# Insurance, biases, discrimination & fairness

Arthur Charpentier

2024

# – Part 2 –

# Models

**Proposition 4.1: Law of Large Numbers (1)**

Consider an infinite collection of i.i.d. random variables $Y, Y_1, Y_2, \cdots, Y_n, \cdots$ in a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$, then

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_i \in \mathcal{A})}_{\text{(empirical) frequency}} \xrightarrow{\text{a.s.}} \underbrace{\mathbb{P}(\{Y \in \mathcal{A}\})}_{\text{probability}} = \mathbb{P}[Y \in \mathcal{A}], \text{ as } n \to \infty.$$

"*law of the unconscious statistician*" (Ross (2014) and Casella and Berger (1990)), "*statisticians make liberal use of conditioning arguments to shorten what would otherwise be long proofs*," Proschan and Presnell (1998)

$$\mathbb{P}(Y \in \mathcal{A}|X = x) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(\{Y \in \mathcal{A}\} \cap \{|X - x| \leq \epsilon\})}{\mathbb{P}(\{|X - x| \leq \epsilon\})} = \lim_{\epsilon \to 0} \mathbb{P}(Y \in \mathcal{A}||X - x| \leq \epsilon).$$

# Statistical Learning

This frequentist approach is unable to make sense of the probability of a "single singular event", as noted by von Mises (1928, 1939).

"*When we speak of the 'probability of death', the exact meaning of this expression can be defined in the following way only. We must not think of an individual, but of a certain class as a whole, e.g., 'all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations'. A probability of death is attached to the class of men or to another class that can be defined in a similar way. We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning for us at all.*"

**Definition 4.1: Loss $\ell$**

A loss function $\ell$ is a function defined on $\mathcal{Y} \times \mathcal{Y}$ such that $\ell(y, y') \geq 0$ and $\ell(y, y) = 0$.

**Definition 4.2: Risk $\mathcal{R}$**

For a fitted model $\widehat{m}$, its risk is

$$\mathcal{R}(\widehat{m}) = \mathbb{E}_{\mathbb{P}}\Big[\ell\big(Y, \widehat{m}(\boldsymbol{X})\big)\Big] = \int \ell(y, \widehat{m}(\boldsymbol{x}))\mathrm{d}\mathbb{P}(y, \boldsymbol{x}).$$

**Definition 4.3: Empirical risk $\widehat{\mathcal{R}}_n$**

Given a sample $\{(y_i, \boldsymbol{x}_i), i = 1, \cdots, n\}$, define the empirical risk

$$\widehat{\mathcal{R}}_n(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{m}(\boldsymbol{x}_i), y_i\right).$$

› Following Vapnik (1991), the "empirical risk minimization principle" states that the learning algorithm $\widehat{m}^*$ is

$$\widehat{m}^* = \underset{\widehat{m}\in\mathcal{M}}{\operatorname{argmin}}\{\widehat{\mathcal{R}}_n(\widehat{m})\}.$$

**Definition 4.4: Minimax Estimator**

An estimator is minimax if its maximal risk is minimal among all estimators. The minimax risk is, for a set $\mathcal{P}$ of distributions,

$$R(\mathcal{P}) = \inf_{\widehat{m}\in\mathcal{M}} \left\{ \sup_{(\boldsymbol{X},Y)\sim\mathbb{P}\in\mathcal{P}} \left\{ \mathbb{E}_{\mathbb{P}}\left[\ell(\widehat{m}(\boldsymbol{X}), Y)\right] \right\} \right\}.$$

**Proposition 4.2: Optimal Decision, "*Bayes decision rule*"**

For each $\boldsymbol{x}$ choose the prediction $m_{\boldsymbol{x}}^\star$ that minimizes the conditional expected loss,

$$m_{\boldsymbol{x}}^\star \in \underset{z\in\mathcal{Y}}{\operatorname{argmin}} \left\{ \int \ell(y,z) \mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) \right\}$$

- It is straightforward since $\mathrm{d}\mathbb{P}_{Y,\boldsymbol{X}}(y,\boldsymbol{x}) = \mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) \cdot \mathrm{d}\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{x})$,

$$\mathcal{R}(\widehat{m}) = \int \Big[ \int \ell(y, \widehat{m}(\boldsymbol{x})) \mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) \Big] \mathrm{d}\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{x}).$$

by definition, $m_{\boldsymbol{x}}^\star$ minimizes the term in blue, i.e., for any $\widehat{m}$

$$\mathcal{R}(\widehat{m}) \geq \int \Big[ \int \ell(y, m_{\boldsymbol{x}}^\star) \mathrm{d}\mathbb{P}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) \Big] \mathrm{d}\mathbb{P}_{\boldsymbol{X}}(\boldsymbol{x}) = \mathcal{R}(m^\star).$$

› It is coined "Bayes decision rule" because the conditional distribution $Y|\boldsymbol{X}$ is sometimes be referred to as the "posterior" distribution of $Y$ given data $\boldsymbol{X}$.

**Definition 4.5: Misclassification loss, $\ell_{0/1}$**

$$\ell_{0/1}(y, \widehat{y}) = \mathbf{1}(y \neq \widehat{y}).$$

In the case of a binary classifier, observe that

$$\begin{aligned}
\mathcal{R}(\widehat{m}) &= \mathbb{E}[\ell(\widehat{m}(\boldsymbol{X}), Y)] = \mathbb{E}\big[\mathbb{E}[\ell(\widehat{m}(\boldsymbol{X}), Y) \mid \boldsymbol{X}]\big] \\
&= \mathbb{E}\big[\ \ell(\widehat{m}(\boldsymbol{X}), 1) \cdot \mathbb{P}(Y = 1 \mid \boldsymbol{X}) + \ell(\widehat{m}(\boldsymbol{X}), 0) \cdot \mathbb{P}(Y = 0 \mid \boldsymbol{X})\big] \\
&= \mathbb{E}\big[\mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] \cdot \mu(\boldsymbol{X}) + \mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 0] \cdot (1 - \mu(\boldsymbol{X}))\ \big] \\
&= \mathbb{E}\big[\mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] \cdot \mu(X) + (1 - \mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1]) \cdot (1 - \mu(\boldsymbol{X}))\big] \\
&= \mathbb{E}\big[\ \mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] \cdot (2\mu(\boldsymbol{X}) - 1) + 1 - \mu(\boldsymbol{X})\big].
\end{aligned}$$

Since $\widehat{m} : \mathcal{X} \to \{0, 1\}$, this expectation is minimized by choosing $\widehat{m} = m^\star$, where

$$m^\star(\boldsymbol{x}) = \mathbf{1}(\mu(\boldsymbol{x}) > 1/2) = \begin{cases} 1 \text{ if } \mu(\boldsymbol{x}) > 1/2 \\ 0 \text{ if } \mu(\boldsymbol{x}) \leq 1/2 \end{cases}$$

The optimal risk ("Bayes risk") is $\mathcal{R}(m^\star) = \inf_m \{\mathcal{R}(m)\}$.

**Definition 4.6: Excess of risk of $\widehat{m}$**

For any model $\widehat{m}$, the excess of risk is $\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star)$.

For a classifier

$$\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star) = \mathbb{E}\big[|2\mu(\boldsymbol{X}) - 1| \cdot \mathbf{1}(\widehat{m}(\boldsymbol{X}) \neq m^\star(\boldsymbol{X}))\big].$$

Since we do not know $\mu$ consider a classifier based on $\widehat{m}$ .....

**Definition 4.7: Plug-in Estimator**

Estimate $\widehat{\mu}$ and use, as a classifier, $\mathbf{1}(\widehat{\mu}(\boldsymbol{x}) > 1/2)$.

**Proposition 4.3**

For any model $\widehat{\mu}$, the risk of the plug-in classifier $\widehat{m}(\boldsymbol{x}) = \mathbf{1}(\widehat{\mu}(\boldsymbol{x}) > 1/2)$ satisfies

$$\mathcal{R}(\widehat{m}) - \mathcal{R}(m^\star) \leq 2\mathbb{E}\big|\mu(\boldsymbol{X}) - \widehat{\mu}(\boldsymbol{X})\big|.$$

**Proof** We have seen that

$$\mathcal{R}(\widehat{m}) - \mathcal{R}\left(m^*\right) = \mathbb{E}\big(1[\widehat{m}(\boldsymbol{X}) \neq 1] - 1\left[m^*(\boldsymbol{X}) \neq 1\right]\big) \cdot (2\mu(\boldsymbol{X}) - 1).$$

# Statistical Learning

But

$$\begin{aligned}
&\left(\mathbf{1}[\widehat{m}(\boldsymbol{X}) \neq 1] - \mathbf{1}\left[m^*(\boldsymbol{X}) \neq 1\right]\right)(2\mu(\boldsymbol{X}) - 1)\\
&= \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]\left(1[\widehat{m}(\boldsymbol{X}) \neq 1] - 1\left[m^*(\boldsymbol{X}) \neq 1\right]\right)(2\mu(\boldsymbol{X}) - 1)\\
&= \begin{cases} \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right](2\mu(\boldsymbol{X}) - 1) & \text{if } 2\mu(\boldsymbol{X}) - 1 > 0, \\ \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right](-1)(2\mu(\boldsymbol{X}) - 1) & \text{if } 2\mu(\boldsymbol{X}) - 1 \leq 0. \end{cases}
\end{aligned}$$

(from the definition of $m^*$ )

$$= \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right] \cdot |2\mu(\boldsymbol{X}) - 1|,$$

$$\mathcal{R}(\widehat{m}) - \mathcal{R}\left(m^*\right) = \mathbb{E}(\mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]) \cdot 2|\mu(\boldsymbol{X}) - 1/2|.$$

If $\widehat{m}(\boldsymbol{x}) \neq m^*(\boldsymbol{x})$, it means that $\widehat{\mu}(\boldsymbol{x})$ and $\mu(\boldsymbol{x})$ lie on opposite sides of $1/2$,

$$|\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})| = |\widehat{\mu}(\boldsymbol{x}) - 1/2| + \underbrace{|1/2 - \mu(\boldsymbol{x})|}_{\geq 0} \geq |\widehat{\mu}(\boldsymbol{x}) - 1/2|$$

i.e.

$$|\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})| \geq |\widehat{\mu}(\boldsymbol{x}) - 1/2| \cdot \mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]$$

which is also valid when $\widehat{m}(\boldsymbol{x}) = m^*(\boldsymbol{x})$, thus

$$\mathcal{R}(\widehat{m}) - \mathcal{R}\left(m^*\right) = 2\mathbb{E}(\mathbf{1}\left[\widehat{m}(\boldsymbol{X}) \neq m^*(\boldsymbol{X})\right]) \cdot |\mu(\boldsymbol{X}) - 1/2| \leq 2\mathbb{E}\big[|\widehat{\mu}(\boldsymbol{X}) - \mu(\boldsymbol{X})|\big].$$

› This $\ell_{0/1}$ loss function may be difficult to directly optimize, as shown in Bartlett et al. (2006). One could consider some surrogate loss $\tilde{\ell}$ which is easier to optimize.

