Problem Set 1

Stats 311/EE 377

Due: Thursday, January 21 in class

Note: If you are familiar with measure theory, you may find it more satisfying to prove every result in this homework in full generality (that is, instead of assuming that all distributions have Lebesgue densities, simply assume that they are all absolutely continuous with respect to some base measure μ). If you are *not* familiar with measure theory, there is essentially no loss of generality in anything in this problem set by proving results assuming all distributions have densities on $\mathbb{R}$.

Our first few questions investigate properties of a divergence between distributions that is weaker than the KL-divergence, but is intimately related to optimal testing. Let P_1 and P_2 be arbitrary distributions on a space $\mathcal{X}$. The *total variation distance* between P_1 and P_2 is defined as

$$\|P_1 - P_2\|_{\mathrm{TV}} := \sup_{A \subset \mathcal{X}} |P_1(A) - P_2(A)| .$$

Question 1: Prove the following identities about total variation. Throughout, let P_1 and P_2 have densities p_1 and p_2 on a (common) set $\mathcal{X}$.

(a) $2\|P_1 - P_2\|_{\mathrm{TV}} = \int |p_1(x) - p_2(x)| dx$.

(b) For functions $f : \mathcal{X} \to \mathbb{R}$, define the supremum norm $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Show that $2\|P_1 - P_2\|_{\mathrm{TV}} = \sup_{\|f\|_\infty \le 1} \int_{\mathcal{X}} f(x)(p_1(x) - p_2(x)) dx$.

(c) $\|P_1 - P_2\|_{\mathrm{TV}} = \int \max\{p_1(x), p_2(x)\} dx - 1$.

(d) $\|P_1 - P_2\|_{\mathrm{TV}} = 1 - \int \min\{p_1(x), p_2(x)\} dx$.

(e) For functions $f, g : \mathcal{X} \to \mathbb{R}$,

$$\inf \left\{ \int f(x) p_1(x) dx + \int g(x) p_2(x) dx : f + g \ge 1, f \ge 0, g \ge 0 \right\} = 1 - \|P_1 - P_2\|_{\mathrm{TV}} .$$

In class, we defined a quantizer as any function $g : \mathcal{X} \to \{1, \ldots, m\}$ for some $m \in \mathbb{N}$, and we noted that g partitions $\mathcal{X}$ into sets $A_1, \ldots, A_m$, where $g(x) = i$ for $x \in A_i$. Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex function satisfying $f(1) = 0$. In class, we defined the f-divergence between two probability measures P and Q; as with the KL-divergence, the most general definition of an f-divergence between P and Q on the space $\mathcal{X}$ may be defined as

$$D_f(P \| Q) := \sup \{ D_f(P \| Q \mid g) : g \text{ quantizes } \mathcal{X} \}, \tag{1}$$

where $D_f(P\|Q \mid g) = \sum_{i=1}^m Q(A_i) f(\frac{P(A_i)}{Q(A_i)})$. Now we show how to extend the data processing inequality in class—which applied to the KL-divergence—to the family of f-divergences.

Question 2 (Generalized "log-sum" inequalities)**:** Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an arbitrary convex function.

(a) Let $a_i, b_i, i = 1, \ldots, n$ be non-negative reals. Prove that

$$\left(\sum_{i=1}^n a_i\right) f\left(\frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n a_i}\right) \le \sum_{i=1}^n a_i f\left(\frac{b_i}{a_i}\right).$$

(b) Generalizing the preceding result, let $a : \mathcal{X} \to \mathbb{R}_+$ and $b : \mathcal{X} \to \mathbb{R}_+$, and let $u : \mathcal{X} \to \mathbb{R}_+$ satisfy $\int u(x)dx < \infty$. Show that

$$\int a(x)u(x)dx f\left(\frac{\int b(x)u(x)dx}{\int a(x)u(x)dx}\right) \le \int a(x) f\left(\frac{b(x)}{a(x)}\right) u(x)dx.$$

(Hint: use the fact that the perspective of a function f, defined by $h(x,t) = tf(x/t)$ for $t > 0$, is jointly convex in x and t [e.g. 1, Chapter 3.2.6].)

