# Homework 2

Due: Friday, October 28, 2016

**Notes:** For positive integers $k$, $[k] := \{1, \ldots, k\}$ denotes the set of the first $k$ positive integers. When $X \sim p$ and $Y \sim q$ are random variables over the same sample space, $D(X||Y)$, $D(X||q)$, and $D(p||Y)$ should all be read as $D(p||q)$. The homework is out of 75 points – 5 points per part.

1. **Maximum Entropy of Independent Bernoulli Sums**

   In this problem, we will show that the binomial and (optionally) Poisson distributions are maximum entropy (MaxEnt) distributions over an appropriate class $\mathcal{P}$ of distributions, and derive several useful properties of KL divergence along the way.

   For any positive integer $n$ and $p \in [0, 1]$, let Binomial$(n, p)$ denote the binomial distribution (the sum of $n$ IID Bernoulli events of probability $p$), which has density function

   $$\text{Binomial}_{n,p}(k) = \binom{n}{k} p^k (1-p)^{1-k}.$$

   For $\lambda \geq 0$, let $\Pi(\lambda)$ denote the mean-$\lambda$ Poisson distribution, which has density function

   $$\text{Poisson}_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \forall k \in \mathbb{N} \cup \{0\}.$$

   The class $\mathcal{P}_\lambda$ of distributions is that of sums $S_n := \sum_{i=1}^n X_i$ of $n$ independent (but not necessarily identically distributed) binary variables $\{X_i\}_{i=1}^n$ constrained such that $\mathbb{E}[S_n] = \lambda$, for some $\lambda \in [0, n]$. Note that any $p \in \mathcal{P}_\lambda$ can be parametrized by $(p_1, \ldots, p_n) \in [0, 1]^n$, with $\sum_{i=1}^n p_i = \lambda$. We will show that the Binomial case $p_1 = \cdots = p_n = \frac{\lambda}{n}$ is the MaxEnt distribution over $\mathcal{P}_\lambda$, and that the Poisson distribution is the limit as $n \to \infty$.

   (a) Derive the maximum likelihood estimate of $\lambda$ under the assumption that you observe $n$ IID samples $X_1, \ldots, X_n$ from a Poisson distribution.

   (b) Define $D(X) := \min_{\lambda \geq 0} D(X||\Pi(\lambda))$. Derive a closed form for $D(X)$ in terms of $X$. [1]

   (c) Show that the KL divergence $D(p||q)$ is convex in $p$.

   (d) Let

   $$\begin{aligned} \mathcal{P}_\lambda(p_3, \ldots, p_n) &= \{q \in \mathcal{P}_\lambda : q_3 = p_3, \ldots, q_n = p_n, \} \\ &= \left\{ (x_1, x_2, p_3, \ldots, p_n) : x_1 + x_2 = \lambda - \sum_{i=3}^n p_i \right\} \end{aligned}$$

   denote the subspace of $\mathcal{P}_\lambda$ with all but two coordinates fixed. Show that $H(S_n)$ is strictly concave on $\mathcal{P}_\lambda(p_3, \ldots, p_n)$. *(Hint: Use parts (b) and (c) to reduce this to showing* $\mathbb{E}[\log(S_n!)]$ *is strictly concave on* $\mathcal{P}_\lambda(p_3, \ldots, p_n)$*. Then, since*

   $$\mathbb{E}[\log(S_n!)] = \mathbb{E}[\mathbb{E}[\log(S_n!)|X_3, \ldots, X_n]],$$

[1] $X$ may have any distribution over $\{0, 1, 2 \ldots\}$, but you may assume any necessary functionals of $X$ are finite.

*which is a linear functional of* $\mathbb{E}[\log(S_n!)|X_3,\ldots,X_n]$*, show that* $\mathbb{E}[\log(S_n!)|X_3,\ldots,X_n]$ *is strictly concave on* $\mathcal{P}_\lambda(p_3,\ldots,p_n)$*, for any values of* $X_3,\ldots,X_n$*.)*

(e) Use part (d) to show that Binomial$(n,\lambda/n)$ is the unique MaxEnt distribution over $\mathcal{P}$.

