Asymptotic Analysis of Generative Semi-Supervised Learning

Joshua V. Dillon   JVDILLON@GATECH.EDU
Krishnakumar Balasubramanian   KRISHNAKUMAR3@GATECH.EDU
Guy Lebanon   LEBANON@CC.GATECH.EDU
College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract

Semi-supervised learning has emerged as a popular framework for improving modeling accuracy while controlling labeling cost. Based on an extension of stochastic composite likelihood we quantify the asymptotic accuracy of generative semi-supervised learning. In doing so, we complement distribution-free analysis by providing an alternative framework to measure the value associated with different labeling policies and resolve the fundamental question of how much data to label and in what manner. We demonstrate our approach with both simulation studies and real world experiments using naive Bayes for text classification and MRFs and CRFs for structured prediction in NLP.

1. Introduction

Semi-supervised learning (SSL) is a technique for estimating statistical models using both labeled and unlabeled data. It is particularly useful when the costs of obtaining labeled and unlabeled samples are different. In particular, assuming that unlabeled data is more easily available, SSL provides improved modeling accuracy by adding a large number of unlabeled samples to a relatively small labeled dataset.

The practical value of SSL has motivated several attempts to mathematically quantify its value beyond traditional supervised techniques. Of particular importance is the dependency of that improvement on the amount of unlabeled and labeled data. In the case of structured prediction the accuracy of the SSL estimator depends also on the specific manner in which sequences are labeled. Focusing on the framework of generative or likelihood-based SSL applied to classification and structured prediction we identify the following questions which we address in this paper.

Q1: Consistency (classification). What combinations of labeled and unlabeled data lead to precise models in the limit of large data.

Q2: Accuracy (classification). How can we quantitatively express the estimation accuracy for a particular generative model as a function of the amount of labeled and unlabeled data. What is the improvement in estimation accuracy resulting from replacing an unlabeled example with a labeled one.

Q3: Consistency (structured prediction). What strategies for sequence labeling lead to precise models in the limit of large data.

Q4: Accuracy (structured prediction). How can we quantitatively express the estimation quality for a particular model and structured labeling strategy. What is the improvement in estimation accuracy resulting from replacing one labeling strategy with another.

Q5: Tradeoff (classification and structured prediction). How can we quantitatively express the tradeoff between the two competing goals of improved prediction accuracy and low labeling cost. What are the possible ways to resolve that tradeoff optimally within a problem-specific context.

Q6: Practical Algorithms. How can we determine how much data to label in practical settings.

The first five questions are of fundamental importance to SSL theory. Recent related work has concentrated on large deviation bounds for discriminative SSL as a response to Q1 and Q2 above. While enjoying broad applicability, such non-parametric bounds are weakened when the model family's worst-case is atypical. By forgoing finite sample analysis, our approach complements these efforts and provides insights which apply to the specific generative models under consideration. In presenting answers to the last question, we reveal the relative merits of asymptotic analysis and how its employ, perhaps surprisingly, renders practical heuristics for controlling labeling cost.

Our asymptotic derivations are possible by extending the recently proposed stochastic composite likelihood

formalism (Dillon & Lebanon, 2009) and showing that generative SSL is a special case of that extension. The implications of this analysis are demonstrated using a simulation study as well as text classification and NLP structured prediction experiments. The developed framework, however, is general enough to apply to any generative SSL problem. As in (Liang & Jordan, 2008), the delta method transforms our results from parameter asymptotics to prediction risk asymptotics. We omit these results due to lack of space.

2. Related Work

Semi-supervised learning has received much attention in the past decade. Perhaps the first study in this area was done by Castelli & Cover (1996) who examined the convergence of the classification error rate as a labeled example is added to an unlabeled dataset drawn from a Gaussian mixture model. Nigam et al. (2000) proposed a practical SSL framework based on maximizing the likelihood of the observed data. An edited volume describing more recent developments is (Chapelle et al., 2006).

The goal of theoretically quantifying the effect of SSL has recently gained increased attention. In (Cohen & Cozman, 2006), the authors use asymptotic bias arguments to analyze generative SSL under various scenarios of labeled and unlabeled data while Zhang (2000) considers the asymptotic efficiency. Sinha & Belkin (2008) examined the effect of using unlabeled samples with imperfect models for mixture models. Balcan & Blum (2010) and Singh et al. (2008) analyze discriminative SSL using PAC theory and large deviation bounds. Additional analysis has been conducted under specific distributional assumptions such as the "cluster assumption", "smoothness assumption" and the "low density assumption."(Chapelle et al., 2006) However, many of these assumptions are criticized by Ben-David et al. (2008).

