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

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

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

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

Solution. By the definition of f-Divergence we have

$$D_f\left(P \| Q\right) = \mathbb{E}_Q\left[f\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right)\right] \geq f\left(\mathbb{E}_Q\left[\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right]\right) \geq f(1) = 0$$

where the first inequality follows from Jensen's inequality and the second from convexity of f. To prove equality, first assume that $P = Q$. Then,

$$D_f\left(P \| Q\right) = \mathbb{E}_Q\left[f(1)\right] = 0.$$

Now assume $D_f\left(P \| Q\right) = 0$. Then $f\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right) = 0 \implies \frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda} = 1$ since f is strongly convex at 1. It follows that $P = Q$. ■

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

Solution. For any function $f : \mathbb{R}^n \to \mathbb{R}$ we can define the perspective of f as the function $g : \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}$ such that

$$g(x, y) = yf\left(\frac{x}{y}\right), \qquad \text{dom } g = \left\{(x, y) \mid \frac{x}{y} \in \text{dom } f, y > 0\right\}$$

For a convex function f, the perspective of f is also convex. That is,

$$\begin{aligned} g(\alpha(x_1, y_1) + (1-\alpha)(x_2, y_2)) &\leq \alpha g(x_1, y_1) + (1-\alpha) g(x_2, y_2) \\ &= \alpha y_1 f\left(\frac{x_1}{y_1}\right) + (1-\alpha) y_2 f\left(\frac{x_2}{y_2}\right) \end{aligned}$$

for all $\alpha \in [0, 1]$ and each (x_i, y_i) in the domain of g.
Define $\tilde{D}_f\left(P, Q\right) = D_f\left(P \| Q\right)$.

Then for $(P_1, Q_1), (P_2, Q_2) \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ such that $P_1, P_2, Q_1, Q_2 \ll \lambda$

$$\begin{aligned}
\tilde{D}_f\left(\alpha(P_1, Q_1) + (1-\alpha)(P_2, Q_2)\right) &= \int_{\mathcal{X}} f\left(\frac{\alpha\frac{\mathsf{d}P_1}{\mathsf{d}\lambda}(x) + (1-\alpha)\frac{\mathsf{d}P_2}{\mathsf{d}\lambda}(x)}{\alpha\frac{\mathsf{d}Q_1}{\mathsf{d}\lambda}(x) + (1-\alpha)\frac{\mathsf{d}Q_2}{\mathsf{d}\lambda}(x)}\right)\mathsf{d}\lambda \\
&\leq \int_{\mathcal{X}} \alpha f\left(\frac{\frac{\mathsf{d}P_1}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q_1}{\mathsf{d}\lambda}(x)}\right)\mathsf{d}\lambda + (1-\alpha) f\left(\frac{\frac{\mathsf{d}P_2}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q_2}{\mathsf{d}\lambda}(x)}\right)\mathsf{d}\lambda \\
&= \alpha\tilde{D}_f(P_1, Q_1) + (1-\alpha)\tilde{D}_f(P_2, Q_2)
\end{aligned}$$

Thus the mapping $(P, Q) \mapsto D_f\left(P \| Q\right)$ is convex. ■

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

$$D_f\left(P_Y \| Q_Y\right) \leq D_f\left(P_{Y|X} \| Q_{Y|X} | P_X\right)$$

Solution.

$$\begin{aligned}
D_f\left(P_Y \| Q_Y\right) &= D_f\left(\mathbb{E}_{P_X}\left[P_{Y|X}(\cdot|X)\right] \| \mathbb{E}_{P_X}\left[Q_{Y|X}(\cdot|X)\right]\right) \\
&\leq \mathbb{E}_{P_X}\left[D_f\left(P_{Y|X}(\cdot|x) \| Q_{Y|X}(\cdot|x)\right)\right] && \text{(Jensen's Inequality)} \\
&= D_f\left(P_{Y|X} \| Q_{Y|X} | P_X\right)
\end{aligned}$$

■

(d) Joint vs. Marginal: For $P_X, Q_X \in \mathcal{P}(\mathcal{X})$ and a transition kernel $P_{Y|X}$, define $P_{XY} := P_X P_{Y|X}$ and $Q_{XY} := Q_X P_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$. Show that

$$D_f\left(P_X \| Q_X\right) = D_f\left(P_{XY} \| Q_{XY}\right)$$

Solution.

