Lecture-19: Divergence

1 Entropy

Definition 1.1. Consider a discrete random vector $X : \Omega \to \mathcal{X}^n$ with joint probability mass function $P_X \in \mathcal{M}(\mathcal{X}^n)$ defined as $P_X(x) \triangleq P\{X = x\}$ for each $x \in \mathcal{X}$. The *entropy* of X is defined as $H(X) \triangleq -\mathbb{E}\log_2 P_X(X)$. Since entropy only depends on the distribution of a random vector, we write $H(P_X)$ in place of $H(X)$.

Remark 1. Entropy measures the intrinsic randomness or uncertainty of a random variable. In the simple setting where X takes values uniformly over a finite set $\mathcal{X}$, the entropy is simply given by log-cardinality, i.e. $H(X) = \log_2 |\mathcal{X}|$. In general, the more spread out or concentrated a probability mass function is, the higher or lower is its entropy,

Definition 1.2. Let $X : \Omega \to \mathcal{X}^n$ be a discrete random vector and $Y : \Omega \to \mathbb{R}$ arbitrary random variable. Let $P_{X|Y} \in \mathcal{M}(\mathcal{X})$ denote the conditional distribution of X given Y. The *conditional entropy* of X given Y is defined as $H(X \mid Y) \triangleq \mathbb{E}H(P_{X|Y})$.

Definition 1.3. Similar to entropy, conditional entropy measures the remaining randomness of a random variable when another is revealed. As such, $H(X \mid Y) = H(X)$ whenever Y is independent of X. But when Y depends on X, observing Y does lower the entropy of X.

2 KL divergence

Definition 2.1. Let $(\mathcal{X}, \mathcal{F})$ be a measurable space, we define the set of probability measures on $\mathcal{X}$ as

$$\mathcal{M}(\mathcal{X}) \triangleq \left\{ P \in [0,1]^{\mathcal{F}} : P \text{ satisfies probability axioms } \right\}.$$

Let $X : \Omega \to \mathcal{X}$ and $P, Q \in \mathcal{M}(\mathcal{X})$. We say P is *absolutely continuous* w.r.t. Q and denoted by $P \ll Q$ if $Q(E) = 0$ implies $P(E) = 0$ for all measurable $E \in \sigma(X)$. If $P \ll Q$, then *Radon-Nikodym theorem* show that there exists a function $g : \mathcal{X} \to \mathbb{R}_+$ alled a *relative density* or a *Radon-Nikodym derivative* of P w.r.t. Q and denoted by $\frac{dP}{dQ} \triangleq g$, such that $P(E) = \int_E g dQ$ for any measurable set $E \in \sigma(X)$.

Remark 2. Note that $\frac{dP}{dQ}$ may not be unique. In the simple cases, $\frac{dP}{dQ}$ is the likelihood ratio.
(a) For discrete distributions, we can just take $\frac{dP}{dQ}(x)$ to be the ratio of probability mass functions.
(b) For continuous distributions, we can take $\frac{dP}{dQ}(x)$ to be the ratio of probability density functions.

Definition 2.2 (Kullback-Leibler (KL) divergence). Adopting the convention $0 \ln 0 = 0$, we can define the *KL divergence* or *relative entropy* between any $P, Q \in \mathcal{M}(\mathcal{X})$ with Q being the reference measure, as

$$D(P \| Q) \triangleq \begin{cases} \mathbb{E}_P \ln \frac{dP}{dQ} = \mathbb{E}_Q \left[\frac{dP}{dQ} \ln \frac{dP}{dQ} \right], & P \ll Q, \\ +\infty, & P \not\ll Q. \end{cases}$$

2.1 Conditional divergence

Definition 2.3 (Conditional divergence). Consider random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ defined on the common probability space $(\Omega, \mathcal{F}, P)$. measurable spaces $(\mathcal{X}, \sigma(X))$ and $(\mathcal{Y}, \sigma(Y))$ and a pair of Markov kernels $P_{Y|X} : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ and $Q_{Y|X} : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$, and also a probability measure $P_X \in \mathcal{M}(\mathcal{X})$. Since $(\mathcal{Y}, \sigma(Y))$ is standard Borel measurable space, i.e. $\sigma(Y) \triangleq \mathcal{B}(\mathcal{Y})$, we define

$$D(P_{Y|X} \| Q_{Y|X} \mid P_X) \triangleq \mathbb{E}_{x \sim P_X}[D(P_{Y|X=x} \| Q_{Y|X=x})].$$

Remark 3. We observe that as usual in Lebesgue integration it is possible that a conditional divergence is finite even though $D(P_{Y|X=x} \| Q_{Y|X=x}) = \infty$ for some x in a P_X-negligible set.

