Homework 2

Due: Friday, October 28, 2016

Notes: For positive integers k, $[k] := \{1, \dots, k\}$ 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-λ 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 $(p_1, \dots, p_n) \in [0, 1]^n$, 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 λ under the assumption that you observe n IID samples $X_1, \dots, X_n$ 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

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

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

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

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

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

(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 λ_1 *and* λ_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*

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

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.)

Solution: (Based on Harremoës [2001], though part (d) is simplified.)

(a) For any $\lambda \in [0,\infty)$, the log-likelihood function is

$$\begin{aligned}\ell(X_1^n|\lambda) &= \sum_{i=1}^n \log\left(\frac{\lambda^{X_i}}{X_i!}e^{-\lambda}\right)\\ &= \sum_{i=1}^n X_i \log\lambda - \log(X_i!) - \lambda = n\bar{X}\log\lambda - n\lambda + \sum_{i=1}^n \log(X_i!)\end{aligned}$$

for $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_1$. This is clearly concave and smooth in λ, so that, at the MLE $\widehat{\lambda}_{MLE} = \operatorname{argmax}_{\lambda\in[0,\infty)} \ell(X_1^n|\lambda)$, we have,

$$0 = \frac{d}{d\lambda}\ell(X_1^n|\lambda)\big|_{\lambda=\widehat{\lambda}_{MLE}} = \frac{n\bar{X}}{\widehat{\lambda}_{MLE}} - n.$$

Hence $\widehat{\lambda}_{MLE} = \bar{X}$ is the empirical mean of X.

(b) Hence, we can plug part (a) into the definition of D, giving First calculate

$$\begin{aligned}D(X||\Pi(\lambda)) &= \sum_{j=0}^\infty p_j \log\left(\frac{p_j}{\frac{\lambda^j}{j!}e^{-\lambda}}\right)\\ &= \lambda + \sum_{j=0}^\infty p_j \log\left(\frac{j!}{\lambda^j}\right) - H(X)\\ &= \lambda - \mathbb{E}[X]\log\lambda + \mathbb{E}\left[\log X!\right] - H(X)\end{aligned}$$

Recalling that the MLE minimizes KL divergence gives, by part (a), that

$$D(X) = D(X||\Pi(\mathbb{E}[X])) = \mathbb{E}[X] - \mathbb{E}[X]\log\mathbb{E}[X] + \mathbb{E}\left[\log X!\right] - H(X).$$

(c) Note that, for any $c > 0$, the function $f_c : [0,\infty) \to \mathbb{R}$ defined by $f_c(x) = x\log\left(\frac{x}{c}\right)$ is convex (since $f_c''(x) = \frac{1}{x} \geq 0$). Thus, $\forall\alpha \in [0,1]$ and probability densities p_1, p_2, q on $\mathcal{X}$,

$$\begin{aligned} D(\alpha p_1 + (1-\alpha)p_2||q) &= \int_{\mathcal{X}} f_{q(x)}(\alpha p_1(x) + (1-\alpha)p_2(x))\,dx \\ &\leq \int_{\mathcal{X}} \alpha f_{q(x)}(p_1(x)) + (1-\alpha)f_{q(x)}(p_2(x))\,dx \\ &= \alpha D(p_1||q) + (1-\alpha)D(p_2||q) \end{aligned}$$

(noting $\alpha p_1(x) + (1-\alpha)p_2(x) = 0$ implies $p_1(x) = 0$ or $p_2(x) = 0$, so that the inequality applied to $f_{q(x)}$ holds trivially when $q(x) = 0$).