$$D_f\left(P_{XY}\|Q_{XY}\right) = \mathbb{E}_{Q_{XY}}\left[f\left(\frac{\mathsf{d}P_{XY}}{\mathsf{d}Q_{XY}}\right)\right] = \mathbb{E}_{Q_{XY}}\left[f\left(\frac{\mathsf{d}P_X P_{Y|X}}{\mathsf{d}Q_X P_{Y|X}}\right)\right]$$

$$= \int_{\mathcal{X}}\int_{\mathcal{Y}} f\left(\frac{\mathsf{d}P_X P_{Y|X}}{\mathsf{d}Q_X P_{Y|X}}\right)\mathsf{d}Q_{X,Y}(x,y) = \int_{\mathcal{X}} f\left(\frac{\mathsf{d}P_X}{\mathsf{d}Q_X}\right)\mathsf{d}Q_X(x)$$

$$= \mathbb{E}_{Q_X}\left[f\left(\frac{\mathsf{d}P_X}{\mathsf{d}Q_X}\right)\right] = D_f\left(P_X\|Q_X\right)$$

■

2. **Example of Data Processing Inequality.**
Let $(\mathcal{X},\mathcal{F})$ be a measurable space. Use the Data Processing Inequality to show that for any two measurable P, Q on $(\mathcal{X},\mathcal{F})$ and any $A \in \mathcal{F}$:

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

Solution. Let $\mathbb{1}^A(\cdot|\cdot)$ denote the Dirac measure, where

$$\mathbb{1}^A(1|x) = \begin{cases} 1, & x\in A \\ 0, & \text{otherwise}\end{cases}$$

By taking the expectation with respect to P_X and Q_X, we get measures P_Y, Q_Y that correspond to the Bernoulli variable $Ber(P(A))$ and $Ber(Q(A))$ respectively. Note that this would be a discrete random variable with sample space $\{0,1\}$ that's absolutely continuous with respect to the counting measure. Computation of the divergence yields

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

Using the data processing inequality, we get

$$\begin{aligned} D_f\left((\|P\right), Q) &\geq D_f\left(P_Y\|Q_Y\right) && (1)\\ &= (1-Q(A))f\left(\frac{1-P(A)}{1-Q(A)}\right) + Q(A)f\left(\frac{P(A)}{Q(A)}\right) \quad \forall A\in\mathcal{F}. && (2)\end{aligned}$$

Since the inequality holds $\forall A \in \mathcal{X}$ we arrive at the conclusion that $D_f\left((\|P\right), Q)$ is an upper bound to

$$\left\{(1-Q(A))f\left(\frac{1-P(A)}{1-Q(A)}\right) + Q(A)f\left(\frac{P(A)}{Q(A)}\right)|A\in\mathcal{F}\right\},$$

and by definition of sup, we get

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

■

3. **f-Divergences, Metrics, and Mismatched Supports.**
For the KL divergence $D_{\mathsf{KL}}\left(\cdot\|\cdot\right)$ and $\chi^2\left(\cdot\|\cdot\right)$ as shown in class, show that:

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

Solution.

i. Identity:

$$\delta_{\mathsf{TV}}\left(P,P\right) = \mathbb{E}_Q\left[\frac{1}{2}\left|\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}P/\mathsf{d}\lambda}-1\right|\right] = 0$$

ii. Symmetry:

$$\begin{aligned}\delta_{\mathsf{TV}}\left(P,Q\right) = \mathbb{E}_Q\left[\frac{1}{2}\left|\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}-1\right|\right] &= \int_{\mathcal{X}}\left|\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)-\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda \\ &= \int_{\mathcal{X}}\left|\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)-\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda \\ &= \mathbb{E}_P\left[\frac{1}{2}\left|\frac{\mathsf{d}Q/\mathsf{d}\lambda}{\mathsf{d}P/\mathsf{d}\lambda}-1\right|\right] \\ &= \delta_{\mathsf{TV}}\left(Q,P\right)\end{aligned}$$

iii. Triangle Inequality: For $P,Q,R\in\mathcal{P}(\mathcal{X})$,

$$\begin{aligned}\delta_{\mathsf{TV}}\left(P,Q\right)+\delta_{\mathsf{TV}}\left(Q,R\right) &= \int_{\mathcal{X}}\left|\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)-\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda+\int_{\mathcal{X}}\left|\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)-\frac{\mathsf{d}R}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda \\ &= \int_{\mathcal{X}}\left|\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)-\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)\right|+\left|\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)-\frac{\mathsf{d}R}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda \\ &\geq \int_{\mathcal{X}}\left|\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)-\frac{\mathsf{d}R}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda \\ &= \delta_{\mathsf{TV}}\left(P,R\right)\end{aligned}$$

■

(b) $D_{\mathsf{KL}}\left(P\|Q\right) = \chi^2\left(P\|Q\right) = \infty$ whenever $P \not\ll Q$.

