Recap

f-divergences

- Divergence: $\delta(P, Q) = 0 \iff P = Q$

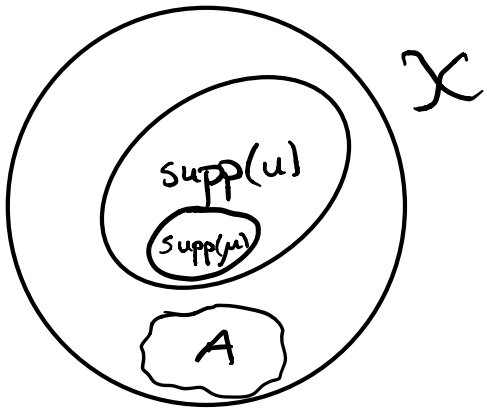
- Metric: Divergence + Symmetry + Triangle Ineq.

- Convexity

Additional Technical Background

Definition (Absolute Continuity): Let $\mu, \nu \in \mathcal{P}(X)$. ($\mathcal{P}$ = set of measures, NOT function; set of non-neg. on X; $\mathcal{P}(X) \to \mathcal{M}_+(X)$)
We say μ is absolutely continuous with respect to ν, denoted by $\mu \ll \nu$, if $\nu(A) = 0 \Rightarrow \mu(A) = 0$.

Illustration:



i.e $\text{supp}(\mu) \subseteq \text{supp}(\nu)$

Theorem (Radon-Nikodym)

Let $\mu, \nu \in \mu_+(X)$ with $\mu \ll \nu$. Then there exists a function $f \in L^1(\nu)$ such that

$$\underbrace{\mu(A)}_{\int_A d\mu(x)} = \int_A f(x)\, d\nu(x) \qquad \forall \text{ measurable sets } A$$

The function f is called the Radon Nikodym derivative of μ w.r.t ν, often denoted by

$$f = \frac{d\mu}{d\nu}$$

Examples - Let $\mu \ll \nu$

① Let $\nu = \#$ be the counting measure

where $\# = \begin{cases} |A|, & \text{if } |A| < \infty \\ \infty, & \text{o/w} \end{cases}$

If $\mu \ll \#$, then the $\text{supp}(\mu)$ is countable and

$$p = \frac{d\mu}{d\#} \text{ is the pmf of } \mu$$

$(p(x) = \mu(\{x\})$, and $\mu(A) = \sum_{x \in A} p(x)$

② Let $\nu = \lambda$ be the Lebesgue measure.

$(\lambda(A) = \text{Vol}(A) = \int_A dx)$

If $\mu << \lambda$ then $p = \frac{d\mu}{d\lambda}$ is the pdf induced by μ, and

$$\mu(A) = \int_A p(x) \underbrace{d\lambda(x)}_{\text{shorthand } dx}$$

▶ f-divergences

Definition (f-divergence): Let

$$f: \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$$

be a convex function such that

① $f(1) = 0$

② f is strictly convex around 1:

$$f(\underbrace{\alpha x + (1-\alpha)y}_{1}) < \alpha f(x) + (1-\alpha) f(y) \quad \forall x, y \in \mathbb{R}_{\geq 0}$$

$\alpha \in [0,1]$ s.t. $\alpha x + (1-\alpha) y = 1$

The f-divergence between two probability measures on the same space $P, Q \in \mathcal{P}(X)$ that are both dominated by the same measure λ i.e $P, Q << \lambda$, is

Such a λ always exists → non-obvious

(doesn't need to be a prob. measure)

RNDs

ratio

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

$$= \int_{\text{supp}(Q)} f\left(\frac{dP/d\lambda\,(x)}{dQ/d\lambda\,(x)} \right) dQ(x)$$

If $P \ll Q$, then

RND

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ} \right) \right]$$

Examples

(1) If $P, Q \ll \#$ then

$$D_f(P\|Q) = \sum_{x \in \text{supp}(Q)} q(x)\, f\left(\frac{p(x)}{q(x)} \right)$$

where p, q are the pmfs of P, Q respectively.

