

Next, we need to define a family of distance measure between distributions. But before that, we need to discuss Jensen's inequality

Detour: Jensen's inequality:

Definition: * let f be a real-valued function. We say that f is convex

$$\text{if } f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \quad \forall x, y, \quad \forall \alpha \in [0,1] \quad (*)$$

* Strictly convex if $\nearrow$ strict inequality.

* Concave if $\geq$

* Strictly concave $>$

We can generalize (*):

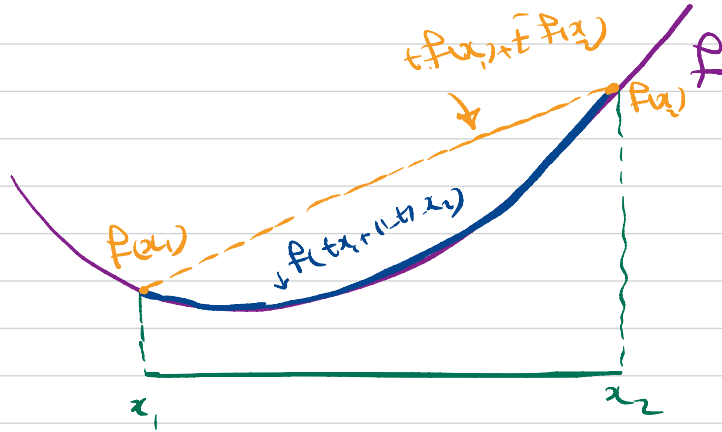
$$f(\sum p_i x_i) \leq \sum p_i f(x_i) \quad \text{for all } p_i \text{ such that } \sum p_i = 1. \quad (**)$$

For example:

$$\begin{aligned} f(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3) &= f(\alpha_1 x_1 + (1-\alpha_1) \underbrace{(\frac{\alpha_2}{1-\alpha_1} x_2 + \frac{\alpha_3}{1-\alpha_1} x_3)}) \\ &\leq \alpha_1 f(x_1) + (1-\alpha_1) f(\frac{\alpha_2}{1-\alpha_1} x_2 + \frac{\alpha_3}{1-\alpha_1} x_3) \\ &\leq \alpha_1 f(x_1) + (1-\alpha_1) \left[\frac{\alpha_2}{1-\alpha_1} f(x_2) + \frac{\alpha_3}{1-\alpha_1} f(x_3) \right] \\ &= \alpha_1 f(x_1) + \alpha_2 f(x_2) + \alpha_3 f(x_3) \end{aligned}$$

More general statement (**) can be proved by induction.

Geometric Interpretation:



line
↙

Any segment connecting any two points
lies above the function.

Theorem. If f is twice differentiable, then:

$$f \text{ is convex} \Leftrightarrow f''(x) \geq 0$$

$$f \text{ is st. convex} \Leftrightarrow f''(x) > 0$$

$$f \text{ is concave} \Leftrightarrow f''(x) \leq 0$$

$$f \text{ is st. concave} \Leftrightarrow f''(x) < 0$$

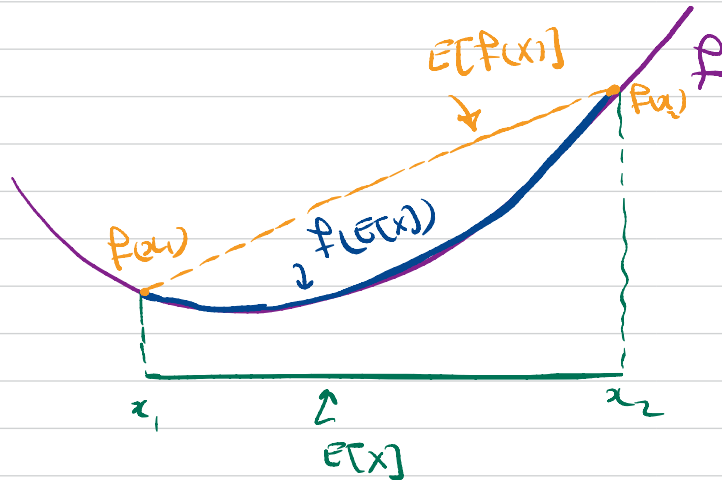
Jensen's Inequality:

Let f be a convex function & X be a random Variable.