(d) By part (b) and the fact that $\mathbb{E}[S_n] = \lambda$,

$$H(S_n) = \lambda - \lambda\log\lambda + \mathbb{E}\left[\log(S_n!)\right] - D(S_n||\Pi(\lambda)).$$

Since λ is fixed and D is convex in its first argument, it remains only to show that $\mathbb{E}\left[\log(S_n!)\right]$ is concave on $\mathcal{P}_\lambda(p_3,\ldots,p_n)$.

$$\mathbb{E}\left[\log S_n!\right] = \mathbb{E}\left[\mathbb{E}\left[\log\left(X_1 + X_2 + \sum_{i=3}^n X_i\right)!|X_3,\ldots,X_n\right]\right]$$

Since this is linear in

$$\mathbb{E}\left[\log\left(X_1 + X_2 + \sum_{i=3}^n X_i\right)!\middle|X_3,\ldots,X_n\right], \tag{2}$$

it suffices to show that, for any fixed values of $X_3,\ldots,X_n$, (2) is concave in p_1 and p_2. Let $T := \sum_{i=3}^n X_i$, and let $r := \lambda - \sum_{i=3}^n p_i$. Using the facts that X_1, X_2, and T are all independent, $X_1 + X_2 \in \{0,1,2\}$, and $p_1 = r - p_2$,

$$\begin{aligned} &\mathbb{E}\left[\log\left(X_1 + X_2 + \sum_{i=3}^n X_i\right)!|X_3,\ldots,X_n\right] \\ &= \mathbb{E}\left[\log\left(X_1 + X_2 + T\right)!|T\right] \\ &= (1-p_1)(1-p_2)\log(T!) + (p_1(1-p_2) + (1-p_1)p_2)\log((T+1)!) + p_1p_2\log((T+2)!) \\ &= \log T! + (p_1 + p_2 - p_1p_2)\log(T+1) + p_1p_2\left(\log(T+2)\right) \\ &= \log T! + (\lambda - r)\log(T+1) + p_1(r-p_1)\left(\log(T+2) - \log(T+1)\right) \end{aligned}$$

The above expression is a quadratic polynomial in p_1, with a negative quadratic coefficient, and is therefore concave. Since, clearly, $p_1(r-p_1)$ is concave and $\log(T+2) > \log(T+1)$, the above expression is convex in p_1 and p_2, along the line $p_1 + p_2 = r$.

(e) Since $\mathcal{P}_\lambda \subseteq \mathbb{R}^n$ is compact and H is continuous, $p^* \in \operatorname{argmax}_{p \in \mathcal{P}_\lambda} H(p)$ exists. Suppose, for sake of contradiction, that, for some $i, j \in [n]$, $p_i^* \neq p_j^*$. Define $q^* \in \mathcal{P}_\lambda$ by

$$q_\ell^* = \begin{cases} \frac{p_i^* + p_j^*}{2} & : \ell \in \{i, j\} \\ p_\ell^* & : \text{ otherwise} \end{cases}.$$

By part (e) and symmetry, $H(q^*) > H(p^*)$, which is a contradiction.

(f) Let $\lambda_X := \operatorname{argmin}_{\lambda > 0} D(X||\Pi(\lambda))$ and $\lambda_Y := \operatorname{argmin}_{\lambda > 0} D(Y||\Pi(\lambda))$. Using the fact that X and Y are independent followed by the General Data Processing Inequality applied the function $(x, y) \mapsto x + y$ and the fact that $\Pi(\lambda_X) + \Pi(\lambda_Y) = \Pi(\lambda_X + \lambda_Y)$,

$$\begin{aligned} D(X) + D(Y) &= D((X, Y)||(\Pi(\lambda_X), \Pi(\lambda_Y))) \\ &\leq D(X + Y||\Pi(\lambda_X + \lambda_Y)) \leq D(X + Y). \end{aligned}$$

(g) Using the inequality $\log(1 - x) \leq -x$,

$$D(X_i) = (1 - p_i) \ln(1 - p_i) + p_i \leq (p_i - 1)p_i + p_i = p_i^2.$$

Thus, for the binomial case $p_1 = \cdots = p_n = \lambda/n$, by (1),

$$D(S_n) \leq \sum_{i=1}^n p_i^2 = \frac{\lambda^2}{n} \to 0 \quad \text{ as } n \to \infty.$$