Theorem 2.4 (Chain rule). *For any pair of measures $P_{X,Y}$ and $Q_{X,Y}$ we have*

$$D(P_{X,Y} \| Q_{X,Y}) = D(P_{Y|X} \| Q_{Y|X} \mid P_X) + D(P_X \| Q_X),$$

regardless of the versions of conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ one chooses.

Proof. Recall that $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X Q_{Y|X}$. If $P_X \not\ll Q_X$ then $P_{X,Y} \not\ll Q_{X,Y}$ and both sides of chain rule equation are infinity. Thus, we can assume $P_X \ll Q_X$ without any loss of generality, and define relative density $\lambda_P \triangleq \frac{dP_X}{dQ_X} \in \mathbb{R}_+^{\mathcal{X}}$. We next define a kernel $R_{Y|X} : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ that is a mixture of kernels $R_{Y|X} \triangleq \frac{1}{2} P_{Y|X} + \frac{1}{2} Q_{Y|X}$, such that $P_{Y|X} \ll R_{Y|X}$ and $Q_{Y|X} \ll R_{Y|X}$. We write the corresponding relative densities for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, as

$$f_P(y \mid x) \triangleq \frac{dP_{Y|X=x}}{dR_{Y|X=x}}(y), \qquad f_Q(y \mid x) \triangleq \frac{dQ_{Y|X=x}}{dR_{Y|X=x}}(y).$$

Defining $R_{X,Y} \triangleq Q_X R_{Y|X}$, we observe that $P_{X,Y} \ll R_{X,Y}$ and $Q_{X,Y} \ll R_{X,Y}$, and we can write down the corresponding relative densities or all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, as

$$\frac{dP_{X,Y}}{dR_{X,Y}}(x,y) = \lambda_P(x) f_P(y \mid x), \qquad \frac{dQ_{X,Y}}{dR_{X,Y}}(x,y) = f_Q(y \mid x).$$

From the linearity of expectation, we can write the following equality

$$D(P_{X,Y} \| Q_{X,Y}) = \mathbb{E}_{P_{X,Y}} \ln \frac{dP_{X,Y}}{dQ_{X,Y}} = \mathbb{E}_{P_{X,Y}} \ln \frac{\lambda_P(X) f_P(Y \mid X)}{f_Q(Y \mid X)} = \mathbb{E}_{P_{X,Y}} \ln \lambda_P(X) + \mathbb{E}_{P_{X,Y}} \ln \frac{f_P(Y \mid X)}{f_Q(Y \mid X)}.$$

The result follows from the observation that $E_{P_{X,Y}} \ln \lambda_P(X) = E_{P_X} \ln \lambda_P(X) = D(P_X \| Q_X)$, and the definition of conditional divergence which implies that

$$\mathbb{E}_{P_{X,Y}} \ln \frac{f_P(Y \mid X)}{f_Q(Y \mid X)} = \mathbb{E}_{x \sim P_X} \mathbb{E}_{P_{Y|X=x}} \ln \frac{dP_{Y|X=x}}{dQ_{Y|X=x}} = D(P_{Y|X} \| Q_{Y|X} \mid P_X).$$

□

2.2 Data processing inequality

Theorem 2.5 (Data processing inequality). *Consider two input distributions $P_X, Q_X \in \mathcal{M}(\mathcal{X})$ and a common Markov kernel $P_{Y|X} : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ such that the joint distributions are $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X P_{Y|X}$, and the corresponding output marginal distributions $P_Y \triangleq \int_{\mathcal{X}} dP_X(x) P_{Y|X=x}$ and $Q_Y \triangleq \int_{\mathcal{X}} dQ_X(x) P_{Y|X=x}$. Then $D(P_Y \| Q_Y) \leqslant D(P_X \| Q_X)$.*