Then:

$$E[f(X)] \geq f(E[X])$$

"=" iff
 $X = E[X]$



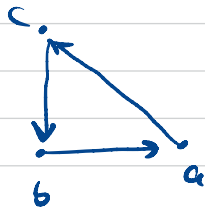
Distance measure among distributions

Question: Given two distributions p & q , how to quantify the "distance" between them?

In math, any appropriate measure of "distance" between any pair of objects, must satisfy the following:

- 1- $d(a, b) = 0 \iff a = b$
can be distributions
- 2- $d(a, b) = d(b, a)$ Symmetry
- 3- Triangle inequality: $d(a, c) \leq d(a, b) + d(b, c)$

Example:
Euclidean distance



We often ignore some of these properties to define "divergence" instead of distance.

I. Total Variation distance (TV)

$$TV(P, Q) := \frac{1}{2} \sum |P(x) - Q(x)| \quad \text{discrete}$$

pmf

$$:= \frac{1}{2} \int |P(x) - Q(x)| dx \quad \text{continuous}$$

pdf

Fact: TV is in fact a distance metric, meaning it satisfies:

1. $TV(P, Q) = 0 \iff P = Q$

2. $TV(P, Q) = TV(Q, P)$

3. $TV(P, R) \leq TV(P, Q) + TV(Q, R)$

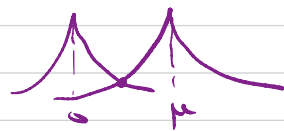
Triangle inequality of 1.1

why? $TV(P, R) = \frac{1}{2} \sum |P(x) - R(x)| \leq \frac{1}{2} \sum |P(x) - Q(x)| + |Q(x) - R(x)|$

Example:

$$\begin{aligned}TV(\text{Bernoulli}(p), \text{Bernoulli}(q)) &= \frac{1}{2} |p(0) - q(0)| + \frac{1}{2} |p(1) - q(1)| \\&= \frac{1}{2} |1-p - (1-q)| + \frac{1}{2} |p - q| \\&= |p - q|\end{aligned}$$

$$TV(L(0, b), L(\mu, b)) = 1 - e^{-\frac{|\mu|}{2b}}$$



$$\uparrow \frac{1}{2} \int \left| \frac{1}{2b} e^{-\frac{|x|}{b}} - \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \right| dx = \frac{1}{4b} \left[\int_{-\infty}^{\mu/2} \cdot + \int_{\mu/2}^{\infty} \cdot \right]$$

assume $\mu > 0$

Exercise:

$$\begin{aligned}TV(N(0, \sigma^2), N(\mu, \sigma^2)) &= \Phi\left(-\frac{\mu}{\sigma}\right) - \Phi\left(\frac{\mu}{\sigma}\right) \\&= 1 - 2\Phi\left(\frac{\mu}{\sigma}\right)\end{aligned}$$

Theorem. For any dist. $P \neq Q$ on support set A , we have:

$$TV(P, Q) = \sup_{A \subseteq \mathcal{A}} [P(A) - Q(A)]$$

* "Variational" expression
for TV

Proof. we need to prove:

$$\textcircled{1} \quad P(A) - Q(A) \leq TV(P, Q) \quad \forall A \in \mathcal{A}$$

$$\textcircled{2} \quad \exists A^* \subseteq \mathcal{A} \text{ s.t. } P(A^*) - Q(A^*) = TV(P, Q).$$