Excluding the works of Cohen & Cozman and Zhang, our work complements the above studies by focusing on generative rather than discriminative SSL. Whereas Cohen & Cozman consider the scenario in which error is dominated by bias rather than variance, we analyze and empirically motivate the reverse. Our work generalizes that of Zhang's by including SSL for structured prediction tasks. Moreover, this leads to heuristics for optimally obtaining partially labeled samples. In contrast to most other studies, we derive model specific asymptotics as opposed to non-parametric large deviation bounds. While such bounds are helpful as they apply to a broad set of cases, they also provide less information than model-based analysis due to their generality. Our analysis, on the other hand, requires

knowledge of the specific model family and an estimate of the model parameters. The resulting asymptotics, apply specifically to the case at hand without the need of potentially loose bounds.

We believe that our work is the first to consider and answer questions Q1-Q6 in the context of generative SSL. In particular, our work provides a new framework for examining the accuracy-cost SSL tradeoff in a way that is quantitative, practical, and model-specific.

3. Stochastic SSL Estimators

Generative SSL (Nigam et al., 2000; Chapelle et al., 2006) estimates a parametric model by maximizing the observed likelihood incorporating L labeled and U unlabeled examples

$$\ell(\theta) = \sum_{i=1}^{L} \log p_\theta(X^{(i)}, Y^{(i)}) + \sum_{i=L+1}^{L+U} \log p_\theta(X^{(i)}) \quad (1)$$

where $p_\theta(X^{(i)})$ above is obtained by marginalizing the latent label $\sum_y p_\theta(X^{(i)}, y)$. A classical example is the naive Bayes model in (Nigam et al., 2000) where $p_\theta(X, Y) = p_\theta(X|Y)p(Y)$, $p_\theta(X|Y = y) = \text{Mult}([\theta_y]_1, \ldots, [\theta_y]_V)$. The framework, however, is general enough to apply to any generative model $p_\theta(X, Y)$.

To analyze the asymptotic behavior of the maximizer of (1) we assume that the ratio between labeled to unlabeled examples $\lambda = L/(L + U)$ is kept constant while $n = L + U \to \infty$. More generally, we assume a stochastic version of (1) where each one of the n samples $X^{(1)}, \ldots, X^{(n)}$ is labeled with probability λ

$$\ell_n(\theta) = \sum_{i=1}^{n} Z^{(i)} \log p_\theta(X^{(i)}, Y^{(i)}) \quad (2)$$

$$+ \sum_{i=1}^{n} (1 - Z^{(i)}) \log p_\theta(X^{(i)}), \quad Z^{(i)} \sim \text{Bin}(1, \lambda).$$

The variable $Z^{(i)}$ above is an indicator taking the value 1 with probability λ and 0 otherwise. Due to the law of large numbers for large n we will have approximately $L = n\lambda$ labeled samples and $U = n(1 - \lambda)$ unlabeled samples thus achieving the asymptotic behavior of (1).

Equation (2) is sufficient to handle the case of classification. However, in the case of structured prediction we may have sequences $X^{(i)}, Y^{(i)}$ where for each i some components of the label sequence $Y^{(i)}$ are missing and some are observed. For example one label sequence may be completely observed, another may be completely unobserved, and a third may have the first half labeled and the second half not.

More formally, we assume the existence of a sequence labeling policy or strategy $\wp$ which maps label sequences $Y^{(i)} = (Y_1^{(i)}, \ldots, Y_m^{(i)})$ to a subset corresponding to the observed labels $\wp(Y^{(i)}) \subset \{Y_1^{(i)}, \ldots, Y_m^{(i)}\}$. To achieve generality we allow the labeling policy $\wp$ to be stochastic, leading to different subsets of $\{Y_1^{(i)}, \ldots, Y_m^{(i)}\}$ with different probabilities. A simple "all or nothing" labeling policy could label the entire sequence with probability λ and otherwise ignore it. Another policy may label the entire sequence, the first half, or ignore it completely with equal probabilities

$$
\wp(Y) = \begin{cases} Y_1^{(i)}, \ldots, Y_m^{(i)} & \text{with probability } 1/3 \\ \emptyset & \text{with probability } 1/3 \\ Y_1^{(i)}, \ldots, Y_{\lfloor m/2 \rfloor}^{(i)} & \text{with probability } 1/3 \end{cases} \quad (3)
$$

We thus have the following generalization of (2) for structured prediction

$$
\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(\wp(Y^{(i)}), X^{(i)}). \quad (4)
$$

Equation (4) generalizes standard SSL from all or nothing labeling to arbitrary labeling policies. The fundamental SSL question in this case is not simply what is the dependency of the estimation accuracy on n and λ. Rather we ask what is the dependency of the estimation accuracy on the labeling policy $\wp$. Of particular interest is the question what labeling policies $\wp$ achieve high estimation accuracy coupled with low labeling cost. Answering these questions leads to a generative SSL theory that quantitatively balances estimation accuracy and labeling cost.

Finally, we note that both (2) and (4) are random variables whose outcomes depend on the random variables $Z^{(1)}, \ldots, Z^{(n)}$ (for (2)) or $\wp$ (for (4)). As a consequence, the analysis of the maximizer $\hat{\theta}_n$ of (2) or (4) needs to be done in a probabilistic manner.

