MATH 582G Homework 4

Cade Ballew #2120804

January 30, 2023

1 Problem 1

Let $(P_1, \ldots, P_n)$ and $(Q_1, \ldots, Q_n)$ be probability distributions and let $P^{1:n}$ and $Q^{1:n}$ be the product measures. Then,

$$D(P^{1:n}||Q^{1:n}) = E_P\left[\log \frac{\prod_{j=1}^n dP_j}{\prod_{j=1}^n dQ_j}\right] = E_P\left[\sum_{j=1}^n (\log dP_j - \log dQ_j)\right]$$
$$= \sum_{j=1}^n E_P\left[\log dP_j - \log dQ_j\right] = \sum_{j=1}^n E_P\left[\log \frac{dP_j}{dQ_j}\right] = \sum_{j=1}^n D(P_j||Q_j).$$

We also have that

$$\tfrac{1}{2}H^2(P^{1:n}||Q^{1:n}) = \frac{1}{2}\int \left(\sqrt{\prod_{j=1}^n dP_j} - \sqrt{\prod_{j=1}^n dQ_j}\right)^2 d\nu = 1 - \frac{1}{2}\int \prod_{j=1}^n \sqrt{dP_j dQ_j}\, d\nu$$

which follows from expanding the square and the fact that these are probability measures. This process also gives us that

$$1 - \prod_{j=1}^n \left(1 - \tfrac{1}{2}H^2(P_j||Q_j)\right) = 1 - \prod_{j=1}^n \left(1 - \left(1 - \tfrac{1}{2}\int \sqrt{dP_j dQ_j}\, d\nu\right)\right)$$
$$= 1 - \frac{1}{2}\prod_{j=1}^n \int \sqrt{dP_j dQ_j}\, d\nu.$$

Independence allows us to push the product inside/outside the integral, so it must hold that

$$\tfrac{1}{2}H^2(P^{1:n}||Q^{1:n}) = 1 - \prod_{j=1}^n \left(1 - \tfrac{1}{2}H^2(P_j||Q_j)\right).$$

2 Problem 2

Now letting H denote entropy, we wish to minimize the negative entropy over X subject to the constraint that

$$\sum_{j=1}^{n} X_j = 1, \quad X_j \geq 0$$

Since H is concave and it and the constraints symmetric in their arguments, it will be maximized when all entries are equal. One can see this by assuming the contrary and noting that any permutation of a solution will achieve the same objective value and satisfy the constraints, so we can consider the average of all such cyclic permutations. Thus,

$$H(X) \leq \sum_{j=1}^{n} -\left(\frac{1}{n}\right) \log \frac{1}{n} = \sum_{j=1}^{n} \log n = \log(|\text{support}(X)|).$$

For the remaining inequalities, we can consider the continuous case.
To prove the contractive property, we note that the KL-divergence is always nonnegative, so by marginalizing and splitting the log, we can write

$$\begin{aligned} 0 \leq D(p_{x,y}||p_x p_y) &= \int_{x,y} p_{x,y} \log \frac{p_{x,y}}{p_x p_y} \\ &= \int_{x,y} p(x,y) \log p_{x|y} - \int_{x,y} p(x,y) \log p_x \\ &= -\int_x p(x) \log p(x) + E_Y \left[\int_x p(x|y) \log p(x|y) \right] = H(X) - H(X|Y), \end{aligned}$$

so

$$H(X|Y) \leq H(X).$$

To see the chain rule, we write

$$\begin{aligned} H(X,Y) &= -\int_{x,y} p_{x,y} \log p_{x,y} d\mu(x) d\nu(y) \\ &= -\int_{x,y} p_{x,y} \log p_y d\mu(x) d\nu(y) - \int_{x,y} p_{x,y} \log p_{x|y} d\mu(x) d\nu(y) \\ &= -\int_y p_y \log p_u d\nu(y) - E_Y \left[\int_x p_{x|y} \log p_{x|y} d\mu(x) \right] \\ &= H(Y) + H(X|Y). \end{aligned}$$

The conditional chain rule follows in a similar manner. Namely, we note that

$$\log \frac{p_{x,y,z}}{p_z} = \log \frac{p_{y,z}}{p_z} + \log \frac{p_{x,y,z}}{p_{y,z}}.$$

Taking negative expectations integrating through variables as necessary and marginalizing, we see that

$$H(X,Y|Z) = H(Y|Z) + H(X|Y,Z)$$

follows by definition.