Proof. The result follows from the chain rule of KL divergence. That is,

$$D(P_{X,Y} \| Q_{X,Y}) = D(P_{X|Y} \| Q_{X|Y} \mid P_Y) + D(P_Y \| Q_Y) = D(P_{Y|X} \| Q_{Y|X} \mid P_X) + D(P_X \| Q_X).$$

Since $Q_{Y|X} = P_{Y|X}$, and KL divergence is always positive, we get the result. □

3 f-divergence

Definition 3.1. Let $f : (0, \infty) \to \mathbb{R}_+$ be a convex function with $f(1) = 0$ and define its value at origin as $f(0) \triangleq \lim_{x \downarrow 0} f(x)$ and derivative at infinity as $f'(\infty) \triangleq \lim_{x \downarrow 0} x f\left(\frac{1}{x}\right)$. We further define

$$0 f\left(\frac{0}{0}\right) = 0, \qquad 0 f\left(\frac{a}{0}\right) = \lim_{x \downarrow 0} x f\left(\frac{a}{x}\right) = a f'(\infty) \text{ for } a > 0.$$

Definition 3.2 (f-divergence). Let $P,Q \in \mathcal{M}(\mathcal{X})$ for a measurable space $(\mathcal{X},\mathcal{F})$ and f be as defined in Definition 3.1. If $P \ll Q$ then the f-divergence is defined as

$$D_f(P\|Q) \triangleq \mathbb{E}_Q f\Big(\frac{dP}{dQ}\Big).$$

Suppose for some common dominating measure μ such that $P \ll \mu$ and $Q \ll \mu$, we have relative densities $q \triangleq \frac{dQ}{d\mu}$ and $p \triangleq \frac{dP}{d\mu}$, then we have

$$D_f(P\|Q) = \int_{q>0} qf\Big(\frac{p}{q}\Big)d\mu + f'(\infty)P\{q=0\}$$

where the last term is taken to be zero when $P\{q=0\}=0$, regardless of the value of $f'(\infty)$ which could be infinite.

Example 3.3 (KL divergence). The map $x \mapsto f(x) \triangleq x\ln x$ results in KL divergence.

Example 3.4 (Total variation). The map $x \mapsto f(x) \triangleq \frac{1}{2}|x-1|$ results in the total variation divergence (distance). For $P,Q \in \mathcal{M}(\mathcal{X})$, we define total variation divergence as

$$\mathrm{TV}(P,Q) \triangleq \frac{1}{2}\mathbb{E}_Q\left|\frac{dP}{dQ}-1\right| = \frac{1}{2}\int_{\mathcal{X}}|dP-dQ|.$$

Exercise 3.5. Show that $\mathrm{TV}(P,Q) = 1 - \int_{\mathcal{X}} d(P\wedge Q)$ for all $P,Q \in \mathcal{M}(\mathcal{X})$.

Example 3.6 (χ^2-divergence). The map $x \mapsto f(x) \triangleq (x-1)^2$ results in the χ^2 divergence. For $P,Q \in \mathcal{M}(\mathcal{X})$, we define χ^2 divergence as

$$\chi^2(P\|Q) \triangleq \mathbb{E}_Q\Big(\frac{dP}{dQ}-1\Big)^2 = \int_{\mathcal{X}}\frac{(dP-dQ)^2}{dQ} = \int_{\mathcal{X}}\frac{dP^2}{dQ}-1.$$

We note that we could have chosen $f(x) \triangleq x^2-1$ as well to get the same χ^2 divergence.

Exercise 3.7. Consider two functions $f,h : (0,\infty) \to \mathbb{R}_+$ differing in a linear term, i.e. $h(x) - f(x) = c(x-1)$ for all $x \in (0,\infty)$ and some $c \in \mathbb{R}$. Show that $D_h = D_f$.