4. A1: Consistency (Classification)

Assuming that the data is generated from $p_{\theta_0}(X, Y)$, consistency corresponds to the convergence of

$$
\hat{\theta}_n = \arg\max_\theta \ell_n(\theta) \quad (5)
$$

to θ_0 with probability 1 as $n \to \infty$ (ℓ_n is defined in (2)). This implies that in the limit of large data our estimator would converge to the truth. Note that large data $n \to \infty$ in this case means that both labeled and unlabeled data increase to ∞ (but their relative sizes remain the constant λ).

We show in this section that the maximizer of (2) is consistent assuming that $\lambda > 0$. This is not an unexpected conclusion but for the sake of completeness we prove it rigorously. The proof technique is also used when we discuss consistency of SSL estimators for structured prediction.

Definition 1. A distribution $p_\theta(X, Y)$ is said to be identifiable if $\theta \neq \eta$ entails that $p_\theta(X, Y) - p_\eta(X, Y)$ is not identically zero.

Proposition 1. *Let $\Theta \subset \mathbb{R}^r$ be a compact set, and $p_\theta(x, y) > 0$ be identifiable and smooth in θ. Then if $\lambda > 0$ the maximizer $\hat{\theta}_n$ of (2) is consistent i.e., $\hat{\theta}_n \to \theta_0$ as $n \to \infty$ with probability 1.*

We omit the proof due to space, but the central idea is to cast the generative SSL estimation problem as an extension of stochastic composite likelihood (Dillon & Lebanon, 2009). The proof follows the consistency proof of (Dillon & Lebanon, 2009) with the exception that it does not assume independence of the indicator functions $Z^{(i)}$ and $(1 - Z^{(i)})$, as is assumed there.

The above proposition is not surprising. As $n \to \infty$ the number of labeled examples increase to ∞ and thus it remains to ensure that adding an increasing number of unlabeled examples does not hurt the estimator. More interesting is the quantitative description of the accuracy of $\hat{\theta}_n$ and its dependency on θ_0, λ, n which we turn to next.

5. A2: Accuracy (Classification)

The proposition below states that the distribution of the maximizer of (2) is asymptotically normal and provides its variance which may be used to characterize the accuracy of $\hat{\theta}_n$ as a function of n, θ_0, λ. As in Section 4 our proof proceeds by casting generative SSL as an extension of stochastic composite likelihood.

In Proposition 2 (below) and in Proposition 4 we use $\mathsf{Var}_{\theta_0}(H)$ to denote the variance matrix of a random vector H under p_{θ_0}. The notations $\xrightarrow{\text{P}}, \rightsquigarrow$ denote convergences in probability and in distribution (Ferguson, 1996) and $\nabla f(\theta), \nabla^2 f(\theta)$ are the $r \times 1$ gradient vector and $r \times r$ matrix of second order derivatives of $f(\theta)$.

Proposition 2. *Under the assumptions of Proposition 1 as well as convexity of Θ we have the following convergence in distribution of the maximizer of (2)*

$$
\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\left(0, \Sigma^{-1}\right) \quad (6)
$$

as $n \to \infty$, where

$$
\Sigma = \lambda \mathsf{Var}_{\theta_0}(V_1) + (1 - \lambda) \mathsf{Var}_{\theta_0}(V_2)
$$
$$
V_1 = \nabla_\theta \log p_{\theta_0}(X, Y), \quad V_2 = \nabla_\theta \log p_{\theta_0}(X).
$$

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

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

Since $\hat{\theta}_n$ maximizes ℓ_n we have $\nabla \ell_n(\hat{\theta}_n) = 0$ and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n} \left(\nabla^2 \ell_n(\theta')\right)^{-1} \left(\nabla \ell_n(\theta_0)\right). \quad (7)$$

By Proposition 1 we have $\hat{\theta}_n \xrightarrow{P} \theta_0$ which implies that $\theta' \xrightarrow{P} \theta_0$ as well. Furthermore, by the law of large numbers and the fact that $W_n \xrightarrow{P} W$ implies $g(W_n) \xrightarrow{P} g(W)$ for continuous g,

$$(\nabla^2 \ell_n(\theta'))^{-1} \xrightarrow{P} (\nabla^2 \ell_n(\theta_0))^{-1} \quad (8)$$
$$\xrightarrow{P} \left(\lambda \mathsf{E}_{\theta_0} \nabla^2 \log p_{\theta_0}(X, Y)\right.$$
$$\left. + (1 - \lambda)\mathsf{E}_{\theta_0} \nabla^2 \log p_{\theta_0}(X)\right)^{-1} = \Sigma^{-1}$$

where in the last equality we used a well known identity concerning the Fisher information.