(f) Given independent random variables $X$ and $Y$ taking values on $\mathbb{N}$, show that

$$D(X+Y) \leq D(X) + D(Y). \tag{1}$$

*(Hint: Use the General Data Processing Inequality from Homework 1 and the fact that the sum of two Poisson-distributed variables with means* $\lambda_1$ *and* $\lambda_2$ *is itself Poisson-distributed with mean* $\lambda_1+\lambda_2$*.)*

(g) Show that $D\left(\text{Binomial}\left(n,\frac{\lambda}{n}\right)\right) \to 0$ as $n\to\infty$. This is (a fairly strong form of) the "Law of Rare Events" (a.k.a. the "Poisson Limit Theorem"), which states that the frequency of a large number of unlikely events is approximately Poisson-distributed and justifies many applications of the Poisson distribution. *(Hint: Show* $D(X_i) \leq p_i^2$ *and apply* (1)*.)*

(h) **(This part is optional.)** Show that $H(\Pi(\lambda)) = \lim_{n\to\infty} H(B(n,\lambda/n))$. *(Hint: Use the equivalence*

$$H(p) + D(p||q) = \mathop{\mathbb{E}}_{X\sim p}[\log q(x)],$$

*discussed in Lecture 1. Note that one step of this proof requires switching a limit and an infinite summation. If you are not familiar with the dominated convergence theorem, you may wish to take this step for granted.)*

2. **Wavelet Denoising with CRM**

In this problem, we will analyze the convergence rate of a wavelet-based denoising estimator.

*Haar wavelets and quantization:* Recall that Haar wavelets over $\mathcal{X} := [0,1)$ are piecewise constant functions $\psi_{j,k} : \mathcal{X} \to \{-2^{j/2}, 0, 2^{j/2}\}$ such that

$$\psi_{j,k}(x) = 2^{j/2}\left(1_{[k2^{-j},(k+1/2)2^{-j})} - 1_{[(k+1/2)2^{-j},(k+1)2^{-j})}\right),$$

for all $j \in \mathbb{N}\cup\{0\}, k\in\{0,\ldots,2^j-1\}, x\in\mathcal{X}$. Since Haar wavelets for a basis for $L^2(\mathcal{X})$, for any $\ell\in\mathbb{N}\cup\{0\}$, if we define the projection

$$f_\ell := \sum_{j=0}^{\ell}\sum_{k=0}^{2^j-1}\langle\psi_{j,k}, f\rangle,$$

of $f$ onto the first $\ell+1$ scales of the Haar basis, then $f_\ell \to f$ as $\ell\to\infty$. To encode the projection $f_\ell$, we also need to quantize the coefficients. Quantized projections lie in the set

$$Q_{\ell,\varepsilon} := \left\{\sum_{j=0}^{\ell}\sum_{k=0}^{2^j-1} a_{j,k}\psi_{j,k} \in L^2(\mathcal{X}) : a_{j,k} = 2b_{j,k}\varepsilon, \text{ for some integer } b_{j,k}\right\},$$

so that their wavelet coefficients are multiples of $\varepsilon$. Our quantized projection of $f$ is then

$$f_{\ell,\varepsilon} := \mathop{\text{argmin}}_{g\in Q_{\ell,\varepsilon}} \|f-g\|_2.$$

Thus, $f_{\ell,\varepsilon}$ is the best (in $L^2$ distance) representation of $f$ in terms of Haar wavelets of scale at most $\ell$ and coefficient precision $\varepsilon$.

*CRM Denoising:* We will assume the true function $f$ lies in the class $\mathcal{F}_{s,M} \subseteq L^2(\mathcal{X})$ of piecewise constant functions with at most $s$ discontinuities and bounded $L^\infty$ norm $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)| \leq M$. We observe $n$ noisy IID pairs $\{(X_i, Y_i)\}_{i=1}^n$, where each $X_1, \ldots, X_n \sim U(\mathcal{X})$ is uniformly distributed and, for $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$, $Y_i = f(X_i) + \varepsilon_i$.

For $\delta \in (0,1)$, the complexity-penalized empirical risk minimizing (CRM) estimator [2] is

$$\widehat{f}_{\ell,\varepsilon,\delta} := \operatorname*{argmin}_{g_{\ell,\varepsilon} \in Q_{\ell,\varepsilon}} \left[ \|g_{\ell,\varepsilon} - f\|_2^2 + \frac{c(g_{\ell,\varepsilon}) - \ln \delta}{n} \right],$$

where $c(g_{\ell,\varepsilon})$ denotes the number of bits required to encode $g_{\ell,\varepsilon}$. In class, we derived the following excess risk bound for CRM estimators:

$$R\left(\widehat{f}_{\ell,\varepsilon,\delta}\right) - R^* = \|\widehat{f}_{\ell,\varepsilon,\delta} - f\|_2^2 \leq \inf_g \left[ \|g_{\ell,\varepsilon} - f\|_2^2 + \frac{c\left(g_{\ell,\varepsilon}\right) - \ln \delta}{n} \right] + \delta. \tag{2}$$

In this problem, we will analyze the terms of (2) to derive a convergence rate bound in terms of the complexity $s$ of $f$ and the sample size $n$.