Exercise 3.8. Show that $D(P\|Q) \leqslant \ln(1+\chi^2(P\|Q))$ for all $P,Q \in \mathcal{M}(\mathcal{X})$.

Example 3.9 (Squared Hellinger distance). The map $x \mapsto f(x) \triangleq (1-\sqrt{x})^2$ results in squared Hellinger distance which is defined for any $P,Q \in \mathcal{M}(\mathcal{X})$ as

$$H^2(P,Q) \triangleq \mathbb{E}_Q\Big(1-\sqrt{\frac{dP}{dQ}}\Big)^2 = \int_{\mathcal{X}}(\sqrt{dQ}-\sqrt{dP})^2 = 2-2\int_{\mathcal{X}}\sqrt{dPdQ}.$$

The quantity $B(P,Q) \triangleq \int_{\mathcal{X}}\sqrt{dPdQ}$ is known as the *Bhattacharyya coefficient* or *Hellinger affinity*. Hellinger distance $H : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$ is defined as $H(P,Q) \triangleq \sqrt{H^2(P,Q)}$ for all $P,Q \in \mathcal{M}(\mathcal{X})$.

Exercise 3.10. Show that Hellinger distance $H(P,Q) \triangleq \sqrt{H^2(P,Q)}$ defines a metric on the space of probability distributions $\mathcal{M}(\mathcal{X})$, and the map $P \mapsto H(P,Q)$ is not convex.

Example 3.11 (Le Cam divergence (distance)). The map $x \mapsto f(x) \triangleq \frac{(1-x)^2}{2x+2}$ results in Le Cam divergence (distance) which is defined for any $P,Q \in \mathcal{M}(\mathcal{X})$ as

$$\mathrm{LC}(P,Q) \triangleq \mathbb{E}_Q \frac{(1-\frac{dP}{dQ})^2}{2(1+\frac{dP}{dQ})} = \frac{1}{2}\int_{\mathcal{X}} \frac{(dQ-dP)^2}{dQ+dP}.$$

Exercise 3.12. Show that square root of Le Cam distance $\sqrt{\mathrm{LC}} : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$ defines a metric on the space of probability distributions $\mathcal{M}(\mathcal{X})$.

Example 3.13 (Jensen-Shannon divergence). The map $x \mapsto f(x) \triangleq x \ln \frac{2x}{x+1} + \ln \frac{2}{x+1}$ results in Jensen-Shannon divergence which is defined for any $P,Q \in \mathcal{M}(\mathcal{X})$ as

$$\begin{aligned}\mathrm{JS}(P,Q) &\triangleq \mathbb{E}_P \ln \frac{2\frac{dP}{dQ}}{1+\frac{dP}{dQ}} + \mathbb{E}_Q \ln \frac{2}{1+\frac{dP}{dQ}} = \mathbb{E}_P \ln \frac{dP}{\frac{1}{2}d(P+Q)} + \mathbb{E}_Q \ln \frac{dQ}{\frac{1}{2}d(P+Q)} \\ &= D(P\|\frac{1}{2}(P+Q)) + D(Q\|\frac{1}{2}(P+Q)).\end{aligned}$$

Exercise 3.14. Show the following maps $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$ define a metric on the space of probability distributions $\mathcal{M}(\mathcal{X})$.
(a) Total variation distance TV.
(b) Hellinger distance H.
(c) Square root of Le Cam divergence $\sqrt{\mathrm{LC}}$.
(d) Square root of Jensen-Shannon divergence $\sqrt{\mathrm{JS}}$.

Proposition 3.15. *Consider functions f, f_1, f_2 from Definition 3.1 and $P,Q \in \mathcal{M}(\mathcal{X})$ such that $P \ll Q$. Then the following properties hold true for f-divergences.*
(a) $D_{f_1+f_2}(P\|Q) = D_{f_1}(P\|Q) + D_{f_2}(P\|Q)$.
(b) $D_f(P\|P) = 0$.
(c) $D_f(P\|Q) = 0$ for all $P \neq Q$ iff $f(x) = c(x-1)$ for some c. For any other f, we have $D_f(P\|Q) = f(0) + f'(\infty) > 0$ for $P \perp Q$.
(d) Let $f_1(x) \triangleq f(x) + c(x-1)$, then $D_{f_1}(P\|Q) = D_f(P\|Q)$ for all measures $P,Q \in \mathcal{M}(\mathcal{X})$. In particular, we can always assume that $f \geqslant 0$ and if f is differentiable at 1 then $f'(1) = 0$.