For the remaining term in the rhs of (7) we have

$$-\sqrt{n}\nabla \ell_n(\theta_0) = -\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}(W^{(i)} + Q^{(i)}) \quad (9)$$

where $W^{(i)} = Z^{(i)}\nabla \log p_{\theta_0}(X^{(i)}, Y^{(i)})$, $Q^{(i)} = (1 - Z^{(i)})\nabla \log p_{\theta_0}(X^{(i)})$. Since (9) is an average of iid random vectors $W^{(i)} + Q^{(i)}$ it is asymptotically normal by the central limit theorem with mean

$$\mathsf{E}_{\theta_0}(Q + W) = \lambda \mathsf{E}_{\theta_0} \nabla \log p_{\theta_0}(X, Y)$$
$$+ (1 - \lambda)\mathsf{E} \nabla \log p_{\theta_0}(X) = \lambda 0 + (1 - \lambda)0.$$

and variance

$$\mathsf{Var}_{\theta_0}(W + Q) = \mathsf{E}_{\theta_0}W^2 + \mathsf{E}_{\theta_0}Q^2 + 2\mathsf{E}_{\theta_0}WQ$$
$$= \lambda \mathsf{Var}_{\theta_0}V_1 + (1 - \lambda)\mathsf{Var}_{\theta_0}V_2$$

where we used $\mathsf{E}(Z(1 - Z)) = \mathsf{E}Z - \mathsf{E}Z^2 = 0$.

We have thus established that

$$-\sqrt{n}\nabla \ell_n(\theta_0) \rightsquigarrow N(0, \Sigma). \quad (10)$$

We finish the proof by combining (7), (14) and (10) using Slutsky's theorem. □

Proposition 2 characterizes the asymptotic estimation accuracy using the matrix Σ. Two convenient one dimensional summaries of the accuracy are the trace and the determinant of Σ. In some simple cases (such as binary event naive Bayes) $\mathrm{tr}(\Sigma)$ can be brought to a mathematically simple form which exposes its dependency on θ_0, n, λ. In other cases the dependency may be obtained using numerical computing.

Figure 1 displays three error measures for the multinomial naive Bayes SSL classifier (Nigam et al., 2000) and the Reuters RCV1 text classification data. In all three figures the error measures are represented as functions of n (horizontal axis) and λ (vertical axis). The error measures are classification error rate (left), trace of the empirical mse (middle), and log-trace of the asymptotic variance (right). The measures were obtained over held-out sets and averaged using cross validation. Figure 3 (middle) displays the asymptotic variance as a function of n and λ for a randomly drawn θ_0.

As expected the measures decrease with n and λ in all the figures. It is interesting to note, however, that the shapes of the contour plots are very similar across the three different measures (top row). This confirms that the asymptotic variance (right) is a valid proxy for the finite sample measures of error rates and empirical mse. We thus conclude that the asymptotic variance is an attractive measure that is similar to finite sample error rate and at the same time has a convenient mathematical expression.

6. A3: Consistency (Structured)

In the case of structured prediction the log-likelihood (4) is specified using a stochastic labeling policy. In this section we consider the conditions on that policy that ensure estimation consistency, i.e., convergence of the maximizer of (4) to θ_0 as $n \to \infty$.

We assume that the labeling policy $\wp$ is a probabilistic mixture of deterministic sequence labeling functions $\chi_1, \ldots, \chi_k$. In other words, $\wp(Y)$ takes values $\chi_i(Y), i = 1, \ldots, k$ with probabilities $\lambda_1, \ldots, \lambda_k$. For example the policy (3) corresponds to $\chi_1(Y) = Y$, $\chi_2(Y) = \emptyset$, $\chi_3(Y) = \{Y_1, \ldots, Y_{\lfloor m/2 \rfloor}\}$ (where $Y = \{Y_1, \ldots, Y_m\}$) and $\lambda = (1/3, 1/3, 1/3)$.

Using the above notation we can write (4) as

$$\ell_n(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{k} Z_j^{(i)} \log p_\theta(\chi_j(Y^{(i)}), X^{(i)}) \quad (11)$$
$$Z^{(i)} \sim \mathrm{Mult}(1, (\lambda_1, \ldots, \lambda_k))$$

which exposes its similarity to the stochastic composite likelihood function in (Dillon & Lebanon, 2009). Note however that (11) is not formally a stochastic composite likelihood since $Z_j^{(i)}, j = 1, \ldots, k$ are not independent and since $\chi_j(Y)$ depends on the length of the sequence Y (see for example χ_1 and χ_3 above). We also use the notation S_j^m for the subset of labels provided by χ_j on length-m sequences

$$\chi_j(Y_1, \ldots, Y_m) = \{Y_i : i \in S_j^m\}.$$

Definition 2. A labeling policy is said to be identifiable if the following map is injective

$$\bigcup_{m:q(m)>0} \bigcup_{j=1}^{k} \{p_\theta(\{Y_r : r \in S_j^m\}, X)\} \to p_\theta(X, Y)$$

where q is the distribution of sequences lengths. In other words, there is at most one collection of probabilities corresponding to the lhs above that does not contradict the joint distribution.