Solution.

$$\chi^2\left(P\|Q\right) = \int_{\mathcal{X}} \left(\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right)^2 - 1\right)\mathsf{d}\lambda$$

which will blow up at values where Q attains 0 and P does not which will happen since $P \not\ll \mathbb{Q}$.

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

will blow up if $P \not\ll Q$ in a similar fashion since the denominator in the log will be zero when P is not zero which prevents our convention of $0f(0/0) = 0$ yielding an infinite divergence. ■

(c) $\delta_{\mathsf{TV}}\left(P, Q\right)$ attains its maximal value of 1 when $\text{supp}(P) \cap \text{supp}(Q) = \phi$.

Solution.

$$\int_{\mathcal{X}} \frac{1}{2}\left|\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x) - \frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda \le \int_{\mathcal{X}} \frac{1}{2}\left|\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)\right| + \left|\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda \le \int_{\mathcal{X}} \frac{1}{2}\left|\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda + \int_{\mathcal{X}}\left|\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)\right|\mathsf{d}\lambda$$

$$\le \frac{1}{2} + \frac{1}{2} = 1$$

with equality if and only if $\text{supp}(P) \cap \text{supp}(Q) = \phi$. ■

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

Solution.
When the supports are disjoint we have a constant distance (which is at its max) between any distribution Q_θ and our data distribution P. This means the model will have the same error throughout training preventing the model from learning what works and what does not in its generations. ■

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

Solution. We compute the second derivative to be

$$f'(x) = \log\left(\frac{2x}{x+1}\right)$$
$$f''(x) = \frac{1}{x^2 + x}$$

which reveals $f''(x) > 0$ for $x > 0$ giving that f is convex. Evaluating $f(1)$ gives

$$f(1) = 1 \cdot \log\left(\frac{2}{2}\right) + \log\left(\frac{2}{2}\right) = 0 + 0 = 0.$$

as desired. Finally, to see that f is strictly convex around 1 observe that the second derivative is positive at all points in our domain. This implies that f is strictly convex. The minimum of f is found to be $x = 1$ after analyzing the first derivative. Thus f is strictly convex around 1. ∎

(b) Let $\mathsf{JSD}\left(P\|Q\right)$ be the f-divergence induced by the above f. Prove that

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

Solution. We iron both expressions out and realize they are the same.

$$\begin{aligned}
\mathsf{JSD}\left(P\|Q\right) &= \mathbb{E}_Q\left[\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\log\left(\frac{2\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}}{\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}+1}\right) + \log\left(\frac{2}{\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}+1}\right)\right] \\
&= \int_{\mathcal{X}}\left[\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\log\left(\frac{2\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}}{\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}+1}\right) + \log\left(\frac{2}{\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}+1}\right)\right]\mathsf{d}Q \\
&= \int_{\mathcal{X}}\frac{\mathsf{d}P}{\mathsf{d}\lambda}\log\left(\frac{2\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}P/\mathsf{d}\lambda+\mathsf{d}Q/\mathsf{d}\lambda}\right)\mathsf{d}\lambda + \int_{\mathcal{X}}\frac{\mathsf{d}Q}{\mathsf{d}\lambda}\log\left(\frac{2\mathsf{d}Q/\mathsf{d}\lambda}{\mathsf{d}P/\mathsf{d}\lambda+\mathsf{d}Q/\mathsf{d}\lambda}\right)\mathsf{d}\lambda
\end{aligned}$$

