

Asymptotic Analysis of Generative SSL

Joshua V. Dillon Krishna Balasubramanian Guy Lebanon

College of Computing
Georgia Institute of Technology
Atlanta, Georgia, USA

June 21-24, 2010

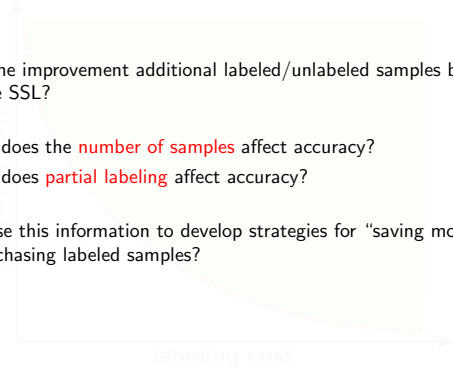


The Problem

What is the improvement additional labeled/unlabeled samples bring in generative SSL?

- How does the **number of samples** affect accuracy?
- How does **partial labeling** affect accuracy?

Can we use this information to develop strategies for “saving money” when purchasing labeled samples?



Motivating Example

- Since we're bored of reading old WSJ articles we decide to build a new “chunked phrase” dataset:

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

- Use Mechanical Turk and pay human labelers **1¢ per token**,
- To label a source of **continually available** sequences, i.e., newfeeds, blogs, tweets.

How do we strike a balance between saving **money** and obtaining enough labels to achieve a desired level of classifier **accuracy**?

The Tools

- **Finite sample analysis.**
- **Asymptotic analysis.**
- Stochastic Composite Likelihood [Dillon, Lebanon 2009]

$$sc\ell(\theta) = \sum_{i=1}^n \sum_{j=1}^k Z_j^{(i)} \beta_j \log P_{\theta}(X_{A_j}^{(i)} | X_{B_j}^{(i)}),$$
$$Z_j^{(i)} \stackrel{iid}{\sim} \text{Bern}(\lambda_j)$$

where weight β_j is used to control the estimator's efficiency.

Outline

- 1 **Consistency**
What combinations of (partially-) unlabeled and labeled samples lead to precise models?
- 2 **Accuracy**
What's the improvement of replacing (partially-) unlabeled samples with (partially-) labeled samples?
- 3 **Tradeoff**
How can we quantitatively express the cost/accuracy tradeoff?
- 4 **Practical Solutions**
How can we determine this tradeoff in realworld scenarios?

First to answer these questions in one framework.

Analytical Setup (classification)

$$\ell(\theta) = \sum_{i=1}^L \log P(X^{(i)}, Y^{(i)}) + \sum_{i=L+1}^{L+U} \log P(X^{(i)})$$
$$= \sum_{i=1}^n \left(Z^{(i)} \log P(X^{(i)}, Y^{(i)}) + (1 - Z^{(i)}) \log P(X^{(i)}) \right), \quad Z^{(i)} \stackrel{iid}{\sim} \text{Bern}(\lambda)$$

- Assume $\lambda = L/(L+U)$ is fixed while $n = L+U \rightarrow \infty$
Hence by LLN: $L \approx n\lambda$ and $U \approx n(1-\lambda)$
- $\ell_n(\theta) \stackrel{def}{=} \frac{1}{n} \ell(\theta)$

Consistency (classification)

Definition

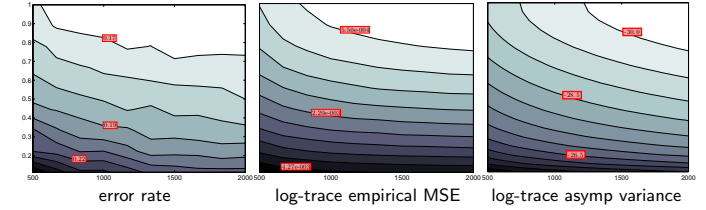
A distribution $P_\theta(X, Y)$ is said to be **identifiable** if $\theta \neq \eta$ entails that $P_\theta(X, Y) - P_\eta(X, Y)$ is not identically zero.

Proposition

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 $\ell_n(\theta)$ is **consistent** i.e., $\hat{\theta}_n \rightarrow \theta_0$ as $n \rightarrow \infty$ w.p. 1.

Accuracy (classification)

- What do we mean by accuracy? **Asymptotic Variance**.
- Is this reasonable? **Yes**; i.e., if bias decays faster.



- Multinomial naive Bayes SSL classifier applied to Reuters RCV1.
- Error is a function of n (x-axis) and λ (y-axis).