3 Problem 3

Let P_θ be the uniform distrbution on $[\theta, \theta+1]$ and define the estimator $\hat{\theta}(z_1 \ldots, z_n) = \min\{z_1, \ldots, z_n\}$. Then,

$$Pr(\hat{\theta} - \theta \leq \delta) = \delta^n,$$

so

$$Pr((\hat{\theta} - \theta)^2 \geq \delta) = (1 - \sqrt{\delta})^n,$$

and

$$E_{P_\theta}(\hat{\theta} - \theta)^2 = \int_0^1 (1 - \sqrt{\delta})^n d\delta = \int_0^1 2x(1-x)^n dx.$$

We use Mathematica to compute this integral which yields that

$$E_{P_\theta}(\hat{\theta} - \theta)^2 = \frac{2}{n^2 + 3n + 2} \leq \frac{2}{n^2}.$$

Because our lower bound for the minimax risk is also $\mathrm{O}(^{-2})$, we can conclude that $\hat{\theta}$ is minimax optimal for learning θ from within the family $\mathcal{P} = \{R_\theta^n : \theta \in \mathbb{R}\}$.

4 Problem 4

4.1 Part a

To see that the KL divergence corresponds to the f-divergence defined by $f(t) = t \log t$, we plug this into the definition to get that

$$D_f(P||Q) = -\int q(x) \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} \nu(dx) = \int p(x) \log \frac{p(x)}{q(x)} \nu(dx) = D(P||Q).$$

4.2 Part b

If we instead take $f(t) = -\log t$, we instead get that

$$D_f(P||Q) = -\int q(x) \log \frac{p(x)}{q(x)} \nu(dx) = \int q(x) \log \frac{q(x)}{p(x)} \nu(dx) = D(Q||P).$$

4.3 Part c

To find an f for which the squared Hellinger distance corresponds to an f-divergence, we equate terms inside the integral to get that

$$qf\left(\frac{p}{q}\right) = (\sqrt{p} - \sqrt{q})^2.$$

Dividing q into the square root, we get that

$$f\left(\frac{p}{q}\right) = \left(\sqrt{\frac{p}{q}} - 1\right)^2,$$

so the choice of

$$f(t) = t - 2\sqrt{t} + 1$$

gives us the squared Hellinger distance which one could confirm by plugging in directly.

4.4 Part d

If we consider $f(t) = 1 - \sqrt{t}$, then we have

$$\begin{aligned} D_f(P||Q) &= \int q(x)\left(1 - \sqrt{\frac{p(x)}{q(x)}}\right)\nu(dx) = \int \left(q(x) - \sqrt{p(x)q(x)}\right)\nu(dx) \\ &= \int \left(\tfrac{1}{2}q(x) - \sqrt{p(x)q(x)} + \tfrac{1}{2}p(x)\right)\nu(dx) - \frac{1}{2}\int (p(x) - q(x))\nu(dx) \\ &= \tfrac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \nu(dx) = \frac{1}{2}H^2(P||Q). \end{aligned}$$

5 Problem 5

5.1 Part a

Let Q_1, Q_2 be d-variate Gaussians with respective means μ_1, μ_2 both with covariance matrix $\Sigma \succ 0$. In this problem, we will make heavy use of the fact that

$$E\left[(x-v)^T A(x-v)\right] = \mathrm{Tr}(A\Sigma) + (\mu - v)^T A(\mu - v)$$

if x has mean vector μ and covariance matrix Σ and A, v are deterministic which follows from expanding the quadratic form. Noting that the distribution of a d-variate Gaussian with mean vector μ and covariance matrix Σ is given by

$$(2\pi)^{-d/2}\det(\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

we get that the KL divergence is given by

$$\begin{aligned} D(Q_1||Q_2) &= E_{Q_1}\left[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right] \\ &= \frac{1}{2}\left(-\operatorname{Tr}(I_d)+(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)+\operatorname{Tr}(I_d)\right) \\ &= \frac{1}{2}\langle \mu_1-\mu_2, \Sigma^{-1}(\mu_1-\mu_2)\rangle \end{aligned}$$

which follows from the linearity of expectation and our identity.

5.2 Part b

Now, we consider the case where Q_1 and Q_2 have respective covariance matrices Σ_1, Σ_2. The KL divergence is now given by

$$\begin{aligned} D(Q_1||Q_2) &= E_{Q_1}\left[\log\frac{\det(\Sigma_1)^{-1/2}}{\det(\Sigma_2)^{-1/2}}-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] \\ &= \frac{1}{2}\left(\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)}-\operatorname{Tr}(I_d)+\operatorname{Tr}(\Sigma_2^{-1}\Sigma_1)+(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right) \\ &= \frac{1}{2}\left(\langle \mu_1-\mu_2, \Sigma_2^{-1}(\mu_1-\mu_2)\rangle+\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)}+\operatorname{Tr}(\Sigma_2^{-1}\Sigma_1)-d\right) \end{aligned}$$

which again follows from the linearity of expectation and our identity.

