ECE 5630 - Solutions Homework Assignment 2

1) a) Using Jensen's inequality

$$D_f(P\|Q) = \mathbb{E}_Q f\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right) \geq f\left(\mathbb{E}_Q \frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right) = 0,$$

where the last equality follows from the fact that $f(1) = 0$ and

$$\mathbb{E}_Q\left[\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right] = \int_{\mathcal{X}} \frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\mathsf{d}Q = \int_{\mathcal{X}} \frac{\mathsf{d}P}{\mathsf{d}\lambda}\mathsf{d}\lambda = 1.$$

Clearly, if $P = Q$ then $D_f(P\|Q) = 0$. By strong convexity of f at 1, it follows that if $D_f(P\|Q) = 0$ then $\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda} = 1$ or equivalently $P = Q$. To see why these two notions are equivalent, one can use the definition of Radon-Nikodym derivative. That is, for any measurable set A,

$$P(A) = \int_A \frac{\mathsf{d}P}{\mathsf{d}\lambda}\mathsf{d}\lambda = \int_A \frac{\mathsf{d}Q}{\mathsf{d}\lambda}\mathsf{d}\lambda = Q(A).$$

b) By convexity of the perspective function of f, for any $P_1, P_2, Q_1, Q_2 \in \mathcal{P}(\mathcal{X})$ and any $\alpha \in [0,1]$ it follows that

$$g\left(\alpha\frac{\mathsf{d}P_1}{\mathsf{d}\lambda} + (1-\alpha)\frac{\mathsf{d}P_2}{\mathsf{d}\lambda},\ \alpha\frac{\mathsf{d}Q_1}{\mathsf{d}\lambda} + (1-\alpha)\frac{\mathsf{d}Q_2}{\mathsf{d}\lambda}\right) \leq \alpha g\left(\frac{\mathsf{d}P_1}{\mathsf{d}\lambda},\ \frac{\mathsf{d}Q_1}{\mathsf{d}\lambda}\right) + (1-\alpha)g\left(\frac{\mathsf{d}P_2}{\mathsf{d}\lambda},\ \frac{\mathsf{d}Q_2}{\mathsf{d}\lambda}\right)$$

Thus by taking the integral of both sides, we get

$$D_f(\alpha P_1 + (1-\alpha)P_2\|\alpha Q_1 + (1-\alpha)Q_2) \leq \alpha D_f(P_1\|Q_1) + (1-\alpha)D_f(P_2\|Q_2).$$

c) Using Jensen's inequality

$$D_f(P_{Y|X}\|Q_{Y|X} \mid P_X) = \mathbb{E}_{P_X} D_f(P_{Y|X}\|Q_{Y|X}) \geq D_f(\mathbb{E}_{P_X} P_{Y|X}\|\mathbb{E}_{P_X} Q_{Y|X}) = D_f(P\|Q).$$

d) $P_X, Q_X \ll \lambda$. Let $\nu = \lambda P_{Y|X}$. Then, $P_{X,Y}, Q_{X,Y} \ll \nu$. We first show that $\mathsf{d}P_{X,Y}/\mathsf{d}\nu = \mathsf{d}P_X/\mathsf{d}\lambda$. For all measurable $A = A_x \times A_y$ where $A_x \in \mathcal{X}$ and $A_y \in \mathcal{Y}$, we have

$$\int_A \frac{\mathsf{d}P_{X,Y}}{\mathsf{d}\nu}\mathsf{d}\nu = \int_A \mathsf{d}P_{X,Y} = \int_{A_y}\left(\int_{A_x}\mathsf{d}P_X\right)\mathsf{d}P_{Y|X} = \int_{A_y}\left(\int_{A_x}\frac{\mathsf{d}P_X}{\mathsf{d}\lambda}\mathsf{d}\lambda\right)\mathsf{d}P_{Y|X} = \int_A \frac{\mathsf{d}P_X}{\mathsf{d}\lambda}\mathsf{d}\nu.$$

Then,

$$D_f(P_{X,Y}\|Q_{X,Y}) = \int_{\mathcal{X}\times\mathcal{Y}} f\left(\frac{\mathsf{d}P_{X,Y}/\mathsf{d}\nu}{\mathsf{d}Q_{X,Y}/\mathsf{d}\nu}\right)\mathsf{d}Q_{X,Y} = \int_{\mathcal{X}} f\left(\frac{\mathsf{d}P_X/\mathsf{d}\lambda}{\mathsf{d}Q_X/\mathsf{d}\lambda}\right)\int_{\mathcal{Y}}\mathsf{d}Q_{X,Y} = \int_{\mathcal{X}} f\left(\frac{\mathsf{d}P_X/\mathsf{d}\lambda}{\mathsf{d}Q_X/\mathsf{d}\lambda}\right)\mathsf{d}Q_X.$$