(h) For convenience, let $p_n = \text{Binomial}(n, \lambda/n)$ and $q = \Pi(\lambda)$. By parts (b) and (g), as $n \to \infty$, $D(p_n) = D(p_n, q) \to 0$. Thus,

$$\begin{aligned} \lim_{n \to \infty} H(p_n) &= \lim_{n \to \infty} \mathop{\mathbb{E}}_{X \sim p_n} [\log q(x)] - D(p_n||q) \\ &= \lim_{n \to \infty} \mathop{\mathbb{E}}_{X \sim p_n} [\log q(x)] \\ &= \lim_{n \to \infty} \sum_{i=0}^\infty p_n(i) \log q(i). \end{aligned} \tag{3}$$

Note that

$$\begin{aligned} p_n(i) &= \binom{n}{i} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} \\ &\leq \binom{n}{i} \left(\frac{\lambda}{n}\right)^i = \frac{n!}{i!(n-i)!} \left(\frac{\lambda}{n}\right)^i \leq \frac{\lambda^i}{i!} = q(i)e^\lambda \end{aligned}$$

and one can easily calculate

$$\sum_{i=0}^\infty q(i)e^\lambda \log q(i) = e^\lambda H(q) < \infty.$$

Hence, by the dominated convergence theorem, the limit and infinite series in (3) commute. Since, as a particular consequence of part (g), $p_n \to q$ pointwise, this gives

$$\lim_{n \to \infty} H(p_n) = \sum_{i=0}^\infty \lim_{n \to \infty} p_n(i) \log q(i) = \sum_{i=0}^\infty q(i) \log q(i) = H(q).$$

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 $j \in \mathbb{N} \cup \{0\}, k \in \{0, \dots, 2^j - 1\}, x \in \mathcal{X}$. 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_ℓ, 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 ε. Our quantized projection of f is then

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

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

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^∞ 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 $X_1, \dots, X_n \sim U(\mathcal{X})$ is uniformly distributed and, for $\varepsilon_1, \dots, \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$, $Y_i = f(X_i) + \varepsilon_i$.

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:

$$R\left(\widehat{f}_{\ell,\varepsilon,\delta}\right) - R^* = \|\widehat{f}_{\ell,\varepsilon,\delta} - f\|_2^2 \leq \inf_g \left[\|g_{\ell,\varepsilon} - f\|_2^2 + \frac{c\left(g_{\ell,\varepsilon}\right) - \ln \delta}{n} \right] + \delta. \tag{4}$$

In this problem, we will analyze the terms of (4) to derive a convergence rate bound in terms of the complexity s of f and the sample size n.

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

(a) Show that the projections f_ℓ 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, ℓ, and ε)?

(d) By choosing $\varepsilon > 0$, $\ell \in \mathbb{N}$, and $\delta > 0$ appropriately, use parts (b) and (c) with the bound(4) 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.

Solution:

(a) If f is constant over the support of some wavelet, then the projection of f onto that wavelet, as well as any child of that wavelet, is clearly 0. Since f changes values only at its (at most s) discontinuities, it is non-constant on the supports of at most s of the wavelets at any scale. Since f_ℓ includes projections onto only the top ℓ scales, it has at most $\boxed{s\ell + 1}$ non-zero coefficients (adding one for the projection onto $\psi_{0,0}$).

Since 0 is a multiple of ε, any non-zero coefficient in $f_{\ell,\varepsilon}$ corresponds to a non-zero coefficient of f_ℓ, and so, by part (a), at most $\boxed{s\ell + 1}$ coefficients of $f_{\ell,\varepsilon}$ are non-zero.