(2) If $P, Q \ll \lambda$ then

$$D_f(P\|Q) = \int_{\text{supp}(Q)} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

where p, q are pdfs of P, Q respectively.

Conventions:

(i) $f(0) = f(0^+)$

(ii) $0 = 0\, f\left(\frac{0}{0}\right)$

Important Special Cases

Shorthands: we'll use

$$\frac{dP}{d\lambda} = dP, \quad \frac{dQ}{d\lambda} = dQ$$

and

$$\int g\left(\frac{dP}{d\lambda}, \frac{dQ}{d\lambda}\right) d\lambda = \int g(dP, dQ)$$

Kullback - Leibler Divergence

$f(x) = x \log x$

$$D_{KL}(P \| Q) = D_{x \log x}(P \| Q)$$

$$= \mathbb{E}_Q \left[\frac{dP}{dQ} \log \left(\frac{dP}{dQ} \right) \right]$$

$$= \mathbb{E}_P \left[\log \left(\frac{dP}{dQ} \right) \right]$$ Radon Theorem

Comments

① If $P, Q \ll \#$: $D_{KL}(P \| Q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$

② If $P, Q \ll \lambda$: $D_{KL}(P \| Q) = \int_X p(x) \log \frac{p(x)}{q(x)}$

③ D_{KL} is a divergence, but definitely NOT a metric. In fact, it is not a metric and it does NOT satisfy triangle ineq.

④ If $P \not\ll Q$, $D_{KL}(P \| Q) = \infty$

⑤ Let $f(x) = -\log x$

$$D_{-\log x}(P \| Q) = \mathbb{E}_Q \left[\log \left(\frac{dQ}{dP} \right) \right] = D_{KL}(Q \| P)$$

Total Variation Distance

$$f(x) = \frac{1}{2}|x-1|$$

$$\delta_{TV}(P,Q) \triangleq D_{\frac{1}{2}|x-1|}(P\|Q)$$

$$= \frac{1}{2}\mathbb{E}_Q\left[\left|\frac{dP}{dQ}-1\right|\right]$$

$$= \frac{1}{2}\int |dP - dQ|$$

Comments:

① $P, Q \ll \#$: $\delta_{TV}(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |p(x)-q(x)|$

$= \frac{1}{2}\|p-q\|_1$

② $P, Q \ll \lambda$: $\delta_{TV}(P,Q) = \frac{1}{2}\|p-q\|_{L^1(\mathbb{R}^d)}$ $\|f\|_{L^1(\mathbb{R}^d)} = \int |f|dx$

③ δ_{TV} is a metric on $\mathcal{P}(\mathcal{X})$

④ If $\text{Supp}(P) \cap \text{Supp}(Q) = \emptyset$, then $\delta_{TV}(P,Q) = 1$

χ^2-Divergence

$$f(x) = (x-1)^2$$

$$\chi^2(P \| Q) = D_{(x-1)^2}(P \| Q)$$

$$= \mathbb{E}_Q\left[\left(\frac{dP}{dQ} - 1\right)^2\right]$$

$$= \int \left(\frac{dP}{dQ}\right)^2 dQ - 2\int \frac{dP}{dQ}\, dQ + \int dQ$$

$$\int \frac{dP}{dQ}\, dQ = 1$$

$$\int dQ = 1 \leftarrow \text{a P measur}$$

$$= \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^2\right] - 1$$

$$= \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^2 - 1\right] = D_{(x-1)^2}(P \| Q)$$

Comments:

① wrt above, mapping (injective).

$f \mapsto D_f$ is NOT one-to-one

② If $P \ll Q$, then $\chi^2(P \| Q) = \infty$

Proposition (properties of f-divergences)

① Non-negativity. $D_f(P\|Q) \geq 0$ w/ equality iff $P = Q$

② Convexity: $(P, Q) \mapsto D_f(P\|Q)$ is jointly convex. In particular, $P \mapsto D_f(P\|Q)$ is convex for fixed Q, and $Q \mapsto D_f(P\|Q)$ is convex for fixed P.