Question 3 (Data processing and f-divergences I)**:** As with the KL-divergence, given a quantizer g of the set $\mathcal{X}$, where g induces a partition $A_1, \ldots, A_m$ of $\mathcal{X}$, we define the f-divergence between P and Q conditioned on g as

$$D_f(P\|Q \mid g) := \sum_{i=1}^m Q(A_i) f\left(\frac{P(A_i)}{Q(A_i)}\right) = \sum_{i=1}^m Q(g^{-1}(\{i\})) f\left(\frac{P(g^{-1}(\{i\}))}{Q(g^{-1}(\{i\}))}\right).$$

Given quantizers g_1 and g_2, we say that g_1 is a *finer* quantizer than g_2 under the following condition: assume that g_1 induces the partition $A_1, \ldots, A_n$ and g_2 induces the partition $B_1, \ldots, B_m$; then for any of the sets B_i, there are exists some k and sets $A_{i_1}, \ldots, A_{i_k}$ such that $B_i = \cup_{j=1}^k A_{i_j}$. We let $g_1 \prec g_2$ denote that g_1 is a finer quantizer than g_2.

(a) Let g_1 and g_2 be quantizers of the set $\mathcal{X}$, and let $g_1 \prec g_2$, meaning that g_1 is a finer quantization than g_2. Prove that

$$D_f(P\|Q \mid g_2) \le D_f(P\|Q \mid g_1).$$

Equivalently, show that whenever $\mathcal{A}$ and $\mathcal{B}$ are collections of sets partitioning $\mathcal{X}$, but $\mathcal{A}$ is a finer partition of $\mathcal{X}$ than $\mathcal{B}$, that

$$\sum_{B \in \mathcal{B}} Q(B) f\left(\frac{P(B)}{Q(B)}\right) \le \sum_{A \in \mathcal{A}} Q(A) f\left(\frac{P(A)}{Q(A)}\right).$$

(*Hint:* Use the result of Question 2(a)).

(b) Suppose that $\mathcal{X}$ is discrete so that P and Q have p.m.f.s p and q. Show that

$$D_f(P\|Q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right).$$

Question 4 (General data processing inequalities)**:** Let f be a convex function satisfying $f(1) = 0$. Let K be a Markov transition kernel from $\mathcal{X}$ to $\mathcal{Z}$, that is, $K(\cdot \mid x)$ is a probability distribution on $\mathcal{Z}$ for each $x \in \mathcal{X}$. (Written differently, we have $X \to Z$, and conditioned on $X = x$, Z has distribution $K(\cdot \mid x)$, so that $K(A \mid x)$ is the probability that $Z \in A$ given $X = x$.)

(a) Define the marginals $K_P(A) = \int_{\mathcal{X}} K(A, x)p(x)dx$ and $K_Q(A) = \int K(A, x)q(x)dx$. Show that

$$D_f\left(K_P \| K_Q\right) \leq D_f\left(P \| Q\right).$$

(Hint: by equation (1), w.l.o.g. we may assume that $\mathcal{Z}$ is finite and $\mathcal{Z} = \{1, \ldots, m\}$.)

(b) Let X and Y be random variables with joint distribution P_{XY} and marginals P_X and P_Y. Define the f-information between X and Y as

$$I_f(X; Y) := D_f\left(P_{XY} \| P_X \times P_Y\right).$$

Use part (a) to show the following general data processing inequality: if we have the Markov chain $X \to Y \to Z$, then

$$I_f(X; Z) \leq I_f(X; Y).$$

Question 5 (Concentration of bounded random variables)**:** Let X be a random variable taking values in $[a, b]$, where $-\infty < a \leq b < \infty$. In this question, we show *Hoeffding's Lemma*, that is, that X is sub-Gaussian: for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

(a) Show that $\mathrm{Var}(X) \leq (\frac{b-a}{2})^2 = \frac{(b-a)^2}{4}$ for any random variable X taking values in $[a, b]$.