**Definition 4.8: Elicitation, Brier (1950), Good (1952)**

A statistical functional $\mathcal{I}(Y)$ is said to be elicitable if it minimizes expected loss for some loss function $s$, in the sense that

$$\mathcal{I}(Y) = \underset{y \in \mathbb{R}}{\text{argmin}} \{\mathbb{E}[s(Y, y)]\}$$

- Important properties for risk measures and backtesting. "*The elicitability of a risk measure means that the risk measure can be obtained by minimizing the expectation of a forecasting objective function. Elicitability is closely related to backtesting, whose objective is to evaluate the performance of a risk forecasting model. If a risk measure is elicitable, then the sample average forecasting error based on the objective function can be used for backtesting the risk measure,*" He et al. (2022)

# Loss Functions

- In a regression problem, a quadratic loss function $\ell_2$ is used

**Definition 4.9: Quadratic loss, $\ell_2$**

$\ell_2(y, \widehat{y}) = (y - \widehat{y})^2$, and the risk is then $\mathcal{R}_2(\widehat{m}) = \mathbb{E}\Big[ (Y - \widehat{m}(\boldsymbol{X}))^2 \Big]$.

- Observe that

$$\mathbb{E}[Y] = \underset{m \in \mathbb{R}}{\operatorname{argmin}} \{\mathcal{R}_2(m)\} = \underset{m \in \mathbb{R}}{\operatorname{argmin}} \Big\{ \mathbb{E}\Big[ \ell_2(Y, m) \Big] \Big\}.$$

The expected value is "ellicitable" (for the $s = \ell_2$ loss).

The empirical risk minimizer is the "least-square" estimate.

# Loss Functions

› See Huttegger (2013), explaining why the expected value is also called "best estimate".

› Up to a monotonic transformation (the square root function), the distance here is the expectation of the quadratic loss function. With the terminology of Angrist and Pischke (2009), the regression function $\mu$ is the function of $\boldsymbol{x}$ that serves as "*the best predictor of $y$, in the mean-squared error sense*."

**Proposition 4.4: Optimal Decision, "*Bayes decision rule*"**

For the quadratic loss $\ell_2$, Bayes decision rule is the (conditional) expected value, $m^\star_{\boldsymbol{x}} = \mathbb{E}\big[Y|\boldsymbol{X}=\boldsymbol{x}\big] = \mu(\boldsymbol{x})$.

# Loss Functions

**Definition 4.10: Inner product**

An inner product on $\mathcal{H}$ is the application $(f, g) \mapsto \langle f, g \rangle_{\mathcal{H}}$ (taking value in $\mathbb{R}$) bilinear, symmetric, definite positive:

- $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
- $\langle \alpha f + \beta g, h \rangle_{\mathcal{H}} = \alpha \langle f, h \rangle_{\mathcal{H}} + \beta \langle g, h \rangle_{\mathcal{H}}$
- $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

**Example** : $\mathcal{H} = \mathbb{R}^n$, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y}$

**Example** : $\mathcal{H} = \mathbb{R}^n$, let $\boldsymbol{\Sigma}$ denote some symmetric $n \times n$ positive definite matrix. Then

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{\Sigma}} = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y} \text{ is an inner product on } \mathbb{R}^n.$$

**Example** : $\mathcal{H} = \ell^2 = \left\{ u : \sum_{i=1}^{\infty} u_i^2 < \infty \right\}$, $\langle u, v \rangle = \sum_{i=1}^{\infty} u_i v_i$

# Loss Functions

**Example** : $\mathcal{H} = L^2(\mu) = \left\{ f : \int f(x)^2 d\mu(x) < \infty \right\}$, $\langle f, g \rangle = \int f(x)g(x)d\mu(x)$

**Example** : Consider the vector space $\mathcal{V}$ that consists of all real-valued random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Given $k \in [1, \infty)$, define

$$\|X\|_k = \left[ \mathbb{E}\left( |X|^k \right) \right]^{1/k}.$$

to go further (for more details on Lebesgue spaces, and $L^2$)

# Loss Functions

A norm $\|\cdot\|$, in $\mathbb{R}^n$, satisfies

- homogeneity, $\|a\vec{\boldsymbol{u}}\| = |a| \cdot \|\vec{\boldsymbol{u}}\|$, $\forall a$
- triangle inequality, $\|\vec{\boldsymbol{u}} + \vec{\boldsymbol{v}}\| \leq \|\vec{\boldsymbol{u}}\| + \|\vec{\boldsymbol{v}}\|$
- positivity, $\|\vec{\boldsymbol{u}}\| \geq 0$
- definiteness, $\|\vec{\boldsymbol{u}}\| = 0 \iff \vec{\boldsymbol{u}} = \vec{\boldsymbol{0}}$

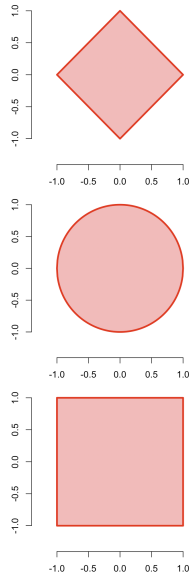
$\ell_1$ norm: $\|\boldsymbol{x}\|_{\ell_1} = |x_1| + \cdots + |x_n|$,

$\ell_2$ norm: $\|\boldsymbol{x}\|_{\ell_2} = \sqrt{x_1^2 + \cdots + x_n^2}$,

$\ell_p$ norm: with $p \geq 1$, $\|\boldsymbol{x}\|_{\ell_p} = \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p}$

$\ell_\infty$ norm: $\|\boldsymbol{x}\|_{\ell_\infty} = \max\{x_i\}$

Unit balls ($\|\boldsymbol{x}\|_{\ell_p} \leq 1$) are convex sets



# Loss Functions

**Proposition 4.5: Gradient of $\ell_p$ norms**

$$\frac{\partial}{\partial x_j}\|\boldsymbol{x}\|_{\ell_p} = \frac{1}{p}\left(\sum_i |x_i|^p\right)^{\frac{1}{p}-1} \cdot p|x_j|^{p-1}\text{sign}(x_j) = \left(\frac{|x_j|}{\|\boldsymbol{x}\|_{\ell_p}}\right)^{p-1}\text{sign}(x_j).$$

$$\begin{aligned}\frac{\partial}{\partial x_j}||\boldsymbol{x}||_{\ell_p} &= \frac{\partial}{\partial x_j}\left(\sum_{i=1}^n |x_i|^p\right)^{1/p} = \frac{1}{p}\left(\sum_{i=1}^n |x_i|^p\right)^{(1/p)-1}\frac{\partial}{\partial x_j}\left(\sum_{i=1}^n |x_i|^p\right)\\ &= \left[\left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}\right]^{1-p}\sum_{i=1}^n |x_i|^{p-1}\delta_{ij}\frac{x_i}{|x_i|} = \left(\frac{|x_j|}{\|\boldsymbol{x}\|_{\ell_p}}\right)^{p-1}\text{sign}(x_j).\end{aligned}$$

# Loss Functions

**Definition 4.11: Quantile loss, $\ell_{\mathbf{q},\alpha}$**

The quantile loss $\ell_{\mathbf{q},\alpha}$ for some $\alpha \in (0,1)$ is

$$\ell_{\mathbf{q},\alpha}(y,\widehat{y}) = \max\{\alpha(y-\widehat{y}),(1-\alpha)(\widehat{y}-y)\} = (y-\widehat{y})(\alpha - \mathbf{1}_{(y<\widehat{y})}).$$

- It is called "quantile" loss since

$$Q(\alpha) = F^{-1}(\alpha) \in \underset{q\in\mathbb{R}}{\operatorname{argmin}}\Big\{\mathbb{E}\Big[\ell_{\mathbf{q},\alpha}(Y,q)\Big]\Big\},$$

(quantiles are also "ellicitable" functionals, elicited by $s(y,\hat{y}) = \alpha(y-\hat{y})_+ + (1-\alpha)(y-\hat{y})_-$)

# Loss Functions

- Indeed, the first order condition of

$$\min_{q\in\mathbb{R}}\left\{(\alpha-1)\int_{-\infty}^{q}(y-q)dF_Y(y)+\alpha\int_{q}^{\infty}(y-q)dF_Y(y)\right\},$$

can be written, using Leibniz integral rule,

$$(1-\alpha)\int_{-\infty}^{q^\star}dF_Y(y)-\alpha\int_{q^\star}^{\infty}dF_Y(y)=0$$

i.e. $F_Y(q^\star)-\alpha=0$.

# Loss Functions

**Definition 4.12: Expectile loss, $\ell_{\mathsf{e},\alpha}$**

The expectile loss $\ell_{\mathsf{e},\alpha}$, for some $\alpha \in (0,1)$ is

$$\ell_{\mathsf{e},\alpha}(y,\widehat{y}) = (y-\widehat{y})^2 \cdot (\alpha - \mathbf{1}_{(y<\widehat{y})})$$

$$E(\alpha) = \underset{e\in\mathbb{R}}{\text{argmin}}\Big\{\mathbb{E}\Big[\ell_{\mathsf{e},\alpha}(Y,e)\Big]\Big\},$$

(expectiles are elicited by $s(x,y) = \alpha(y-x)_+^2 + (1-\alpha)(y-x)_-^2$).

"*Expectiles have properties that are similar to quantiles*" Newey and Powell (1987)

# Loss Functions

Portnoy and Koenker (1997), "*The Gaussian Hare and the Laplacian Tortoise*"


to go further (for more details on optimization issues)

# Loss and Generalized Linear Models

- In GLM, the scaled deviance ($-2\times$ the log-likelihood) of the exponential model is

$$D^\star = \sum_{i=1}^{n} d^\star\left(y_i, \widehat{y}_i\right), \text{ where } d^\star\left(y_i, \widehat{y}_i\right) = 2\big(\log \mathcal{L}_i(y_i) - \log \mathcal{L}_i(\widehat{y}_i)\big).$$

that can be related to in-sample empirical risk

$$\widehat{\mathcal{R}}_n(\widehat{m}) = \sum_{i=1}^{n} \ell\left(y_i, \widehat{m}(\boldsymbol{x}_i)\right),$$

- For the Poisson distribution (with a log-link), the loss would be

$$\ell(y_i, \widehat{y}_i) = \begin{cases} 2\left(y_i \log y_i - y_i \log \widehat{y}_i - y_i + \widehat{y}_i\right) & y_i > 0 \\ 2\widehat{y}_i & y_i = 0, \end{cases}$$

while for a logistic regression, we have the standard binary cross-entropy loss

$$\ell(y_i, \widehat{y}_i) = -\big(y_i \log[\widehat{y}_i] + (1 - y_i) \log[1 - \widehat{y}_i]\big).$$

# Distance Between Distributions

**Definition 4.13: Distance (or metric)**

A distance $d$ on a set $E$ is a function $E \times E \to \mathbb{R}_+$ such that

- $d$ is symmetric, $\forall (a,b) \in E^2,\ d(a,b) = d(b,a)$,
- $d$ is separable, $\forall (a,b) \in E^2,\ d(a,b) = 0 \Leftrightarrow a = b$,
- $d$ satisfies $\forall (a,b,c) \in E^3,\ d(a,c) \leq d(a,b) + d(b,c)$

In a vector space, with norm $\|\cdot\|$ the induced distance is $d(x,y) = \|y - x\|$.
Conversely, if

- $d$ invariant by translation, $d(x,y) = d(x+a, y+a)$
- $d$ is homogeneous, $d(\alpha x, \alpha y) = |\alpha| d(x,y)$

# Distance Between Distributions

then $\|x\| = d(x, 0)$ is a norm.