(a) Show that the projections $f_\ell$ and $f_{\ell,\varepsilon}$ can each have at most $C_0 s\ell + 1$ nonzero coefficients, for some constant $C_0$.

(b) Bound the approximation errors $\|f - f_\ell\|_2^2$ and $\|f - f_{\ell,\varepsilon}\|_2^2$.

(c) How many bits $c(f)$ are required to encode $f_{\ell,\varepsilon}$ (for known $s$, $M$, $\ell$, and $\varepsilon$)?

(d) By choosing $\varepsilon > 0$, $\ell \in \mathbb{N}$, and $\delta > 0$ appropriately, use parts (b) and (c) with the bound(2) show [3]

$$\|\widehat{f} - f\|_2^2 \in O\left(\frac{s \log^2 n}{n}\right).$$

Note that, up to log factors, this is a parametric rate with $s$ parameters.

[2] Recall that $\widehat{f}_{\ell,\varepsilon,\delta}$ can be easily computed by hard-thresholding.

[3] Here, treat $M$ as a constant.

3. **Universal Prediction with Exponential Weights**

Fix a (potentially infinite) countable class of predictors $\mathcal{F}$. Recall that, in the universal prediction setting, at each time point $t \in \{1, \dots, T\}$ up to a predetermined time horizon $T$, we see some data $x_t$ and choose a predictor $\widehat{f}_t \in \mathcal{F}$, before then seeing a true label $y_t$ and suffering loss $\ell\left(\widehat{f}_t(x_t), y_t\right) \in [0, 1]$. Since we are allowing, for example, adversarial sequences $\{(x_t, y_t)\}_{t=1}^T$, a randomized algorithm is needed to provide any guarantees. Given a learning rate $\eta > 0$ and prior $\pi$ over $\mathcal{F}$, the exponential weights algorithm proposes to draw $\widehat{f}_t$ according to a distribution $q_t$ defined such that $q_1 = \pi$ and each

$$q_{t+1}(f) \propto q_t(f) \exp\left(-\eta \ell\left(f(x_t), y_t\right)\right).$$

For each $f \in \mathcal{F}$ and $t \in [T]$, let

$$L_t(f) := \sum_{\tau=1}^{t} \ell\left(f(x_\tau), y_\tau\right) \quad \text{and} \quad L_t(\widehat{f}) := \sum_{\tau=1}^{t} \ell\left(\widehat{f}_\tau(x_\tau), y_\tau\right)$$

denote the cumulative losses of $f$ and our predictions, respectively, at time $t$. Define

$$W_t = \mathop{\mathbb{E}}_{f \sim \pi}\left[\exp\left(-\eta L_t(f)\right)\right], \quad \forall t \in \{1, \dots, T\}.$$

(a) Show that $\ln W_T \geq -\inf_{f \in \mathcal{F}}\left[\eta L_T(f) - \log \pi(f)\right]$.

(b) Show that

$$\frac{W_{t+1}}{W_t} = \mathop{\mathbb{E}}_{f \sim q_{t+1}}\left[\exp\left(-\eta \ell\left(f(x_{t+1}), y_{t+1}\right)\right)\right].$$

(c) Use part (b) to show that

$$\ln W_T \leq -\eta \sum_{t=1}^{T} \mathop{\mathbb{E}}_{f \sim q_t}\left[\ell\left(f_t(x_t), y_t\right)\right] + \frac{\eta^2 T}{8}.$$

*Hint: Recall Hoeffding's Lemma: for a random variable $X$ with $X \in [a, b]$ a.s.,*

$$\ln \mathbb{E}\left[e^{sX}\right] \leq s\,\mathbb{E}\left[X\right] + \frac{s^2(b-a)^2}{8}.$$

(d) Use parts (a) and (c) and a convenient choice of $\eta$ to bound the expected loss of the exponential weights algorithm by

$$\mathbb{E}\left[L_T(\widehat{f})\right] \leq \inf_{f \in \mathcal{F}}\left[L_T(f) + (1 - \log \pi(f))\sqrt{\frac{T}{8}}\right].$$

If $\mathcal{F}$ is finite, give a simple sufficient condition on the prior $\pi$ such that the regret

$$\mathbb{E}\left[L_T(\widehat{f})\right] - \inf_{f \in \mathcal{F}} L_T(f) \in O\left(T^{1/2}\right).$$