Statistical Distances

Heuna Kim

Examples of statistical distances

- Kolomogorov-Smirnov Statistic
- Kullback Leibler Divergence
- F-divergence and Bergmann divergence
- Jensen-Shannon Divergence
- Wasserstein Distance (Earthmover’s distance)
- Mahalanobis Distance and Cook’s distance

Definition of Metrics

A metric on a set X is a function (called *distance function* or simply *distance*)

$$d : X \times X \to [0, \infty),$$

where $[0, \infty)$ is the set of non-negative real numbers and for all $x, y, z \in X$, the following three axioms are satisfied

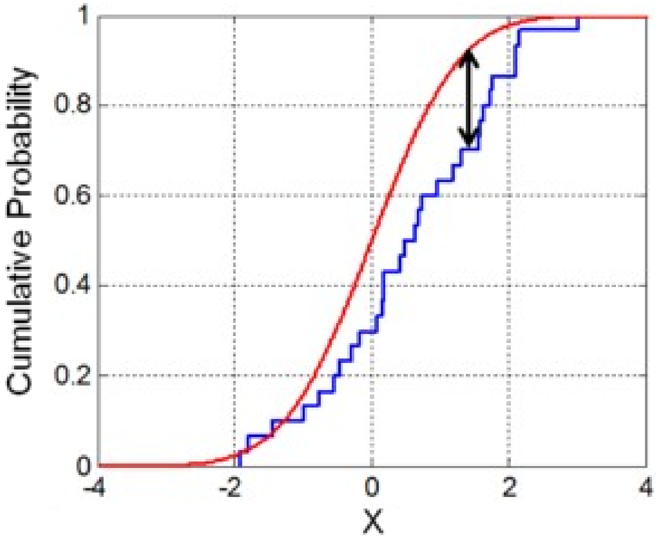
1. $d(x, y) = 0 \Leftrightarrow x = y$ — identity of indiscernibles
2. $d(x, y) = d(y, x)$ — symmetry
3. $d(x, y) \leq d(x, z) + d(z, y)$ — triangle inequality

Statistical distances don't satisfy these all.

e.g.) cosine distance is a metric.

Kolomogorov-Smirnov Statistic

- K-S Test: are two samples drawn from populations of the same distribution (univariate)?



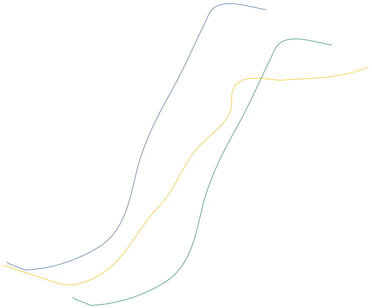
The Kolmogorov–Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)|$$

$$\sqrt{n} D_n > K_\alpha,$$

where K_α is found from

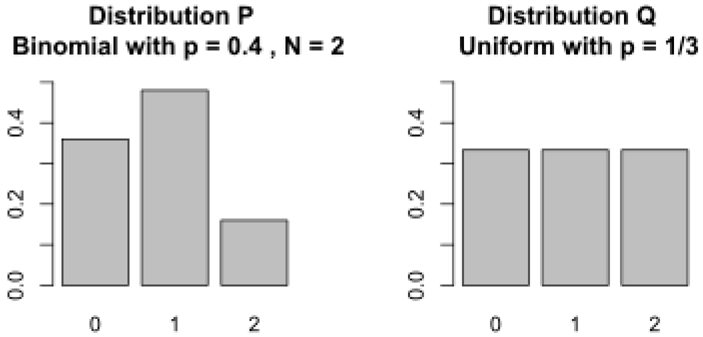
$$\Pr(K \leq K_\alpha) = 1 - \alpha.$$



Total Variation Distance

$$\delta(P,Q)=\sup_{A\in\mathcal{F}}\left|P(A)-Q(A)\right|.$$

$\mathcal{F}$: all possible events (subsets of a sample space)



x	0	1	2
Distribution *P*(x)	$9/25$	$12/25$	$4/25$
Distribution *Q*(x)	$1/3$	$1/3$	$1/3$