The importance of Definition 2 is that it ensures the recovery of θ_0 from the sequences partially labeled using the labeling policy. For example, a labeling policy characterized by $\chi_1(Y) = Y_1$, $\lambda_1 = 1$ (always label only the first sequence element) is non-identifiable for most interesting p_θ as the first sequence component is unlikely to provide sufficient information to characterize the parameters associated with transitions $Y_t \to Y_{t+1}$.

Proposition 3. *Assuming the conditions of Proposition 1, and $\lambda_1, \ldots, \lambda_k > 0$ with identifiable $\chi_1, \ldots, \chi_k$, the maximizer of (11) is consistent i.e., $\hat{\theta}_n \to \theta_0$ as $n \to \infty$ with probability 1.*

We omit the proof due to lack of space. As with Proposition 1, the proof follows the consistency proof of (Dillon & Lebanon, 2009) with the exception that it does not assume independence of the indicator functions $Z^{(i)}$ and $(1 - Z^{(i)})$.

Ultimately, the precise conditions for consistency will depend on the parametric family p_θ under consideration. For many structured prediction models such as Markov random fields the consistency conditions are mild. Depending on the precise feature functions, consistency is generally satisfied for every policy that labels contiguous subsequences with positive probability. However, some care need be applied for models like HMMs which contain parameters associated with start and/or end labels and with models asserting higher order Markov assumptions.

7. A4: Accuracy (Structured)

We consider in this section the dependency of the estimation accuracy in structured prediction SSL (4) on n, θ_0 but perhaps most interestingly on the labeling policy $\wp$. Doing so provides insight into not only how much data to label but also in what way.

Proposition 4. *Under the assumptions of Proposition 3 as well as convexity of Θ we have the following convergence in distribution of the maximizer of (11)*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N\left(0, \Sigma^{-1}\right) \qquad (12)$$

as $n \to \infty$, where

$$\Sigma = E_{q(m)} \left\{ \sum_{j=1}^{k} \lambda_j \mathsf{Var}_{\theta_0}(\nabla V_{jm}) \right\}$$

$$V_{jm} = \log p_{\theta_0}(\{Y_i : i \in S_j^m\}, X).$$

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

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

Since $\hat{\theta}_n$ maximizes ℓ, $\nabla \ell_n(\hat{\theta}_n) = 0$ and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n}(\nabla^2 \ell_n(\theta'))^{-1} \nabla \ell_n(\theta_0). \qquad (13)$$

By Proposition 3 we have $\hat{\theta}_n \xrightarrow{\text{P}} \theta_0$ 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,

$$(\nabla^2 \ell_n(\theta'))^{-1} \xrightarrow{\text{P}} (\nabla^2 \ell_n(\theta_0))^{-1} \qquad (14)$$

$$\xrightarrow{\text{P}} \left(\sum_{m>0} q(m) \sum_{j=1}^{k} \lambda_j \mathsf{E}_{\theta_0}(\nabla^2 V_{jm}) \right)^{-1}$$

$$= -\left(\sum_{m>0} q(m) \sum_{j=1}^{k} \lambda_j \mathsf{Var}_{\theta_0}(\nabla V_{jm}) \right)^{-1}.$$

where in the last equality we used a well known identity concerning the Fisher information.

For the remaining term on the rhs of (13) we have

$$\sqrt{n} \nabla \ell_n(\theta_0) = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} W_i \qquad (15)$$

where the random vectors

$$W_i = \sum_{m>0} 1_{\{\text{length}(Y^{(i)})=m\}} \sum_{j=1}^{k} Z_j^{(i)} \nabla V_{jm}^{(i)}$$

have expectation 0 due to the fact that the expectation of the score is 0. The variance of W_i is

$$\mathsf{Var}_{\theta_0} W_i = \mathsf{E}_{\theta_0} \sum_{m>0} 1_{\{\text{length}(Y^{(i)})=m\}} \sum_{j=1}^{k} Z_j^{(i)} \nabla V_{jm}^{(i)} \nabla V_{jm}^{(i)\top}$$

$$= \sum_{m>0} q(m) \sum_{j=1}^{k} \lambda_j \mathsf{E}\left(\nabla V_{jm}^{(i)} \nabla V_{jm}^{(i)\top}\right)$$

where in the first equality we used the fact that $Y^{(i)}$ can have only one length and only one of $\chi_1, \ldots, \chi_k$

is chosen. Using the central limit theorem we thus conclude that

$$\sqrt{n}\,\nabla\ell_n(\theta_0) \rightsquigarrow N\left(0, \Sigma\right)$$

and finish the proof by combining (13), (14), and (10) using Slutsky's theorem. ☐

Figure 2 (left, middle) displays the test-set log-perplexity for the CoNLL2000 chunking task as a function of the total number of labeled tokens. We used the Boltzmann chain MRF model that is the MRF corresponding to HMM (though not identical e.g., (MacKay, 1996)). We consider labeling policies $\wp$ that label the entire sequence with probability λ and otherwise label contiguous sequences of length 5 (left) or leave the sequence fully unlabeled (middle). Lighter nodes indicate larger n and unsurprisingly show a decrease in the test-set perplexity as n is increased. Interestingly, the middle figure shows that labeling policies using a smaller amount of labels may outperform other policies. This further motivates our analysis and indicates that naive choices of $\wp$ may inflate labeling cost with negligible improvement to accuracy (cf. Sec. 8 for avoiding this pitfall).

