ECE 5630 - Homework Assignment 2

February 21st 2020

Due to: Tuesday, March 5th, 2020 (at the beginning of the lecture)

Instructions: Submission in pairs is allowed. Prove and explain every step in your answers.

1) **Properties of f-divergences:** For any $P, Q \in \mathcal{P}(\mathcal{X})$ probability measures on the same probability space, dominated by a common measure $P, Q \ll \lambda$, recall that

$$D_f(P\|Q) := \mathbb{E}_Q f\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right),$$

where f is a convex function satisfying the assumption given in class and $\mathsf{d}\mu/\mathsf{d}\lambda$ is the Radon-Nikodym derivative of μ with respect to λ. Prove the following properties:

a) Non-negativity: $D_f(P\|Q) \geq 0$ with equality if and only if $P = Q$.

b) Joint convexity: The map $(P, Q) \mapsto D_f(P\|Q)$ is (jointly) convex.

Hint: Use the 'perspective' of f, defined by $g(x, y) = yf\left(\frac{x}{y}\right)$, which is convex in (x, y) if and only if f is convex.

c) Conditioning increases f-divergence: For $P_X \in \mathcal{P}(\mathcal{X})$ and two transition kernels (channels) $P_{Y|X}$ and $Q_{Y|X}$ from $\mathcal{X}$ to $\mathcal{Y}$, consider the probability measures $P_{X,Y} := P_X P_{Y|X}$ and $Q_{X,Y} := P_X Q_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$. Denoting by P_Y and Q_Y their marginals on $\mathcal{Y}$, show that

$$D_f(P_Y\|Q_Y) \leq D_f(P_{Y|X}\|Q_{Y|X}|P_X) =: \int D_f(P_{Y|X=x}\|Q_{Y|X=x})\mathsf{d}P_X(x). \tag{1}$$

d) Joint vs. marginal: For $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ and a transition kernel $P_{Y|X}$, define $P_{X,Y} := P_X P_{Y|X}$ and $Q_{X,Y} := Q_X P_{Y|X}$ (measures on the product space, as before). Show that

$$D_f(P_X\|Q_X) = D_f(P_{X,Y}\|Q_{X,Y}).$$

2) **Example of Data Processing Inequality:** Let $(\mathcal{X}, \mathcal{F})$ be a measurable space ($\mathcal{X}$ is the sample set and $\mathcal{F}$ the σ-algebra). Use the Data Processing Inequality to show that for any two probability measures P, Q on $(\mathcal{X}, \mathcal{F})$ and any $E \in \mathcal{F}$:

$$D_f(P\|Q) \geq \sup_{A \in \mathcal{F}} \left\{ \big(1 - Q(A)\big) f\left(\frac{1 - P(A)}{1 - Q(A)}\right) + Q(A) f\left(\frac{P(A)}{Q(A)}\right) \right\}.$$

3) **f-divergences, metrics, and mismatched support:** Recall the definitions of Kullback-Leibler (KL) divergence $D_{\mathsf{KL}}(\cdot\|\cdot)$ and χ^2-divergence $\chi^2(\cdot\|\cdot)$ provided in class. Show that:

a) $\delta_{\mathsf{TV}}(\cdot, \cdot)$ is a metric on $\mathcal{P}(\mathcal{X})$.

Hint: Use relation to L^1 norm. You may assume probability measures have densities, but a general proof is preferable.

b) $D_{\mathsf{KL}}(P,Q) = \chi^2(P,Q) = \infty$ whenever $P \not\ll Q$ (i.e., P is not absolutely continuous with respect to Q).

c) $\delta_{\mathsf{TV}}(P,Q)$ attains its maximal value of 1, whenever $\mathrm{supp}(P) \cap \mathrm{supp}(Q) = \emptyset$.

d) Explain why the previous property is undesired when performing generative modeling $\inf_{\theta\in\Theta} \delta(P, Q_\theta)$ of a data distribution P via a parametrized family $\{Q_\theta\}_{\theta\in\Theta}$ under divergence δ.

4) **Jensen-Shannon divergence:** Let $f(x) = x\log\left(\frac{2x}{x+1}\right) + \log\left(\frac{2}{x+1}\right)$. Show that:

a) Shown that $f : (0,\infty) \to \mathbb{R}$ is a convex function, with $f(1) = 0$, which is strictly convex around 1.

b) Let $\mathsf{JSD}(P\|Q)$ be the f-divergence induced by the above f. This is known as the *Jensen-Shannon divergence* (JSD). Prove that

i) $\mathsf{JSD}(P\|Q) = \frac{1}{2}D_{\mathsf{KL}}\left(P\middle\|\frac{P+Q}{2}\right) + \frac{1}{2}D_{\mathsf{KL}}\left(Q\middle\|\frac{P+Q}{2}\right)$.

Note: This is why JSD is sometimes referred to as symmetrized KL divergence.

ii) $\mathsf{JSD}(P\|Q)$ is maximized at $2\log 2$ but pairs (P,Q) with $\mathrm{supp}(P) \cap \mathrm{supp}(Q) = \emptyset$.

Note: It can be shown that $\sqrt{\mathsf{JSD}(P\|Q)}$ is a metric on the space of probability measures. This is non-trivial.

5) ***f*-divergences variational formula:** The convex conjugate of a function $f : I \to \mathbb{R}$ is $f^\star(y) = \sup_{x\in I} yx - f(x)$. We saw the following variational representation of f-divergences:

$$D_f(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\big[f^\star\big(g(X)\big)\big],$$

where the supremum is over all measurable g for which the expectations are finite. Show that

a) $D_f(P\|Q) \geq \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\big[f^\star\big(g(X)\big)\big]$, when supremising over all g as above.

Note: You may follow the argument given in class but must precisely justify each step.

b) Derive the following variational formulas by computing convex conjugates:

i) $D_{\mathsf{KL}}(P\|Q) = 1 + \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\big[e^{g(X)}\big]$

ii) $\delta_{\mathsf{TV}}(P,Q) = \sup_{\|g\|_\infty\leq 1} \frac{1}{2}\big(\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)]\big)$

iii) $\chi^2(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[g(X) + \frac{g^2(x)}{4}\right]$

Hint: Consider the change of variables $h(x) = \frac{g(x)}{2} + 1$.

6) **Inequalities between *f*-divergences:** We examine how some f-divergences relate to one another. Prove the following:

a) For any distributions $P, Q \in \mathcal{P}(\mathcal{X})$, it holds that

$$D_{\mathsf{KL}}(P\|Q) \leq \log\big(1 + \chi^2(P\|Q)\big) \leq \chi^2(P\|Q).$$

Hint: For all $x > -1$, it holds that $x \geq \log(1+x)$.

b) Assume that $P = \mathsf{Ber}(p)$ and $Q = \mathsf{Ber}(q)$ where $p, q \in (0,1)$. Show that

$$\delta_{\mathsf{TV}}(P,Q)^2 \leq \frac{\ln(2)}{2} D_{\mathsf{KL}}(P\|Q).$$

Hint: Define $g(p,q) := D_{\mathsf{KL}}(P\|Q) - \frac{2}{\ln(2)}\delta_{\mathsf{TV}}(P,Q)^2$ and consider its derivative.

c) Assume that P and Q have finite supports. Show that

$$\delta_{\mathsf{TV}}(P,Q)^2 \leq \frac{1}{2} D_{\mathsf{KL}}(P\|Q).$$

This results is known as *Pinker's Inequality*.

Hint: Define $h(x) = x\log(x) - x + 1$. Start by showing that $(4+2x)h(x) \geq 3(x-1)^2$, $\forall\, x \geq 0$.