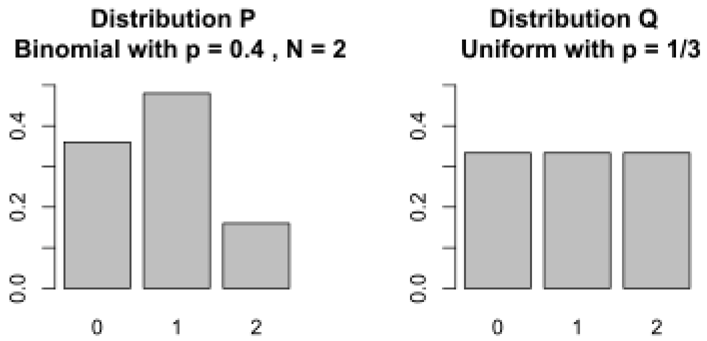
Kullback Leibler Divergence

- Relative entropy or information gain

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

$$= -\sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

$$= \mathrm{H}(P,Q) - \mathrm{H}(P)$$



$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

x	0	1	2
Distribution *P*(x)	$9/25$	$12/25$	$4/25$
Distribution *Q*(x)	$1/3$	$1/3$	$1/3$

To Note about KL-Divergence

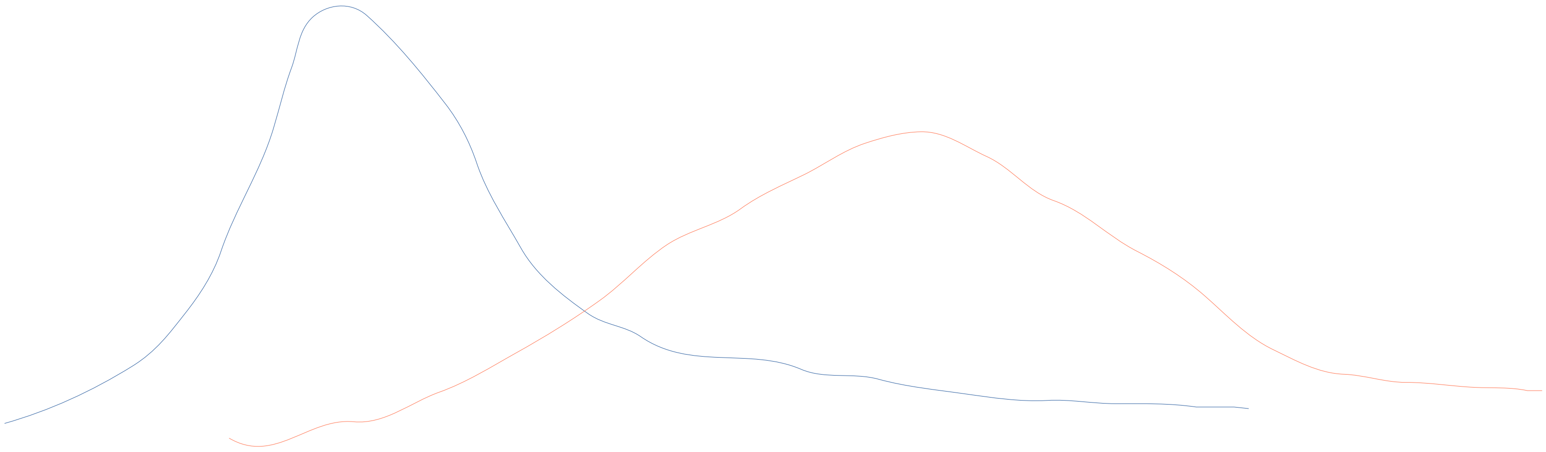
- Not a metric
  - Why?
- f-divergence

$$D_f(P \parallel Q) \equiv \int_{\Omega} f\left(\frac{dP}{dQ}\right) dQ.$$

- Bergman-divergence

  - **Convexity:** $D_F(p,q)$ is convex in its first argument, but not necessarily in the second argument