7.1. Conditional Structured Prediction

Thus far our discussion on structured prediction has been restricted to generative models such as HMM or Boltzmann chain MRF. Similar techniques, however, can be used to analyze SSL for conditional models such as CRFs that are estimated by maximizing the conditional likelihood. The key to extending the results in this paper to CRFs is to express conditional SSL estimation in a form similar to (4)

$$\hat{\theta}_n = \arg\max \sum_{i=1}^{n} \log p_\theta(\wp(Y^{(i)})|X^{(i)})$$

and to proceed with an asymptotic analysis that extends the classical conditional MLE asymptotics. We omit further discussion due to lack of space but include some experimental results for CRFs.

Figure 3 (left) depicts a similar experiment to the one described in the previous section for conditional estimation in CRF models. The figure displays log-perplexity as a function of n (x axis) and λ_1 (y axis). We observe a trend nearly identical to that of the Boltzmann chain MRF (Figure 2, left, middle).

8. A5: Tradeoff

As the figures in the previous sections display, the estimation accuracy increases with the total number of labels. The Cramer-Rao lower bound states that the highest accuracy for unbiased estimators is obtained by the maximum likelihood operating on fully observed data. However, assuming that a certain cost is associated with labeling data SSL resolves a fundamental accuracy-cost tradeoff. A decrease in estimation accuracy is acceptable in return for decreased labeling cost.

Our ability to mathematically characterize the dependency of the estimation accuracy on the labeling cost leads to a new quantitative formulation of this tradeoff. Each labeling policy (λ, n in classification and $\wp$ in structured prediction) is associated with a particular estimation accuracy via Propositions 2 and 4 and with a particular labeling cost. The precise way to measure labeling cost depends on the situation at hand, but we assume in this paper that the labeling cost is proportional to the number of labeled samples (classification) and of labeled sequence elements (structured prediction). This assumption may be easily relaxed by using other labeling cost functions, e.g, obtaining unlabeled data may incur some cost as well.

Geometrically, each labeling policy may thus be represented in a two dimensional scatter plot where the horizontal and vertical coordinates correspond to labeling cost and estimation error respectively. Three such scatter plots appear in Figure 2 (see Section 7 for a description of the left and middle panels). The right panel corresponds to multinomial naive Bayes SSL classifier and the 20-newsgroups classification dataset. Each point in that panel corresponds to different n, λ.

The origin corresponds to the most desirable (albeit unachievable) position in the scatter plot representing zero error at no labeling cost. The cloud of points obtained by varying n, λ (classification) and $\wp$ (structured prediction) represents the achievable region of the diagram. Most attractive is the lower and left boundary of that region which represents labeling policies that dominate others in both accuracy and labeling cost. The non-achievable region is below and to the left of that boundary (see shaded region in Figure 2, right). The precise position of the optimal policy on the boundary of the achievable region depends on the relative importance of minimizing estimation error and minimizing labeling cost. A policy that is optimal in one context may not be optimal in a different context.