**Proposition 4.6**

If $d$ is a distance on $E$, and if $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function such that $\psi(0) = 0$ and $\psi(t) > 0$ for all $t > 0$. If $\psi$ if subadditive ($\psi(s+t) \leq \psi(s)+\psi(t)$), then $\delta(a, b) = \psi(d(a, b))$ is also a distance on $E$.

**Proposition 4.7:** ⚠

If $d$ is a distance on $E$, then $d^2$ is not necessarily a distance.

# Distance Between Distributions

- Consider the Euclidean distance in $E = \mathbb{R}^2$, i.e. $d(\boldsymbol{z}_1, \boldsymbol{z}_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. $d^2$ is not a distance, see

$$\begin{cases} d^2(-\mathbf{1}, +\mathbf{1}) = 2^2 + 2^2 = 8 \\ d^2(-\mathbf{1}, \mathbf{0}) = 1^2 + 1^2 = 2 \\ d^2(\mathbf{0}, +\mathbf{1}) = 1^2 + 1^2 = 2 \end{cases}$$

i.e. $d^2$ does not satisfy the triangular inequality

$$d^2(-\mathbf{1}, +\mathbf{1}) > d^2(-\mathbf{1}, \mathbf{0}) + d^2(\mathbf{0}, +\mathbf{1}),$$

while

$$d(-\mathbf{1}, +\mathbf{1}) \leq d(-\mathbf{1}, \mathbf{0}) + d(\mathbf{0}, +\mathbf{1}).$$

(functions that generalize squared distance are sometimes referred to as divergences)

# Distance Between Distributions

- In addition to "distance", similar terms are used, including "dissimilarity", "deviance", "deviation", "discrepancy", "discrimination", and "divergence"

(... all denoted "$d$", or "$D$")

- A fundamental problem in statistics and machine learning is to come up with useful measures of "distance" between pairs of probability distributions. Two desirable properties of a distance function are symmetry and the triangle inequality.

Unfortunately, many notions of "distance" between probability distributions do not satisfy these properties. Weaker notions of distance are often used, such as dissimilarity measures and divergences.
See Cha (2007) for a comprehensive list of distances...

# Distance Between Distributions

**Definition 4.14: Dissimilarity measure**

A dissimilarity measure $D$ on a set $E$ is a function $E \times E \to \mathbb{R}_+$ such that $D$ is positive and separable, i.e., $\forall (a, b) \in E^2,\ D(a, b) = 0 \Leftrightarrow a = b$,

**Definition 4.15: Divergence on $\mathbb{R}^n$**

A divergence $D$ on a set $E \subset \mathbb{R}^n$ is a function $E \times E \to \mathbb{R}_+$ such that

- $D$ is separable, $\forall (\boldsymbol{x}, \boldsymbol{y}) \in E^2,\ D(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$,
- $D$ admits development
  $\forall (\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\epsilon}) \in E^2,\ D(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{\epsilon}) = \frac{1}{2} \sum A_{i,j}(\epsilon) \epsilon_i \epsilon_j + O(|\epsilon|^3)$,
  where $A(\epsilon)$ is definite positive.

# Distance Between Distributions

**Definition 4.16: Scale sensitive divergence, Zolotarev (1976)**

A divergence $D$ is scale sensitive (of order $\beta > 0$) if $D(c\boldsymbol{x}, c\boldsymbol{y}) \leq |c|^{\beta} D(\boldsymbol{x}, \boldsymbol{y})$

**Definition 4.17: Bregman Divergence, Bregman (1967)**

Let $\psi : \mathcal{X} \to \mathbb{R}$ be a strictly convex function that is continuously differentiable. Then the Bregman divergence $D_{\psi}(\boldsymbol{x}, \boldsymbol{y})$ is defined as

$$D_{\psi}(\boldsymbol{x}, \boldsymbol{y}) = \psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle \nabla\psi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.$$

- If $\psi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ (strictly convex), then $D_{\psi}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$.

(recall that $\nabla\|\boldsymbol{x}\|^2 = 2\boldsymbol{x}$)

# Distance Between Distributions

**Proposition 4.8: Bregman Divergence**

Let $\psi : \mathcal{X} \to \mathbb{R}$ be a strictly convex function that is continuously differentiable. Then Bregman divergence $D_\psi(\boldsymbol{x}, \boldsymbol{y})$ is

- strictly convex in $\boldsymbol{x}$,
- (generally) non-convex in $\boldsymbol{y}$,
- non-negative $D_\psi(\boldsymbol{x}, \boldsymbol{y}) \geq 0$,
- separable, $D_\psi(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$,
- (generally) asymmetric.

# Distance Between Distributions

- If $\mathcal{X} = \mathbb{R}^n$, and $\psi(\boldsymbol{x}) = \frac{1}{2}\sum_{ij} A_{ij}x_i x_j = \frac{1}{2}\boldsymbol{x}^\top A($$\boldsymbol{x}$ for some $n \times n$ matrix $A$ definite positive, then

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\sum_{ij} A_{ij}(x_i - y_i)(x_j - y_j) = (\boldsymbol{x} - \boldsymbol{y})^\top A(\boldsymbol{x} - \boldsymbol{y})$$

(see Mahalanobis distance).

- If $\mathcal{X} = \mathbb{R}^n$, and $\psi(\boldsymbol{x}) = -\sum_i \log(x_i)$ then

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \frac{x_i}{y_i} - \log\frac{x_i}{y_i} - 1$$

See Banerjee et al. (2005) for more examples.

# Distance Between Distributions

We have defined norms

- on $\mathbb{R}^n$, e.g.,

$$\|\boldsymbol{x}\|_{\ell_2} = \left(|x_1|^2 + \cdots + |x_n|^2\right)^{1/2} = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

that could be extended

- on $\mathbb{R}$-valued random variables, e.g.,

$$\|X\|_2 = \left(\mathbb{E}\left[|X|^2\right]\right)^{1/2} = \left(\sum |x|^2 p(x)\right)^{1/2} = \left(\int |x|^2 f(x)\mathrm{d}x\right)^{1/2}$$

We can also define "distances", "dissimilarity" measures, and "divergences"

- on $\mathbb{R}^n$, e.g.,

$$\ell_2(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{y}) = \left(|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2\right)^{1/2} = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$

# Distance Between Distributions

that could be extended

- on $\mathbb{R}$-valued random variables as components of a random vector, e.g.,

$$D(X,Y)=\left(\mathbb{E}\left[|X-Y|^2\right]\right)^{1/2}=\left(\sum|x-y|^2p(x,y)\right)^{1/2}=\left(\int|x-y|^2f(x,y)\mathrm{d}x\mathrm{d}y\right)^{1/2}$$

where $p$ or $f$ is the joint distribution of $(X,Y)$, e.g., for a Gaussian vector

$$D(X,Y)=(\mu_x-\mu_y)^2+(\sigma_x-\sigma_y)^2+2\sigma_x\sigma_y(1-\rho).$$

- on $\mathbb{R}$-valued random variables assuming that random variables are independent, e.g.,

$$D_\perp(X,Y)=\left(\sum|x-y|^2p_x(x)p_y(y)\right)^{1/2}=\left(\int|x-y|^2f_x(x)f_y(y)\mathrm{d}x\mathrm{d}y\right)^{1/2}$$

e.g., for two Gaussian distributions

$$D_\perp(X,Y)=(\mu_x-\mu_y)^2+\sigma_x^2+\sigma_y^2.$$

# Distance Between Distributions

and one can consider some distance

› on $\mathbb{R}$-valued distributions, e.g.,

$$D(\mathcal{N}(\mu_x, \sigma_x^2), \mathcal{N}(\mu_y, \sigma_y^2)) = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2.$$

In the context of "probabilistic forecasts" (as in Gneiting et al. (2007)), a "distance"

› on pairs $\mathbb{R} \times \mathbb{R}$-valued distributions, e.g.,

$$D(x, \mathcal{N}(\mu_y, \sigma_y^2)) = (x - \mu_y)^2 + \sigma_y^2.$$

# Distance Between Distributions

**Definition 4.18: Sum invariant divergence, Zolotarev (1976)**

A divergence $D$ is sum invariant if $D(X+Z, Y+Z) \leq D(X,Y)$ whenever $Z \perp\!\!\!\perp X, Y$

**Example**: if $D$ is 1-scale sensitive, $D(\mathbf{1}_0, \mathbf{1}_1) \leq \frac{1}{2} D(\mathbf{1}_0, \mathbf{1}_2)$

**Example**: if $D$ is sum invariant, $D(\mathbf{1}_0, \mathbf{1}_1) = D(\mathbf{1}_1, \mathbf{1}_2)$

See Bellemare et al. (2017a).

# Distance Between Distributions

Consider sample $\{x_1, \cdots, x_n\}$ an i.i.d. sample, with empirical measure $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i}$

**Definition 4.19: Divergence based inference**

Consider some parametric family $\mathcal{Q} = \{q_\theta, \theta \in \Theta\}$. Given a divergence $D$, we want to find

$$\theta^\star = \underset{\theta\in\Theta}{\operatorname{argmin}}\{D(\, p \,,\, q_\theta \,)\}$$

unknown $p$ ; $q_\theta \in \mathcal{Q}$

or its empirical version

$$\widehat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}}\{D(\, \widehat{p}_n \,,\, q_\theta \,)\}$$

estimated $\widehat{p}_n$

# Distance Between Distributions

**Definition 4.20: Unbiased sample gradients, Bellemare et al. (2017a)**

A divergence $D$ has unbiased sample gradients when the expected gradient of the sample loss equals the gradient of the true loss for all $p$ and $n$,

$$\mathbb{E}\left(\nabla_\theta D(\widehat{p}_n, q_\theta)\right) = \nabla_\theta D(p, q_\theta).$$

- Then $D$ is a proper scoring rule (see Gneiting and Raftery (2007)).
- If this is not satisfied, stochastic gradient descent may not converge...