Exercise



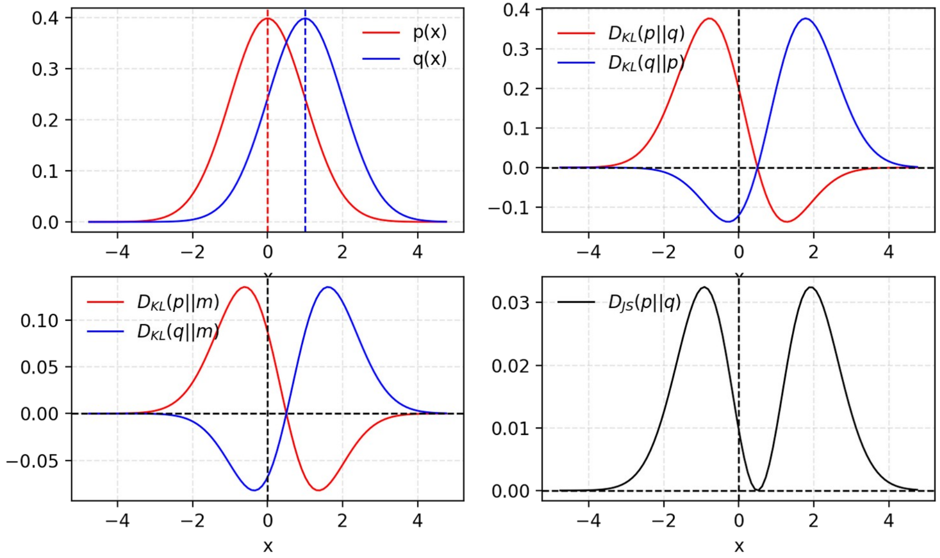
Jensen-Shannon divergence

- A symmetrized version of KL divergence

$$D_{\mathrm{JS}} = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M)$$

where M is the average of the two distributions,

$$M = \frac{1}{2}(P + Q).$$



Adversarial loss and JS div

$$
\begin{aligned}
\min_G \max_D L(D, G) &= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \\
&= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x)]
\end{aligned}
$$

When both G and D are at their optimal values, we have $p_g = p_r$ and $D^*(x) = 1/2$ and the loss function becomes:

$$
\begin{aligned}
L(G, D^*) &= \int_x \Big(p_r(x) \log(D^*(x)) + p_g(x) \log(1 - D^*(x)) \Big) dx \\
&= \log \frac{1}{2} \int_x p_r(x) dx + \log \frac{1}{2} \int_x p_g(x) dx \\
&= -2 \log 2
\end{aligned}
$$

$$
\begin{aligned}
D_{JS}(p_r \| p_g) =& \frac{1}{2} D_{KL}(p_r || \frac{p_r + p_g}{2}) + \frac{1}{2} D_{KL}(p_g || \frac{p_r + p_g}{2}) \\
=& \frac{1}{2} \Big(\log 2 + \int_x p_r(x) \log \frac{p_r(x)}{p_r + p_g(x)} dx \Big) + \\
& \frac{1}{2} \Big(\log 2 + \int_x p_g(x) \log \frac{p_g(x)}{p_r + p_g(x)} dx \Big) \\
=& \frac{1}{2} \Big(\log 4 + L(G, D^*) \Big)
\end{aligned}
$$