It is interesting to note that even in the case of naive Bayes classification (Figure 2, right) some labeling policies (corresponding to specific choices of n, λ) are suboptimal. These policies correspond to points in the interior of the achievable region. A similar conclusion holds for Boltzmann chain MRF. For example, some of the points in Figure 2 (left) denoted by the label 700 are dominated by the more lightly shaded points.
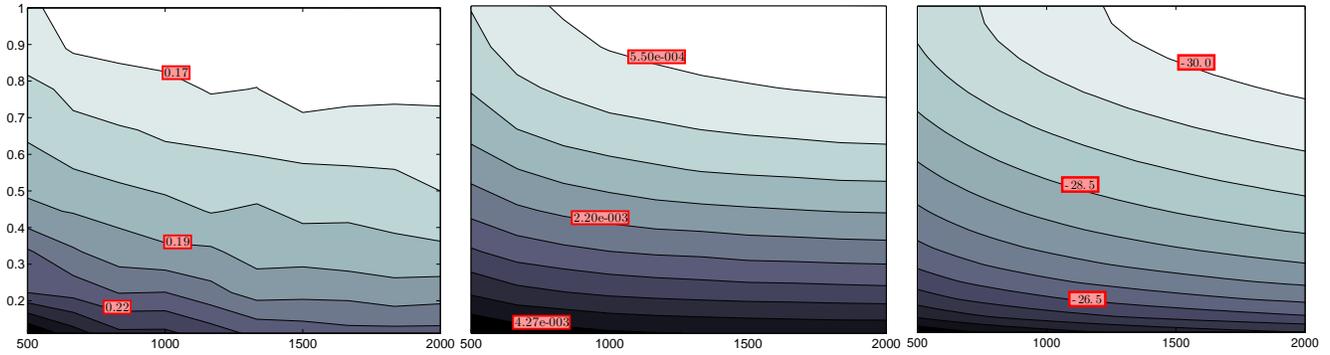
Figure 1. Three error measures for the multinomial naive Bayes SSL classifier applied to Reuters RCV1 text data. In each, error is a function of n (horizontal axis) and λ (vertical axis). The left depicts classification error rate, the middle depicts the trace of empirical mse, and right depicts the log-trace of the asymptotic variance. Results were obtained using held-out sets and averaged using cross validation. Particularly noteworthy is a striking correlation among all three figures, justifying the use of asymptotic variance as a surrogate for classification error, even for relatively small values of n.
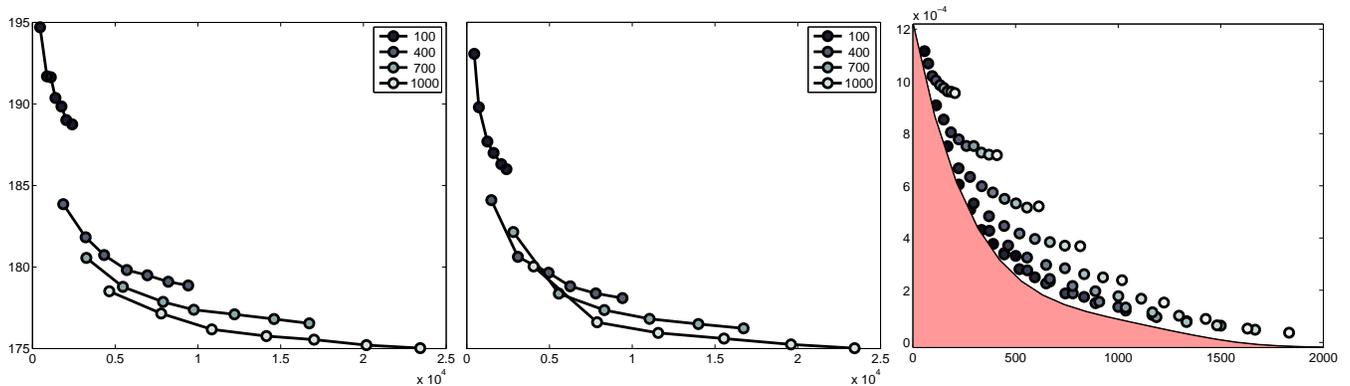

Figure 2. Test-set results for two policies of unlabeled data for Boltzmann chain MRFs applied to the CoNLL 2000 text-chunking dataset (left, middle). The shaded portion of the right panel depicts the empirically unachievable region for naive Bayes SSL classifier on the 20-newsgroups dataset. The left two share a common log-perplexity scale (vertical axis) while the vertical axis of the right panel corresponds to trace of the empirical MSE; the horizontal axis indicates labeling cost. As above, results were obtained using held-out sets and averaged using cross validation. Collectively these figures represent the application and effect of various labeling policies. The left figure depicts the consequence of partially missing samples for various n,λ while the middle and right represent SSL in the more traditional all or nothing sense: either labeled or unlabeled samples. See text for more details.
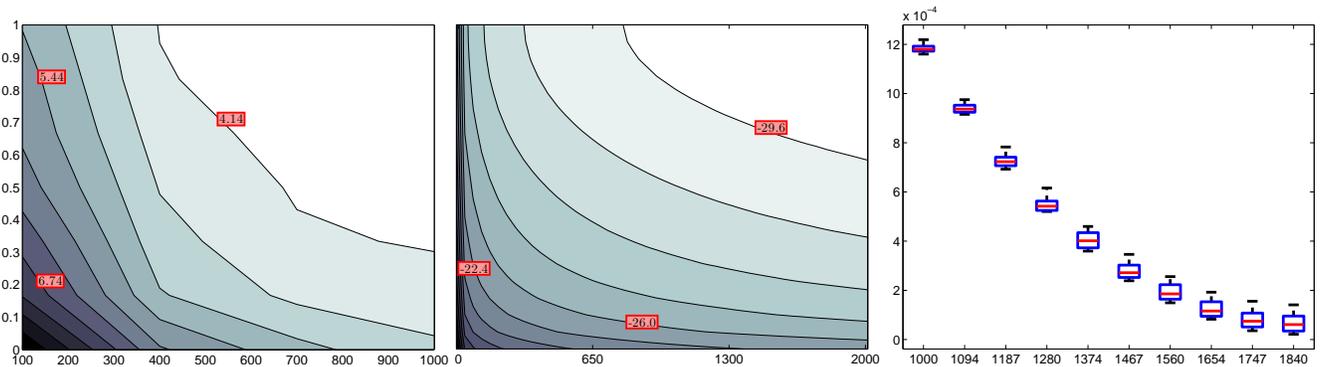

Figure 3. Left figure depicts sentence-wise log-perplexity for CRFs under the same policy and experimental design of the above Boltzmann chain. Center figure represents log-trace of the theoretical variance and demonstrates phenomena under a simplified scenario, i.e., a mixture of two 1000-dim multinomials with unbalanced prior. Rightmost figure demonstrates the practical applicability of utilizing asymptotic analysis to characterize parameter error as a function of size of training-set partition. The training-set is fixed at 2000 samples and split for training and validating. As the proportion used for training is increased, we see a decrease in error. See text for more details.

