ECE 5630 - Solutions Homework Assignment 2

1) a) Using Jensen’s inequality

\[ D_f(P||Q) = E_Q f \left( \frac{dP}{d\lambda} \right) \geq f \left( E_Q \frac{dP}{d\lambda} \right) = 0, \]

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

\[ E_Q \left[ \frac{dP}{d\lambda} \right] = \int_X \frac{dP}{d\lambda} dQ = \int_X \frac{dP}{d\lambda} d\lambda = 1. \]

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 \( \frac{dP}{d\lambda} = \frac{dQ}{d\lambda} \). Thus, by taking the integral of both sides, we get

\[ \int_A \frac{dP}{d\lambda} d\lambda = \int_A \frac{dQ}{d\lambda} d\lambda = Q(A). \]

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

\[ g \left( \alpha \frac{dP_1}{d\lambda} + (1-\alpha) \frac{dP_2}{d\lambda}, \alpha \frac{dQ_1}{d\lambda} + (1-\alpha) \frac{dQ_2}{d\lambda} \right) \leq \alpha g \left( \frac{dP_1}{d\lambda}, \frac{dQ_1}{d\lambda} \right) + (1-\alpha) g \left( \frac{dP_2}{d\lambda}, \frac{dQ_2}{d\lambda} \right). \]

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

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

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

\[ \int_A \frac{dP_{X,Y}}{d\nu} d\nu = \int_A \frac{dP_{X,Y}}{d\lambda} d\lambda = \int_{A_x} \left( \int_{A_y} \frac{dP_X}{d\lambda} d\nu \right) dP_{Y|X} = \int_{A_x} \left( \int_{A_y} \frac{dP_X}{d\lambda} d\lambda \right) dP_{Y|X} = \int_A \frac{dP_X}{d\lambda} d\nu. \]

Then,

\[ D_f(P_{X,Y}||Q_{X,Y}) = \int_{X \times Y} f \left( \frac{dP_{X,Y}/d\nu}{dQ_{X,Y}/d\nu} \right) dQ_{X,Y} = \int_X f \left( \frac{dP_X/d\lambda}{dQ_X/d\lambda} \right) \int_Y dQ_{X,Y} = \int_X f \left( \frac{dP_X/d\lambda}{dQ_X/d\lambda} \right) dQ_X. \]

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

\[ D_f(P||Q) \geq D_f(P_Y||Q_Y) = D_f(\text{Bern}(P(A))||\text{Bern}(Q(A))) = (1 - Q(A)) f \left( \frac{1 - P(A)}{1 - Q(A)} \right) + Q(A) f \left( \frac{P(A)}{Q(A)} \right). \]

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

\[ \delta_{TV}(P, Q) = \frac{1}{2} \int_X |dP - dQ|. \]

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

\[
\delta_{TV}(P_1, P_3) = \frac{1}{2} \int_X |dP_1 - dP_3| \\
= \frac{1}{2} \int_X |dP_1 - dP_2 + dP_2 - dP_3| \\
\leq \frac{1}{2} \int_X |dP_1 - dP_2| + \frac{1}{2} \int_X |dP_2 - dP_3| \\
= \delta_{TV}(P_1, P_2) + \delta_{TV}(P_2, P_3).
\]

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

\[ D_{KL}(P||Q) = \int_X \log \left( \frac{dP/d\lambda}{dQ/d\lambda} \right) dP \geq \int_A \log \left( \frac{dP/d\lambda}{dQ/d\lambda} \right) dP = \infty. \]

It holds that

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

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

c) We have

\[ \delta_{TV}(P||Q) = \frac{1}{2} \int_X |dP - dQ| \leq \frac{1}{2} \left( \int_X dP + \int_X dQ \right) = 1, \]

