mathcal{R}}$ vanishes and we obtain the $O(n^{-1})$ rate. This is in concordance with a related observation made by Wainwright (2006) that it is better to use the same inference procedure at both training and test time.

When the model is well-specified, there is another appealing property that holds if the training and test criterion are the same up to constants: the asymptotic distribution of the risk depends on only the dimensionality of the exponential family, not the actual structure of the model. In particular, for composite likelihood estimators with one component, $\Sigma = \Gamma^{-1} = (-\mathbb{E}\ddot{m}_{\theta^\circ})^{-1} = \ddot{\mathcal{R}}^{-1}$. Therefore, $\ddot{\mathcal{R}}^{\frac{1}{2}}\Sigma\ddot{\mathcal{R}}^{\frac{1}{2}} = I_d$ and so $n(\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^\circ)) \xrightarrow{d} \frac{1}{2}\operatorname{tr}\mathcal{W}(I_d, 1) \stackrel{d}{=} \frac{1}{2}\chi^2_d$, where d is the number of parameters. This result is essentially another way of looking at the fact that the likelihood ratio test statistic is asymptotically distributed as χ^2.

3.3. Comparing Estimation Errors

In the previous section, we analyzed the asymptotics of a single estimator. Now, given two estimators, we would like to be able to tell which one is better. In order to compare two estimators, it would be convenient if they converged to the same limit. In this section, we ensure this by assuming that the model is well-specified and that our estimators are consistent.

Since all parameter estimates are used in the same way for prediction, it suffices to analyze the relative efficiencies of the parameter estimates. The following theorem says that coarser partitionings of $\mathcal{Z}$ generally lead to more efficient estimators:

Theorem 3 (Asymptotic relative efficiency)**.** *Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two consistent estimators with asymptotic variances Σ_1 and Σ_2 as defined in (15). Assume that either R_1 or R_2 is constant (that is, either $\hat{\theta}_1$ or $\hat{\theta}_2$ has exactly one component) and $R_1(z) \supset R_2(z)$ for all $z \in \mathcal{Z}$. If the model is well-specified, then $\Sigma_1 \preceq \Sigma_2$ ($\hat{\theta}_1$ is no worse than $\hat{\theta}_2$).*

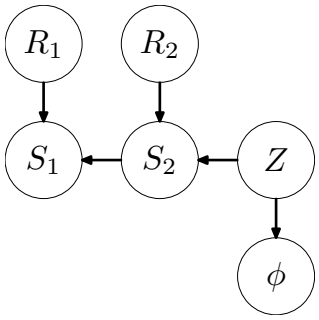


Figure 2. Graphical model showing the dependencies between the random variables in the proof of Theorem 3. We define $S_1 = \cup_{z \in S_2} R_1(z)$.

Proof. In order to compare the two estimators, it will be useful to define the random variables associated with each on the same probability space. We do this in Figure 2. The idea is that since $S_1 \supset S_2$ and R_1, R_2 are equivalence relations, $S_1 = \cup_{z \in S_2} R_1(z) = R_1(Z)$. In other words, from the point of view of the first estimator, R_2, S_2 might as well not have been there. Intuitively, S_2 preserves strictly more information about Z than S_1, so applying S_2 first is harmless. Furthermore, since the model is well-specified, Z and Z^{m} have the same distribution conditioned on either S_1 or S_2 ($\phi^{\mathrm{m}} \stackrel{d}{=} \phi \mid S_k$, $k = 1, 2$). Operationally, this allows us to replace ϕ^{m} with ϕ whenever we condition on S_1 or S_2.

The proof is carried out in two parts. In the first part, we prove the theorem for the case where R_2 is constant. We will show that $\Gamma_1^{-1} \preceq \Gamma_2^{-1}$, where Γ_1 and Γ_2 are the sensitivities of the two estimators. Note $\Gamma_k = \mathbb{E}\,\mathrm{var}(\phi \mid S_k)$. We use Lemma 1 (with $X = \phi$ and $Y = S_2$ conditioned on S_1) to decompose the variance:

$$\Gamma_1 = \mathbb{E}\,\mathrm{var}(\phi \mid S_2, S_1) + \mathbb{E}\,\mathrm{var}[\mathbb{E}(\phi \mid S_2, S_1) \mid S_1].$$

Since ϕ is conditionally independent of S_1 given S_2, $\mathrm{var}(\phi \mid S_2, S_1) = \mathrm{var}(\phi \mid S_2)$, which is exactly Γ_2. The second term is positive semidefinite, so $\Gamma_1 \succeq \Gamma_2$, which implies $\Gamma_1^{-1} \preceq \Gamma_2^{-1}$.