We consider three different ways to define an optimal labeling policy (i.e., determining how much data to label) on the boundary of the achievable region:

$$(\lambda^*, n^*)_1 = \underset{(\lambda,n):\lambda n \leq C}{\arg\min}\ \text{tr}(\Sigma^{-1}) \qquad (16)$$

$$(\lambda^*, n^*)_2 = \underset{(\lambda,n):\text{tr}(\Sigma^{-1})\leq C}{\arg\min}\ \lambda n \qquad (17)$$

$$(\lambda^*, n^*)_3 = \underset{(\lambda,n)}{\arg\min}\ \lambda n + \alpha\ \text{tr}(\Sigma^{-1}). \qquad (18)$$

The first applies in situations where the labeling cost is bounded by a certain available budget. The second applies when a certain estimation accuracy is acceptable and the goal is to minimize the labeling cost. The third considers a more symmetric treatment of the estimation accuracy and labeling cost.

Equations (16)-(18) may be easily generalized to arbitrary labeling costs $f(n,\lambda)$. Equations (16)-(18) may also be generalized to the case of structured prediction with $\wp$ replacing (λ, n) and cost$(\wp)$ replacing λn.

9. A6: Practical Algorithms

Choosing a policy (λ, n) or $\wp$ resolves the SSL tradeoff of accuracy vs. cost. Such a resolution is tantamount to answering the basic question of how many labels should be obtained (and in the case of structured prediction also which ones). Resolving the tradeoff via (16)-(18) or in any other way, or even simply evaluating the asymptotic accuracy $\text{tr}(\Sigma)$ requires knowledge of the model parameter θ_0 that is generally unknown in practical settings.

We propose in this section a practical two stage algorithm for computing an estimate $\hat{\theta}_n$ within a particular accuracy-cost tradeoff. Assuming we have n unlabeled examples, the algorithm begins the first stage by labeling r samples. It then estimates θ' by maximizing the likelihood over the r labeled and $n - r$ unlabeled samples. The estimate $\hat{\theta}'$ is then used to obtain a plug-in estimate for the asymptotic accuracy $\text{tr}(\Sigma)$. In the second stage the algorithm uses the estimate $\widehat{\text{tr}(\Sigma)}$ to resolve the tradeoff via (16)-(18) and determine how many more labels should be collected. Note that the labels obtained at the first stage may be used in the second stage as well with no adverse effect.

The two-stage algorithm spends some initial labeling cost in order to obtain an estimate for the quantitative tradeoff parameters. The final labeling cost, however, is determined in a principled way based on the relative importance of accuracy and labeling cost via (16)-(18). The selection of the initial number of labels r is important and should be chosen carefully. In particular it should not exceed the total desirable labeling cost.

We provide some experimental results on the performance of this algorithm in Figure 3 (right). It displays box-plots for the differences between $\text{tr}(\Sigma)$ and $\widehat{\text{tr}(\Sigma)}$ as a function of the initial labeling cost r for naive Bayes SSL classifier and 20-newsgroups data. The figure illustrates that the two stage algorithm provides a very accurate estimation of $\text{tr}(\Sigma)$ for $r \geq 1000$ which becomes almost perfect for $r \geq 1300$.

References

Balcan, M. F. and Blum, A. A discriminative model for semi-supervised learning. *Journal of the Association for Computing Machinery*, 2010.

Ben-David, S., Lu, T., and Pal, D. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *International Conference on Learning Theory*, 2008.

Castelli, V. and Cover, T. M. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, 1996.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. MIT Press, 2006.

Cohen, I. and Cozman, F. G. Risks of semi-supervised learning. In *Semi-Supervised Learning*. MIT press, 2006.

Dillon, J. and Lebanon, G. Statistical and computational tradeoffs in stochastic composite likelihood. In *Proc. of the International Conference on Artificial Intelligence and Statistics*, 2009.

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

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

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

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2):103–134, 2000.

Singh, A., Nowak, R., and Zhu, X. Unlabeled data: Now it helps, now it doesnt. In *Advances in Neural Information Processing Systems*, 2008.

Sinha, K. and Belkin, M. The value of labeled and unlabeled examples when the model is imperfect. In *Advances in Neural Information Processing Systems*, 2008.

Zhang, Tong. The value of unlabeled data for classification problems. In *Proc. of the International Conference on Machine Learning*, 2000.