

General Properties for f -divergence:

1- $D_f(p||q) \geq 0$

2- IF f is strictly convex at 1, then $D_f(p||q) = 0 \Rightarrow p=q$.

$\uparrow$
* 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

* Note that $p=q \Rightarrow D_f(p||q) = 0$

$x_i \neq x_j$

$\downarrow$
IF $\sum \lambda_i \cdot x_i = 1$

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

proof: $D_f(p||q) = \sum q(x_i) \cdot f\left(\frac{p(x_i)}{q(x_i)}\right) \geq f(1) \quad (*)$

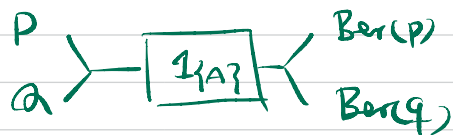
IF f is strictly convex at 1 $\Rightarrow f(1) < \sum q(x_i) \cdot f\left(\frac{p(x_i)}{q(x_i)}\right)$
(because $\sum q(x_i) \cdot \frac{p(x_i)}{q(x_i)} = 1$)

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

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 DP5 (which is about to be proved in Item 5), we have

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

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

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

↑
contradicts the
definition of
strict convexity.

3- $(p, q) \mapsto D_F(p \| q)$ is jointly convex

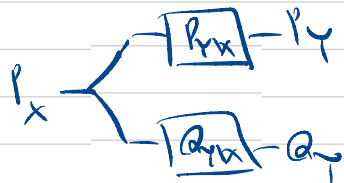
$$D_F(\lambda p_1 + \bar{\lambda} p_2 \| \lambda q_1 + \bar{\lambda} q_2) \leq \lambda D(p_1 \| q_1) + \bar{\lambda} D(q_2 \| q_2)$$

Define: $g(a, b) = b \cdot f(\frac{a}{b})$ [perspective of f]

If f is convex g is jointly convex

$$\begin{aligned} D_F(\lambda p_1 + \bar{\lambda} p_2 \| \lambda q_1 + \bar{\lambda} q_2) &= \underbrace{\sum (\lambda q_1 + \bar{\lambda} q_2) f\left(\frac{\lambda p_1 + \bar{\lambda} p_2}{\lambda q_1 + \bar{\lambda} q_2}\right)}_{\substack{\uparrow \\ \text{convex}}} \\ &= (\lambda b_1 + \bar{\lambda} b_2) f\left(\frac{\lambda a_1 + \bar{\lambda} a_2}{\lambda b_1 + \bar{\lambda} b_2}\right) \\ &= (\lambda b_1 + \bar{\lambda} b_2) f\left(\frac{\lambda b_1}{\lambda b_1 + \bar{\lambda} b_2} \frac{a_1}{b_1} + \frac{\bar{\lambda} b_2}{\lambda b_1 + \bar{\lambda} b_2} \frac{a_2}{b_2}\right) \\ &\leq (\lambda b_1 + \bar{\lambda} b_2) \left[\frac{\lambda b_1}{\lambda b_1 + \bar{\lambda} b_2} f\left(\frac{a_1}{b_1}\right) + \frac{\bar{\lambda} b_2}{\lambda b_1 + \bar{\lambda} b_2} f\left(\frac{a_2}{b_2}\right) \right] \\ &= \lambda \sum g(p_{i1}, q_{i1}) + \bar{\lambda} \sum g(p_{i2}, q_{i2}) \end{aligned}$$

4. Conditioning increase f -divergence

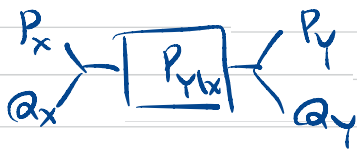


$$D_f(P_Y || Q_Y) \leq \sum_x P_X D_f(P_{Y|X=x} || Q_{Y|X=x})$$

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

$$\begin{aligned} \sum_x P_X(x) D_f(P_{Y|X} || Q_{Y|X}) &\stackrel{f}{\geq} D_f\left(\sum_x P_X(x) \cdot P_{Y|X} \parallel \sum_x P_X(x) \cdot Q_{Y|X}\right) \\ &= D_f(P_Y || Q_Y) \end{aligned}$$

