General Properties for f_divergence:

1- $D_f(P\|Q) \geq 0$

2- If f is strictly convex at 1, then $D_f(P\|Q) = 0 \Rightarrow P = Q$.

$*$ if $\forall x \neq y$ & $\lambda \in (0,1)$ such that $\lambda x + \bar{\lambda} y = 1$

then: $f(1) < \lambda f(x) + \bar{\lambda} f(y)$.

or equivalently

$x_i \neq x_j$

if $\sum \lambda_i x_i = 1$

$f(1) < \sum \lambda_i f(x_i)$

$*$ Note that $P = Q \Rightarrow D_f(P\|Q) = 0$

proof: $D_f(P\|Q) = \sum Q(x) \cdot f\left(\frac{P(x)}{Q(x)}\right) \geq f(1)$ $(*)$

If f is strictly convex at $1 \Rightarrow f(1) < \sum Q(x) \cdot f\left(\frac{P(x)}{Q(x)}\right)$

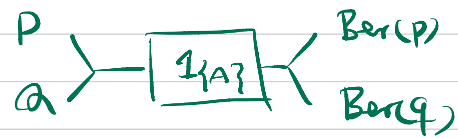
(because $\sum Q(x) \cdot \frac{P(x)}{Q(x)} = 1$)

Thus, equality in $(*)$ happens only if $\frac{P(x)}{Q(x)} = c \quad \forall x$.

A more direct proof:

Suppose, for the sake of a contradiction, that $D_f(P\|Q)=0$ for some $P\neq Q$. Then there exists some set A such that $P(A)\neq Q(A)$.

Let $p := P(A)$ & $q := Q(A)$.



By DPI (which is about to be proved in Item 5), we have

$$D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) \leq D_f(P\|Q) = 0$$

$$\Rightarrow D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) = 0$$

$$D_f(\mathrm{Ber}(p)\|\mathrm{Ber}(q)) = q\, f(p/q) + \bar{q}\, f(\bar{p}/\bar{q}) = f(1)$$

contradicts the definition of strict convexity.

3. $(P,Q) \mapsto D_f(P\|Q)$ is jointly convex

$$D_f(\lambda P_1 + \bar\lambda P_2 \| \lambda Q_1 + \bar\lambda Q_2) \le \lambda D(P_1\|P_2) + \bar\lambda D(Q_1\|Q_2)$$

Define: $g(a,b) = b \cdot f(a/b)$ [Perspective of f]

If f is convex g is jointly convex

$$D_f(\lambda P_1 + \bar\lambda P_2 \| \lambda Q_1 + \bar\lambda Q_2) = \sum \underbrace{(\lambda Q_1 + \bar\lambda Q_2) f\left(\frac{\lambda P_1 + \bar\lambda P_2}{\lambda Q_1 + \bar\lambda Q_2}\right)}_{g(\lambda P_1 + \bar\lambda P_2,\, \lambda Q_1 + \bar\lambda Q_2)}$$

convex

$$\le \lambda \cdot \sum g(P_1(x), Q_1(x)) + \bar\lambda \sum g(P_2(x), Q_2(x))$$

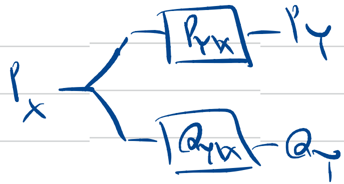
$$(\lambda b_1 + \bar\lambda b_2) f\left(\frac{\lambda a_1 + \bar\lambda a_2}{\lambda b_1 + \bar\lambda b_2}\right)$$

$$= (\lambda b_1 + \bar\lambda b_2) f\left(\frac{\lambda b_1}{\lambda b_1 + \bar\lambda b_2} \frac{a_1}{b_1} + \frac{\bar\lambda b_2}{\lambda b_1 + \bar\lambda b_2} \frac{a_2}{b_2}\right)$$

$$\le (\lambda b_1 + \bar\lambda b_2) \left[\frac{\lambda b_1}{\lambda b_1 + \bar\lambda b_2} f\left(\frac{a_1}{b_1}\right) + \frac{\bar\lambda b_2 f\left(\frac{a_2}{b_2}\right)}{\lambda b_1 + \bar\lambda b_2}\right]$$

$$= \lambda b_1 f(a_1/b_1) + \bar\lambda b_2 f(a_2/b_2)$$

4. conditioning increase f-divergence



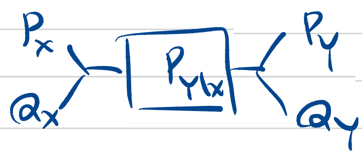
$$D_f(P_Y\|Q_Y) \le \sum_x P_x D_f(P_{Y|X=x}\|Q_{Y|X=x})$$