↓ Note $P(A) - Q(A) = \frac{1}{2} (P(A) - Q(A)) + \frac{1}{2} (Q(A) - P(A))$

$$\begin{aligned} \textcircled{1} \quad \text{Take arbitrary } A: P(A) - Q(A) &= \frac{1}{2} (P(A) - Q(A)) + \frac{1}{2} \underbrace{(Q(A) - P(A))}_{-P(A)} = \frac{1}{2} \sum_{x \in A} (P(x) - Q(x)) + \frac{1}{2} \sum_{A^c} (Q(x) - P(x)) \\ &\leq \frac{1}{2} \sum_A |P(x) - Q(x)| + \frac{1}{2} \sum_{A^c} |Q(x) - P(x)| \\ &= \frac{1}{2} \sum |P(x) - Q(x)| = TV(P, Q) \end{aligned}$$

What is the maximizing A ? $A^* = \{x: p(x) \geq q(x)\}$

$$\begin{aligned} \text{TV}(p, q) &= \frac{1}{2} \sum_x |p(x) - q(x)| = \frac{1}{2} \left[\sum_{x \in A^*} |p(x) - q(x)| + \sum_{x \in A^{*c}} |p(x) - q(x)| \right] \\ &= \frac{1}{2} \left[\sum_{x \in A^*} (p(x) - q(x)) + \sum_{x \in A^{*c}} (q(x) - p(x)) \right] \\ &= p(A^*) - q(A^*) \end{aligned} \quad \square$$

Another equivalent expression for TV:

$$\text{TV}(p, q) = \frac{1}{2} \sup_{f: \mathbb{R} \rightarrow [0, 1]} E_p[f(X)] - E_q[f(X)]$$

Proof is similar!

There are two other equivalent expressions for TV:

Supported on finite set

* Theorem: For any P, Q , we have:

$$TV(P, Q) = \inf_{\substack{P_{XY} \\ \begin{cases} P_X = P \\ P_Y = Q \end{cases}}} P_r(X \neq Y)$$

← minimization is over all "coupling" of P & Q .

* Theorem: For any P, Q , we have:

$$\begin{aligned} TV(P, Q) &= 1 - \int \min\{p(x), q(x)\} dx \\ &= 1 - \sum_x \min\{p(x), q(x)\} \end{aligned}$$

II. kullback-Leibler ^{KL} Divergence:

$$D(P \parallel Q) := E_P \left[\underset{\substack{\uparrow \\ \text{natural} \\ \text{log}}}{\log \frac{P(x)}{Q(x)}} \right] \overset{\substack{\text{discrete} \\ \downarrow}}{=} \sum_x P(x) \cdot \log \frac{P(x)}{Q(x)} \overset{\substack{\text{cont.} \\ \downarrow}}{=} \int_x p(x) \cdot \log \frac{p(x)}{q(x)}.$$

Example:

$$\begin{aligned} - D(\text{Bernoulli}(p) \parallel \text{Bernoulli}(q)) &= P(1) \cdot \log \frac{P(1)}{Q(1)} + P(0) \cdot \log \frac{P(0)}{Q(0)} \\ &= p \log \frac{p}{q} + (1-p) \cdot \log \frac{1-p}{1-q}. \end{aligned}$$

$$- D(L(0, b) \parallel L(\mu, b)) = \frac{b e^{-|\mu|/b} - b - |\mu|}{b}$$

$$\frac{1}{2b} \int e^{-\frac{|x|}{b}} \cdot \log \frac{e^{-|x|/b}}{e^{-|x+\mu|/b}} dx \overset{\substack{\uparrow \\ \text{assume } \mu > 0}}{=} \int_{-\infty}^0 + \int_0^{\mu} + \int_{\mu}^{\infty}$$

$$\begin{aligned}
 D(N(a, \sigma^2) || N(b, \sigma^2)) &= E_{N(a, \sigma^2)} \left[\log \frac{e^{-\frac{(x-a)^2}{2\sigma^2}}}{e^{-\frac{(x-b)^2}{2\sigma^2}}} \right] \\
 &= \frac{1}{2\sigma^2} \cdot E[(a-b)(2x-a-b)] \log e \\
 &= \frac{1}{2\sigma^2} (a-b) \cdot \log e \underbrace{E[(2x-a-b)]}_{a-b} \\
 &= \frac{(a-b)^2}{2\sigma^2}
 \end{aligned}$$

