Homework 1
CS229T/STATS231 (Winter 2015–2016)

Please structure your writeups hierarchically: convey the overall plan before diving into details. You should justify with words why something’s true (by algebra, convexity, etc.). There’s no need to step through a long sequence of trivial algebraic operations. Be careful not to mix assumptions with things which are derived. Up to two additional points will awarded for especially well-organized and elegant solutions.

Due date: Wednesday, Jan. 20 (at the beginning of class)

1. Value of labeled data (15 points)

In many applications, labeled data is expensive and therefore limited, while unlabeled data is cheap and therefore abundant. For example, there are tons of images on the web, but getting labeled images is much harder. But what is the statistical value of having labeled data versus unlabeled data? This problem will explore this formally using asymptotics.

Specifically, suppose we have an exponential family model over a discrete latent variable $h$ and a discrete observed variable $x$:

$$p_{\theta}(h, x) = \exp\{\theta \cdot \phi(h, x) - A(\theta)\},$$

where $A(\theta) = \sum_{h, x} \exp\{\theta \cdot \phi(h, x)\}$ is the usual log-partition function.

Suppose that $n$ examples $(h^{(1)}, x^{(1)}), \ldots, (h^{(n)}, x^{(n)})$ are drawn i.i.d. from some true distribution $p_{\theta^*}$. Define the following two estimators:

$$\hat{\theta}_{\text{sup}} = \arg \max_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log p_{\theta}(h^{(i)}, x^{(i)})$$

(1)

$$\hat{\theta}_{\text{unsup}} = \arg \max_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{h} p_{\theta}(h, x^{(i)}).$$

(2)

The supervised estimator $\hat{\theta}_{\text{sup}}$ uses the variable $h^{(i)}$ and maximizes the joint likelihood, while the unsupervised estimator $\hat{\theta}_{\text{unsup}}$ marginalizes out the latent variable $h$.

One important caveat: our results will hold when we assume that data is actually generated from our model family and that unsupervised learning is possible. Otherwise, labeled data is worth a lot more.

a. (2 points) (supervised asymptotic variance) Compute the asymptotic variance of $\hat{\theta}_{\text{sup}}$: that is, given that $\sqrt{n}(\hat{\theta}_{\text{sup}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, V_{\text{sup}})$, write an expression for $V_{\text{sup}}$ that depends on expectations/variances involving $\phi$.

b. (3 points) (unsupervised asymptotic variance) Compute the asymptotic variance of $\hat{\theta}_{\text{unsup}}$: that is, given that $\sqrt{n}(\hat{\theta}_{\text{unsup}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, V_{\text{unsup}})$, write an expression for $V_{\text{unsup}}$ that depends on expectations/variances involving $\phi$.

c. (5 points) (comparing estimators) First, prove that $\hat{\theta}_{\text{sup}}$ has lower (or equal) asymptotic variance compared to $\hat{\theta}_{\text{unsup}}$.

Second, suppose that you were offered the choice of having $n$ supervised examples or $\alpha n$ unsupervised examples—i.e.,

$$\hat{\theta}_{\text{unsup}} = \arg \max_{\theta \in \mathbb{R}^d} \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} \log \sum_{h} p_{\theta}(h, x^{(i)}).$$
Give conditions on $\alpha$ such that (i) $\hat{\theta}_{\text{sup}}$ has lower asymptotic variance, and (ii) $\hat{\theta}_{\text{unsup}}$ has lower asymptotic variance. Note: the asymptotic variance is always scaled up by $\sqrt{n}$.

d. (5 points)
Consider the exponential family

$$p_\theta(h, x) \propto \exp(\theta hx),$$

where $h, x \in \{0, 1\}$. Essentially, $(h, x)$ is a pair of correlated biased coin flips, where $p_\theta(1, 1) \propto \exp(\theta^*)$ and $p_\theta(0, 0) = p_\theta(0, 1) = p_\theta(1, 0) \propto 1$. Compute $V_{\text{sup}}$ and $V_{\text{unsup}}$. Assuming a choice between $n$ supervised examples and $\alpha n$ unsupervised examples, find $\Delta$ such that $\hat{\theta}_{\text{unsup}}$ has lower asymptotic variance than $\hat{\theta}_{\text{sup}}$ iff $\alpha \geq \Delta$. 