$(P,Q) \mapsto D_f(p\|q)$

$$\sum P_X(x) D_f(P_{Y|x}\|Q_{Y|x}) \ge D_f\left(\sum_x P_X(x)\cdot P_{Y|x} \,\Big\|\, \sum_x P_X(x)\cdot Q_{Y|X=x}\right)$$

$$= D_f(P_Y\|Q_Y)$$

5. DPI



$$D_f(P_Y\|Q_Y) \le D_f(P_X\|Q_X)$$

*Recall proof of DPI for KL depends on chain rule which doesn't hold for f-divergence!

$$D_f(P_X \| Q_X) = \sum_x Q_X(x) \cdot f\left(\frac{P_X(x)}{Q_X(x)}\right)$$

$$= \sum_{x,y} P_{Y|X}(y|x) \cdot Q_X(x) \cdot f\left(\frac{P_X(x)}{Q_X(x)} \cdot \frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x)}\right)$$

backward channel

$$= \sum_{x,y} Q_Y(y) \cdot Q_{X|Y}(x|y) \, f\left(\frac{P_{XY}(x,y)}{Q_Y(y) \cdot Q_{X|Y}(x|y)}\right)$$

$$= \sum_y Q_Y(y) \cdot \underbrace{\sum_x Q_{XY}(x,y) \cdot f\left(\frac{P_{XY}(x,y)}{Q_Y(y) \cdot Q_{X|Y}(x|y)}\right)}_{\text{Jensen}}$$

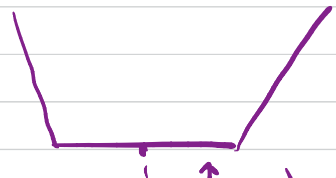
$$\geq \sum_y Q_Y(y) \cdot f\left(\sum_x \cancel{Q_{X|Y}(x|y)} \cdot \frac{P_{XY}(x,y)}{Q_Y(y) \cdot \cancel{Q_{X|Y}(x|y)}}\right)$$

$$= \sum_y Q_Y(y) \cdot f\left(\sum_x \frac{P_{XY}(x,y)}{Q_Y(y)}\right)$$

$$= \sum_y Q_Y(y) \cdot f\left(\frac{P_Y(y)}{Q_Y(y)}\right) \checkmark$$

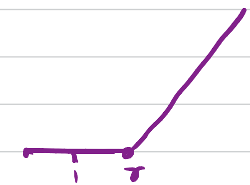
Remark:

1) As we proved: $D_f(P\|Q)=0 \Rightarrow P=Q$ only if f is strictly convex at 1



↑ not strict convexity

For example: For hockey-stick divergence, for $\lambda>1$



↑ not strictly convex

Hence, in general $E_\gamma(P\|Q)=0 \not\Rightarrow P=Q$.

Example: $E_\gamma\left(\mathrm{Ber}\left(\frac{\gamma-1}{2\gamma}\right) \,\Big\|\, \mathrm{Ber}\left(1-1/\gamma\right)\right) = \left(\frac{\gamma-1}{2\gamma} - \overbrace{\gamma(1-1/\gamma)}^{(\gamma-1)}\right)_+ + \left(\frac{\gamma+1}{2\gamma} - \overbrace{\gamma\cdot\frac{1}{\gamma}}^{1}\right)_+$

for $\gamma>1$

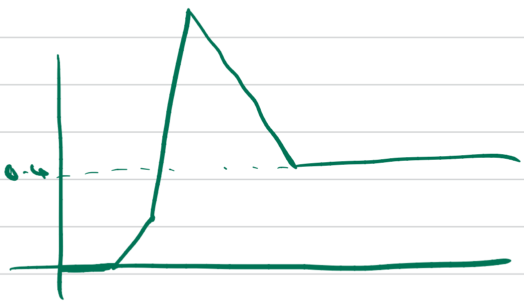
$= \left(\underbrace{\frac{\gamma-1}{2\gamma} - (\gamma-1)}_{<0}\right)_+ + \left(\underbrace{\frac{1-\gamma}{2\gamma}}_{<0}\right)_+ = 0$