Exercise:

$$D(N(a, \sigma_1^2) || N(b, \sigma_2^2)) = \frac{1}{2} \log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2 + (a-b)^2}{2\sigma_2^2} - \frac{1}{2}$$

Properties of KL divergence:

1) $D(P||Q) \geq 0 \quad \forall P, Q$, " $=$ " if & only if $P=Q$

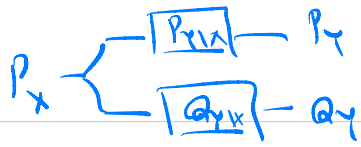
2) $D(P_{XY}||Q_{XY}) = D(P_X||Q_X) + \sum P_X(x) D(P_{Y|X=x}||Q_{Y|X=x})$ (chain rule)

$\downarrow$ P_X $\leftarrow$ Q_X $Q_{Y|X}$ $\downarrow$ marginal

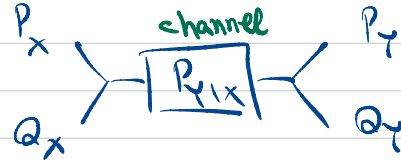
3) $D(P_{XY}||Q_{XY}) \geq D(P_X||Q_X)$ $\leftarrow$ or $D(Y||Q_Y)$ monotonicity

4) let P_Y & Q_Y be output dist. of $P_{Y|X}$ & $Q_{Y|X}$ with common input P_X . Then

$$D(P_Y||Q_Y) \leq \sum P_X(x) \cdot D(P_{Y|X=x}||Q_{Y|X=x})$$



5) Data Processing Inequality:



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

* Outputs of a channel are closer than the corresponding inputs.

Proofs. 1. $D(P \| Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

$$= \sum_x q(x) \cdot \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} = \sum q(x) \cdot f\left(\frac{p(x)}{q(x)}\right)$$

Jensen's ineq $\geq f\left(\sum_x q(x) \cdot \frac{p(x)}{q(x)}\right) = f(1) = 0$

$t \mapsto t \log t$

$$\begin{aligned}
2. \quad D(P_{XY} \| Q_{XY}) &= \sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{Q(x,y)} = \sum_{x,y} P(x) \cdot P(y|x) \cdot \log \frac{P(x) \cdot P(y|x)}{Q(x) \cdot Q(y|x)} \\
&= \sum_{x,y} P(x) \cdot P(y|x) \cdot \log \frac{P(x)}{Q(x)} + \sum_{x,y} P(x) \cdot P(y|x) \cdot \log \frac{P(y|x)}{Q(y|x)} \\
&= \sum_x P(x) \cdot \log \frac{P(x)}{Q(x)} + \sum_x P(x) \cdot \sum_y P(y|x) \cdot \log \frac{P(y|x)}{Q(y|x)}
\end{aligned}$$

$$3. \quad \text{Note } D(P_{XY} \| Q_{XY}) = D(P_X \| Q_X) + \underbrace{\sum P(x) \cdot D(P_{Y|X=x} \| Q_{Y|X=x})}_{\geq 0}$$

$$\Rightarrow D(P_{XY} \| Q_{XY}) \geq D(P_X \| Q_X)$$

4.

consider joint dist. $P_{XY} = P_X \cdot P_{Y|X}$ & $Q_{XY} = P_X \cdot Q_{Y|X}$

$$D(P_{XY} \| Q_{XY}) = \underbrace{D(P_X \| P_X)}_0 + \sum P(x) \cdot D(P_{Y|X=x} \| Q_{Y|X=x})$$

↑
chain rule

By monotonicity: $D(P_Y \| Q_Y) \leq D(P_{XY} \| Q_{XY}) = \sum p(x) D(P_{Y|X=x} \| Q_{Y|X=x})$