2) Let $A \in \mathcal{F}$. Define the transition kernel as $P_{Y|X}(A|x) = \delta_x(A)$. Let $P_{X,Y} = PP_{X|Y}$ and $Q_{X,Y} = QP_{X|Y}$. Then $P_Y = \mathbb{E}_P P_{Y|X} = \mathsf{Bern}(P(A))$ and $Q_Y = \mathbb{E}_Q P_{Y|X} = \mathsf{Bern}(Q(A))$. By data processing inequality, we get

$$D_f(P\|Q) \geq D_f(P_Y\|Q_Y) = D_f(\mathsf{Bern}(P(A))\|\mathsf{Bern}(Q(A))) = (1-Q(A))\, f\left(\frac{1-P(A)}{1-Q(A)}\right) + Q(A) f\left(\frac{P(A)}{Q(A)}\right).$$

The above inequality holds for all measurable sets A. By taking the supremum over all measurable sets, we get the desired inequality.

3) a) Recall the definition of Total Variation distance

$$\delta_{\mathsf{TV}}(P,Q)=\frac{1}{2}\int_{\mathcal{X}}|\mathsf{d}P-\mathsf{d}Q|.$$

Clearly, $\delta_{\mathsf{TV}}(P,Q)\geq 0$ with equality if and only if $P=Q$ and $\delta_{\mathsf{TV}}(P,Q)=\delta_{\mathsf{TV}}(Q,P)$. We show the triangle inequality for $P_1,P_2,P_3\in\mathcal{P}(\mathcal{X})$:

$$\begin{aligned}\delta_{\mathsf{TV}}(P_1,P_3)&=\frac{1}{2}\int_{\mathcal{X}}|\mathsf{d}P_1-\mathsf{d}P_3|\\&=\frac{1}{2}\int_{\mathcal{X}}|\mathsf{d}P_1-\mathsf{d}P_2+\mathsf{d}P_2-\mathsf{d}P_3|\\&\leq\frac{1}{2}\int_{\mathcal{X}}|\mathsf{d}P_1-\mathsf{d}P_2|+\frac{1}{2}\int_{\mathcal{X}}|\mathsf{d}P_2-\mathsf{d}P_3|\\&=\delta_{\mathsf{TV}}(P_1,P_2)+\delta_{\mathsf{TV}}(P_2,P_3).\end{aligned}$$

b) If P is not absolutely continuous with respect to Q, then there exists a measurable set A such that $Q(A)=0$ while $P(A)>0$. The KL-Divergence is then given by

$$D_{\mathsf{KL}}(P\|Q)=\int_{\mathcal{X}}\log\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right)\mathsf{d}P\geq\int_{A}\log\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right)\mathsf{d}P=\infty.$$

It holds that

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

Thus, if $D_{\mathsf{KL}}(P\|Q)=\infty$ then $\chi^2(P,Q)=\infty$.

c) We have

$$\delta_{\mathsf{TV}}(P\|Q)=\frac{1}{2}\int_{\mathcal{X}}|\mathsf{d}P-\mathsf{d}Q|\leq\frac{1}{2}\left(\int_{\mathcal{X}}\mathsf{d}P+\int_{\mathcal{X}}\mathsf{d}Q\right)=1,$$

with equality if $\operatorname{supp}(P)\cap\operatorname{supp}(Q)=\emptyset$.

d) We approximate the statistical distance between P and Q_θ using samples from the respective distributions. Thus, $\operatorname{supp}(\widehat{P})\cap\operatorname{supp}(\widehat{Q}_\theta)=\emptyset$. As a result, the statistical divergence between the two (empirical) distributions is not informative, which, in turn, makes the optimization problem $\inf_{\theta\in\Theta}\delta(\widehat{P},\widehat{Q}_\theta)$ challenging. For example, one cannot rely on gradient descent methods for the optimization problem as the gradient is 0 a.s.