Example: $E_\gamma(\mathrm{Ber}(p) \| \mathrm{Ber}(q)) = (p-\gamma q)_+ + (\bar{p}-\gamma\bar{q})_+ - (1-\gamma)_+$

Example 2: $P = [0.1, 0.2, 0.3, 0.4]$ $\quad Q = [0.4, 0.5, 0.1, 0]$

$$E_\gamma(P \| Q) = (0.1-\gamma \times 0.4)_+ + (0.2-\gamma \times 0.5)_+ + (0.3-\gamma \times 0.1)_+ + (0.4-\gamma \cdot 0)_+$$

$$= (0.1-0.4\gamma)_+ + (0.2-0.5\gamma)_+ + (0.3-0.1\gamma)_+ + 0.4$$



$=$

$$Q(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-u^2/2}\, du$$

Exercise: $E_\gamma(\mathcal{N}(0,\sigma^2) \| \mathcal{N}(\mu,\sigma^2)) = Q\left(\frac{\log \gamma}{\beta} - \frac{\beta}{2}\right) - \gamma Q\left(\frac{\log \gamma}{\beta} + \beta/2\right)$

where $\beta = \frac{|\mu|}{\sigma}$

the same result holds even

$\mathcal{N}(\mu_1, \sigma^2 I_d)$ $\quad \mathcal{N}(\mu_2, \sigma^2 I_d)$

with $\beta = \|\mu_1 - \mu_2\| / \sigma$

Properties of E_γ-divergence

$$E_\gamma(P\|Q) := \sum_x (P(x) - \gamma Q(x))_+ - (1-\gamma)_+$$

What is $E_1(P\|Q)$?

- $E_\gamma(P\|Q) \neq E_\gamma(Q\|P)$

Equivalent forms:

1- $E_\gamma(P\|Q) = \frac{1}{2}\sum |P(x) - \gamma Q(x)| - \frac{1}{2}|1-\gamma|$

$2(a)_+ = |a| + a$

so: $E_\gamma(P\|Q) = \sum (P(x) - \gamma Q(x))_+ - (1-\gamma)_+$

$$= \sum \frac{|P(x) - \gamma Q(x)| + (P(x) - \gamma Q(x))}{2} - \frac{|1-\gamma| + (1-\gamma)}{2}$$

$$= \frac{1}{2}\sum |P(x) - \gamma Q(x)| + \cancel{\frac{1-\gamma}{2}} - \frac{1}{2}|1-\gamma| - \cancel{\frac{1-\gamma}{2}}$$

$$2-\quad E_\gamma(P\|Q) \overset{\text{For } \gamma\geq 1}{=} \sup_A \left(P(A) - \gamma Q(A) \right) = P(A^*) - \gamma Q(A^*)$$

For $\gamma \leq 1$: $\sup_A \left(\gamma Q(A) - P(A) \right)$

where

$A^* = \{x : P(x) \geq \gamma Q(x)\}$

Proof requires two steps:

1. For any A: $P(A) - \gamma Q(A) \leq E_\gamma(P\|Q)$.

2. $\exists A^*$ such that $P(A^*) - \gamma Q(A^*) = E_\gamma(P\|Q)$.

$$1-\quad P(A) - \gamma Q(A) = \sum_{x\in A} P(x) - \gamma Q(x) \leq \sum_{x\in A} \left(P(x) - \gamma Q(x)\right)_+$$