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

d) We approximate the statistical distance between \( P \) and \( Q_\theta \) using samples from the respective distributions. Thus, \( \text{supp}(\hat{P}) \cap \text{supp}(\hat{Q}_\theta) = \emptyset \). As a result, the statistical divergence between the two (empirical) distributions is not informative, which, in turn, makes the optimization problem \( \inf_{\theta \in \Theta} \delta(\hat{P}, \hat{Q}_\theta) \) 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, \( f''(x) = \frac{1}{x(x+1)^2} > 0 \) for all \( x > 0 \). So \( f \) is strictly convex. b) We use the shorthand notation \( \frac{dP}{dQ} = \frac{dP/d\lambda}{dQ/d\lambda} \).

i) Consider:

\[
JSD(P||Q) = \int_X \frac{dP}{dQ} \log \left( \frac{dP}{dQ} + 1 \right) dQ + \int_X \log \left( \frac{dQ}{dP} + 1 \right) dQ \\
= \int_X \log \left( \frac{dP}{d(P+Q)/2} \right) dP + \int_X \log \left( \frac{dQ}{d(P+Q)/2} \right) dQ \\
= D_{KL}(P\parallel\frac{P+Q}{2}) + D_{KL}(Q\parallel\frac{P+Q}{2}).
\]

where we have used the fact that \( dP/d\lambda + dQ/d\lambda = d(Q + P)/d\lambda \), which follows from the definition of the Radon-Nikodym derivative and linearity of the expectation operator.
ii) We have
\[ D_{\text{KL}} \left( P \left\| P + \frac{Q}{2} \right\| \right) = \int_X \log \left( \frac{dP}{dP/2} \right) dP = \int_{\text{supp}(P)} \log \left( \frac{dP}{dP/2} \right) dP \leq \int_{\text{supp}(P)} \log \left( \frac{dP}{dP/2} \right) dP = \log(2) , \]
with equality if \( Q(\text{supp}(P)) = 0 \). Similarly, \( D_{\text{KL}} \left( P \left\| \frac{P + Q}{2} \right\| \right) \leq \log(2) \) with equality if \( P(\text{supp}(Q)) = 0 \). So, \( \text{JSD}(P||Q) \) is maximized at \( 2\log(2) \) if \( \text{supp}(P) \cap \text{supp}(Q) = 0 \).

5) We use the shorthand notation \( dP/dQ = \frac{dP}{d\lambda} \frac{d\lambda}{dQ} \).

a) \( f^{**} = f \) by convexity of \( f \). Thus,
\[ D_f(P||Q) = \int_X \sup_{y \in \text{dom}(f^{**})} \left( \frac{dP(x)}{dQ(x)} - f^{**}(y) \right) dQ(x) \geq \int_X \left( g(x) \frac{dP(x)}{dQ(x)} - f^{**}(g(x)) \right) dQ(x), \]
for all measurable \( g : X \to \mathbb{R} \). Notice that for each \( x \) the suprimizer \( y \) may be different. Finally, for all \( g : X \to \mathbb{R} \), it holds that
\[ D_f(P||Q) \geq \int_X g(x)dP(x) - \int_X f^{**}(g(x))dQ(x) = E_P[g(X)] - E_Q[f^{**}(g(X))]. \]
Thus,
\[ D_f(P||Q) \geq \sup_{g : X \to \mathbb{R}} \left[ E_P[g(X)] - E_Q[f^{**}(g(X))] \right]. \]

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 \( f^{**}(y) = \sup_x h(x, y) \).

i) \( f(x) = x \log(x) \) and \( h(x, y) = xy - x \log(x) \). From the first order optimality condition \( dh/dx = 0 \) it follows that \( x^* = \arg\max_{x > 0} h(x, y) = e^{y-1} \) and
\[ f^{**}(y) = ye^{y-1} - (y - 1)e^{y-1} = e^{y-1}. \]
Then,
\[ D_f(P||Q) = \sup_{g : X \to \mathbb{R}} \left[ E_P[g(X)] - E_Q[e^{g(X)}] \right] + \sup_{g : X \to \mathbb{R}} \left[ E_P[g(X)] - E_Q[e^{g(X)}] \right], \]
where the last equation follows from a change of variable of the form \( g(x) = g(x) - 1 \).