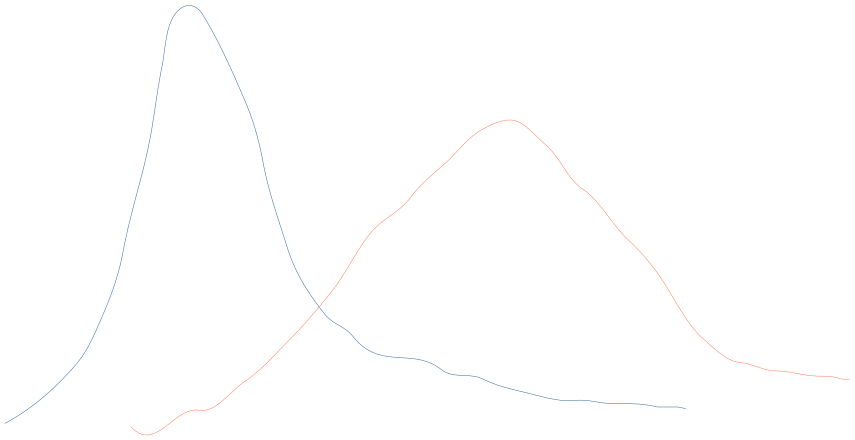
Wasserstein Metric

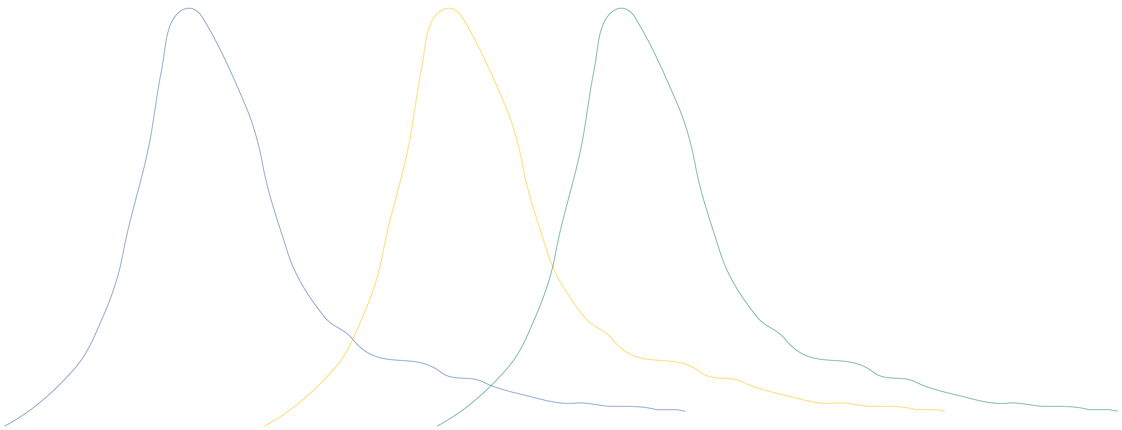
- Earthmover's distance and optimal transport plan

The p^{th} **Wasserstein distance** between two probability measures μ and ν in $P_p(M)$ is defined as

$$W_p(\mu,\nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x,y)^p \, \mathrm{d}\gamma(x,y) \right)^{1/p},$$

where $\Gamma(\mu,\nu)$ denotes the collection of all measures on $M \times M$ with marginals μ and ν on the first and second factors respectively.





Kantorovich-Rubinstein duality

If $\|f\|_L \leq K,$ and both distributions have bounded supports

$$W(p_r, p_g) = \frac{1}{K} \sup_{\|f\|_L \leq K} \mathbb{E}_{x \sim p_r}[f(x)] - \mathbb{E}_{x \sim p_g}[f(x)]$$

To note about Wasserstein distance

- Metric
- If two distribtions are multidimentional gaussian, equivalent to frechet inception distance.

$$\mathrm{FID} = |\mu - \mu_w|^2 + \mathrm{tr}(\Sigma + \Sigma_w - 2(\Sigma\Sigma_w)^{1/2}).$$

- Cf. frechet distance == dogwalker's distance
  - For similarity of two curves

Mahalanobis distance

- Unit-less scale-invariant metric

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^{\mathsf{T}} \mathbf{S}^{-1} (\vec{x} - \vec{y})}.$$

- Bergmann divergence
- Assume bell shapes
- cf. Cook’s distance:

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{ps^2}$$

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}} \quad \underset{p\times 1}{\boldsymbol{\beta}} \quad + \quad \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}\right)$ is the error term, $\boldsymbol{\beta} = [\beta_0 \, \beta_1 \ldots \beta_{p-1}]$

where $\widehat{y}_{j(i)}$ is the fitted response value obtained when excluding i, and $s^2 = \dfrac{\mathbf{e}^\top \mathbf{e}}{n - p}$