4) a) Clearly, $f(1)=0$. Also, $f''(x)=\frac{1}{x(x+1)}>0$ for all $x>0$. So f is strictly convex.

b) We use the shorthand notation $\frac{\mathsf{d}P}{\mathsf{d}Q}=\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}$.

i) Consider:

$$\begin{aligned}\mathsf{JSD}(P\|Q)&=\int_{\mathcal{X}}\frac{\mathsf{d}P}{\mathsf{d}Q}\log\left(\frac{\frac{\mathsf{d}P}{\mathsf{d}Q}}{\frac{\mathsf{d}P}{\mathsf{d}Q}+1}\right)\mathsf{d}Q+\int_{\mathcal{X}}\log\left(\frac{2}{\frac{\mathsf{d}P}{\mathsf{d}Q}+1}\right)\mathsf{d}Q\\&=\int_{\mathcal{X}}\log\left(\frac{\mathsf{d}P}{\mathsf{d}(P+Q)/2}\right)\mathsf{d}P+\int_{\mathcal{X}}\log\left(\frac{\mathsf{d}Q}{\mathsf{d}(P+Q)/2}\right)\mathsf{d}Q\\&=D_{\mathsf{KL}}\left(P\middle\|\frac{P+Q}{2}\right)+D_{\mathsf{KL}}\left(Q\middle\|\frac{P+Q}{2}\right),\end{aligned}$$

where we have used the fact that $\mathsf{d}P/\mathsf{d}\lambda+\mathsf{d}Q/\mathsf{d}\lambda=\mathsf{d}(Q+P)/\mathsf{d}\lambda$, which follows from the definition of the Radon-Nikodym derivative and linearity of the expectation operator.

ii) We have

$$D_{\mathsf{KL}}\left(P \middle\| \frac{P+Q}{2}\right) = \int_{\mathcal{X}} \log\left(\frac{\mathsf{d}P}{\mathsf{d}P/2+\mathsf{d}Q/2}\right)\mathsf{d}P = \int_{\mathrm{supp}(P)} \log\left(\frac{\mathsf{d}P}{\mathsf{d}P/2+\mathsf{d}Q/2}\right)\mathsf{d}P$$
$$\leq \int_{\mathrm{supp}(P)} \log\left(\frac{\mathsf{d}P}{\mathsf{d}P/2}\right)\mathsf{d}P = \log(2),$$

with equality if $Q(\mathrm{supp}(P)) = 0$. Similarly, $D_{\mathsf{KL}}\left(Q \middle\| \frac{P+Q}{2}\right) \leq \log(2)$ with equality if $P(\mathrm{supp}(Q)) = 0$. So, $\mathsf{JSD}(P\|Q)$ is maximized at $2\log(2)$ if $\mathrm{supp}(P) \cap \mathrm{supp}(Q) = \emptyset$.

5) We use the shorthand notation $\frac{\mathsf{d}P}{\mathsf{d}Q} = \frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}$.

a) $f^{\star\star} = f$ by convexity of f. Thus,

$$D_f(P\|Q) = \int_{\mathcal{X}} \sup_{y\in\mathsf{dom}(f^\star)} \left(y\frac{\mathsf{d}P(x)}{\mathsf{d}Q(x)} - f^\star(y)\right)\mathsf{d}Q(x) \geq \int_{\mathcal{X}} \left(g(x)\frac{\mathsf{d}P(x)}{\mathsf{d}Q(x)} - f^\star(g(x))\right)\mathsf{d}Q(x),$$

for all measurable $g : \mathcal{X} \to \mathbb{R}$. Notice that for each x the suprimizer y may be different. Finally, for all $g : \mathcal{X} \to \mathbb{R}$, it holds that

$$D_f(P\|Q) \geq \int_{\mathcal{X}} g(x)\mathsf{d}P(x) - \int_{\mathcal{X}} f^\star(g(x))\mathsf{d}Q(x) = \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^\star(g(X))].$$

Thus,

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

b) We need to find convex conjugate of respective f functions. Let $h(x, y) = xy - f(x)$. Notice that $h(x, y)$ is concave in x as $f(x)$ is convex. So we can use the first-order optimality condition to find $f^\star(y) = \sup_x h(x, y)$.

i) $f(x) = x\log(x)$ and $h(x, y) = xy - x\log(x)$. From the first order optimality condition $\mathsf{d}h/\mathsf{d}x = 0$ it follows that $x^* = \mathrm{argmax}_{x>0}\, h(x, y) = e^{y-1}$ and

$$f^\star(y) = ye^{y-1} - (y-1)e^{y-1} = e^{y-1}.$$

Then,

$$D_f(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[e^{g(X)-1}\right] = 1 + \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[e^{g(X)}\right],$$

where the last equation follows from a change of variable of the form $\tilde{g}(x) = g(x) - 1$.

c) $f(x) = \frac{1}{2}|x-1|$ and $h(x, y) = xy - \frac{1}{2}|x-1|$. So,

$$f^\star(y) = \begin{cases} y, & \text{if } |y| \leq \frac{1}{2}, \\ \infty, & \text{if } |y| > \frac{1}{2}. \end{cases}$$