Because we ignore all negative terms

$$\leq \sum_{x\in \mathcal{X}} \left(P(x) - \gamma Q(x)\right)_+$$

$$= E_\gamma(P\|Q)$$

because on A^*, we always $p(x) \geq \gamma Q(x)$

2. $$P(A^*) - \gamma Q(A^*) = \sum_{x \in A^*} p(x) - \gamma Q(x) = \sum_{x \in A^*} (p(x) - \gamma Q(x))_+$$

$$= \sum_{x \in A^*} (p(x) - \gamma Q(x))_+ + \underbrace{\sum_{x \in A^{*c}} (p(x) - \gamma Q(x))_+}_{=0}$$

$$= \sum_{x} (p(x) - \gamma Q(x))_+ = E_\gamma(P \| Q).$$

$\square$

Properties of E_γ - divergence :

1) Given P, Q, $\gamma \mapsto E_\gamma(P\|Q)$ is non-increasing & convex for $\gamma \geq 1$

& is non-decreasing & convex

- monotinicity is obvious: Note that for $r<1$ use the other def.

$$E_r(P\|Q) \underset{\uparrow \atop r<1}{=} \sum (rQ(x) - P(x))_+$$

- convexity: $E_r(P\|Q) = \sup \underbrace{P(A) - rQ(A)}_{\text{linear}}$ $\longrightarrow$ Supremum of linear function is convex

Direct approach:

$$E_{tr_1 + \bar{t}r_2}(P\|Q) = \sup \left[P(A) - (tr_1 + \bar{t}r_2) Q(A) \right]$$

$$= \sup \left[P(A) - tr_1 Q(A) - \bar{t} r_2 Q(A) \right]$$

$$= \sup \left[t(P(A) - r_1 Q(A)) + \bar{t}(P(A) - r_2 Q(A)) \right]$$

$$\leq t \cdot \sup P(A) - r_1 Q(A) + \bar{t} \sup P(A) - r_2 Q(A)$$

2) Reciprocity: $E_\gamma(P\|Q) = \gamma \cdot E_{\gamma^{-1}}(Q\|P)$

suppose $\gamma \geq 1$: $E_\gamma(P\|Q) = \sum (P - \gamma Q)_+$

& for $\gamma < 1$: $E_\gamma(P\|Q) = \sum (\gamma Q - P)_+$

$$\gamma \cdot E_{\gamma^{-1}}(Q\|P) \overset{\gamma \geq 1}{=} \gamma \left[\sum \left(\tfrac{1}{\gamma} P - Q\right)_+ \right] = \sum (P - \gamma Q)_+$$

3) $\lim_{\gamma \to 0} E_\gamma(P\|Q) = 0$ $\lim_{\gamma \to \infty} E_\gamma(P\|Q) = 0$

If P & Q are supported on the same alphabet

Strong DPI

We saw that all f-divergences satisfies DPI. But in many cases & for many channels, DPI is **strict**.

$$D_f(P_Y \| Q_Y) \leq D_f(P_X \| Q_X)$$

we wish to improve this

So, we wish to find smallest $\lambda \leq 1$ such that

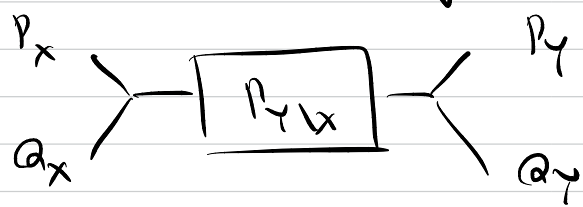
$$D_f(P_Y \| Q_Y) \leq \lambda \, D_f(P_X \| Q_X) \quad \forall \, P_X \,\&\, Q_X$$

What's the best λ?

Smallest λ is given by $\eta_f(P_{Y|X}) := \sup_{P_X \neq Q_X} \dfrac{D_f(P_Y \| Q_Y)}{D_f(P_X \| Q_X)}$

We call this number: contraction coefficient of channel $P_{Y|X}$ under f-divergence,

By definition



$$D_f(P_Y \| Q_Y) \leq \eta_f(P_{Y|X}) \cdot D_f(P_X \| Q_X)$$

If $\eta_f(P_{Y|X}) < 1$ then this inequality is strictly better than DPI $\Rightarrow$ Strong DPI

we the call the channel $P_{Y|X} \Rightarrow$ Contractive.

Computing $\eta_f(P_{Y|X})$ is not easy: Infinite_dimensional optimization.