Let $C_{\mathrm{c}1}$ and $C_{\mathrm{c}2}$ be the component corrections of the two estimators. Note that $C_{\mathrm{c}2} = 0$ because R_2 is constant, so $C_{\mathrm{c}1} \succeq C_{\mathrm{c}2}$. The misspecification corrections are both zero. Putting these results together, we get that $\Sigma_1 \preceq \Sigma_2$.

Now, for the second part of the proof, let us assume that R_1 is constant. The goal is to rewrite Σ_1^{-1} and Σ_2^{-1} and apply Lemma 2 to conclude that $\Sigma_1^{-1} \succeq \Sigma_2^{-1}$. Let U_1 and U_2 be $\dot{m}_{\theta^\circ}$ corresponding to the two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. Since R_1 is a constant and the model is well-specified, for the first estimator, we have:

$$\Sigma_1^{-1} = \Gamma_1 \Gamma_1^{-1} \Gamma_1 = \Gamma_1 = \mathbb{E}(U_1^{\otimes}).$$

For the second estimator, we have

$$\Sigma_2^{-1} = \mathbb{E}\mathrm{var}(\phi \mid S_2)\mathbb{E}(U_2^{\otimes})^{-1}\mathbb{E}\mathrm{var}(\phi \mid S_2).$$

Our next step is to show $\mathbb{E}\mathrm{var}(\phi \mid S_2) = \mathbb{E}(U_1 U_2^\top)$. We can rewrite the definition of U_k (10) for $k = 1, 2$:

$$U_k = \phi - \mathbb{E}(\phi^{\mathrm{m}} \mid Z) = \phi - \sum_r P_k^{\mathrm{r}}(r)\mathbb{E}(\phi \mid r(Z)),$$

where we used the fact that $\mathbb{E}(\phi \mid r(Z)) = \mathbb{E}(\phi^{\mathrm{m}} \mid r(Z))$ for any r (since the model is well-specified). Plugging in definitions for U_1 and U_2 and simplifying:

$$\begin{aligned}
&\mathbb{E}(U_1 U_2^\top) \qquad (30)\\
=&\mathbb{E}[(\phi - \mathbb{E}(\phi \mid S_1))(\phi - \sum_r P_2^{\mathrm{r}}(r)\mathbb{E}(\phi \mid r(Z)))^\top]\\
=&\sum_r P_2^{\mathrm{r}}(r)\mathbb{E}[(\phi - \mathbb{E}(\phi \mid S_1))(\phi - \mathbb{E}(\phi \mid r(Z)))^\top]\\
=&\sum_r P_2^{\mathrm{r}}(r)\mathbb{E}\mathrm{var}(\phi \mid r(Z))\\
=&\mathbb{E}\mathrm{var}(\phi \mid S_2),
\end{aligned}$$

where in the second equality, we used the key fact that conditioned on $r(Z)$, $\mathbb{E}(\phi \mid S_1)$ is constant because R_1 is constant and $S_1 \supset r(Z)$. This allows us to write

$$\Sigma_2^{-1} = \mathbb{E}(U_1^{\otimes}) \succeq \mathbb{E}(U_1 U_2^\top)\mathbb{E}(U_2^{\otimes})\mathbb{E}(U_2 U_1^\top) = \Sigma_2^{-1}.$$

Applying Lemma 2 yields $\Sigma_1^{-1} \succeq \Sigma_2^{-1}$. □

One might wonder if we really need one of R_1 or R_2 to be constant. Is it not enough to just assume that $R_1(z) \supset R_2(z)$ (for some coupling of the random variables R_1 and R_2)? The answer is in general no, as the following counterexample shows:

Counterexample Let $\mathcal{Z} = \{1, 2, 3\}$. The general shape of the distribution is given by the single feature $\phi(1) = 1, \phi(2) = 3, \phi(3) = 2$ and a scalar parameter θ controls the peakiness of the distribution. Let the true parameter be $\theta^* = 1$. Consider two estimators: $\hat{\theta}_1$ has two components, $r_{1a} = \{\{1, 2\}, \{3\}\}$

and $r_{1b} = \{\{1\}, \{2, 3\}\}$; $\hat{\theta}_2$ also has two components, $r_{2a} = \{\{1, 2\}, \{3\}\}$ and $r_{2b} = \{\{1\}, \{2\}, \{3\}\}$. Both estimators place $\frac{1}{2}$ probability on each component.