# Distance Between Distributions

**Definition 4.21: Integral probability metric, Müller (1997)**

Integral probability metrics (IPMs) are distances on the space of distributions over a set $\mathcal{X}$, defined by a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$ as

$$D_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(X)] - \mathbb{E}[f(Y)] \right|.$$

$X \sim p$ $\qquad$ $Y \sim q$

Discussed also in Dedecker and Merlevède (2007)

# Distance Between Distributions

- Note that it is still possible to define projections with deviance (that will not be "orthogonal" projections since divergence are not related to inner products)

**Definition 4.22: Projection, Bregman (1967), Bauschke et al. (1997)**

Given a strictly convex function continuously differentiable $\psi$ and the associated Bregman divergence $D_\psi$, a closed closed convex $K \subset \mathcal{X}$ and a point $\boldsymbol{x} \in \mathcal{X}$. The Bregman projection of $\boldsymbol{x}$ onto $K$ is

$$\boldsymbol{x}^\star = \underset{\boldsymbol{y} \in K}{\operatorname{argmin}} \{ D_\psi(\boldsymbol{x}, \boldsymbol{y}) \}$$

- If $\psi(\boldsymbol{x}) = \|\boldsymbol{x}\|_{\ell_2}^2$, Bregman projection is the standard orthogonal projection onto a convex set,

$$\boldsymbol{x}^\star = \underset{\boldsymbol{y} \in K}{\operatorname{argmin}} \{ \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}^2 \}$$

# Distance Between Distributions

› With Bregman divergence $D_\psi$, we have a generalized version of the Pythagorean theorem

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) \geq D_\psi(\boldsymbol{x}, \boldsymbol{x}^\star) + D_\psi(\boldsymbol{x}^\star, \boldsymbol{y})$$

Numerically, one can use (cyclical) Dykstra algorithm with Bregman projections (see Censor and Reich (1998), Bauschke and Lewis (2000)) to compute $\boldsymbol{x}^\star$: suppose that $K = \bigcap_{i=1}^{m} K_i$ where $K_i$'s are convex sets (e.g. half-planes - when $K$ is some polyhedral). Let $P_i^\psi$ denote the orthogonal on $K_i$ based on $D_\psi$,

$$P_i^\psi : \boldsymbol{x} \mapsto \underset{\boldsymbol{y} \in K_i}{\operatorname{argmin}} \left\{ \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}^2 \right\}$$

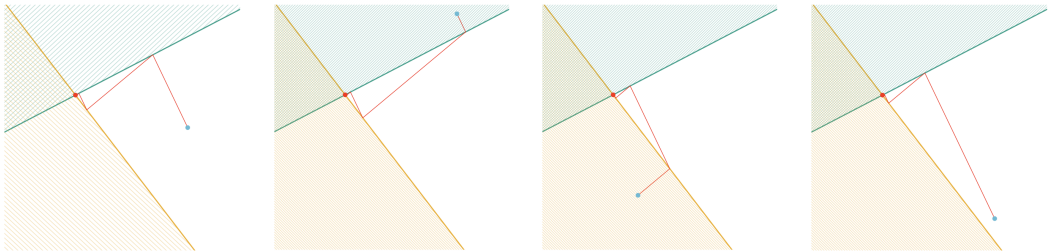
and consider the following iterative sequence of projections, until some $\boldsymbol{x}_j \in K$,

$$\boldsymbol{x}_0 \xrightarrow{P_1^\psi} \boldsymbol{x}_1 \xrightarrow{P_2^\psi} \boldsymbol{x}_2 \xrightarrow{P_3^\psi} \cdots \xrightarrow{P_{m-1}^\psi} \boldsymbol{x}_{m-1} \xrightarrow{P_m^\psi} \boldsymbol{x}_m \xrightarrow{P_1^\psi} \boldsymbol{x}_{m+1} \xrightarrow{P_2^\psi} \boldsymbol{x}_{m+2} \cdots$$

# Distance Between Distributions

$$\boldsymbol{x}_0 \xrightarrow{P_1^{\psi}} \boldsymbol{x}_1 \xrightarrow{P_2^{\psi}} \boldsymbol{x}_2 \xrightarrow{P_3^{\psi}} \cdots \xrightarrow{P_{m-1}^{\psi}} \boldsymbol{x}_{m-1} \xrightarrow{P_m^{\psi}} \boldsymbol{x}_m \xrightarrow{P_1^{\psi}} \boldsymbol{x}_{m+1} \xrightarrow{P_2^{\psi}} \boldsymbol{x}_{m+2} \cdots$$

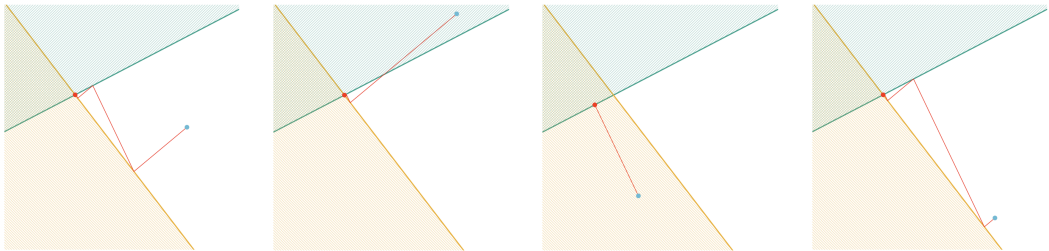
until some $\boldsymbol{x}_j \in K$,see Boyle and Dykstra (1986) for the original idea



# Distance Between Distributions

$$\boldsymbol{x}_0 \xrightarrow{P_1^\psi} \boldsymbol{x}_1 \xrightarrow{P_2^\psi} \boldsymbol{x}_2 \xrightarrow{P_3^\psi} \cdots \xrightarrow{P_{m-1}^\psi} \boldsymbol{x}_{m-1} \xrightarrow{P_m^\psi} \boldsymbol{x}_m \xrightarrow{P_1^\psi} \boldsymbol{x}_{m+1} \xrightarrow{P_2^\psi} \boldsymbol{x}_{m+2} \cdots$$

until some $\boldsymbol{x}_j \in K$,see Boyle and Dykstra (1986) for the original idea



with half spaces, $\|\boldsymbol{x}_j - \boldsymbol{x}^\star\|_{\ell_2} \leq cr^j \|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\ell_2}$ for some $r \in (0,1)$ and $c > 0$ (or "linear convergence" since $\|\boldsymbol{x}_j - \boldsymbol{x}^\star\|_{\ell_2} \leq r\|\boldsymbol{x}_{j-1} - \boldsymbol{x}^\star\|_{\ell_2}$).

# Distance Between Distributions

**Definition 4.23: Hellinger distance, Hellinger (1909)**

For two discrete distributions $p$ and $q$, Hellinger distance is

$$d_{\mathrm{H}}(p,q)^2 = \frac{1}{2}\sum_i \left(\sqrt{p(i)} - \sqrt{q(i)}\right)^2 = 1 - \sum_i \sqrt{p(i)q(i)} \in [0,1],$$

and for absolutely continuous distributions, if $p$ and $q$ are densities,

$$d_{\mathrm{H}}(p,q)^2 = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x \text{ or } \frac{1}{2}\int_{\mathbb{R}^k} \left(\sqrt{p(\boldsymbol{x})} - \sqrt{q(\boldsymbol{x})}\right)^2 \mathrm{d}\boldsymbol{x}.$$

See Pardo (2018).

# Distance Between Distributions

**Proposition 4.9: Distance between Beta variables**

Consider two Beta distributions, then $d_{\mathrm{H}}^2\left(\mathcal{B}(a_1, b_1), \mathcal{B}(a_2, b_2)\right)$ is

$$1 - \frac{1}{\sqrt{B(a_1, b_1)B(a_2, b_2)}} B\left(\frac{a_1 + a_2}{2}, \frac{b_1 + b_2}{2}\right)$$

**Proof**

$$1 - \int_0^1 \sqrt{f_1(t)f_2(t)}\mathrm{d}t = 1 - \frac{1}{\sqrt{B(a_1, b_1)B(a_2, b_2)}} \int_0^1 t^{(a_1+a_2)/2-1}(1-t)^{(b_1+b_2)/2-1}\mathrm{d}t,$$

then use $B(a, b) = \mathrm{B}(a, b) = \displaystyle\int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$.

# Distance Between Distributions

**Proposition 4.10: Distance between Gaussian vectors**

Consider two Gaussian distributions, then $d_{\mathrm{H}}^2\left(\mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right), \mathcal{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)\right)$ is

$$2 - 2\frac{\left|\boldsymbol{\Sigma}_1\right|^{\frac{1}{4}}\left|\boldsymbol{\Sigma}_2\right|^{\frac{1}{4}}}{\left|\bar{\boldsymbol{\Sigma}}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{8}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^\top \bar{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right)$$

where $\bar{\boldsymbol{\Sigma}} = \frac{1}{2}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$.

Note that it is a Bregman divergence $D_\psi$ with $\psi(\boldsymbol{x}) = \sum_{i=1} x_i^2$

# Distance Between Distributions

**Definition 4.24: Pearson/Neyman $\chi$-square divergences Nielsen and Nock (2013)**

For two discrete distributions $p$ and $q$, Pearson chi-square divergence is

$$d_{\mathrm{P}\chi}(p\|q)^2 = \sum_i \frac{\left[p(i) - q(i)\right]^2}{q(i)},$$

while Neyman chi-square divergence is

$$d_{\mathrm{N}\chi}(p\|q)^2 = \sum_i \frac{\left[(i) - q(i)\right]^2}{p(i)} = d_{\mathrm{P}\chi}(q\|p),$$

# Distance Between Distributions

- Note that both are Bregman divergences $D_\psi$ with $\psi_{\mathrm{P}}(\boldsymbol{x}) = -2\sum_{i=1}\sqrt{x_i}$ and $\psi_{\mathrm{N}}(\boldsymbol{x}) = \sum_{i=1} {x_i}^{-1}$.
- $d_\chi$ can be extended to the case of continuous distributions, e.g.,

$$d_{\mathrm{P}\chi}(p\|q)^2 = \int \left(\frac{p(x)}{q(x)} - 1\right)^2 p(x)\mathrm{d}x$$

# Distance Between Distributions

**Definition 4.25: Total Variation, Jordan (1881); Rudin (1966)**

For two univariate distributions $p$ and $q$, the total variation distance between $p$ and $q$ is

$$d_{\text{TV}}(p, q) = \sup_{\mathcal{A} \subset \mathbb{R}^k} \{|p(\mathcal{A}) - q(\mathcal{A})|\}.$$

**Proposition 4.11: Total Variation**

For two univariate distributions $p$ and $q$, the total variation distance between $p$ and $q$ is

$$d_{\text{TV}}(p, q) = \frac{1}{2} \sum_i |p(i) - q(i)| = \frac{1}{2} \|p - q\|_{\ell_1} = \sum_{i: p(i) \geq q(i)} (p(i) - q(i))$$

See Proposition 4.2 in Levin and Peres (2017).

# Distance Between Distributions

- Equivalently,

$$d_{\text{TV}}(p,q) = \frac{1}{2} \sup_{f:\mathbb{R}^k \to \{0,1\}} \left\{ \int f \mathrm{d}p - \int f \mathrm{d}q \right\}$$

(see e.g. https://djalil.chafai.net/blog/, with $f : \mathbb{R}^k \to \{-1, 1\}$, $f = \mathbf{1}_{\mathcal{A}} - \mathbf{1}_{\mathcal{A}^c}$)

- It is an IPM with $\mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$, so that $\mathcal{F}$ is a set of indicator functions for any event.

- For Gaussian distributions, the distance has no explicit formula, see, e.g., Devroye et al. (2018).

# Distance Between Distributions

**Definition 4.26: Entropy, Shannon (1948)**

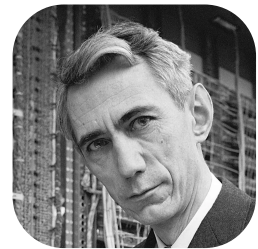
The entropy associated with distribution $p$ is

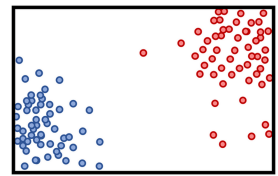
$$\mathcal{E}_p(p) = -\sum_i p(i)\log p(i) = \mathbb{E}_p\big[-\log p(X)\big].$$

and define cross-entropy (of $q$ relative to $p$) as

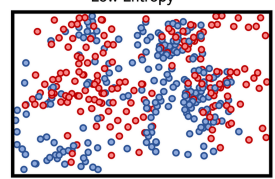
$$\mathcal{E}_q(p) = -\sum_i p(i)\log q(i) = \mathbb{E}_p\big[-\log q(X)\big].$$

See Amari (2016) or Chambert-Loir (2023) for more details.