5. $D(P_{XY} \| Q_{XY}) = D(P_X \| Q_X) + \sum p(x) D(P_{Y|X=x} \| Q_{Y|X=x})$

By monotonicity: $D(P_Y \| Q_Y) \leq D(P_{XY} \| Q_{XY}) = D(P_X \| Q_X)$

* Note that

1) $D(P \| Q) \neq D(Q \| P)$ * Not symmetric *

2) $D(P \| R) \leq D(P \| Q) + D(Q \| R)$ No triangle inequality

* So KL-divergence is not a distance metric

Theorem. (Pinsker's Inequality.)

$$TV(P, Q) \leq \sqrt{\frac{D(P||Q)}{2}}$$

* You may see different constant in RHS, but 2 is optimal. Note that this doesn't mean that this is the best inequality connecting TV & D.

* Note that $0 \leq TV(P, Q) \leq 1$

↑
"=" $P=Q$

↑
if $P \neq Q$ are mutually singular; i.e.,

However;

$0 \leq D(P||Q) \leq \infty$
↑
"=" $P=Q$

they have non-overlapping support.

if & only if
 $\exists x \in \text{Supp}(P)$

but $x \notin \text{Supp}(Q)$.

Proof: We first show that it suffices to prove this inequality

for Bernoulli dist. Then, we prove it assuming P & Q are Bernoulli.

Equivalently

$\text{Supp}(Q) \subseteq \text{Supp}(P)$

I. consider the following channel:



Note that when $X \sim P \Rightarrow Y \sim \text{Ber}(P(E))$

where $E = \{x: P(x) \geq Q(x)\}$

$X \sim Q \Rightarrow Y \sim \text{Ber}(Q(E))$

Because

$$P_{Y=1} = \sum_x P(X=x) P(Y=1|X=x)$$

$$= \sum_x P(x) \cdot \underbrace{P_{Y=1|X=x}}_{1_{\{x \in E\}}} = P(E)$$

DPI $\downarrow$

$$D(P||Q) \geq D(\text{Ber}(P(E)) || \text{Ber}(Q(E)))$$

Pinsker's ine
for Bernoulli $\geq$

$$2 \text{TV}^2(\text{Ber}(P(E)), \text{Ber}(Q(E))) = 2 \cdot \frac{(P(E) - Q(E))^2}{2} = 2 \text{TV}^2(P, Q)$$

So we only need to prove it for binary.

II. Suppose $P = \text{Ber}(p)$ & $Q = \text{Ber}(q)$. WLOG, we assume $p \geq q$.

[Note that $D(\text{Ber}(p) || \text{Ber}(q)) = D(\text{Ber}(\bar{p}) || \text{Ber}(\bar{q}))$ so we can always assume $p \geq q$]

$$D(P||Q) = P \log \frac{P}{Q} + \bar{P} \log \frac{\bar{P}}{\bar{Q}} \quad \& \quad TV(P, Q) = (P - Q)$$

Define,

$$g(P, q) := P \log \frac{P}{q} + \bar{P} \log \frac{\bar{P}}{\bar{q}} - 2(P - q)^2. \quad \text{we need to show that } g(P, q) \geq 0 \quad \forall P, q.$$

$$\frac{\partial g}{\partial q} = -\frac{P}{q} + \frac{\bar{P}}{\bar{q}} - 4(P - q) = \underbrace{\frac{q - P}{q(1 - q)}}_{\leq 0} - 4(P - q) \leq 0$$

So $q \mapsto g(P, q)$ is decreasing for $q \leq P \Rightarrow g(P, q) \geq g(P, P) = 0$



* Important Observation: a relationship between TV & KL
was obtained just by looking at the binary case!

* We discussed two measures of "distance": TV & KL.