Coupling r_{1a} with r_{2a} and r_{1b} with r_{2b}, we have $R_1(z) \supset R_2(z)$. However, we computed and found that $\Sigma_1 \cong 2.36$ and $\Sigma_2 \cong 3.15$, so $\hat{\theta}_2$ actually has higher asymptotic variance although it has finer partitionings.

To explain this, note that the contribution of r_{2b} to the criterion function is zero, so the second estimator is equivalent to just using the single component r_{2a} ($= r_{1a}$), so the first estimator actually suffers by using the additional component r_{1b}. In general, while we would still expect coarser partitionings to be better even for estimators with many components, this counterexample shows that we must exercise caution.

3.4. Comparing Estimators

Finally, we use Theorem 3 to compare the estimation and approximation errors of the generative ($\hat{\theta}_{\mathrm{g}}$), fully discriminative ($\hat{\theta}_{\mathrm{d}}$), and pseudolikelihood discriminative ($\hat{\theta}_{\mathrm{p}}$) estimators. The subscripts $\mathrm{g}, \mathrm{d}, \mathrm{p}$ will be attached to other variables to refer to the quantities associated with the corresponding estimators. In the following corollaries, we use the word "lower" loosely to mean "no more than," although in general we expect the inequality to be strict.

Corollary 1 (Generative versus fully discriminative)**.** *(1) If the model is well-specified, $\hat{\theta}_{\mathrm{g}}$ has lower asymptotic estimation error than $\hat{\theta}_{\mathrm{d}}$; both have zero approximation error. (2) If the model is misspecified, $\hat{\theta}_{\mathrm{d}}$ has lower approximation and asymptotic estimation errors than $\hat{\theta}_{\mathrm{g}}$.*

Proof. For (1), since $R_{\mathrm{d}}(z) \subset R_{\mathrm{g}}(z)$, we have $\Sigma_{\mathrm{g}} \preceq \Sigma_{\mathrm{d}}$ by Theorem 3. Zero approximation error follows from consistency. For (2), since the discriminative estimator achieves the minimum risk in the model family, it has the lowest approximation error. Also, by Theorem 2 and the ensuing discussion, it always converges at a $O(n^{-1})$ rate, whereas the generative estimator will in general converge at a $O(n^{-\frac{1}{2}})$ rate. □

Note that there is a qualitative change of asymptotics in going from the well-specified to the misspecified scenario. This discontinuity demonstrates one weakness of asymptotic analyses: we would expect that for a very minor model misspecification, the generative estimator would still dominate the discriminative estimator for moderate sample sizes, but even a small misspecification is magnified in the asymptotic limit.

In the following toy example where the model is well-specified, we see concretely that the generative estimator has smaller asymptotic estimation error:

Example Consider a model where x and y are binary variables: $\phi(x, y)^\top \theta = \theta_0 \mathbb{I}[x = 0, y = 1] + \theta_1 \mathbb{I}[x = 1, y = 1]$, where the true parameters are $\theta^* = (0, 0)$. We can compute $\Gamma_{\mathrm{g}} = \mathrm{var}(\phi) = \frac{1}{16}\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$ and $\ddot{\mathcal{R}}(\theta^*) = \Gamma_{\mathrm{d}} = \mathbb{E}\,\mathrm{var}(\phi \mid X) = \frac{1}{16}\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. The mean asymptotic estimation error (scaled by n) of the generative estimator is $\frac{1}{2}\mathrm{tr}(\Gamma_{\mathrm{d}}\Gamma_{\mathrm{g}}^{-1}) = \frac{3}{4}$ while that of the discriminative estimator is $\frac{1}{2}\mathrm{tr}(\Gamma_{\mathrm{d}}\Gamma_{\mathrm{d}}^{-1}) = 1$.

We now show that fully discriminative estimators are statistically superior to pseudolikelihood discriminative estimators in all regimes, but of course pseudolikelihood is computationally more efficient.

