ECE 6970 - Homework Assignment 3

November 15th 2019

Due to: Tuesday, November 26th, 2019 (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) Same channel $\implies$ same divergence: 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 $E \in \mathcal{F}$:

$$D_f(P\|Q) \geq D_f\big(\mathsf{Bern}\big(P(E)\big)\big\|\mathsf{Bern}\big(Q(E)\big)\big),$$

where $\mathsf{Bern}(p)$, for $p \in [0, 1]$, is a Bernoulli p distribution.

3) **f-divergences, metrics, and mismatched support:** Recall the definitions of Kullback-Leibler (KL) divergence $D_{\mathsf{KL}}(\cdot\|\cdot)$, χ^2-divergence $\chi^2(\cdot\|\cdot)$, Total Variations Distance $\delta_{\mathsf{TV}}(\cdot, \cdot)$, Squared Hellinger Distance $\mathsf{H}^2(\cdot, \cdot)$, and Jensen-Shannon Divergence $\mathsf{JSD}(\cdot\|\cdot)$ provided in class. Show that:

a) $\sqrt{H^2(\cdot,\cdot)}$ is a metric on $\mathcal{P}(\mathcal{X})$.

Hint: Use relation to L^2 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) $\delta_{\mathsf{TV}}(P,Q)$, $\mathsf{H}^2(P,Q)$ and $\mathsf{JSD}(P,Q)$ attain their maximal values, 1, 2, and $2\log 2$, respectively, whenever $\mathrm{supp}(P) \cap \mathrm{supp}(Q) = \emptyset$.

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

4) ***f*-divergences variational formula:** The convex conjugate of a function f on $\mathbb{R}$ is $f^\star(y) = \sup_{x\in\mathsf{dom}(f)} xy - f(x)$, where $\mathsf{dom}(f)$ is the domain of f. We saw the following variational representation of f-divergences:

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

where the supremum is over all measurable g for which the expectations are finite. In random variable notation, the right-hand side is written as $\sup_g \mathbb{E}_P\big[g(x)\big] - \mathbb{E}_Q\big[f^\star\big(g(X)\big)\big]$, with the law of X specified in the subscript. Show that

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

Hint: The convex conjugate is a bicunjugation, i.e., $(f^\star)^\star = f$. and for any $y \in \mathsf{dom}(f^\star)$, $f(x) \geq yx - f^\star(y)$.

b) **Bonus:** Assuming f is differentiable, equality in the supremum is attained by $g(x) = f'\left(\frac{\mathsf{d}P}{\mathsf{d}Q}(x)\right)$, where f' is the derivative of f. Prove this fact (not mandatory).

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

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

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

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 $h(x) = \frac{g(x)}{2} + 1$.

5) **Entropy (full) chain rule:** Let $(X_1,\ldots,X_k) \sim P_{X_1,\ldots,X_n}$. Show that:

a) If $(X_1,\ldots,X_k)$ is discrete, then its Shannon entropy decomposes as $H(X_1,\ldots,X_k) = \sum_{i=1}^k H(X_i|X_1,\ldots,X_{i-1})$, where $H(X_1|X_0) = H(X_1)$.

b) If $(X_1,\ldots,X_k)$ is jointly continuous, then its differential entropy decomposes as $h(X_1,\ldots,X_k) = h(X_k) + \sum_{i=1}^{k-1} h(X_{k-i}|X_k,\ldots,X_{k-(i-1)})$.

6) **Properties of mutual information:** Let $(X,Y,Z) \sim P_{X,Y,Z}$. Use properties learned in class to show that:

a) <u>Mutual information and conditional KL divergence:</u> $I(X;Y) = D_{\mathsf{KL}}(P_{Y|X}\|P_Y|P_X)$, where $P_{X,Y} = P_X P_{Y|X}$ and P_Y is its Y-marginal. The conditional KL divergence is defined in (1).

b) <u>More data $\implies$ more information:</u> $I(X;Y) \leq I(X;Y,Z)$.

c) <u>Mutual information and functions:</u> $I(X;Y) \geq I\big(X;f(Y)\big)$ for any deterministic function f. Furthermore, if f is continuous and one-to-one, then $I\big(X;f(X)\big) = H(X)$ for discrete X, and $I\big(X;f(X)\big) = \infty$ for continuous X. Do <u>not</u> use mutual information Data Processing Inequality in your proof.