Accuracy (classification)

Proposition

Under the assumptions of Proposition (Consistency) as well as convexity of Θ we have the following convergence in distribution of the maximizer of $\ell_n(\theta)$

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

as $n \rightarrow \infty$, where

$$\Sigma = \lambda \text{Var}_{\theta_0}(V_1) + (1 - \lambda) \text{Var}_{\theta_0}(V_2)$$

$$V_1 = \nabla_\theta \log P_{\theta_0}(X, Y), \quad V_2 = \nabla_\theta \log P_{\theta_0}(X)$$

Analytical Setup (structured prediction)

Assume the existence of a **labeling policy**:
For example,

$$\varphi(Y) = \begin{cases} Y_1^{(i)}, \dots, Y_m^{(i)} & \text{w.p. } 1/3 \\ \emptyset & \text{w.p. } 1/3 \\ Y_1^{(i)}, \dots, Y_{\lfloor m/2 \rfloor}^{(i)} & \text{w.p. } 1/3 \end{cases}$$

Resulting the generalization:

$$\ell(\theta) = \sum_{i=1}^n \log P_\theta(\varphi(Y^{(i)}), X^{(i)})$$

$$= \sum_{i=1}^n \sum_{j=1}^k Z_j^{(i)} \log P_\theta(\chi_j(Y^{(i)}), X^{(i)}), \quad Z^{(i)} \stackrel{\text{i.i.d.}}{\sim} \text{Mult}(1, \vec{\lambda})$$

Consistency (structured prediction)

Definition

A labeling policy is said to be **identifiable** if the following map is injective

$$\bigcup_{m:q(m)>0} \bigcup_{j=1}^k P_\theta(\chi_j(Y; m), X) \rightarrow P_\theta(X, Y)$$

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.

Proposition

Assuming the conditions of Proposition (Consistency, classif.), and $\lambda_1, \dots, \lambda_k > 0$ with identifiable χ_1, \dots, χ_k , the maximizer of $\ell_n(\theta)$ is **consistent**, i.e., $\hat{\theta}_n \rightarrow \theta_0$ as $n \rightarrow \infty$ w.p. 1.

Accuracy (structured prediction)

Proposition

Under the consistency assumptions as well as convexity of Θ we have the following convergence in distribution of the maximizer of $\ell_n(\theta)$,

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

as $n \rightarrow \infty$, where

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

$$V_{jm} = \nabla_\theta \log P_{\theta_0}(\chi_j(Y; m), X)$$

Analytical Setup—Extensions

- Results can be extended for conditional distributions, such as CRFs.

$$\ell(\theta) = \sum_{i=1}^n \log P_{\theta}(\varphi(Y^{(i)})|X^{(i)})$$

- Risk analysis follows from the assumed existence of its derivative and application of the delta method, i.e., something along the lines of:

$$\sqrt{n}(\mathcal{R}(\hat{\theta}_n) - \mathcal{R}(\theta_0)) \rightsquigarrow N(0, \hat{\mathcal{R}}\Sigma^{-1}\hat{\mathcal{R}})$$

(Note however, our work omits this result.)

Tradeoff

- Classification: (λ, n) determines estimation accuracy.
- Structured Prediction: $\varphi(Y; m)$ determines estimation accuracy.

Tradeoff, continued

Assume cost is linearly proportional to number of labels.

- Cost is bounded by an available budget:

$$(\lambda^*, n^*) = \arg \min_{(\lambda, n): \lambda n \leq C} \{ \text{tr}(\Sigma^{-1}) \}$$

- Certain estimation accuracy is acceptable and goal to minimize cost:

$$(\lambda^*, n^*) = \arg \min_{(\lambda, n): \text{tr}(\Sigma^{-1}) \leq C} \{ \lambda n \}$$

- Strike a κ -balance between the two goals:

$$(\lambda^*, n^*) = \arg \min_{(\lambda, n)} \{ \lambda n + \kappa \text{tr}(\Sigma^{-1}) \}$$

(With structured prediction analogues.)

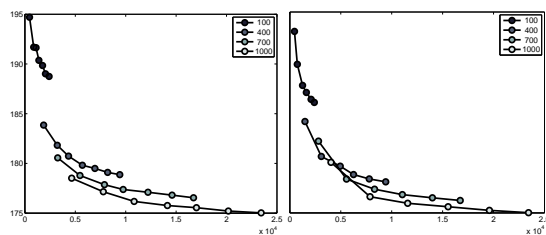
Practical Algorithms

- Two-stage approach.

- 1 Purchase some labeled samples.
- 2 Estimate θ .
- 3 Plug-in estimate for $\text{tr}(\Sigma^{-1})$.
- 4 Use plugin to resolve tradeoff (previous).
- 5 Repeat.