$$
\begin{aligned}
D_{\mathsf{KL}}\left(P\Big\|\frac{P+Q}{2}\right)+D_{\mathsf{KL}}\left(Q\Big\|\frac{P+Q}{2}\right) &= \mathbb{E}_{\frac{P+Q}{2}}\left[\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\log\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\right)\right] \\
&\quad+\mathbb{E}_{\frac{P+Q}{2}}\left[\frac{\mathsf{d}Q/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\log\left(\frac{\mathsf{d}Q/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\right)\right] \\
&= \int_{\mathcal{X}}\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\log\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\right)\mathsf{d}\left(\frac{P+Q}{2}\right) \\
&\quad+\int_{\mathcal{X}}\frac{\mathsf{d}Q/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\log\left(\frac{\mathsf{d}Q/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\right)\mathsf{d}\left(\frac{P+Q}{2}\right) \\
&= \int_{\mathcal{X}}\frac{\mathsf{d}P}{\mathsf{d}\lambda}\log\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\right)\mathsf{d}\lambda+\int_{\mathcal{X}}\frac{\mathsf{d}Q}{\mathsf{d}\lambda}\log\left(\frac{\mathsf{d}Q/\mathsf{d}\lambda}{\mathsf{d}\left(\frac{P+Q}{2}\right)/\mathsf{d}\lambda}\right)\mathsf{d}\lambda \\
&= \int_{\mathcal{X}}\frac{\mathsf{d}P}{\mathsf{d}\lambda}\log\left(\frac{2\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}P/\mathsf{d}\lambda+\mathsf{d}Q/\mathsf{d}\lambda}\right)\mathsf{d}\lambda+\int_{\mathcal{X}}\frac{\mathsf{d}Q}{\mathsf{d}\lambda}\log\left(\frac{\mathsf{d}Q/\mathsf{d}\lambda}{\mathsf{d}P/\mathsf{d}\lambda+\mathsf{d}Q/\mathsf{d}\lambda}\right)\mathsf{d}\lambda
\end{aligned}
$$

Thus the expressions are equivalent. ■

ii. $\mathsf{JSD}\left(P\|Q\right)$ is maximized at $2\log 2$.

Solution. We note that

$$
\begin{aligned}
D_{\mathsf{KL}}\left(P\Big\|\frac{P+Q}{2}\right) = \int_{\mathcal{X}}\frac{\mathsf{d}P}{\mathsf{d}\lambda}\log\left(\frac{2\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}P/\mathsf{d}\lambda+\mathsf{d}Q/\mathsf{d}\lambda}\right)\mathsf{d}\lambda &\le \int_{\mathcal{X}}\frac{\mathsf{d}P}{\mathsf{d}\lambda}\log\left(\frac{2\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}P/\mathsf{d}\lambda+0}\right)\mathsf{d}\lambda \\
&= \log(2)\int_{\mathcal{X}}\frac{\mathsf{d}P}{\mathsf{d}\lambda}\mathsf{d}\lambda = \log(2).
\end{aligned}
$$

Note that by construction of the proof that equality only holds when

$$\mathrm{supp}(P)\cap\mathrm{supp}(Q)=\phi.$$

We similarly conclude that

$$D_{\mathsf{KL}}\left(Q\Big\|\frac{P+Q}{2}\right)\le\log(2).$$

Therefore

$$\mathsf{JSD}\left(P\|Q\right) = D_{\mathsf{KL}}\left(P\|\frac{P+Q}{2}\right) + D_{\mathsf{KL}}\left(Q\|\frac{P+Q}{2}\right) \leq \log(2) + \log(2) = 2\log(2)$$

■

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

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

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

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

Solution. For $P, Q \in \mathcal{P}(\mathcal{X})$, $I^* = \{y \in R \mid yx - f(x) < \infty\}$,

$$D_f\left(P\|Q\right) = \int\limits_{x\in\mathcal{X}} f\left(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)}\right)\mathsf{d}Q(x) = \int\limits_{x\in\mathcal{X}} \sup_{y\in I^*}\left(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)}y - f^*(y)\right)\mathsf{d}Q(x).$$

Pick any $g : \mathcal{X} \to I^*$ and set $y = g(x)$ to get a lower bound:

$$\begin{aligned}
\int\limits_{x\in\mathcal{X}} \sup_{y\in I^*}\left(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)}y - f^*(y)\right)\mathsf{d}Q(x) &\geq \int\limits_{x\in\mathcal{X}} \left(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)}g(x) - f^*(g(x))\right)\mathsf{d}Q(x) \\
&= \int\limits_{x\in\mathcal{X}} \frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)}g(x)\mathsf{d}Q(x) - \int\limits_{x\in\mathcal{X}} f^*(g(x))\mathsf{d}Q(x) \\
&= \int\limits_{x\in\mathcal{X}} \frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)g(x)\mathsf{d}\lambda - \int\limits_{x\in\mathcal{X}} f^*(g(x))\mathsf{d}Q(x) \\
&= \mathbb{E}_P\left[g(x)\right] - \mathbb{E}_Q\left[f^*\left(g(x)\right)\right]
\end{aligned}$$