Proof. We will show these properties individually.
(a) This follows from linearity of expectation.
(b) This follows from the fact that $\log_2 1 = 0$.
(c) We verify that $D_f(P\|Q) = 0$ for $f = c(x-1)$, since $c(\mathbb{E}_Q \frac{dP}{dQ} - 1) = 0$. For a general f and orthogonal measures $P \perp Q$, we have $pq = 0$, and hence by definition

$$D_f(P\|Q) = f(0)\int_{q>0} q d\mu + f'(\infty)P\{p \geqslant 0\} = f(0) + f'(\infty).$$

This divergence is well-defined (i.e., $\infty - \infty$ is not possible) since by convexity $f(0) > -\infty$ and $f'(\infty) > -\infty$. So all we need to verify is that $f(0) + f'(\infty) = 0$ if and only if $f = c(x-1)$ for some $c \in \mathbb{R}$. Since $f(1) = 0$ and f is convex, we obtain $f'(x)(1-x) \leqslant -f(x)$ for each $x \in \mathbb{R}_+$. We define a map $x \mapsto g(x) \triangleq \frac{f(x)}{x-1}$, and observe that $g'(x) = \frac{f'(x)}{x-1} - \frac{f(x)}{(x-1)^2} \geqslant 0$ for each $x \in \mathbb{R}_+$. That is, g is nondecreasing in $\mathbb{R}_+$. Further, we have

$$g(0) = -\lim_{x \downarrow 0} f(0), \qquad g(\infty) = \lim_{x \downarrow 0} g(\frac{1}{x}) = \lim_{x \downarrow 0} \frac{f(1/x)}{\frac{1}{x} - 1} = \lim_{x \downarrow 0} x f\Big(\frac{1}{x}\Big) = f'(\infty) = -f(0).$$

By assumption, we have $g(0) = g(\infty)$ and hence $g(x)$ is a constant on $x > 0$, as desired.

□

3.1 Conditional f-divergence

Definition 3.16 (Conditional f-divergence). Consider measurable spaces $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ and a pair of Markov kernels $P_{Y|X} : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$ and $Q_{Y|X} : \mathcal{X} \to \mathcal{M}(\mathcal{Y})$, a probability measure P_X on $\mathcal{X}$, and convex function $f : (0, \infty) \to \mathbb{R}$ with $f(1) = 0, f(0) \triangleq \lim_{x \downarrow 0} f(x), f'(\infty) \triangleq \lim_{x \downarrow 0} x f(\frac{1}{x})$. Assuming $(\mathcal{Y}, \mathcal{G})$ is standard Borel measurable space, i.e. $\mathcal{G} \triangleq \mathcal{B}(\mathcal{Y})$, we define

$$D_f(P_{Y|X} \| Q_{Y|X} \mid P_X) \triangleq \mathbb{E}_{x \sim P_X}[D_f(P_{Y|X=x} \| Q_{Y|X=x})]. \tag{1}$$

We observe that as usual in Lebesgue integration it is possible that a conditional f-divergence is finite even though $D_f(P_{Y|X=x} \| Q_{Y|X=x}) = \infty$ for some x in a P_X-negligible set.

Exercise 3.17. Consider f from Definition 3.1 and $P, Q \in \mathcal{M}(\mathcal{X})$ such that $P \ll Q$. Show that the following properties hold true for f-divergences.

(a) If $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = P_X Q_{Y|X}$ then the function $x \mapsto D_f(P_{Y|X=x} \| Q_{Y|X=x})$ is measurable and $D_f(P_{X,Y} \| Q_{X,Y}) = D_f(P_{Y|X} \| Q_{Y|X} \mid P_X)$.