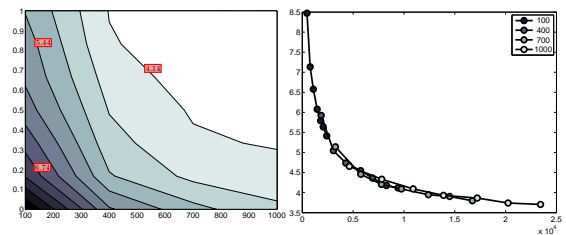
- Multi-stage approaches, etc.

Experiments: Structured Prediction (Chain MRF)



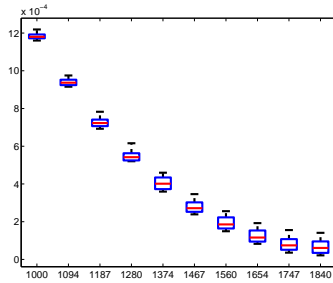
- Test-set log-perplexity (y -axis) for a given labeling cost (x -axis) of two policies on CoNLL2000 using Boltzmann chain model.
- (left) Partially missing samples for a particular φ .
- (right) SSL in the more traditional all or nothing sense: either labeled or unlabeled samples.

Experiments: Structured Prediction (Chain CRF)



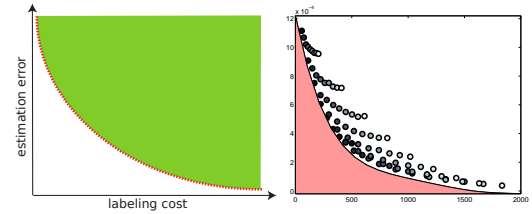
- Test-set log-perplexity for CoNLL2000 with CRF.
- (left) Traditional SSL, i.e., number of samples (n) vs. percentage labeled (λ).
- (right) Partially labeled samples as a function of cost (x -axis).
- Note that the discriminative model exhibits a much tighter tradeoff than its generative counterpart.

Two-Stage Algorithm



- Two-stage log-perplexity using 20News.
- Training-set fixed at 2000 samples and split for training and validating.
- As the proportion used for training is increased (x-axis), we see a decrease in error at the expense of an increase in the *estimate* of variance.

Conclusion



(empirical tradeoff for 20-News classification)

- **Characterize** the asymptotics of generative SSL.
- Use these results to optimize the cost/accuracy **tradeoff**.
- Develop **practical algorithms** for building datasets.

Consistency (classification)

Proof Sketch. Consider $\theta \in S = B(\theta_0, c_2) - B(\theta_0, c_1)$, $c_2 \geq c_1$.

- 1 Write likelihood as stochastic sum using $Z^{(i)} \stackrel{iid}{\sim} \text{Bern}(\lambda)$,

$$n\ell_n(\theta) = \sum_{i=1}^n \left(Z^{(i)} \log P(X^{(i)}, Y^{(i)}) + (1 - Z^{(i)}) \log P(X^{(i)}) \right)$$

- 2 Augment with a constant; denote as $\tilde{\ell}_n$.

$$-\lambda \log P_{\theta_0}(X^{(i)}, Y^{(i)}) - (1 - \lambda) \log P_{\theta_0}(X^{(i)})$$

- 3 SLLN: $\tilde{\ell}_n(\theta) \xrightarrow[n \rightarrow \infty]{} \mu(\theta)$ w.p. 1.

$$\mu(\theta) = -\lambda D_{\text{KL}}(P_{\theta_0}(X, Y) \| P_{\theta}(X, Y)) - (1 - \lambda) D_{\text{KL}}(P_{\theta_0}(X) \| P_{\theta}(X))$$

- 4 $\mu(\theta) < 0$ but $\tilde{\ell}_n$ can be made arbitrarily close to zero (i.e., $\theta = \theta_0$) so $\hat{\theta}_n \notin S$ for $n > N$.
- 5 Since c_1, c_2 were arbitrarily chosen, we have $\hat{\theta}_n \rightarrow \theta_0$.

Accuracy (classification)

Proof Sketch.

- 1 Consider first-order Taylor series of $\nabla \ell_n(\hat{\theta}_n)$.

$$\begin{aligned} \nabla \ell_n(\hat{\theta}_n) &= \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\theta')(\hat{\theta}_n - \theta_0) \\ \theta' &= \theta_0 + \eta(\hat{\theta}_n - \theta_0) \end{aligned}$$

- 2 Rewrite as:

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

- 3 $(\nabla^2 \ell(\theta))^{-1} \xrightarrow[n \rightarrow \infty]{P} \Sigma^{-1}$
- 4 $-\sqrt{n} \nabla \ell_n(\theta_0) \rightsquigarrow N(0, \Sigma)$