(b) Linear combinations of wavelets of scale at most ℓ can exactly fit f except on the at most s intervals of lengths $2^{-\ell}$ on which f is discontinuous. That is, the measure of the set $E \subseteq [0, 1]$ on which the wavelet approximation is not exactly equal to f is at most $s2^{-\ell}$. Since $\|f\|_\infty \leq M$, if the wavelet approximation is 0 whenever it is not exactly f, the error of the approximation for any $x \in E$ is at most M. Since the wavelet basis is orthonormal, f_ℓ as defined minimizes the error of approximating f by wavelets of scale at most ℓ, and thus, $\boxed{\|f - f_\ell\|_2^2 \leq M^2 s 2^{-\ell}.}$

Note, for any j, k, $a \in \mathbb{R}$, if a_ε denotes a rounded to the nearest multiple of ε, then

$$\|a\psi_{j,k} - a_\varepsilon \psi_{j,k}\|_2^2 = |a - a_\varepsilon|^2 (2^{j/2})^2 \cdot 2^{-j} \leq \varepsilon^2.$$

(since $a\psi_{j,k}$ and $a_\varepsilon \psi_{j,k}$ disagree by at most $\varepsilon 2^{j/2}$ on an interval of length at most 2^{-j}). Thus, $\|f_\ell - (f_\ell)_{\ell,\varepsilon}\|_2^2 \leq 2s\ell\varepsilon^2$, where k is the number of nonzero coefficients of f_ℓ. Using the definition of $f_{\ell,\varepsilon}$, the Pythagorean Theorem, and parts (a) and (b),

$$\|f - f_{\ell,\varepsilon}\|_2^2 \leq \|f - (f_\ell)_{\ell,e}\|_2^2 = \|f - f_\ell\|_2^2 + \|f_\ell - (f_\ell)_{\ell,e}\|_2^2 \leq \boxed{M^2 s 2^{-\ell} + 2s\ell\varepsilon^2.}$$

(c) Since each coefficient of $f_{\ell,\varepsilon}$ has $2M/\varepsilon$ possible nonzero values, any given non-zero coefficient can be specified with at most $\log_2(2M/\varepsilon)$. Since there are at most $2s\ell$ nonzero coefficients, we can encode $f_{\ell,\varepsilon}$ using $\boxed{c(f_{\ell,\varepsilon}) \leq 2s\ell \log_2(2M/\varepsilon)}$ bits.

[3] Here, treat M as a constant.

(d) Since $\widehat{f}_{\ell,\varepsilon} \in Q_{\ell,\varepsilon}$, the CRM bound and parts (b) and (c) give

$$\begin{aligned}
\|\widehat{g} - \widehat{f}\|_2^2 &\leq \min_{g \in Q_{\ell,\varepsilon}} \left\{ \|g - \widehat{f}\|_2^2 + \frac{c(g) + \ln(1/\delta)}{n} \right\} + \delta \\
&\leq \|\widehat{f}_{\ell,\varepsilon} - \widehat{f}\|_2^2 + \frac{c(f^*_{\ell,\varepsilon}) + \ln(1/\delta)}{n} + \delta \\
&\leq M^2 s 2^{-\ell} + 2s\ell\varepsilon^2 + \frac{2s\ell \log_2(2M/\varepsilon) + \ln(1/\delta)}{n} + \delta \\
&\leq \frac{s(2\log_2 n + M)}{n} + \frac{s \log_2^2(2Mn) + \ln n}{n} + \frac{1}{n} \in O\left(\frac{s \log^2 n}{n}\right),
\end{aligned}$$

for $\varepsilon = n^{-1/2}, l = \log_2 n, \delta = 1/n$.

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 π 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 η 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 π such that the regret

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

Solution:

(a) Since a max of non-negative elements is at most their sum,

$$\begin{aligned}
\ln W_T &= \ln\left(\sum_{f \in \mathcal{F}} e^{\ln \pi(f) - \eta L_T(f)}\right) \\
&\geq \ln\left(\max_{f \in \mathcal{F}} e^{\ln \pi(f) - \eta L_T(f)}\right) = \max_{f \in \mathcal{F}} \left[\ln \pi(f) - \eta L_t(f)\right] = -\min_{f \in \mathcal{F}} \eta L_t(f) - \ln(\pi(f)).
\end{aligned}$$