2. Learning with privacy (15 points)

Suppose we are conducting a survey across \( n \) users. For each \( i = 1, \ldots, n \), user \( i \) has a true response \( x^{(i)} \sim p_{\theta^*} \), where \( p_{\theta^*}(x) = \exp\{\theta \cdot \phi(x) - A(\theta)\} \) is a standard exponential family. Assume all users are independent.

However, for privacy reasons, we do not want to collect the actual responses from the user. Instead, each user is going to return \( y^{(i)} \) which is equal to \( x^{(i)} \) with probability \( 1 - \epsilon \) and equal to a draw from some known noise distribution \( p_{\theta_0}^* \) with probability \( \epsilon \). Again, all these choices are made independently.

Our goal is to estimate \( \theta^* \) given \( y^{(1)}, \ldots, y^{(n)} \).

a. (5 points) (constructing an estimator)

Construct a consistent estimator \( \hat{\theta} \), one that converges in probability to \( \theta^* \). Your estimator must be computationally efficient, only requiring basic vector arithmetic and convex optimization. In particular, your estimator cannot be maximum marginal likelihood, which would involve a non-convex optimization problem.

b. (5 points) (asymptotic variance)

Compute the asymptotic variance of \( \hat{\theta} \). Write your answer in terms of expectations and/or covariances of \( \phi(x) \).

c. (5 points) (concrete example)

Suppose that \( x \in \{0, 1\} \), and \( p_{\theta^*}(x) \propto \exp(\theta^* x) \). Equivalently, \( p_{\theta^*} = \text{Bernoulli}(p) \), where

\[
p = \frac{1}{1 + \exp(-\theta^*)}.
\]

Also, suppose that \( p_{\theta_0}^*(x) = \text{Bernoulli}\left(\frac{1}{2}\right) \). Compute the following two quantities:

1. \( V \), the asymptotic variance of \( \hat{\theta} \).

2. \( D(p) = \frac{dv}{d\epsilon}\big|_{\epsilon=0} \), which measures how adding a small amount of noise (i.e., \( \epsilon \approx 0 \)) affects the asymptotic variance. Comment on the dependence of \( D(p) \) on \( p \).

(Hint: it may be easier to write your expressions in terms of \( p \) and \( q = 1 - p \).)
3. **Importance sampling (10 points)**

Estimating the expectation of a function with respect to an intractable distribution is a fundamental problem in statistics and machine learning, especially when doing Bayesian inference and graphical model inference. Importance sampling is a classic technique for approximating expectations. Let’s analyze how well it does asymptotically.

Let $p$ be a distribution over $\mathcal{X}$ (assumed to be discrete), and let $f : \mathcal{X} \rightarrow \mathbb{R}$ be some function of interest. Furthermore, suppose we have:

$$p(x) = \frac{\tilde{p}(x)}{Z},$$

where we assume we can evaluate $\tilde{p}(x)$ at any $x$, but we do not know the normalization constant $Z$ (because it is intractable to compute). Our goal is to estimate the expectation:

$$\mathbb{E}_p[f(x)].$$

A standard approach to this problem is to use importance sampling. In importance sampling, we draw $n$ i.i.d. samples from an tractable proposal distribution $q$ over $\mathcal{X}$:

$$x_1, \ldots, x_n \sim q. \tag{3}$$

Define the importance estimator to be:

$$\hat{f} = \frac{\sum_{i=1}^{n} w(x_i) f(x_i)}{\sum_{i=1}^{n} w(x_i)}, \tag{4}$$

where $w(x) = \frac{\tilde{p}(x)}{q(x)}$ is the importance weighting function.

One can check that $\hat{f} \xrightarrow{P} f^*$, where $f^* = \mathbb{E}_p[f(x)]$.

**a. (5 points) (asymptotic variance)**

Compute the asymptotic variance of $\hat{f}$.

**b. (5 points) (optimal proposal)**

Compute the optimal proposal distribution $q$, the one that minimizes the asymptotic variance $V$. Of course, $q$ will not be efficiently computable; this question is just to get a sense of the best thing we could hope for.

4. **Feedback (0 points)**

These questions are only to help us calibrate and improve the assignments, and the responses will not impact your grade.

(a) How many hours did it take you to do this assignment?

(b) On a scale of 1 to 10, how useful was this assignment (1 was not useful, 10 was very useful)?