But we know a lot about them for specific D_f:

1- $\eta_{\chi^2}(P_{Y|X}) := \sup_{P_X \neq Q_X} \frac{\chi^2(P_Y \| Q_Y)}{\chi^2(P_X \| Q_X)}$ is related to

Rényi maximal correlation: $\sup_{f,g} \rho(f(X), g(Y))$

2- $\eta_{KL}(P_{Y|X})$ is related propogation of information over network.

3- $\eta_{TV}(P_{Y|X})$ has been a central quantity in studying Markov chain, graph theory, dynamical system:

$$\eta_{TV}(P_{Y|X}) := \sup_{P_X \neq Q_X} \frac{TV(P_Y, Q_Y)}{TV(P_X, Q_X)}$$

Theorem: (Dobrushin's two point characterization)

For any channel $P_{Y|X}$, we have:

$$\eta_{TV}(P_{Y|X}) = \max_{x_1, x_2} TV(P_{Y|X=x_1}, P_{Y|X=x_2})$$

↑ Dobrushin coefficient

Proof: Strategy: ① $TV(P_{Y|X=x_1}, P_{Y|X=x_2}) \leq \eta_{TV}(P_{Y|X})$ for any pair of x_1 & x_2

② $\sup_{x_1, x_2} TV(P_{Y|X=x_1}, P_{Y|X=x_2}) \geq \eta_{TV}(P_{Y|X})$

① Pick P_X & Q_X degenerate distribution on some arbitrary points:

$$P_X(x_1)=1 \quad \text{with } x_1 \neq x_2 \quad \Rightarrow \quad TV(P_X, Q_X) = 1$$

$$Q_X(x_2)=1$$

Corresponding $P_Y = P_{Y|X=x_1}$ & $Q_Y = P_{Y|X=x_2}$

So: $TV(P_Y, Q_Y) = TV(P_{Y|X=x_1}, P_{Y|X=x_2})$

Thus: $$\eta_r(P_{Y|X}) = \sup_{P_X \neq Q_X} \frac{TV(P_Y, Q_Y)}{TV(P_X, Q_X)} \geq TV(P_{Y|X=x_1}, P_{Y|X=x_2}) \quad \forall x_1, x_2$$

② Fix two dist. P_X & Q_X & define $A = \{x : P_X(x) \geq Q_X(x)\}$.

Now, we can write: for any $B \subseteq \mathcal{Y}$

$\uparrow$ output support

$$P_Y(B) - Q_Y(B) = \sum_x P_{Y|X}(B|x)\, P_X(x) - P_{Y|X}(B|x) \cdot Q_X(x)$$

$$= \sum_x P_{Y|X}(B|x) \cdot [P_X(x) - Q_X(x)]$$

$$= \sum_{x \in A} P_{Y|X}(B|x) \cdot [P_X(x) - Q_X(x)] + \sum_{x \in A^c} P_{Y|X}(B|x) \cdot [P_X(x) - Q_X(x)]$$

$$= \sum_{x \in A} P_{Y|X}(B|x) \cdot [P_X(x) - Q_X(x)] - \sum_{x \in A^c} P_{Y|X} [Q_X(x) - P_X(x)]$$

$$= TV(P,Q) \sum_{x \in A} P_{Y|X}(B|x) \underbrace{\frac{[P_X(x) - Q_X(x)]}{TV(P,Q)}}_{U(x) =:} - \sum_{x \in A^c} P_{Y|X} \underbrace{\frac{[Q_X(x) - P_X(x)]}{TV(P,Q)}}_{V(x) =:}$$

note that $\sum_{x \in A} u(x) = TV(P_X, Q_X)$

& $\sum_{x \in A^c} v(x) = TV(P_X, Q_X)$

$$= TV(P,Q) \cdot \left[\sum_A u(x) \cdot P_{Y|X}(B|x) - \sum_{x' \in A^c} v(x')\, P_{Y|X}(B|x') \right]$$