Low Entropy



High Entropy

# Distance Between Distributions

**Definition 4.27: Kullback–Leibler, Kullback and Leibler (1951)**

For two discrete distributions $p$ and $q$, Kullback–Leibler divergence of $p$, with respect to $q$ is

$$D_{\mathrm{KL}}(p\|q) = \sum_i p(i) \log \frac{p(i)}{q(i)},$$

and for absolutely continuous distributions,

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} \,\mathrm{d}x \text{ or } \int_{\mathbb{R}^k} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x},$$

in higher dimension.

Also called relative entropy, since $D_{\mathrm{KL}}(p\|q) = \mathcal{E}_q(p) - \mathcal{E}_p(p)$.

# Distance Between Distributions

**Proposition 4.12: Divergence for Gaussian vectors**

Consider two Gaussian distributions, then $D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \| \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))$ is

$$\frac{1}{2}\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - \log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - k\right]$$

where $k$ is the dimension, see Polyanskiy and Wu (2022).

# Distance Between Distributions

The entropy of $X$ according to $p$ is smaller than or equal to the cross-entropy of $p$ and $q$, or equivalently

**Proposition 4.13: Gibbs' inequality**

$D_{\mathrm{KL}}(p\|q)$ is positive and separable, i.e. $D_{\mathrm{KL}}(p\|q) \geq 0$ and $D_{\mathrm{KL}}(p\|q) = 0$ if and only if $p = q$.

**Proof**: $\sum_{x \in I} p(x) \log \frac{p(x)}{q(x)} \geq 0$ where $I$ is the set of all $x$ for which $p(x) > 0$. Recall that $\log x \leq x - 1$ (with equality only when $x = 1$), thus $\log(1/x) \geq 1 - x$, and

$$\sum_{x \in I} p(x) \ln \frac{p(x)}{q(x)} \geq \sum_{x \in I} p(x) \left(1 - \frac{q(x)}{p(x)}\right) = \underbrace{\sum_{x \in I} p(x)}_{=1} - \underbrace{\sum_{x \in I} q(x)}_{\leq 1} \geq 0.$$

# Distance Between Distributions

**Proposition 4.14: Additivity for independence distributions**

$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = D_{\mathrm{KL}}(p_x\|q_x) + D_{\mathrm{KL}}(p_y\|q_y)$ if $\boldsymbol{p}(x,y) = p_x(x)p_y(y)$ and $\boldsymbol{q}(x,y) = q_x(x)q_y(y)$.

**Proof** By definition

$$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \int_{\mathcal{X}}\int_{\mathcal{Y}} p(x,y)\cdot\log\frac{p(x,y)}{q(x,y)}\,\mathrm{d}y\,\mathrm{d}x\,.$$

and since $\boldsymbol{p}(x,y) = p_x(x)p_y(y)$ and $\boldsymbol{q}(x,y) = q_x(x)q_y(y)$,

$$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \int_{\mathcal{X}}\int_{\mathcal{Y}} p_1(x)\,p_2(y)\cdot\log\frac{p_1(x)\,p_2(y)}{q_1(x)\,q_2(y)}\,\mathrm{d}y\,\mathrm{d}x\,.$$

# Distance Between Distributions

and

$$\begin{aligned}
D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} p_x(x)\,p_y(y)\cdot\left(\log\frac{p_x(x)}{q_x(x)}+\log\frac{p_y(y)}{q_y(y)}\right)\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} p_x(x)\,p_y(y)\cdot\log\frac{p_x(x)}{q_x(x)}\,\mathrm{d}y\,\mathrm{d}x+\int_{\mathcal{X}}\int_{\mathcal{Y}} p_x(x)\,p_y(y)\cdot\log\frac{p_y(y)}{q_y(y)}\,\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathcal{X}} p_x(x)\cdot\log\frac{p_x(x)}{q_x(x)}\int_{\mathcal{Y}} p_y(y)\,\mathrm{d}y\,\mathrm{d}x+\int_{\mathcal{Y}} p_y(y)\cdot\log\frac{p_y(y)}{q_y(y)}\int_{\mathcal{X}} p_x(x)\,\mathrm{d}x\,\mathrm{d}y\\
&= \int_{\mathcal{X}} p_x(x)\cdot\log\frac{p_x(x)}{q_x(x)}\,\mathrm{d}x+\int_{\mathcal{Y}} p_y(y)\cdot\log\frac{p_y(y)}{q_y(y)}\,\mathrm{d}y\\
&= D_{\mathrm{KL}}(p_x\|q_x)+D_{\mathrm{KL}}(p_y\|q_y).
\end{aligned}$$

# Distance Between Distributions

- It is only defined in this way if, for all $x$, $q(x) = 0$ implies $p(x) = 0$ ("absolute continuity" with respect to $p$).

**Proposition 4.15**

The KL divergence has unbiased sample gradients, but is not scale sensitive.

**Proof** Bellemare et al. (2017b).

- In a Bayesian setting, $D_{\mathrm{KL}}(p\|q)$ is a measure of the information gained by revising one's beliefs from the prior probability distribution $q$ to the posterior probability distribution $p$ (it is the amount of information lost when $q$ is used to approximate $p$).

- If $\psi(\boldsymbol{x}) = \sum x_i \log(x_i)$ (strictly convex), then Bregman divergence is

$$D_\psi(\boldsymbol{x}, \boldsymbol{y}) = \sum x_i \log \frac{x_i}{y_i} = D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{y})$$

# Distance Between Distributions

$$D_{\mathrm{KL}}(\mathcal{B}(p)\|\mathcal{B}(q)) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

$$D_{\mathrm{KL}}(\mathcal{B}(n,p)\|\mathcal{B}(n,q)) = np\log\frac{p}{q} + n(1-p)\log\frac{1-p}{1-q} = nD_{\mathrm{KL}}(\mathcal{B}(p)\|\mathcal{B}(q))$$

$$D_{\mathrm{KL}}(\mathcal{U}([a_1,b_1])\|\mathcal{U}([a_2,b_2])) = \log\frac{b_2-a_2}{b_1-a_1}$$

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1,\sigma_1^2)\|\mathcal{N}(\mu_2,\sigma_2^2)) = \frac{1}{2}\left[\frac{(\mu_1-\mu_2)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \log\frac{\sigma_1^2}{\sigma_2^2} - 1\right]$$

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\|\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) = \frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - \log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - n\right]$$

# Distance Between Distributions

- Consider some distribution $p_\theta$, as in Nielsen (2022). Using Taylor expansion,

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta+\mathrm{d}\theta}) = \frac{1}{2}\mathrm{d}\theta^\top I(\theta)\mathrm{d}\theta \approx \frac{1}{2}\mathrm{d}s_\theta^2.$$

**Definition 4.28: Jeffreys (symmetric) divergence** Jeffreys (1946)

The Jeffrey divergence is a symmetric divergence induced by Kullback-Liebler divergence,

$$D_{\mathrm{J}}(p_1, p_2) = \frac{1}{2}D_{\mathrm{KL}}(p_1 \| p_2) + \frac{1}{2}D_{\mathrm{KL}}(p_2 \| p_1).$$

# Distance Between Distributions

**Definition 4.29: Jensen-Shannon, Lin (1991)**

The Jensen-Shannon divergence is a symmetric divergence induced by Kullback-Liebler divergence,

$$D_{\text{JS}}(p_1, p_2) = \frac{1}{2} D_{\text{KL}}(p_1 \| q) + \frac{1}{2} D_{\text{KL}}(p_2 \| q),$$

where $q = \frac{1}{2}(p_1 + p_2)$.

Endres and Schindelin (2003) proved that $\sqrt{D_{\text{JS}}(p_1, p_2)}$ is a proper distance.

- See `philentropy` package.

# Distance Between Distributions

**Definition 4.30: $f$-divergence, Rényi (1961), Ali and Silvey (1966)**

Given a continuous convex function $f : [0, \infty) \to \overline{\mathbb{R}}$, define

$$D_f(p\|q) = \sum_i q(i) \cdot f\left(\frac{p(i)}{q(i)}\right)$$

and for absolutely continuous function

$$D_f(p\|q) = \int_{\mathbb{R}} q(x) f\left(\log \frac{p(x)}{q(x)}\right) \mathrm{d}x \text{ or } \int_{\mathbb{R}^k} q(\boldsymbol{x}) f\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x},$$

- $D_f(p\|q)$ is properly defined when $p \ll q$, see also Csiszár (1964, 1967).

If $f(u) = u \log u$, $D_f(p\|q) = D_{\mathrm{KL}}(p, q)$
If $f(u) = |u - 1|$, $D_f(p\|q) = d_{\mathrm{TV}}(p, q)$

# Distance Between Distributions

If $f(u) = \frac{1}{2}(\sqrt{u}-1)^2$, $D_f(p\|q) = d_{\mathrm{H}}(p,q)^2$

If $f(u) = \frac{1}{2}\left(u\log u - (u+1)\log\left(\frac{u+1}{2}\right)\right)$, $D_f(p\|q) = d_{\mathrm{JS}}(p,q)$

- One can define $D_f(p\|q)$ when $p \not\ll q$: Since $f$ is convex, and $f(1)=0$, the function $\frac{f(x)}{x-1}$ must nondecrease, so there exists $f'(\infty) := \lim_{x\to\infty} f(x)/x$, taking value in $(-\infty, +\infty]$. And since for any $p(x) > 0$, we have $\lim_{q(x)\to 0} q(x) f\left(\frac{p(x)}{q(x)}\right) = p(x)f'(\infty)$.

**Proposition 4.16**

$D_f(p\|q)$ is linear in $f$, $D_{af+bg}(p\|q) = aD_f(p,q) + bD_g(p\|q)$.

# Distance Between Distributions

**Proposition 4.17**

$D_f = D_g$ if and only if $f(x) = g(x) + c(x-1)$ for some $c \in \mathbb{R}$.

- The only $f$-divergence that is also a Bregman $\psi$-divergence is the KL divergence
- The only $f$-divergence that is also an integral probability metric is the total variation.
- There is a variational representation of $D_f$, in Polyanskiy and Wu (2022).