(b) If $P_{X,Y} = P_X P_{Y|X}$ and $Q_{X,Y} = Q_X P_{Y|X}$ then $D_f(P_{X,Y} \| Q_{X,Y}) = D_f(P_X \| Q_X)$. In particular, $D_f(P_X P_Y \| Q_X P_Y) = D_f(P_X \| Q_X)$.

3.2 Data processing inequality

Theorem 3.18 (Monotonicity). $D_f(P_{X,Y} \| Q_{X,Y}) \geqslant D_f(P_X \| Q_X)$.

Proof. Note that in the case $P_{X,Y} \ll Q_{X,Y}$ and thus $P_X \ll Q_X$, the proof is a simple application of Jensen's inequality to definition (1)

$$D_f(P_{X,Y} \| Q_{X,Y}) = \mathbb{E}_{X \sim Q_X} \mathbb{E}_{Y \sim Q_{Y|X}} f\Big(\frac{dP_{Y|X} P_X}{dQ_{Y|X} Q_X}\Big) \geqslant \mathbb{E}_{X \sim Q_X} f\Big(\mathbb{E}_{Y \sim Q_{Y|X}} \frac{dP_{Y|X} P_X}{dQ_{Y|X} Q_X}\Big) = \mathbb{E}_{X \sim Q_X} f(\frac{dP_X}{dQ_X}).$$

To prove the general case we define

$$R_X \triangleq \frac{1}{2}(P_X + Q_X), \qquad R_{Y|X} \triangleq \frac{1}{2} P_{Y|X} + \frac{1}{2} Q_{Y|X}, \qquad R_{X,Y} \triangleq R_X R_{Y|X}.$$

It follows that $P_{X,Y}, Q_{X,Y} \ll R_{X,Y}$ and that $P_{Y|X=x}, Q_{Y|X=x} \ll R_{Y|X=x}$ for every x. By Theorem 2.12 there exist measurable functions p_1, p_2, q_1, q_2 so that

$$\frac{dP_{X,Y}}{dR_{X,Y}} = p_1(x) p_2(y|x), \qquad \frac{dQ_{X,Y}}{dR_{X,Y}} = q_1(x) q_2(y|x) \qquad \frac{dP_{Y|X=x}}{dR_{Y|X=x}} = p_2(y|x), \qquad \frac{dQ_{Y|X=x}}{dR_{Y|X=x}} = q_2(y|x).$$

We also denote $p(x,y) = p_1(x) p_2(y|x)$ and $q(x,y) = q_1(x) q_2(y|x)$. Fix $t > 0$ and by convexity of f, we consider a supporting line to f at t with slope μ, so that for all $u \geqslant 0$

$$f(u) \geqslant f(t) + \mu(u - t).$$

Thus, $f'(\infty) \geqslant \mu$ and taking $u = \lambda t$ for any $\lambda \in [0,1]$ we have shown for all $t \geqslant 0$ and $\lambda \in [0,1]$

$$f(\lambda t) + \bar{\lambda} t f'(\infty) \geqslant f(\lambda t) + \bar{\lambda} t \mu \geqslant f(t). \tag{2}$$

Note that we added $t = 0$ case as well, since for $t = 0$ the statement is obvious (recall, though, that $f(0) \triangleq \lim_{x\downarrow 0} f(0)$ can be equal to ∞). Next, fix some x with $q_1(x) > 0$ and consider the chain

$$\begin{aligned}
&\int_{y:q_2(y|x)>0} dR_{Y|X=x} q_2(y \mid x) f\Big(\frac{p_1(x)p_2(y \mid x)}{q_1(x)q_2(y \mid x)}\Big) + \frac{p_1(x)}{q_1(x)} P_{Y|X=x}\{q_2(Y \mid x) = 0\} f'(\infty) \\
&\geqslant f\Big(\frac{p_1(x)}{q_1(x)} P_{Y|X=x}\{q_2(Y \mid x) > 0\}\Big) + \frac{p_1(x)}{q_1(x)} P_{Y|X=x}\{q_2(Y \mid x) = 0\} f'(\infty) \geqslant f\Big(\frac{p_1(x)}{q_1(x)}\Big).
\end{aligned}$$