Corollary 2 (Fully discriminative versus pseudolikelihood discriminative)**.** *(1) If the model is well-specified, $\hat{\theta}_{\mathrm{d}}$ has lower asymptotic estimation error than $\hat{\theta}_{\mathrm{p}}$; both have zero approximation error. (2) If the model is misspecified, $\hat{\theta}_{\mathrm{d}}$ has lower approximation and asymptotic estimation errors than $\hat{\theta}_{\mathrm{p}}$.*

Proof. For (1), since $R_{\mathrm{p}}(z) \subset R_{\mathrm{d}}(z)$ and R_{d} is constant, $\Sigma_{\mathrm{d}} \preceq \Sigma_{\mathrm{p}}$ by Theorem 3. Zero approximation error follows from consistency. For (2), the same arguments as the corresponding part of the proof of Corollary 1 apply. □

4. Experiments

In this section, we validate our theoretical analysis empirically. First, we evaluate the three estimators on a simple graphical model which allows us to plot the real asymptotics of the estimation error (Section 4.1). Then we show that in the non-asymptotic regime, the qualitative predictions of the asymptotic analyses are also valid (Section 4.2).

4.1. A Simple Graphical Model

Consider a four-node binary-valued graphical model where $z = (x_1, x_2, y_1, y_2)$. The true model family p^* is an Markov random field parametrized by $\theta^* = (\alpha^*, \beta^*, \gamma^*)$ as follows:

$$\begin{aligned}\phi(z)^\top \theta \quad = \quad & \alpha \mathbb{I}[y_1 = y_2] + \beta(\mathbb{I}[x_1 = y_1] + \mathbb{I}[x_2 = y_2]) + \\ & \gamma(\mathbb{I}[x_1 = y_2] + \mathbb{I}[x_2 = y_1]).\end{aligned}$$

To emulate misspecification, we set γ^* to be nonzero and force $\gamma = 0$ during parameter estimation.

In the first experiment, we estimated the variance (by running 10K trials) of the estimation error as we increased the number of data points. We set $\alpha^* = \beta^* = 1$ for the true model. When $\gamma^* = 0$ (the model is well-specified), Figures 3(a)–(c) show that scaling the variance by n yields a constant; this implies that all three estimators achieve $O(n^{-1})$ convergence.

When the model is misspecified with $\gamma^* = 0.5$ (Figures 3(d)–(f)), there is a sharp difference between the rates of the generative and discriminative estimators. The fully discriminative estimator still enjoys the $O(n^{-1})$ convergence; scaling by n reveals that the generative and pseudolikelihood discriminative estimators are only attaining a $O(n^{-\frac{1}{2}})$ rate as predicted by Theorem 2 (Figure 3(f)). Note that the generative estimator is affected most severely.

Figures 3(g)–(h) demonstrate the non-asymptotic impact of varying the parameters of the graphical model in terms of the total error. In (g), as we increase the amount of misspecification γ, the error increases for all estimators, but most sharply for the generative estimator. In (h), as we increase the strength of the edge potential α, the pseudolikelihood discriminative estimator suffers, the fully discriminative estimator is unaffected, and the generative estimator actually improves.

4.2. Part-of-speech Tagging

In this section, we present experiments on part-of-speech (POS) tagging. In POS tagging, the input is a sequence of words $x = (x_1, \ldots, x_\ell)$ and the output is a sequence of POS tags $y = (y_1, \ldots, y_\ell)$, e.g., noun, verb, etc. (There are 45 tags total.) We consider the following model, specified by the following features (roughly 2 million total):

$$\phi(x, y) = \sum_{i=1}^{\ell} \phi_{\text{node}}(y_i, x_i) + \sum_{i=1}^{\ell-1} \phi_{\text{edge}}(y_i, y_{i+1}), \quad (31)$$

where the node features $\phi_{\text{node}}(y_i, x_i)$ are a vector of indicator functions of the form $\mathbb{I}[y_i = a, x_i = b]$, and the edge features $\phi_{\text{edge}}(y_i, y_{i+1})$ are a vector of indicator functions of the form $\mathbb{I}[y_i = a, y_{i+1} = b]$. Trained generatively, this model is essentially an HMM, but slightly more expressive. Trained (fully) discriminatively, this model is a CRF.

We used the Wall Street Journal (WSJ) portion of the Penn Treebank, with sections 0–21 for training (38K sentences) and 22–24 for testing (5.5K sentences). Table 2(a) shows that the discriminative estimators perform better than the generative one. This is not surprising given that the model is misspecified (language does not come from an HMM).