③ Conditioning Increased f-Divergences

Let $P_{Y|X}$ and $Q_{Y|X}$ be transition kernels and $P_X \in \mathcal{P}(\mathcal{X})$.

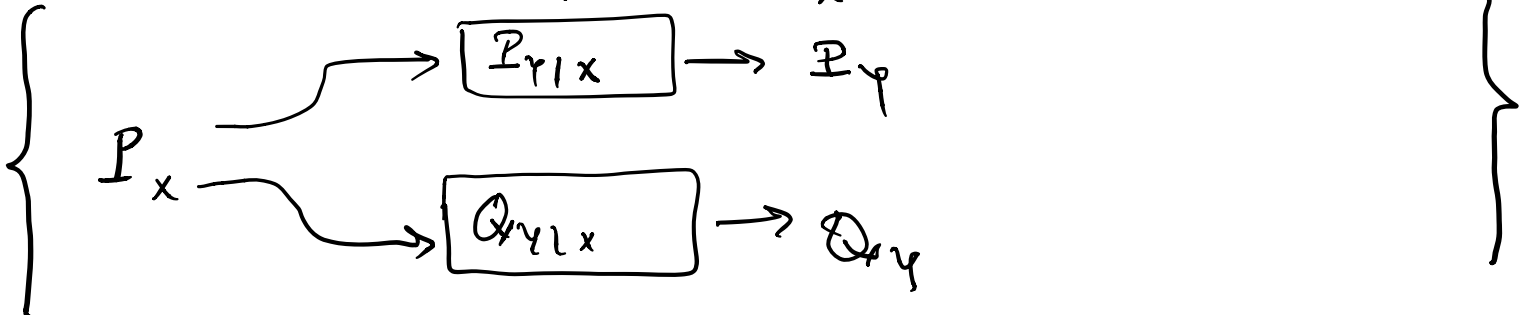
Define the conditional f-divergence

$$D_f(P_{Y|X}\|Q_{Y|X}|P_X) := \int_{\mathcal{X}} D_f(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x))\, dP_X(x)$$

$$= \mathbb{E}_{P_X}\left[D_f(P_{Y|X}(\cdot|X)\|Q_{Y|X}(\cdot|X))\right]$$

and recall

$$P_Y(\cdot) = \mathbb{E}_{P_X}[P_{Y|X}(\cdot|X)]$$
$$Q_Y(\cdot) = \mathbb{E}_{P_X}[Q_{Y|X}(\cdot|X)]$$



then

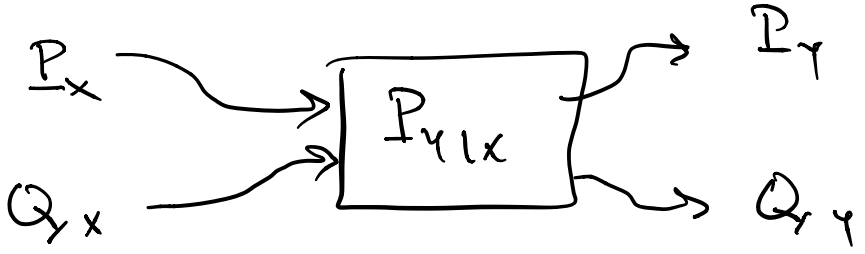
$$D_f(P_Y \| Q_Y) = D_f(\mathbb{E}_{P_X}[P_{Y|X}(\cdot|X)] \| \mathbb{E}_{P_X}[Q_{Y|X}(\cdot|X)])$$

$$\leq D_f(P_{Y|X} \| Q_{Y|X} | P_X)$$

f-divergence grows!

④ Let $P_{XY} = P_X P_{Y|X}$

$Q_{XY} = Q_X P_{Y|X}$



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