5. DPI



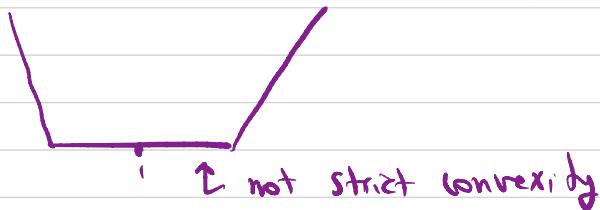
$$D_f(P_Y || Q_Y) \leq D_f(P_X || Q_X)$$

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

$$\begin{aligned}
D_f(p_x \| q_x) &= \sum_x q_x(x) \cdot f\left(\frac{p_x(x)}{q_x(x)}\right) \\
&= \sum_{x(y)} p_{Y|X}(y|x) \cdot q_x(x) \cdot f\left(\frac{p_x(x)}{q_x(x)} \cdot \frac{p_{Y|X}(y|x)}{p_{Y|X}(y|x)}\right) \\
&= \sum_{x(y)} q_Y(y) \cdot \overset{\substack{\text{backward} \\ \text{channel}}}{\downarrow} q_{X|Y}(x|y) \cdot f\left(\frac{p_{XY}(x,y)}{q_Y(y) \cdot q_{X|Y}(x|y)}\right) \\
&= \sum_y q_Y(y) \cdot \underbrace{\sum_x q_{XY}(x,y)}_{\text{Jensen}} \cdot f\left(\frac{p_{XY}(x,y)}{q_Y(y) \cdot q_{X|Y}(x|y)}\right) \\
&\geq \sum_y q_Y(y) \cdot f\left(\sum_x q_{X|Y}(x|y) \cdot \frac{p_{XY}(x,y)}{q_Y(y) \cdot q_{X|Y}(x|y)}\right) \\
&= \sum_y q_Y(y) \cdot f\left(\sum_x \frac{p_{XY}(x,y)}{q_Y(y)}\right) \\
&= \sum_y q_Y(y) \cdot f\left(\frac{p_Y(y)}{q_Y(y)}\right) \checkmark
\end{aligned}$$

Remark:

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



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



Hence, in general $E_r(p||q) = 0 \Rightarrow p = q$.

Example: $E_r(\text{Ber}(\frac{r-1}{2r}) || \text{Ber}(\frac{1}{r})) = \left(\frac{r-1}{2r} - \widetilde{r} \overbrace{(\frac{1}{r})}^{(r-1)} \right)_+ + \left(\frac{r-1}{2r} - \widetilde{r} \cdot \frac{1}{r} \right)_+$

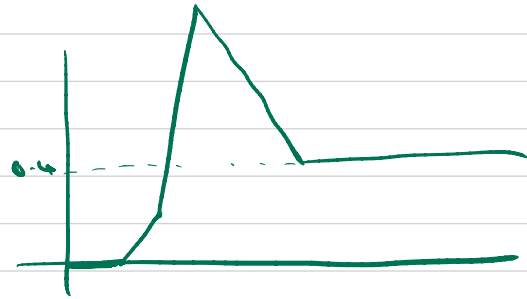
for $r > 1$

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

Example: $E_r(\text{Ber}(p) \parallel \text{Ber}(q)) = (p - rq)_+ + (\bar{p} - r\bar{q})_+ - (1-r)_+$

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

$$\begin{aligned} E_r(P \parallel Q) &= (0.1 - r \times 0.4)_+ + (0.2 - r \times 0.5)_+ + (0.3 - r \times 0.1)_+ \\ &\quad + (0.4 - r \times 0)_+ \\ &= (0.1 - 0.4r)_+ + (0.2 - 0.5r)_+ + (0.3 - 0.1r)_+ + 0.4 \end{aligned}$$



Exercise: $E_r(N(0, \sigma^2) \parallel N(\mu, \sigma^2)) = Q\left(\frac{\log r}{\beta} - \frac{\beta}{2}\right) - rQ\left(\frac{\log r}{\beta} + \frac{\beta}{2}\right)$