Thus U is a pmf on A

& V is a pmf on A^c

$$= TV(P,Q)\left[\sum_{x\in A}\sum_{x'\in A^c} u(x).V(x')\,P_{Y|X}(B|x) - \sum_{x\in A}\sum_{x'\in A^c} V(x').u(x)\,P_{Y|X}(B|x')\right]$$

$$= TV\left[\sum_{x\in A}\sum_{x'\in A^c} u(x).V(x').\underbrace{\left(P_{Y|X}(B|x) - P_{Y|X}(B|x')\right)}_{\leq TV(P_{Y|X=x'},P_{Y|X=x})}\right]$$

$$\leq TV(P_X,Q_X).\sup_{x,x'} TV(P_{Y|X=x},P_{Y|X=x'})$$

$$\Rightarrow \quad P_Y(B) - Q_Y(B) \leq TV(P_X,Q_X).\sup_{x,x'} TV(P_{Y|X=x},P_{Y|X=x'})$$

Since it holds for all $B\subseteq \mathcal{Y}$, we can take supremum over B. $\square$

For more than 60 years, TV was a only f-divergence whose contraction coefficient was known in closed-form.

However, it was proved recently, that a similar two-point characterization can be proved for hockey-stick divergence

Notation: let $\eta_\gamma(P_{Y|X})$ denote the contraction coefficient of $P_{Y|X}$ under E_γ-divergence. That is, we have:

$$\eta_\gamma(P_{Y|X}) := \sup_{\substack{P_X, Q_X \\ E_\gamma(P_X \| Q_X) \neq 0}} \frac{E_\gamma(P_Y \| Q_Y)}{E_\gamma(P_X \| Q_X)}$$

Theorem: For any channel $P_{Y|X}$ & $\gamma \geq 1$, we have

$$\eta_\gamma(P_{Y|X}) = \sup_{x_1, x_2} E_\gamma(P_{Y|X=x_1} \| P_{Y|X=x_2})$$

* You'll prove this result in HW 1.

* Note that setting $\gamma = 1$ in this result, we can obtain Dobrushin's two-point characterization result; thus this result is a natural generalization of Dobrushin's.

* **Remark:** This result gives $\eta_\gamma(P_{Y|X})$ Only for $\gamma \geq 1$.

How should we compute $\eta_\gamma(P_{Y|X})$ for $\gamma < 1$?

Lemma. Let $P_{Y|X}$ be a channel & $\gamma < 1$. Then we have

$$\eta_\gamma(P_{Y|X}) = \eta_{1/\gamma}(P_{Y|X}).$$

Proof. Use reciprocity property of E_γ.

Example: consider channel $P_{Y|X}$ with binary input:

$$P_{Y|X=0} = \text{Bernoulli}(\delta) \qquad P_{Y|X=1} = \text{Bernoulli}(1-\delta)$$

for some $\delta \in (0, 1/2)$.

Compute $\eta_r(P_{Y|X})$ for $r \geq 1$ & $r < 1$.

* This channel is often referred to as Binary Symmetric Channel with crossover probability δ; denoted by $BSC(\delta)$.

we know from the above theorem that

$$\eta_r(P_{Y|X}) = \sup_{x_1, x_2} E_r(P_{Y|X=x_1} \| P_{Y|X=x_2}) \quad \text{for } r \geq 1.$$

Since input of channel is binary; this supremum is over all binary x_1 & x_2. Thus;

$$\eta_r(P_{Y|X}) = \max\left\{ E_r(P_{Y|X=0} \| P_{Y|X=1}),\ E_r(P_{Y|X=1} \| P_{Y|X=0}) \right\}$$

$$= \max\{\underbrace{E_\gamma(\mathrm{Bern}(\delta)\,\|\,\mathrm{Bern}(\bar\delta))}_{=(\bar\delta-\gamma\delta)_+},\ \underbrace{E_\gamma(\mathrm{Bern}(\bar\delta)\,\|\,\mathrm{Bern}(\delta))}_{=(\bar\delta-\gamma\delta)_+}\}$$

note that we assume $\gamma\geq 1$ & $\bar\delta \geq \delta$

$$= (\bar\delta-\gamma\delta)_+$$

Thus $\eta_\gamma(P_{Y|X}) = (\bar\delta-\gamma\delta)_+$ for $\gamma\geq 1$. This implies that:

- $\eta_{TV}(P_{Y|X}) = 1-2\delta$

- $\eta_\gamma(P_{Y|X}) = \frac{1}{\gamma}(\gamma\bar\delta-\delta)_+$ for $\gamma<1$.