# Distance Between Distributions

- Since $f$ is convex, let $f^\star$ be the convex conjugate of $f$. Let $\text{effdom}(f^\star)$ be the effective domain of $f^\star$ (i.e., $\text{effdom}(f^*) = \{y : f^*(y) < \infty\}$)

$$D_f(p; q) = \sup_{g:\Omega \to \text{effdom}(f^\star)} \mathbb{E}_p[g] - \mathbb{E}_q[f^\star \circ g]$$

- For example, with the total variation, $f(x) = \frac{1}{2}|x - 1|$, its convex conjugate is $f^*(x^*) = \begin{cases} x^* \text{ on } [-1/2, 1/2], \\ +\infty \text{ else.} \end{cases}$, and we obtain

$$d_{\text{TV}}(p, q) = \sup_{|g| \leq 1/2} \mathbb{E}_p[g(X)] - \mathbb{E}_q[g(X)].$$

# Distance Between Distributions

- Extending Rényi entropy of order $\alpha$, $\mathrm{H}_\alpha(X) = \frac{1}{1-\alpha}\log\left(\sum_i p(i)^\alpha\right)$, define

**Definition 4.31: Rényi $\alpha$-divergence, Rényi (1961)**

Given $\alpha \in (0, \infty)$, define

$$D_\alpha(p\|q) = \frac{1}{\alpha - 1}\log\left(\sum_i \frac{p(i)^\alpha}{q(i)^{\alpha-1}}\right)$$

and for absolutely continuous function

$$D_\alpha(p\|q) = \frac{1}{\alpha - 1}\log\left(\int_{\mathbb{R}} \frac{p(x)^\alpha}{q(x)^{\alpha-1}}\mathrm{d}x\right) \text{ or } \frac{1}{\alpha - 1}\log\left(\int_{\mathbb{R}^k} \frac{p(\boldsymbol{x})^\alpha}{q(\boldsymbol{x})^{\alpha-1}}\mathrm{d}\boldsymbol{x}\right).$$

# Distance Between Distributions

- Recall that

$$D_\alpha(p\|q) = \frac{1}{\alpha - 1} \log \left( \sum_i \frac{p(i)^\alpha}{q(i)^{\alpha-1}} \right) \text{ when } \alpha \in (0, \infty).$$

- One can define limiting cases, $D_0(P\|Q) = -\log Q(\{i : p_i > 0\})$ and $D_\infty(P\|Q) = \log \sup_i \frac{p_i}{q_i}$
- Observe also that $D_1(p\|q) = D_{\mathrm{KL}}(p\|q)$

# Distance Between Distributions

**Definition 4.32: Cramér, Cramér (1928a,b) and Székely (2003)**

Consider two measures on $p$ and $q$ on $\mathbb{R}$. Then define Cramér distance

$$C_k(p,q) = \left( \int_{-\infty}^{\infty} |F_p(x) - F_q(x)|^k \right)^{1/k}.$$

- $C_2$ is named "energy-distance" in Székely (2003) and Rizzo and Székely (2016), and "continuous ranked probability score" in Gneiting et al. (2007).
- It is an Integral Probability Metrics (IPM), since

$$C_k(p,q) = \sup_{f \in \mathcal{F}_{k'}} |\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]|.$$

where $k^{-1} + k'^{-1} = 1$, $X \sim p$, $Y \sim q$

where $\mathcal{F}_{k'}$ is the set of absolutely continuous functions such that $\|\nabla f\|_{k'} \leq 1$.

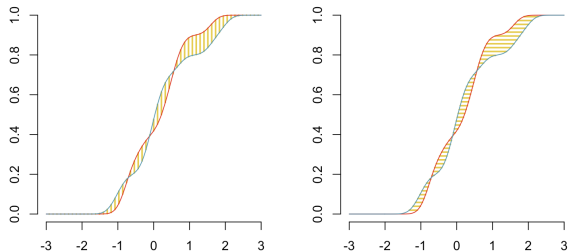
- For example, if $k = 1$, $\|\nabla f\|_\infty \leq 1$ (corresponding to 1-Lipschitz functions).

# Distance Between Distributions

**Definition 4.33: Wasserstein, Wasserstein (1969)**

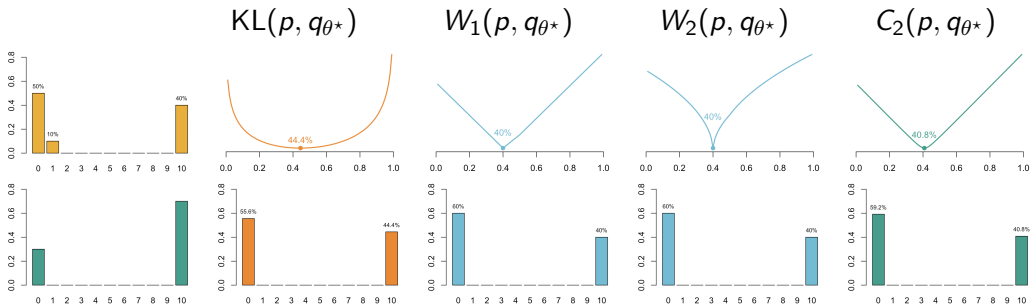
Consider two measures on $p$ and $q$ on $\mathbb{R}$. Then define Wasserstein distance

$$W_k(p,q) = \Big( \int_0^1 |F_p^{-1}(u) - F_q^{-1}(u)|^k \Big)^{1/k}.$$



# Distance Between Distributions

- $\mu$: multinomial distribution on $\{0, 1, 10\}$, with $\boldsymbol{p} = (.5, .1, .4)$
- $\nu_\theta$: binomial type distribution on $\{0, 10\}$, with $\boldsymbol{q}_\theta = (1 - \theta, \theta)$
- Let $\theta^\star = \text{argmin}\{d(p, q_\theta)\}$



# Distance Between Distributions

**Proposition 4.18**

The Wasserstein metric is scale and sum invariant, but does not have unbiased sample gradients.

**Proof** Bellemare et al. (2017b)

**Example** If $x_i$ are drawn from a Bernoulli distribution

- Non-vanishing minimax bias:
$\forall n,\ \exists p, q_\theta,\ \left|\mathbb{E}(\nabla_\theta W_k^k(\widehat{p}_n, q_\theta)) - \nabla_\theta W_k^k(p, q_\theta)\right| \geq 2e^{-2}$
- Wrong minimum: in general,
$\widehat{\theta}_n = \text{argmin}\left\{\mathbb{E}((W_k^k(\widehat{p}_n, q_\theta)))\right\} \neq \text{argmin}\left\{W_k^k(\mathbb{P}, \mathbb{Q}_\theta)))\right\} = \theta$

# Distance Between Distributions

**Proposition 4.19**

The Cramér metric is scale and sum invariant.

- $C_k(X+Z, Y+Z) \leq C_k(X,Y)$ whenever $Z \perp\!\!\!\perp X, Y$ and $k \geq 1$, and $C_k(cX, cY) \leq |c|^{1/k} C_k(X,Y)$.

**Proposition 4.20**

$C_2$ has unbiased sample gradients (only $k=2$),

$$\mathbb{E}\left(\nabla_\theta C_2(\widehat{p}_n, q_\theta)\right) = \nabla_\theta C_2(p, q_\theta).$$

# Distance Between Distributions

- Consider first $W_1$ (earth mover's distance), which was the only distance discussed in Wasserstein (1969). See also Vallender (1974) for an extensive review.
- $W_1$ is an IPM where $\mathcal{F}$ the set of 1-Lipschitz functions, Kantorovich and Rubinstein (1958), i.e., if $p$ and $q$ have bounded support,

$$W_1(p,q) = \sup_{f\in\mathcal{F}} \left\{ \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}(p-q)(x) \right\},$$

$\mathcal{F}$ being the class of 1-Lipschitz functions

**Proposition 4.21: $W_1$ and First Order Dominance**

Suppose that $X_1 \preceq X_2$ (first order dominance, $F_2^{-1}(u) \geq F_1^{-1}(u)$, $\forall u \in (0,1)$),

$$W_1(p_1,p_2) = \mathbb{E}[X_2] - \mathbb{E}[X_1].$$

# Distance Between Distributions

**Proof**

$$W_1(p_1,p_2)=\int_0^1 |\underbrace{F_2^{-1}(u)-F_1^{-1}(u)}_{\geq 0}|\mathrm{d}u=\underbrace{\int_0^1 F_2^{-1}(u)\mathrm{d}u}_{\mathbb{E}[X_2]}-\underbrace{\int_0^1 F_1^{-1}(u)\mathrm{d}u}_{\mathbb{E}[X_1]}$$

then (property discussed later)

$$W_1(p_1,p_2)=\inf_C\int\int |x_2-x_1|\mathrm{d}C(F_1(x_1),F_2(x_2))=\inf_C\underbrace{\int\int |F_2^{-1}(v)-F_1^{-1}(u)|\mathrm{d}C(u,v)}_{\mathbb{E}[|X_1-X_2|]}$$

As discussed in Vallender (1974),

$$\begin{aligned}\mathbb{E}[|X_1-X_2|] &= \int \big[\mathbb{P}[X_1<t,X_2\geq t]+\mathbb{P}[X_1\geq t,X_2<t]\big]\mathrm{d}t\\ &= \int \big[\mathbb{P}[X_1<t]+\mathbb{P}[X_2<t]-2\mathbb{P}[X_1<t,X_2<t]\big]\mathrm{d}t\end{aligned}$$

# Distance Between Distributions

$$\mathbb{E}[|X_1 - X_2|] = [F_1(t) + F_2(t) - 2C(F_1(t), F_2(t))]\mathrm{d}t$$

From Fréchet-Hoeffding bounds, $C(u, v) \leq M(u, v) = \min\{u, v\}$ and

$$F_1(t) + F_2(t) - 2C(F_1(t), F_2(t)) \geq F_1(t) + F_2(t) - 2M(F_1(t), F_2(t))$$

$$\mathbb{E}[|X_1 - X_2|] \geq \int\int |F_2^{-1}(v) - F_1^{-1}(u)|\mathrm{d}M(u, v)$$

$$\int |F_2^{-1}(u) - F_1^{-1}(u)|\mathrm{d}u$$

**Example** let $p_1 \leq p_2$

$$W_1(\mathcal{B}(p_1), \mathcal{B}(p_2)) = p_2 - p_1.$$

# Distance Between Distributions

- We can also consider $W_2$

**Example** let $p_1 \leq p_2$

$$W_2(\mathcal{B}(p_1), \mathcal{B}(p_2)) = \sqrt{p_2 - p_1}.$$

**Proposition 4.22: $W_2$ for Gaussian vectors**

Consider two Gaussian distributions, then

$$W_2(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))^2 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \operatorname{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2})^{1/2})$$

**Proof**: Let $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\boldsymbol{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, and $\Gamma$ define the covariance matrix of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$,

$$\Gamma = \begin{pmatrix} \Sigma_1 & C \\ C^\top & \Sigma_2 \end{pmatrix}$$

# Distance Between Distributions

where (generally), $C$ is some $n_1 \times n_2$ matrix. Recall that $n_1 \times n_2$ matrices can have a pseudo-inverse, in the sense that (Penrose conditions)

$$\begin{cases} AA^-A = A \\ A^-AA^- = A^- \\ (AA^-)^\top = AA^- \\ (A^-A)^\top = A^-A, \end{cases}$$

Observe that $\mathbb{E}(\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_{\ell_2}^2) = \mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2C)$. Recall that $C$ must satisfy the Schur complement constraint, $\boldsymbol{\Sigma}_1 - C\boldsymbol{\Sigma}_2^{-1}C^\top \succeq 0$, so that we want to solve

$$C^\star = \mathrm{argmin}\{-2\mathrm{tr}(C)\} \text{ s.t. } \boldsymbol{\Sigma}_1 - C\boldsymbol{\Sigma}_2^{-1}C^\top \succeq 0,$$

as studied in Olkin and Pukelsheim (1982), where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive ($\succeq 0$) matrices.

# Distance Between Distributions

Let $\mathcal{G} = \{C, n_1 \times n_2 : \ \boldsymbol{\Sigma}_1 - C\boldsymbol{\Sigma}_2^{-1}C^\top \succeq 0\}$, $\mathcal{S} = \{S : \ SS^-\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_2\}$, one can prove (standard duality and convexity arguments) that

$$\max_{C\in\mathcal{G}}\{2\text{tr}(C)\} = \max_{S\in\mathcal{S}}\{\text{tr}(\boldsymbol{\Sigma}_1 S + \boldsymbol{\Sigma}_2 S^-)\} = 2\text{tr}(\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{1/2})$$

with respective (unique) solutions

$$\begin{cases} C^\star = \boldsymbol{\Sigma}_1 S^\star \\ S^\star = \boldsymbol{\Sigma}_2^{1/2}[(\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{1/2})^{1/2}]\boldsymbol{\Sigma}_2^{1/2} \end{cases}$$

See Olkin and Pukelsheim (1982), Givens and Shortt (1984) and Knott and Smith (1984), or more recently Takatsu (2008) and Takatsu and Yokota (2012), with more geometric interpretations.