where first inequality is by Jensen's inequality applied to convex function f, and second inequality by taking $t = \frac{p_1(x)}{q_1(x)}$ and $\lambda = P_{Y|X=x}\{q_2(Y|x) > 0\}$ in (2). Now multiplying the obtained inequality by $q_1(x)$ and integrating over $\{x : q_1(x) > 0\}$ we get

$$\int_{q>0} dR_{X,Y} q(x,y) f\Big(\frac{p(x,y)}{q(x,y)}\Big) + f'(\infty) P_{X,Y}\{q_1(X) > 0, q_2(Y|X) = 0\} \geqslant \int_{x:q_1(x)>0} dR_X q_1(x) f\Big(\frac{p_1(x)}{q_1(x)}\Big).$$

Adding $f'(\infty)P_X\{q_1(X) = 0\}$ to both sides we obtain the result since both sides evaluate to definition of f divergence $D_f(P\|Q)$. □

Theorem 3.19 (Data processing). *Consider a channel that produces Y given X based on the conditional law $P_{Y|X}$. Let P_Y and Q_Y denote the distribution of Y when X is distributed as P_X and Q_X respectively. For any f-divergence $D_f(\cdot\|\cdot)$,*

$$D_f(P_Y\|Q_Y) \leqslant D_f(P_X\|Q_X). \tag{3}$$

Proof. This follows from the Theorem 3.18 on monotonicity and the fact that $D_f(P_{X,Y}\|Q_{X,Y}) = D_f(P_X\|Q_X)$ since $P_{Y|X} = Q_{Y|X}$. □

Next we discuss some of the more useful properties of f-divergence that parallel those of KL divergence in Theorem 2.16.

Theorem 3.20 (Properties of f-divergences). *Consider two measures $P, Q \in \mathcal{M}(\mathcal{X})$.*

(a) ***Non-negativity.*** *$D_f(P\|Q) \geqslant 0$. If f is strictly convex*[1] *at 1, then $D_f(P\|Q) = 0$ if and only if $P = Q$.*

(b) ***Joint convexity.*** *$(P, Q) \mapsto D_f(P\|Q)$ is a jointly convex function. Consequently, $P \mapsto D_f(P\|Q)$ and $Q \mapsto D_f(P\|Q)$ are also convex.*

(c) ***Conditioning increases f-divergence.*** *Let $P_Y = P_X P_{Y|X}$ and $Q_Y = P_X Q_{Y|X}$, then*

$$D_f(P_Y\|Q_Y) \leqslant D_f(P_{Y|X}\|Q_{Y|X} \mid P_X).$$

Proof. (a) Non-negativity follows from monotonicity by taking X to be unary. To show strict positivity, suppose for the sake of contradiction that $D_f(P\|Q) = 0$ for some $P \neq Q$. Then there exists some measurable A such that $P(A) = p \neq q = Q(A) > 0$. Applying the data processing inequality with $Y = \mathbb{1}_{\{X\in A\}}$, we obtain $D_f(Ber(p)\|Ber(q)) = 0$. Consider two cases.

(i) $0 < q < 1$: Then $D_f(Ber(p)\|Ber(q)) = qf(\frac{p}{q}) + (1-q)f(\frac{p}{q}) = f(1)$;

(ii) $q = 1$: Then $p < 1$ and $D_f(Ber(p)\|Ber(q)) = f(p) + (1-p)f'(\infty) = 0$, i.e. $f'(\infty) = \frac{f(p)}{p-1}$. Since $x \mapsto \frac{f(x)}{x-1}$ is non-decreasing, we conclude that f is affine on $[p, \infty)$.

Both cases contradict the assumed strict convexity of f at 1.

(b) Convexity follows from the data processing inequality.

(c) Recall that the conditional divergence was defined in (1) and hence the inequality follows from the monotonicity. Another way to see the inequality is as result of applying Jensen's inequality to the jointly convex function $D_f(P\|Q)$.

□

[1] By strict convexity at 1, we mean for all $s, t \in [0, \infty)$ and $\alpha \in (0, 1)$ such that $\alpha s + (1-\alpha)t = 1$, we have $\alpha f(s) + (1-\alpha)f(t) > f(1)$.