c) \( f(x) = \frac{1}{2}|x - 1| \) and \( h(x, y) = xy - \frac{1}{2}|x - 1| \). So,
\[ f^{**}(y) = \begin{cases} y, & \text{if } |y| \leq \frac{1}{2} \\ \infty, & \text{if } |y| > \frac{1}{2} \end{cases}. \]
Then,
\[ \delta_{TV}(P, Q) = \sup_{\|g\|_{\infty} \leq \frac{1}{2}} \left( E_P[g(X)] - E_Q[g(X)] \right) = \sup_{\|g\|_{\infty} \leq 1} \left( 1 - \frac{1}{2} \left( E_P[g(X)] - E_Q[g(X)] \right) \right), \]
where the uniform norm (sup norm) \( \|g\|_{\infty} \) is defined as \( \|g\|_{\infty} = \sup_{y \in \text{dom}(g)} |g(y)| \).

d) \( f(x) = (x - 1)^2 \) and \( h(x, y) = xy - (x - 1)^2 \). From the first order optimality condition \( dh/dx = 0 \) it follows that
\[ x^* = \arg\max_{x > 0} h(x, y) = \frac{y}{2} + 1 \text{ and } f^*(y) = \frac{y^2}{4} + y. \]

So,

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

6) a) By Jensen’s inequality

\[ D_{KL}(P\|Q) = \mathbb{E}_P \left[ \log \left( \frac{dP}{dQ} \right) \right] \leq \log \left( \mathbb{E}_P \left[ \frac{dP}{dQ} \right] \right) = \log \left( \mathbb{E}_Q \left[ \left( \frac{dP}{dQ} \right)^2 \right] \right) = \log \left( 1 + \mathbb{E}_Q \left[ \left( \frac{dP}{dQ} \right)^2 - 1 \right] \right). \]

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

b) First notice that

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

Then,

\[ \frac{dq}{dp} = \frac{p - q}{\ln(2)} \left( 4 - \frac{1}{q(1 - q)} \right) \]

It follows that \( \frac{dq}{dp} \leq 0 \) if \( p \geq q \) and \( \frac{dq}{dp} \geq 0 \) otherwise. Notice that \( g(p, q) = 0 \) at \( p = q \). So \( g(p, q) \geq 0 \), which implies the desired inequality.

c) Let \( g(x) = (x - 1)^2 - \left( \frac{4}{3} + \frac{2}{9} x \right) h(x) \). Notice that \( g(1) = 0, g'(1) = 0, g''(x) = -4h(x)/(3x) \). By convexity of \( h \) it follows that \( g''(x) \leq 0 \) for all \( x \geq 0 \). By Taylor’s theorem, there exists \( z \) such that \( |z - 1| < |x - 1| \)

\[ g(x) = g(1) + g'(1)(x - 1) + \frac{g''(z)}{2}(x - 1)^2 \leq 0. \]

Thus, for all \( x \geq 0 \)

\[ |x - 1| \leq \sqrt{\left( \frac{4}{3} + \frac{2}{9} x \right) h(x)}. \quad (1) \]

Using inequality (1), we get

\[ \delta_{TV}(P, Q) = \frac{1}{2} \int_X |p(x) - q(x)| dx = \frac{1}{2} \int_X \left| \frac{p(x)}{q(x)} - 1 \right| q(x) dx \leq \frac{1}{2} \int_X \sqrt{\left( \frac{4}{3} + \frac{2p(x)}{3q(x)} \right) h \left( \frac{p(x)}{q(x)} \right) q(x) dx}
\]

Using Cauchy-Schwarz inequality, we get

\[ \frac{1}{2} \int_X \sqrt{\left( \frac{4}{3} + \frac{2p(x)}{3q(x)} \right) h \left( \frac{p(x)}{q(x)} \right) q(x) dx} \leq \frac{1}{2} \int_X \left( \frac{4}{3} + \frac{2p(x)}{3q(x)} \right) q(x) dx \sqrt{h \left( \frac{p(x)}{q(x)} \right) q(x) dx} = \frac{1}{2} \sqrt{2} D_{KL}(p\|q), \]

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