Then,

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

where the uniform norm (sup norm) $\|g\|_\infty$ is defined as $\|g\|_\infty = \sup_{y\in\mathsf{dom}(g)} |g(y)|$.

d) $f(x) = (x-1)^2$ and $h(x, y) = xy - (x-1)^2$. From the first order optimality condition $\mathsf{d}h/\mathsf{d}x = 0$ it follows that

$x^* = \operatorname{argmax}_{x>0} h(x,y) = \frac{y}{2}+1$ and

$$f^\star(y) = \frac{y^2}{4} + y.$$

So,

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

6) a) By Jensen's inequality

$$D_{\mathsf{KL}}(P\|Q) = \mathbb{E}_P\left[\log\left(\frac{\mathsf{d}P}{\mathsf{d}Q}\right)\right] \le \log\left(\mathbb{E}_P\left[\frac{\mathsf{d}P}{\mathsf{d}Q}\right]\right) = \log\left(\mathbb{E}_Q\left[\left(\frac{\mathsf{d}P}{\mathsf{d}Q}\right)^2\right]\right) = \log\left(1 + \mathbb{E}_Q\left[\left(\frac{\mathsf{d}P}{\mathsf{d}Q}\right)^2 - 1\right]\right).$$

So $D_{\mathsf{KL}}(P\|Q) \le \log\left(1+\chi^2(P\|Q)\right) \le \chi^2(P\|Q)$. The last inequality follows from the hint and the fact that $\chi^2(P\|Q) \ge 0$ for all $p, Q \in \mathcal{P}(\mathcal{X})$.

b) First notice that

$$g(p,q) = D_{\mathsf{KL}}(P\|Q) - \frac{2}{\ln(2)}\delta_{\mathsf{TV}}(P,Q)^2 = (1-p)\log\left(\frac{1-p}{1-q}\right) + p\log\left(\frac{p}{q}\right) - \frac{2}{\ln(2)}(p-q)^2.$$

Then,

$$\frac{\mathsf{d}g}{\mathsf{d}q} = \frac{p-q}{\ln(2)}\left(4 - \frac{1}{q(1-q)}\right)$$

It follows that $\frac{\mathsf{d}g}{\mathsf{d}q} \le 0$ if $p \ge q$ and $\frac{\mathsf{d}g}{\mathsf{d}q} \ge 0$ otherwise. Notice that $g(p,q) = 0$ at $p = q$. So $g(p,q) \ge 0$, which implies the desired inequality.

c) Let $g(x) = (x-1)^2 - \left(\frac{4}{3} + \frac{2}{3}x\right)h(x)$. Notice that $g(1) = 0$, $g'(1) = 0$, $g''(x) = -4h(x)/(3x)$. By convexity of h it follows that $g''(x) \le 0$ for all $x \ge 0$. By Taylor's theorem, there exists z such that $|z-1| < |x-1|$

$$g(x) = g(1) + g'(1)(x-1) + \frac{g''(z)}{2}(x-1)^2 \le 0.$$

Thus, for all $x \ge 0$

$$|x-1| \le \sqrt{\left(\frac{4}{3} + \frac{2}{3}x\right)h(x)}. \tag{1}$$

Using inequality (1), we get

$$\delta_{\mathsf{TV}}(P,Q) = \frac{1}{2}\int_{\mathcal{X}} |p(x)-q(x)|\mathsf{d}x = \frac{1}{2}\int_{\mathcal{X}}\left|\frac{p(x)}{q(x)} - 1\right| q(x)\mathsf{d}x \le \frac{1}{2}\int_{\mathcal{X}}\sqrt{\left(\frac{4}{3} + \frac{2p(x)}{3q(x)}\right)h\left(\frac{p(x)}{q(x)}\right)}q(x)\mathsf{d}x$$

Using Cauchy-Schwarz inequality, we get

$$\frac{1}{2}\int_{\mathcal{X}}\sqrt{\left(\frac{4}{3} + \frac{2p(x)}{3q(x)}\right)h\left(\frac{p(x)}{q(x)}\right)}q(x)\mathsf{d}x \le \frac{1}{2}\sqrt{\int_{\mathcal{X}}\left(\frac{4}{3} + \frac{2p(x)}{3q(x)}\right)q(x)\mathsf{d}x}\sqrt{h\left(\frac{p(x)}{q(x)}\right)q(x)\mathsf{d}x} \quad = \frac{1}{2}\sqrt{2}\sqrt{D_{\mathsf{KL}}(p\|q)},$$

where the last equality follows from the definition of KL divergence and h.