(b) Since each $q_1(f) = \pi(f)$ and each $q_{t+1}(f) = q_t(f)e^{-\eta l(f(x_t), y_t)}$, it is easy to see by induction on t that each

$$q_t(f) = \frac{\pi(f)e^{-\eta \sum_{s=1}^{t-1} l(f(x_s), y_s)}}{\sum_{f \in \mathcal{F}} q_t(f)} = \frac{\pi(f)e^{-\eta L_{t-1}(f)}}{\sum_{f \in \mathcal{F}} \pi_j e^{-\eta L_{t-1}(f)}}.$$

Thus,

$$\begin{aligned}
\frac{W_t}{W_{t-1}} = \sum_{f \in \mathcal{F}} \frac{\pi(f)e^{-\eta L_t(f)}}{\sum_{g \in \mathcal{F}} \pi(g)e^{-\eta L_{t-1}(g)}} &= \sum_{f \in \mathcal{F}} e^{-\eta l(f(x_t), y_t)} \frac{\pi_i e^{-\eta L_{t-1}(f)}}{\sum_{g \in \mathcal{F}} \pi(g)e^{-\eta L_{t-1}(g)}} \\
&= \sum_{f \in \mathcal{F}} e^{-\eta l(f(x_t), y_t)} q_t(i) = \mathop{\mathbb{E}}_{f \sim q_t}\left[e^{-\eta l(f(x_t), y_t)}\right].
\end{aligned}$$

(c) Expanding $\ln(W_T)$ as a telescoping sum, applying part (b), and using the given bound (with $a = 0, b = 1, X = l(f(x_t), y_t), s = -\eta$),

$$\begin{aligned}
\ln(W_T) = \ln(W_0) + \sum_{t=1}^{T} \ln(W_t) - \ln(W_{t-1}) &\leq \sum_{t=1}^{T} \ln \mathop{\mathbb{E}}_{f \sim q_t}\left[e^{-\eta l(f(x_t, y_t)}\right] \\
&\leq \sum_{t=1}^{T} -\eta \mathop{\mathbb{E}}_{f \sim q_t}\left[l(f(x_t), y_t)\right] + \frac{\eta^2}{8} \\
&\leq -\eta\left(\sum_{t=1}^{T} \mathbb{E}\left[l(f_t(x_t), y_t)\right]\right) + \frac{\eta^2 T}{8},
\end{aligned}$$

since $\ln(W_0) = 0$.

(d) By parts (a) and (c),

$$-\eta \min_{f\in\mathcal{F}} L_t(f) - \ln \pi(f) \le -\eta \sum_{t=1}^{T} \mathop{\mathbb{E}}_{f\sim q_t} l(f(x_t), y_t) + \frac{\eta^2 T}{8}.$$

Dividing through by T and solving for $L_T(\widehat{f})$, gives

$$\begin{aligned} L_T(\widehat{f}) &\le \inf_{f\in\mathcal{F}} \left\{ L_T(f) + \frac{\eta}{8} + \frac{\ln(1/\pi(f))}{\eta} \right\} \\ &\le \inf_{f\in\mathcal{F}} \left\{ L_T(f) + (1 + \ln(1/\pi(f))) \sqrt{\frac{T}{8}} \right\}, \end{aligned} \tag{5}$$

for $\eta = \sqrt{\frac{8}{T}}$. If $\mathcal{F}$ is finite and $\pi_* := \min_{f\in\mathcal{F}} \pi(f) > 0$, then, letting $f_* := \operatorname{argmin}_{f\in\mathcal{F}} L_T(f)$, (5) implies

$$L_T(\widehat{f}) - L_T(f_*) \le (1 - \ln \pi_*) \sqrt{\frac{T}{8}} \in O\left(T^{1/2}\right).$$

References

Peter Harremoës. Binomial and poisson distributions as maximum entropy distributions. *IEEE Transactions on information theory*, 47(5):2039–2041, 2001.