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

the same result holds even

$$N(\mu_1, \sigma^2 I_d) \parallel N(\mu_2, \sigma^2 I_d)$$

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:

$$1. E_\gamma(p||q) = \frac{1}{2} \sum |p(x) - \gamma q(x)| - \frac{1}{2} |1-\gamma|$$

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

$$\begin{aligned} \text{so: } E_\gamma(p||q) &= \sum (p(x) - \gamma q(x))_+ - (1-\gamma)_+ \\ &= \sum \frac{|p(x) - \gamma q(x)| + (p(x) - \gamma q(x))}{2} - \frac{|1-\gamma| + (1-\gamma)}{2} \\ &= \frac{1}{2} \sum |p(x) - \gamma q(x)| + \cancel{\frac{1-\gamma}{2}} - \frac{1}{2} |1-\gamma| - \cancel{\frac{1-\gamma}{2}} \end{aligned}$$

$$2- E_r(p||q) \stackrel{\text{For } r \geq 1}{=} \sup_A (p(A) - r q(A)) = p(A^*) - r q(A^*)$$

$$\text{For } \underline{r \leq 1}: \sup_A (r q(A) - p(A))$$

where

$$A^* = \{x: p(x) \geq r q(x)\}$$

proof requires two steps:

$$\textcircled{1} \text{ For any } A: p(A) - r q(A) \leq E_r(p||q).$$

$$\textcircled{2} \exists A^* \text{ such that } p(A^*) - r q(A^*) = E_r(p||q).$$

$$\begin{aligned} \text{L } p(A) - r q(A) &= \sum_{x \in A} (p(x) - r q(x)) \leq \sum_{x \in A} (p(x) - r q(x))_+ \\ &\quad \uparrow \text{Because we ignore all negative terms} \\ &\leq \sum_{x \in X} (p(x) - r q(x))_+ \\ &= E_r(p||q) \end{aligned}$$

because on A^* , we always $p(x) \geq r q(x)$

$$\begin{aligned}
 2- \quad p(A^*) - r q(A^*) &= \sum_{x \in A^*} (p(x) - r q(x)) = \sum_{x \in A^*} (p(x) - r q(x))_+ \\
 &= \sum_{x \in A^*} (p(x) - r q(x))_+ + \underbrace{\sum_{x \in A^{*c}} (p(x) - r q(x))}_{{=0}} \\
 &= \sum_x (p(x) - r q(x))_+ = E_r(p \| q).
 \end{aligned}$$

□

Properties of E_r -divergence:

- 1) Given p, q , $r \mapsto E_r(p \| q)$ is non-increasing & convex for $r \geq 1$
- 2 is non-decreasing & convex

- monotonicity is obvious:

Note that for $r < 1$
use the other def.

$$E_r(p||q) = \sum_{r < 1} (r q_i - p_i)_+$$

- convexity: $E_r(p||q) = \sup \underbrace{p(A) - r q(A)}_{\text{linear}} \rightarrow$ Supremum of linear function is convex

Direct approach:

$$\begin{aligned} E_{t r_1 + \bar{t} r_2}(p||q) &= \sup [p(A) - (t r_1 + \bar{t} r_2) q(A)] \\ &= \sup [p(A) - t r_1 q(A) - \bar{t} r_2 q(A)] \\ &= \sup [t (p(A) - r_1 q(A)) \\ &\quad + \bar{t} (p(A) - r_2 q(A))] \\ &\leq t \sup [p(A) - r_1 q(A)] \\ &\quad + \bar{t} \sup [p(A) - r_2 q(A)] \end{aligned}$$

2) Reciprocity: $E_r(p||q) = r \cdot E_{r^{-1}}(q||p)$

suppose $r \geq 1$: $E_r(p||q) = \sum (p - rq)_+$

& for $r < 1$: $E_r(p||q) = \sum (rq - p)_+$

$r \cdot E_{r^{-1}}(q||p) \stackrel{r^{-1}}{\rightarrow} r \cdot \left[\sum \left(\frac{1}{r} p - q \right)_+ \right] = \sum (p - rq)_+$

3) $\lim_{r \rightarrow 0} E_r(p||q) = 0$

$\lim_{r \rightarrow \infty} E_r(p||q) = 0$

$\downarrow$ If p & q are supported on the same alphabet