	Accuracy		Log-loss	
	Train	Test	Train	Test
Gen.	0.940	0.935	4.628	4.945
Fully dis.	**0.977**	**0.956**	**1.480**	**3.120**
Pseudo dis.	0.975	0.955	1.562	3.170

(a) Real data (misspecified)

	Accuracy		Log-loss	
	Train	Test	Train	Test
Gen.	0.989	**0.898**	0.570	**7.297**
Full dis.	**0.992**	0.879	**0.407**	12.431
Pseudo dis.	0.990	0.891	0.469	10.840

(b) Synthetic data (well-specified)

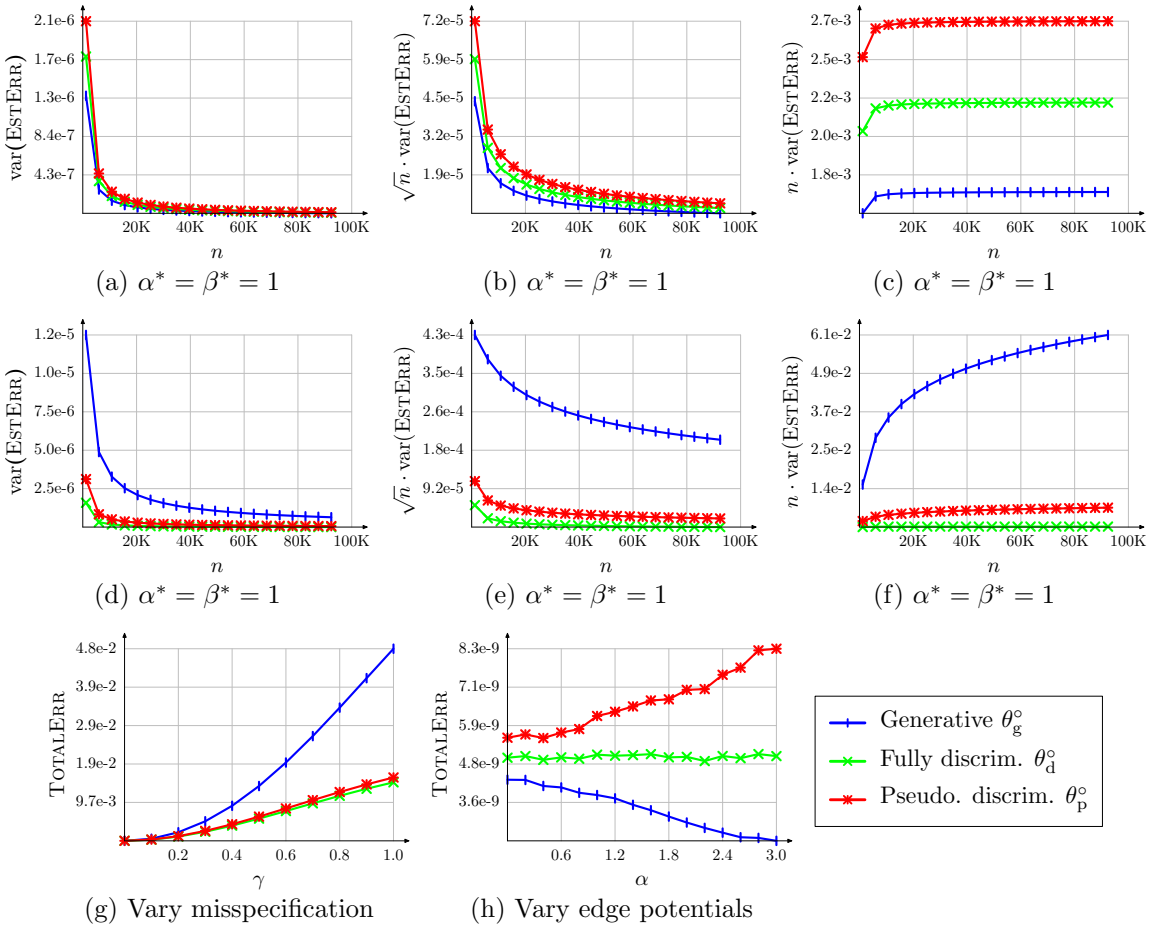
Table 2. Part-of-speech tagging results. Discriminative estimators outperform the generative estimator (on both test accuracy and log-loss) when the model is misspecified, but the reverse is true when the model is well-specified.

To verify that the generative estimator is superior when the model is well-specified, we used the learned generative model in the previous experiment to sample 1000 synthetic training and 1000 synthetic test examples. We then applied the estimators as before on this artificial data. Table 2(b) shows that the generative estimator has an advantage over the fully discriminative estimator, and both are better than the pseudolikelihood estimator.