(b) Let

$$\varphi(\lambda) = \log \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))].$$

Assuming that $\mathbb{E}[X] = 0$ (convince yourself that this is no loss of generality) show that

$$\varphi(0) = 0, \quad \varphi'(0) = 0, \quad \varphi''(t) = \frac{\mathbb{E}[X^2 e^{tX}]}{\mathbb{E}[e^{tX}]} - \frac{\mathbb{E}[Xe^{tX}]^2}{\mathbb{E}[e^{tX}]^2}.$$

(You may assume that derivatives and expectations commute, which they do in this case.)

(c) Construct a random variable Y_t, defined for $t \in \mathbb{R}$, such that $Y_t \in [a, b]$ and

$$\mathrm{Var}(Y_t) = \varphi''(t).$$

(You may assume X has a density for simplicity.)

(d) Using the result of part (c), show that $\varphi(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}$ for all $\lambda \in \mathbb{R}$.

Question 6 (Variational forms of KL divergence)**:** Let P and Q be arbitrary distributions on a common space $\mathcal{X}$. Prove the following variational representation of the KL divergence:

$$D_{\rm kl}\left(P\|Q\right) = \sup_{f:\mathbb{E}_Q[e^{f(X)}]<\infty} \left\{\mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp(f(X))]\right\}.$$

You may assume that P and Q have densities. *Hint: the technique of Problem 5*(c) *may be useful. It might not be, too.*

Question 7 (Getting information the right way)**:** In this question, we study the differences in recovery of a random signal when we have sequential measurements, chosen optimally, and a block of random measurements performed *a priori*. We assume the following model: we have

$$X \sim \mathsf{N}(0, \Sigma),$$

where $\Sigma \succ 0$ is a positive definite matrix in $\mathbb{R}^{d\times d}$. We would like to estimate X based on a sequence of random observations of the form

$$Y_i = w_i^\top X + Z_i, \quad Z_i \stackrel{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2),$$

where Z_i are independent of everything else in the problem and W_i are constrained to satisfy $\|w_i\|_2 \le 1$.

(a) Give the joint distribution of the vector

$$(X, Y_1, \ldots, Y_k) \in \mathbb{R}^{d+k}.$$

(b) What is the distribution of X conditional on observing $Y_1, \ldots, Y_k$ (assuming that $w_1, \ldots, w_k$ are fixed vectors)?

(c) Prove that for *any* distribution on X (including non-Gaussian), if w_i is a function of $Y_1, \ldots, Y_{i-1}$, we have

$$I(X; Y_i \mid Y_1, \ldots, Y_{i-1}) \le \frac{1}{2} \log \left(1 + \frac{\mathbb{E}[\mathrm{Var}(w_i^\top X \mid Y_1, \ldots, Y_{i-1})]}{\sigma^2}\right).$$

(d) Consider the following sequential observation strategy: at each iteration $i = 1, 2, \ldots, k$, we choose the measurement vector w_i to approximately maximize the information $I(X; Y_i \mid Y_1, \ldots, Y_{i-1})$ by choosing w_i to maximize the upper bound in part (c). In the case that $X \sim \mathsf{N}(0, \Sigma)$ and we know Σ, what choice should we make for w_i? Do the previous observations $Y_1, \ldots, Y_{i-1}$ affect this choice?

(e) Using the data online at http://web.stanford.edu/class/stats311/data/information-gathering.jl or http://web.stanford.edu/class/stats311/data/InformationGathering.m, implement and compare the procedure of part (d) with a naive strategy of choosing k vectors w_i uniformly at random from the set $\{w \in \mathbb{R}^d : \|w\|_2 = 1\}$. Repeat this experiment 400 times each for each k in the file. Using the plotting code in the file, plot the ratio of the mean squared errors $\mathbb{E}[\|X - \mathbb{E}[X \mid Y_1, \ldots, Y_k]\|_2^2]$ for the random w to that of the greedy choice of w for each $k \in \{2^i, i = 1, \ldots, 8\}$.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.