Supremizing over all measurable $g : \mathcal{X} \to I^*$, we get

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

as desired. ■

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

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

Solution. For KL divergence we have

$$\begin{aligned} f(x) &= x\log(x), \\ f^*(y) &= \sup_{x\in(0,\infty)} xy - x\log(x). \end{aligned}$$

Letting $s(x) = xy - x\log(x)$, we compute the location of the maxima of $s(x)$ by

$$\begin{aligned} \left.\frac{\mathsf{d}s(x)}{\mathsf{d}x}\right|_{x>0} &= y - 1 - \log(x) \\ &\to x = e^{y-1}. \end{aligned}$$

Thus

$$\begin{aligned} f^*(y) &= ye^{y-1} - e^{y-1}\log(e^{y-1}) \\ &= ye^{y-1} - (y-1)e^{y-1} \\ &= e^{y-1}. \end{aligned}$$

It follows that

$$D_{\mathsf{KL}}\left(P\|Q\right) = \sup_{g:\mathcal{X}\to\mathbb{R}} \left(\mathbb{E}_P[g(x)] - \mathbb{E}_Q[f^*(g(x))]\right) \quad = \sup_{g:\mathcal{X}\to\mathbb{R}} \left(\mathbb{E}_P[g(x)] - \mathbb{E}_Q[e^{g(x)-1}]\right).$$

Letting $h(x) = g(x) - 1$ we get the desired expression

$$\begin{aligned} D_{\mathsf{KL}}\left(P\|Q\right) &= \sup_{h:\mathcal{X}\to\mathbb{R}} \left(\mathbb{E}_P[h(x)+1] - \mathbb{E}_Q[e^{h(x)}]\right) \\ &= 1 + \sup_{h:\mathcal{X}\to\mathbb{R}} \left(\mathbb{E}_P[h(x)] - \mathbb{E}_Q[e^{h(x)}]\right). \end{aligned}$$

■

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

Solution. For TV distance we have that

$$\begin{aligned} f(x) &= \frac{1}{2}|x-1|, \\ f^*(y) &= \sup_{x\in(0,\infty)} xy - \frac{1}{2}|x-1|. \end{aligned}$$

Letting $s(x) = xy - \frac{1}{2}|x-1|$, we compute the location of the maxima of $s(x)$ by

$$\left.\frac{\mathsf{d}s(x)}{\mathsf{d}x}\right|_{x>0} = \begin{cases} y + \frac{1}{2}, & x \le 1 \\ y - \frac{1}{2}, & x > 1 \end{cases}$$

Note that for $f(y)$ to be bounded, we need the derivative for when $x > 1$ to be negative as otherwise the function will explode to infinity. This means that we need $|y| \leq \frac{1}{2}$. And for when $x \leq 1$, the max is attained at $x = 1$ and it is bounded by y. So supremizing over x we have

$$f^*(y) = \begin{cases} y, & |y| \leq \frac{1}{2} \\ \infty, & \text{otherwise} \end{cases}$$

Thus, we have

$$\begin{aligned} \delta_{\mathsf{TV}}(P, Q) &= \sup_{\|g\|_\infty \leq \frac{1}{2}} \mathbb{E}_P[g(x)] + \mathbb{E}_Q[f^*(g(x))] \\ &= \sup_{\|g\|_\infty \leq 1} \frac{1}{2}(\mathbb{E}_P[g(x)] + \mathbb{E}_Q[g(x)]) \end{aligned}$$

■

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

Solution. For Chi-squared distance we have that

$$\begin{aligned} f(x) &= (x-1)^2, \\ f^*(y) &= \sup_{x\in(0,\infty)} xy - (x-1)^2. \end{aligned}$$

Letting $s(x) = xy - (x-1)^2$ we compute the location of the maxima by

$$\begin{aligned} \left.\frac{\mathsf{d}s(x)}{\mathsf{d}x}\right|_{x>0} &= y - 2(x-1) \\ &\to x = \frac{y}{2} + 1 \end{aligned}$$

Thus

$$\begin{aligned} f^*(y) &= \frac{y^2}{2} + y - \left(\frac{y}{2} + 1 - 1\right)^2 \\ &= \frac{y^2}{4} + y \end{aligned}$$

and

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

■

6. **Inequalities Between f-Divergences.**
Prove the following:

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

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

Solution. Note that $\chi^2\left(P \| Q\right)$ is non-negative since it is an f-divergence. So we have $\chi^2\left(P \| Q\right) > -1$ and thus $\log(1 + \chi^2\left(P \| Q\right)) \leq \chi^2\left(P \| Q\right)$. Now we expand the log term:

$$\begin{aligned}
\log(1 + \chi^2\left(P \| Q\right)) &= \log(1 + \int \frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)}^2 - 1)\mathsf{d}\lambda) \\
&= \log(1 + \int \frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)})^2\mathsf{d}\lambda - 1) \\
&= \log(\int \frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)})\mathsf{d}\lambda)
\end{aligned}$$

Now we expand and rewrite the KL divergence using the alternative form $D_{-log(x)}(Q \| P)$:

$$\int \frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)(-\log(\frac{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)})\mathsf{d}\lambda) = \int \frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)(\log(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)})\mathsf{d}\lambda)$$

Finally, we use Jansen's equality. Since log is concave, we have $\mathbb{E}[\log(f(x))] \leq \mathbb{E}[\int f(x)]$. So we have

$$\begin{aligned}
D_{\mathsf{KL}}\left(P \| Q\right) &= \int \frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)(\log(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)})\mathsf{d}\lambda) \\
&\leq \log(\int \frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)(\frac{\frac{\mathsf{d}P}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q}{\mathsf{d}\lambda}(x)})\mathsf{d}\lambda) \\
&= \log(1 + \chi^2\left(P \| Q\right))
\end{aligned}$$

as desired. ■

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

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

Solution. We expand both sides under the assumption that P and Q are Bernoulli measures. We get

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

Taking the derivative, we get that the partials

$$\frac{\delta f}{\delta p} = \frac{-\ln\left(\frac{p-1}{q-1}\right) + \ln\left(\frac{p}{q}\right) - 4p + 4q}{\ln(2)}$$
$$\frac{f}{\delta q} = \frac{(1-2q)^2(p-q)}{(q-1)q\ln(2)}$$

Note that the critical point is when $p = q$, which gives us zero for $f(p,p) = f(q,q) = 0$. For the inequality to hold, we need the critical point to be that of a local maxima. However, by taking the determinant of the hessian we would get that it's 0 so the second derivative test is inconclusive. So instead, we evaluate f by perturbing p, and have $p = q + \epsilon$, and $p = q - \epsilon$ and verify that $f(p,q) < f(q+\epsilon, q)$ and $f(p,q) < f(q-\epsilon, q)$. ■

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

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

Solution. First note that the derivative of $(4+2x)h(x)-3(x+1)^2$ is $-8x+4(x+1)\log(x)+8$, which is 0 at 1, and achieve a minimum value of 0.
observe that for $h(x) = x\log(x)+x-1$, we have $\mathbb{E}_Q[h(\frac{\frac{dP}{d\lambda}(x)}{\frac{dQ}{d\lambda}(x)})] = \mathbb{E}_Q[\frac{\frac{dP}{d\lambda}(x)}{\frac{dQ}{d\lambda}(x)}\log(\frac{\frac{dP}{d\lambda}(x)}{\frac{dQ}{d\lambda}(x)}) + \frac{\frac{dP}{d\lambda}(x)}{\frac{dQ}{d\lambda}(x)} - 1] = \mathbb{E}_Q[\frac{\frac{dP}{d\lambda}(x)}{\frac{dQ}{d\lambda}(x)}\log(\frac{\frac{dP}{d\lambda}(x)}{\frac{dQ}{d\lambda}(x)})] = D_{\mathsf{KL}}\left(P\|Q\right)$. Denote $X = \frac{\frac{dP}{d\lambda}(x)}{\frac{dQ}{d\lambda}(x)}$ Using the inequality in the hint, we have

$$\begin{aligned}
D_{KL}(F) &\geq \frac{3}{2}\mathbb{E}_Q[\frac{(F-1)^2}{2+F}] \\
&= \frac{3}{2}\mathbb{E}_Q[\frac{(F-1)^2}{2+F}]\frac{1}{3}\mathbb{E}_Q[2+F] \quad \text{since } \frac{1}{3}\mathbb{E}_Q[2+F] = 1 \\
&\geq \frac{1}{2}\mathbb{E}_Q[\sqrt{\frac{(x-1)^2}{(2+x)(2+x)}}] \quad \text{using Cauchy-Shcwarz} \\
&= \frac{1}{2}\mathbb{E}_Q[F-1]^2 \\
&= \frac{1}{2}\delta_{\mathsf{TV}}\left(P,Q\right)
\end{aligned}$$

■