5. Discussion and Extensions

We believe our analysis captures the essence of the generative-discriminative distinction: by modeling the input, we reduce the variance of the parameter estimates. In related work, Ng and Jordan (2002) showed that Naive Bayes requires exponentially fewer examples than logistic regression to obtain the same estimation error. The key property needed in their proof was that the Naive Bayes estimator decouples into d independent closed form optimization problems, which does not seem to be the defining property of generative estimation. In particular, this property does not apply to general globally-normalized generative models, but one would still expect those models to have the advantages of being generative.

Given that the generative and discriminative estimators are complementary, one natural question is how to interpolate between the two to get the benefits of both. Our framework naturally suggests two ways to go about this. First, we could vary the coarseness of the partitioning. Generative and discriminative estimators differ only in this coarseness and there is a



(a) $\alpha^* = \beta^* = 1$ (b) $\alpha^* = \beta^* = 1$ (c) $\alpha^* = \beta^* = 1$

(d) $\alpha^* = \beta^* = 1$ (e) $\alpha^* = \beta^* = 1$ (f) $\alpha^* = \beta^* = 1$

(g) Vary misspecification (h) Vary edge potentials

Figure 3. Asymptotics of the simple four-node graphical model. In (a)–(c), $\alpha^* = \beta^* = 1$ and $\gamma^* = 0$; we plot the asymptotic variance of the estimation error, scaled by 1, $\sqrt{n}$, and n. In (d)–(f), we repeat with $\gamma^* = 0.5$. In (g), we take $n = 20000$ examples, $\alpha^* = \beta^* = 1$ and vary γ. In (h), we take $n = 20000$, $\beta^* = 1, \gamma^* = 0$ and vary α.

range of intermediate choices corresponding to conditioning on more or fewer of the input variables. Second, we could take a weighted combination of estimators (e.g., Bouchard and Triggs (2004); McCallum et al. (2006)). For one-parameter models, Lindsay (1988) derived the optimal weighting of the component likelihoods, but unfortunately these results cannot be applied directly in practice.

It would also be interesting to perform a similar asymptotic analysis on other estimators used in practice, for example marginal likelihoods with latent variables, tree-reweighted belief propagation (Wainwright et al., 2003; Wainwright, 2006), piecewise training (Sutton & McCallum, 2005), etc. Another important extension is to curved exponential families, which account for many of the popular generative models based on directed graphical models.

6. Conclusion

We have analyzed the asymptotic distributions of composite likelihood estimators in the exponential family. The idea of considering different partitionings of the outcome space allows a clean and intuitive characterization of the asymptotic variances, which enables us to compare the commonly used generative, discriminative, and pseudolikelihood estimators as special cases. Our work provides new theoretical support for existing intuitions and a basis for developing new estimators which balance the tradeoff between computational and statistical efficiency.

Acknowledgments We thank Peter Bartlett for useful discussions and Simon Lacoste-Julien for comments. We also wish to acknowledge NSF grant 0509559 and a grant from Microsoft Research.

References

Besag, J. (1975). The analysis of non-lattice data. *The Statistician*, *24*, 179–195.

Bouchard, G., & Triggs, B. (2004). The trade-off between generative and discriminative classifiers. *International Conference on Computational Statistics* (pp. 721–728).

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling data. *International Conference on Machine Learning (ICML)*.

Lasserre, J. A., Bishop, C. M., & Minka, T. P. (2006). Principled hybrids of generative and discriminative models. *Computer Vision and Pattern Recognition (CVPR)* (pp. 87–94).

Liang, P., Klein, D., & Jordan, M. I. (2008). Agreement-based learning. *Advances in Neural Information Processing Systems (NIPS)*.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*, 221–239.

McCallum, A., Pal, C., Druck, G., & Wang, X. (2006). Multi-conditional learning: Generative/discriminative training for clustering and classification. *Association for the Advancement of Artificial Intelligence (AAAI)*.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems (NIPS)*.

Sutton, C., & McCallum, A. (2005). Piecewise training of undirected models. *Uncertainty in Artificial Intelligence (UAI)*.

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

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, *92*, 1–28.

Wainwright, M. (2006). Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, *7*, 1829–1859.

Wainwright, M., Jaakkola, T., & Willsky, A. (2003). Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. *Artificial Intelligence and Statistics (AISTATS)*.

Wainwright, M., & Jordan, M. I. (2003). *Graphical models, exponential families, and variational inference* (Technical Report). Department of Statistics, University of California at Berkeley.