- Let us spend some time on the Wasserstein distance... and associated concepts.

# Optimal transport

**Definition 4.34: Wasserstein, Wasserstein (1969)**

Consider two measures on $p$ and $q$ on $\mathbb{R}^k$, with a norm $\|\cdot\|$ (on $\mathbb{R}^k$). Then define Wasserstein distance

$$W_k(p,q) = \left( \inf_{\pi \in \Pi(p,q)} \int_{\mathbb{R}^k \times \mathbb{R}^k} \|\boldsymbol{x} - \boldsymbol{y}\|^k \mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}) \right)^{1/k},$$

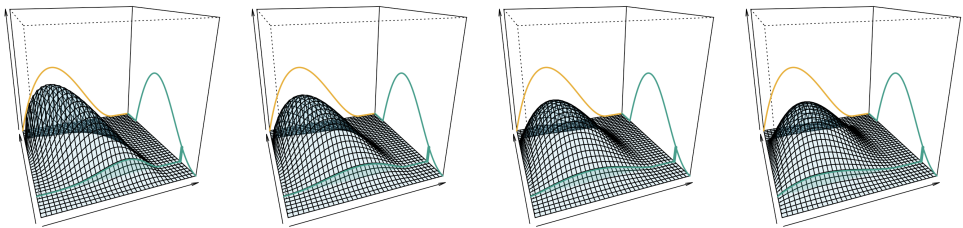
where $\Pi(p,q)$ is the set of all couplings of $p$ and $q$.

› One can prove that $W_1(p,q) \leq W_k(p,q)$ for all $k \geq 1$. And because of Jensen's inequality, if $k \leq k'$,

$$\left( \inf_{\pi \in \Pi(p,q)} \int_{\mathbb{R}^k \times \mathbb{R}^k} \|\boldsymbol{x} - \boldsymbol{y}\|^k \mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}) \right)^{1/k} \leq \left( \inf_{\pi \in \Pi(p,q)} \int_{\mathbb{R}^k \times \mathbb{R}^k} \|\boldsymbol{x} - \boldsymbol{y}\|^{k'} \mathrm{d}\pi(\boldsymbol{x},\boldsymbol{y}) \right)^{1/k'}$$

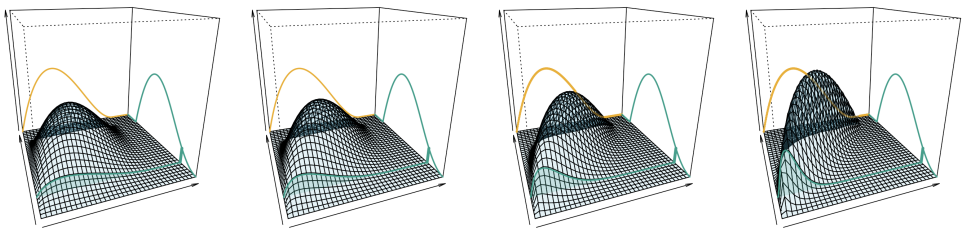
# Optimal transport

- Wasserstein/Monge-Kantorovich distance between probability measures $\mathbb{P}_A$ and $\mathbb{P}_B$: how much (kinetic) energy does it require to move a mass from $\mathbb{P}_A$ to $\mathbb{P}_B$ ?
- $\Pi(\mathbb{P}_A, \mathbb{P}_B)$ denotes the set of joint probabilities with "marginals" $\mathbb{P}_A$ and $\mathbb{P}_B$



# Optimal transport

- Wasserstein/Monge-Kantorovich distance between probability measures $\mathbb{P}_A$ and $\mathbb{P}_B$: how much (kinetic) energy does it require to move a mass from $\mathbb{P}_A$ to $\mathbb{P}_B$ ?
- $\Pi(\mathbb{P}_A, \mathbb{P}_B)$ denotes the set of joint probabilities with "marginals" $\mathbb{P}_A$ and $\mathbb{P}_B$



# Optimal transport

- If $P \in \Pi(\mathbb{P}_{\mathtt{A}}, \mathbb{P}_{\mathtt{B}})$, for all $\mathcal{A}$ and $\mathcal{B}$,

$$P(\mathcal{A} \times \mathcal{Y}) = \mathbb{P}_{\mathtt{A}}(\mathcal{A}) \text{ and } P(\mathcal{X} \times \mathcal{B}) = \mathbb{P}_{\mathtt{B}}(\mathcal{B})$$

or, put differently, for all $\psi$ and $\varphi$,

$$\int_{\mathcal{X}\times\mathcal{Y}} (\psi(\boldsymbol{x}) + \varphi(\boldsymbol{y}))\, dP(\boldsymbol{x}, \boldsymbol{y}) = \int_{\mathcal{X}} \psi(\boldsymbol{x}) d\mathbb{P}_{\mathtt{A}}(\boldsymbol{x}) + \int_{\mathcal{Y}} \varphi(\boldsymbol{y}) d\mathbb{P}_{\mathtt{B}}(\boldsymbol{y}),$$

or equivalently

$$\sup_{\psi,\varphi} \left\{ \int_{\mathcal{X}} \psi(\boldsymbol{x}) d\mathbb{P}_{\mathtt{A}}(\boldsymbol{x}) + \int_{\mathcal{Y}} \varphi(\boldsymbol{y}) d\mathbb{P}_{\mathtt{B}}(\boldsymbol{y}) - \int_{\mathcal{X}\times\mathcal{Y}} (\psi(\boldsymbol{x}) + \varphi(\boldsymbol{y}))\, dP(\boldsymbol{x}, \boldsymbol{y}) \right\} = 0.$$

Wasserstein-$k$ distance (with $k \geq 1$) is

$$W_k(\mathbb{P}_{\mathtt{A}}, \mathbb{P}_{\mathtt{B}}) = \left( \inf_{P \in \Pi(\mathbb{P}_{\mathtt{A}}, \mathbb{P}_{\mathtt{B}})} \mathbb{E}_P\left[c(X, Y)^k\right] \right)^{1/k} \quad \text{where } (X, Y) \sim P.$$

# Optimal transport

**Definition 4.35: Kantorovich Problem**

Kantorovich Problem is defined as

$$W_c(\mathbb{P}_A, \mathbb{P}_B) = \inf_{P \in \Pi(\mathbb{P}_A, \mathbb{P}_B)} \int_{\mathcal{X}\times\mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}P(\boldsymbol{x}, \boldsymbol{y}),$$

for cost function $c$ (or loss function).

- We can rewrite using a Lagrangian form

$$W_c(\mathbb{P}_A, \mathbb{P}_B) = \inf_{P} \Big\{ \sup_{\psi,\varphi} \{L(P, \psi, \varphi)\} \Big\},$$

where

$$L(P, \psi, \varphi) = \int_{\mathcal{X}\times\mathcal{Y}} (c(\boldsymbol{x}, \boldsymbol{y}) - (\psi(\boldsymbol{x}) + \varphi(\boldsymbol{y})) \mathrm{d}P(\boldsymbol{x}, \boldsymbol{y}) + \int_{\mathcal{X}} \psi(\boldsymbol{x}) d\mathbb{P}_A(\boldsymbol{x}) + \int_{\mathcal{Y}} \varphi(\boldsymbol{y}) \mathrm{d}\mathbb{P}_B(\boldsymbol{y}).$$

# Optimal transport

$$L(P,\psi,\varphi)=\int_{\mathcal{X}\times\mathcal{Y}}(c(\boldsymbol{x},\boldsymbol{y})-(\psi(\boldsymbol{x})+\varphi(\boldsymbol{y}))\mathrm{d}P(\boldsymbol{x},\boldsymbol{y})+\int_{\mathcal{X}}\psi(\boldsymbol{x})d\mathbb{P}_{\mathtt{A}}(\boldsymbol{x})+\int_{\mathcal{Y}}\varphi(\boldsymbol{y})\mathrm{d}\mathbb{P}_{\mathtt{B}}(\boldsymbol{y}).$$

- The dual problem becomes

$$\sup_{\psi,\varphi}\Big\{\inf_{P}\{L(P,\psi,\varphi)\}\Big\},$$

clearly

$$\inf_{P}\{L(P,\psi,\varphi)\}=\int_{\mathcal{X}}\psi(\boldsymbol{x})\mathrm{d}\mathbb{P}_{\mathtt{A}}(\boldsymbol{x})+\int_{\mathcal{Y}}\varphi(\boldsymbol{y})\mathrm{d}\mathbb{P}_{\mathtt{B}}(\boldsymbol{y})\text{ if }c(\boldsymbol{x},\boldsymbol{y})\geq\psi(\boldsymbol{x})+\varphi(\boldsymbol{y}).$$

# Optimal transport

**Theorem 4.1: Minimax theorem, von Neumann (1928)**

Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be compact convex sets. If $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a continuous function that is concave-convex, i.e.

$$\begin{cases} f(\cdot, y) : \mathcal{X} \to \mathbb{R} \text{ is concave for fixed } y \\ f(x, \cdot) : \mathcal{Y} \to \mathbb{R} \text{ is convex for fixed } x, \end{cases}$$

Then we have that

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f(x, y) = \min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} f(x, y).$$

- For example, $f(x, y) = x^\top A y$ for a finite matrix $A \in \mathbb{R}^{n \times m}$,

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} x^\top A y = \min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} x^\top A y.$$

# Optimal transport

- Using a generalized minimax theorem, Parthasarathy (1970),

$$W_c(\mathbb{P}_A, \mathbb{P}_B) = \sup_{\psi,\varphi} \left\{ \int_{\mathcal{X}} \psi(\boldsymbol{x}) \mathrm{d}\mathbb{P}_A(\boldsymbol{x}) + \int_{\mathcal{Y}} \varphi(\boldsymbol{y}) \mathrm{d}\mathbb{P}_B(\boldsymbol{y}) \right\}, \text{ s.t. } c(\boldsymbol{x},\boldsymbol{y}) \geq \psi(\boldsymbol{x}) + \varphi(\boldsymbol{y}),$$

the constraint can be rewritten

$$\psi(\boldsymbol{x}) \leq \overbrace{\min_{\boldsymbol{z}} \{c(\boldsymbol{x},\boldsymbol{z}) - \psi(\boldsymbol{z})\}}^{=\varphi^c(\boldsymbol{x})}$$

where $\varphi^c$ is the $c$-transform of $\varphi$, and

$$W_c(\mathbb{P}_A, \mathbb{P}_B) = \sup_{\varphi} \left\{ \int_{\mathcal{X}} \varphi^c(\boldsymbol{x}) \mathrm{d}\mathbb{P}_A(\boldsymbol{x}) + \int_{\mathcal{Y}} \phi(\boldsymbol{y}) \mathrm{d}\mathbb{P}_B(\boldsymbol{y}) \right\}$$

called Kantorovich duality formula.

# Optimal transport

**Definition 4.36: Bures metric on positive-definite matrices, Bures (1969)**

Bures distance between two positive-definite matrices,

$$d_B(\boldsymbol{A}, \boldsymbol{B}) = \sqrt{\text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B}) - 2\text{tr}((\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{\frac{1}{2}})^{\frac{1}{2}})}$$

is a metric on the space of positive semidefinite matrices

- It is related to Hellinger distance if $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal.