6 Problem 6

Consider

$$P_\theta(y_1,\ldots,y_n)=\prod_{j=1}^{n}\left(h(y_j)\exp\left(\frac{y_j\langle x_i,\theta\rangle-\Phi(\langle x_i,\theta\rangle)}{s(\sigma)}\right)\right)$$

where $s(\sigma)>0$ is a known scale factor and $\Phi:\mathbb{R}\to\mathbb{R}$ is the cumulant function of the generalized linear model.

6.1 Part a

We compute the KL divergence by definition immediately simplifies to

$$D(P_\theta||P_{\theta'})=\frac{1}{s(\sigma)}\sum_{j=1}^{n}\left(E_{P_\theta}[y_j]\left(\langle x_i,\theta\rangle-\langle x_i,\theta'\rangle\right)-\Phi(\langle x_i,\theta\rangle)+\Phi(\langle x_i,\theta'\rangle)\right).$$

The fact that Φ is the partition function gives that $E_{P_\theta}[y_j]=\Phi'(\langle x_i,\theta\rangle)$ which follows from elementary properties of GLMs, so we find that

$$D(P_\theta||P_{\theta'})=\frac{1}{s(\sigma)}\sum_{j=1}^{n}\left(\Phi(\langle x_i,\theta'\rangle)-\Phi(\langle x_i,\theta\rangle)+\Phi'(\langle x_i,\theta\rangle)\left(\langle x_i,\theta\rangle-\langle x_i,\theta'\rangle\right)\right).$$

6.2 Part b

Assuming that $\|\Phi''\|_\infty \leq L < \infty$, we use the mean value form of the Taylor expansion to find that

$$\Phi(\langle x_i, \theta' \rangle) = \Phi(\langle x_i, \theta \rangle) + \Phi'(\langle x_i, \theta \rangle)(\langle x_i, \theta' \rangle - \langle x_i, \theta \rangle) + \tfrac{1}{2}\Phi''(\xi)(\langle x_i, \theta' \rangle - \langle x_i, \theta \rangle)^2.$$

Plugging this into our expression from part a, with our assumption, we bound

$$D(P_\theta || P_{\theta'}) \leq \frac{L}{2s(\sigma)} \sum_{i=1}^{n} (\langle x_i, \theta' \rangle - \langle x_i, \theta \rangle)^2 = \frac{L}{2s(\sigma)} \|X(\theta - \theta')\|_2^2 \leq \frac{L\|X\|_2^2}{2s(\sigma)} \|\theta - \theta'\|_2^2$$

where X is the design matrix, so we have an upper bound which scales quadratically up to a constant times $1/s(\sigma)$.

6.3 Part c

To build a minimax lower bound, we apply Fano's method via conditions (15.35a) and (15.35b) and in the vein of example 15.14 all in Wainwright. To choose δ consider a set $\{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 4\delta\}$ and let $\theta_1, \ldots, \theta_M$ be a 2δ-packing in the 2-norm. From this, we can find a packing with $\log M \geq d \log 2$. Then,

$$\|\theta_j - \theta_k\| \leq 8\delta,$$

so we have from part b that

$$D(P_\theta || P_{\theta'}) \leq \frac{32L\eta_{\max}^2 n}{s(\sigma)} \delta^2.$$

Letting $c^2 = \frac{32L\eta_{\max}^2}{s(\sigma)}$, the condition (15.35b) to apply Fano's method can be satisfied by setting $\delta^2 = \frac{d}{2c^2}$[1]. Now, this is just a direct application of Fano's inequality

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} E\left[\|\hat{\theta} - \theta\|_2^2\right] \geq \frac{1}{64} \frac{s(\sigma)}{L\eta_{\max}^2} \frac{d}{n}$$

Since we wish to instead consider $\theta \in \mathbb{B}_2^d(1)$, we can tighten our bound by considering that the worst case distance is 1 since we are in the unit ball. Thus,

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} E\left[\|\hat{\theta} - \theta\|_2^2\right] \geq \min\left\{1, \frac{1}{64} \frac{s(\sigma)}{L\eta_{\max}^2} \frac{d}{n}\right\}.$$

6.4 Part d

To reduce this to linear regression, we note that

$$\frac{1}{n}\|X(\hat{\theta} - \theta)\|_2^2 \geq \eta_{\min}^2 \|\hat{\theta} - \theta\|_2^2$$

[1] Note that we reverse engineered this from example 15.14, but this can be verified to satisfy the condition.

and $s(\sigma) = \sigma^2$ and $L = 1$ since $\Phi(t) = \frac{1}{2}t^2$, so our bound becomes

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} E\left[\frac{1}{n}\|X(\hat{\theta} - \theta)\|_2^2\right] \geq \frac{1}{64}\frac{\eta_{\min}^2}{\sigma^2 \eta_{\max}^2}\frac{d}{n} \geq \frac{\sigma^2}{128}\frac{\operatorname{rank}(X)}{n}.$$