We can construct a family of other divergences:

Def. Let $f: (0, \infty) \rightarrow \mathbb{R}$ be a **convex** function satisfying **$f(1) = 0$** . Given two distributions p & q , their f -divergence is:

$$D_f(p \parallel q) := E_Q \left[f \left(\frac{p(x)}{q(x)} \right) \right]$$

$$\stackrel{\text{dis}}{=} \sum_x q(x) \cdot f \left(\frac{p(x)}{q(x)} \right) \quad \begin{array}{l} \text{pmf of} \\ p \& q \end{array}$$

$$= \int q(x) \cdot f \left(\frac{p(x)}{q(x)} \right) dx \quad \begin{array}{l} \text{pdf of} \\ p \& q \end{array}$$

Examples:

$$\begin{aligned} 1. \quad F(t) &= \frac{1}{2} |t-1| & : \quad D_F(p||q) &= E_a \left[\frac{1}{2} \left| \frac{p(x)}{q(x)} - 1 \right| \right] \\ & & \text{discrete} \quad \Downarrow & \\ & & &= \frac{1}{2} \cdot \sum q(x) \cdot \left| \frac{p(x)}{q(x)} - 1 \right| \\ & & &= \frac{1}{2} \sum |p(x) - q(x)| \\ & & &= TV(p, q). \end{aligned}$$

$$2. \quad F(t) = t \log t \quad : \quad D_F(p||q) = D(p||q)$$

$$\begin{aligned}
 3- \quad f(t) &= (\sqrt{t}-1)^2 : \quad D_f(p||q) = \sum_{\alpha} q(\alpha) \cdot \left(\sqrt{\frac{p(\alpha)}{q(\alpha)}} - 1 \right)^2 \\
 &= \sum \left(p(\alpha) + q(\alpha) - 2\sqrt{p(\alpha) \cdot q(\alpha)} \right) \\
 &= 2 - 2 \sum_{\alpha} \sqrt{p(\alpha) \cdot q(\alpha)} \\
 &= H^2(p, q)
 \end{aligned}$$

Squared Hellinger distance

can be obtained by $f(t) = 2(1 - \sqrt{t})$

$$4- \quad f(t) = (t-1)^2 \quad \chi^2 \text{ divergence}$$

$$5- \quad f(t) = \frac{t^\alpha - 1}{\alpha - 1} \quad \text{for } \alpha \in (0, 1) \cup (1, \infty)$$

χ^α divergence

Same thing
for $f(t) = t^2 - 1$.

$$\text{or } f(t) = \frac{1}{t} - t$$

$$\chi^\alpha(p||q) = \frac{1}{\alpha-1} \left[\sum_x p(x)^\alpha \cdot q(x)^{1-\alpha} - 1 \right]$$

$\alpha = 1/2 \rightarrow$ Squared Hellinger distances

We are often interested in a function of χ^α :

$$D_\alpha(p||q) := \frac{1}{\alpha-1} \left[\log \left(1 + (\alpha-1) \cdot \chi^\alpha(p||q) \right) \right]$$

* Rényi divergence *

$$6 \quad f(t) = t \log t - (1+t) \cdot \log \left(\frac{1+t}{2} \right)$$

$$D_F(p||q) = D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})$$

$\uparrow$ Symmetric

7. Hockey-Stick divergence:
or E_r -divergence

$$E_r(p||q) := D_{f_r}(p||q)$$

Hence,
$$E_r(p||q) \stackrel{r \geq 1}{=} \sum q(x) \cdot \left(\frac{p(x)}{q(x)} - r \right)_+ = \sum (p(x) - r q(x))_+$$

&

$$\begin{aligned} E_r(p||q) &\stackrel{r < 1}{=} \sum q(x) \left(\frac{p(x)}{q(x)} - r \right)_+ - 1 + r = \sum (p(x) - r q(x))_+ - \underbrace{(1-r)}_{= \sum (p(x) - q(x))_+} \\ &= \sum q(x) (r - \frac{p(x)}{q(x)})_+ = \sum (r q(x) - p(x))_+ \end{aligned}$$

we could equivalently define:
$$f_r(t) = \begin{cases} (t-r)_+ & r \geq 1 \\ (r-t)_+ & r < 1 \end{cases}$$