For two Gaussian distributions

$$W_2(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))^2 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + d_B(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)^2.$$

# Optimal transport and Monge mapping

**Definition 4.37: Push-Forward and Transport Map**

Given two metric spaces $\mathcal{X}$ and $\mathcal{Y}$, a measurable map $T : \mathcal{X} \to \mathcal{Y}$ and a measure $\mu$ on $\mathcal{X}$. The push-forward of $\mu$ by $T$ is the measure $\nu = T_{\#}\mu$ on $\mathcal{Y}$ defined by

$$\forall B \subset \mathcal{Y}, \ T_{\#}\mu(B) = \mu(T^{-1}(B)).$$

- By the change-of-variable formula

**Proposition 4.23: Push-Forward and Transport Map**

For all measurable and bounded $\varphi : \mathcal{Y} \to \mathbb{R}$,

$$\int_{\mathcal{Y}} \varphi(y) \mathrm{d} T_{\#}\mu(y) = \int_{\mathcal{X}} \varphi(T(x)) \mathrm{d}\mu(x).$$

# Optimal transport and Monge mapping

- If $\mathcal{Y}$ is a finite set $\{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$,

$$T_{\#}\mu = \sum_{i=1}^{n} \mu\big(T^{-1}(\{\boldsymbol{y}_i\})\big) \cdot \delta_{\{\boldsymbol{y}_i}$$

- If $\mathcal{X}$ is a single atom, $\{\boldsymbol{x}\}$, $\mu = \delta_{\boldsymbol{x}}$ and $T_{\#}\mu(B) = \mu\big(T^{-1}(B)\big) = \delta_{T(\boldsymbol{x})}$. If $\text{Card}(\text{support}(\nu)) > 1$, there is no transport map.
- One solution is to allow mass to split, leading to Kantorovich's relaxation of Monge's problem

**Proposition 4.24: Existence of a map**

If $\mathcal{X} = \mathcal{Y}$ is a compact subset of $\mathbb{R}^k$, if $\mu$ and $\nu$ are two measures, and if $\mu$ is atomless, then there exists $T$ such that $\nu = T_{\#}\mu$.

(see Santambrogio (2015).

# Optimal transport and Monge mapping

- If $\mathcal{X}$ and $\mathcal{Y}$ are two sets of $\mathbb{R}^k$, and if measures $\mu$ and $\nu$ are absolutely continuous, with densities $f$ and $g$ (w.r.t. Lebesgue measure),

$$\int_{\mathcal{Y}} \varphi(\boldsymbol{y}) g(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} = \int_{\mathcal{X}} \varphi(T(\boldsymbol{x})) \cdot \underbrace{g(T(\boldsymbol{x})) \det \nabla T(\boldsymbol{x})}_{=f(\boldsymbol{x})} \cdot \mathrm{d}\boldsymbol{x}.$$

**Definition 4.38: Monge Problem**

Monge problem

$$\inf_{T_{\#}\mathbb{P}_A = \mathbb{P}_B} \int_{\mathcal{X}} c(\boldsymbol{x}, T(\boldsymbol{x})) \mathrm{d}\mathbb{P}_A(\boldsymbol{x}),$$

for cost function $c$.

- Note that the constraint and the objective function are non-convex.

# Optimal transport and Monge mapping

Gangbo (1999) proved, when $\mathcal{X} = \mathcal{Y}$ is a compact subset of $\mathbb{R}$, the infimum in Monge problem and the minimum in Kantorovich problem coincide, if $\mu$ is atomless,

**Proposition 4.25: Monge/Kantorovich Problems**

$\mathcal{X} = \mathcal{Y}$ is a compact subset of $\mathbb{R}^k$ and if $\mu$ is atomless,

$$\min\{\text{Monge problem, see Def. 4.38}\} = \min\{\text{Kantorovich problem, see Def. 4.35}\}.$$

# Optimal transport and Monge mapping

- One can consider optimal transport for empirical measures, $\mathbb{P} = \sum_{i=1}^{n} \omega_i \delta_{\boldsymbol{x}_i}$.
- With uniform weights and $n$ points for $\mathbb{P}_{\mathtt{A}}$ and $\mathbb{P}_{\mathtt{B}}$, $W_k^k$ is the optimal matching cost (Hungarian algorithm, Kuhn (1955, 1956)), cast as a linear program

$$W_k(\mathbb{P}_{\mathtt{A}}, \mathbb{P}_{\mathtt{B}}) = \left( \min_{s \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_{s(i)})^k \right)^{1/k},$$

where $\mathcal{S}_n$ is the set of permutations on $\{1, 2, \cdots, n\}$.

# Optimal transport (discrete)

- In a very general setting (with $n_{\mathtt{A}} \neq n_{\mathtt{B}}$), if $\boldsymbol{a}_{\mathtt{A}} \in \mathbb{R}_+^{n_{\mathtt{A}}}$ and $\boldsymbol{a}_{\mathtt{B}} \in \mathbb{R}_+^{n_{\mathtt{B}}}$ satisfy $\boldsymbol{a}_{\mathtt{A}}^\top 1_{n_{\mathtt{A}}} = \boldsymbol{a}_{\mathtt{B}}^\top 1_{n_{\mathtt{B}}}$ (identical sums), define

$$U(\boldsymbol{a}_{\mathtt{A}}, \boldsymbol{a}_{\mathtt{B}}) = \left\{ M \in \mathbb{R}_+^{n_{\mathtt{A}} \times n_{\mathtt{B}}} : M\mathbf{1}_{n_{\mathtt{B}}} = \boldsymbol{a}_0 \text{ and } M^\top \mathbf{1}_{n_{\mathtt{A}}} = \boldsymbol{a}_1 \right\}.$$

This set of matrices is a convex transportation polytope (see Brualdi (2006)).

- In our case, let $U_{n_{\mathtt{A}}, n_{\mathtt{B}}}$ denote $U\left(\mathbf{1}_{n_{\mathtt{A}}}, \frac{n_{\mathtt{A}}}{n_{\mathtt{B}}}\mathbf{1}_{n_{\mathtt{B}}}\right)$ ($U_{n,n}$ is the set of permutation matrices associated with $\mathcal{S}_n$). Let $C$ denote the cost matrix, $C_{i,j} = d(x_i, y_j)^k$.

$$W_k^k(\boldsymbol{x}, \boldsymbol{y}) = \underset{P \in U_{n_{\mathtt{A}}, n_{\mathtt{B}}}}{\operatorname{argmin}} \left\{ \langle P, C \rangle \right\}, \text{ where } \langle P, C \rangle = \sum_{i=1}^{n_{\mathtt{A}}} \sum_{j=1}^{n_{\mathtt{B}}} P_{i,j} C_{i,j} \tag{1}$$

and "optimal transport"

$$P^* \in \underset{P \in U_{n_A, n_B}}{\operatorname{argmin}} \left\{ \langle P, C \rangle \right\} \tag{2}$$

# Optimal transport (discrete)

- From Kantorovich (1942), one can use the dual linear programming problem

$$W_k^k(\boldsymbol{a}, \boldsymbol{b}) = \begin{cases} \text{primal}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{C}) = \min\limits_{P \in U_{\boldsymbol{a},\boldsymbol{b}}} \{\langle P, \boldsymbol{C} \rangle\} \\ \quad \text{or} \\ \text{dual}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{C}) = \max\limits_{(\boldsymbol{u},\boldsymbol{v}) \in M_{\boldsymbol{C}}} \{\boldsymbol{u}^\top \boldsymbol{a} + \boldsymbol{v}^\top \boldsymbol{b}\} \end{cases}$$

where $M_{\boldsymbol{C}} = \{(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{n_A + n_B} | u_i + v_j \leq \boldsymbol{C}_{i,j}\}$.

- If $n_A \sim n_B \sim n$, $O(n^3 \log(n))$ problem.

Set $\psi_{\boldsymbol{b}}(\boldsymbol{a}, \boldsymbol{C}) = \max\limits_{(\boldsymbol{u},\boldsymbol{v}) \in M_{\boldsymbol{C}}} \{\boldsymbol{u}^\top \boldsymbol{a} + \boldsymbol{v}^\top \boldsymbol{b}\}$, $\boldsymbol{a} \mapsto \psi_{\boldsymbol{b}}(\boldsymbol{a}, \boldsymbol{C})$ is a convex non-smooth map.

- The dual optimum $\boldsymbol{u}^\star$ is subgradient of $\boldsymbol{a} \mapsto \psi_{\boldsymbol{b}}(\boldsymbol{a}, \boldsymbol{C})$.
- If $k = 2$ (Euclidean distance), convex quadratic problem.

# Optimal transport (discrete)

- Given $P \in U_{n_A, n_B}$, define the entropy as

$$\mathcal{E}(P) = -\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} P_{i,j} \log P_{i,j} \text{ or } \mathcal{E}'(P) = -\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} P_{i,j}\big[\log P_{i,j} - 1\big]$$

and consider the $\gamma$-regularized optimal transport problem

$$P_\gamma^* = \underset{P \in U_{n_A, n_B}}{\operatorname{argmin}} \Big\{ \langle P, C\rangle - \gamma \mathcal{E}(P) \Big\} \tag{3}$$

since the problem is strictly convex.

- The Lagrangian is here

$$\mathcal{L}(P, \lambda_A, \lambda_B) = \langle P, C\rangle - \gamma\mathcal{E}(P) - \langle \lambda_A, P\mathbf{1}_{n_B} - \mathbf{1}_{n_A}\rangle - \langle \lambda_B, P^\top \mathbf{1}_{n_A} - \mathbf{1}_{n_B}\rangle$$

# Optimal transport (discrete)

and the first order conditions are

$$C_{i,j} + \gamma \log(P_{i,j}) - \lambda_{\mathtt{A},i} - \lambda_{\mathtt{B},j} = 0,$$

i.e.

$$P_{i,j} = \exp[\lambda_{\mathtt{A},i} - C_{i,j} + \lambda_{\mathtt{B},j}] \text{ or } P = D_{\mathtt{A}} \exp[-C] D_{\mathtt{B}}$$

where $D_{\mathtt{A}}$ and $D_{\mathtt{B}}$ are diagonal matrices.

- This can be related to the Doubly Stochastic Scaling Problem: let $A$ be some $n \times n$ matrix with positive coefficients, we want to find $D_{\mathtt{A}}$ and $D_{\mathtt{B}}$ two positive diagonal matrices ($n \times n$) such that $D_{\mathtt{A}} A D_{\mathtt{B}}$ is doubly stochastic (see Parlett and Landis (1982))
- More generally, this corresponds to the Matrix Scaling Problem: Let $A$ be some $n_{\mathtt{A}} \times n_{\mathtt{B}}$ matrix with positive coefficients, we want to find $D_{\mathtt{A}}$ and $D_{\mathtt{B}}$ two positive diagonal matrices (respectively $n_{\mathtt{A}} \times n_{\mathtt{A}}$ and $n_{\mathtt{B}} \times n_{\mathtt{B}}$) such that $D_{\mathtt{A}} A D_{\mathtt{B}}$ is in $U(\boldsymbol{a}_{\mathtt{A}}, \boldsymbol{a}_{\mathtt{B}})$.

# Optimal transport (discrete)

**Theorem 4.2: Sinkhorn - Matrix Scaling, Sinkhorn (1962)**

For any matrix $\boldsymbol{A}$ $n \times m$ with positive entries, for any $\boldsymbol{a}$ and $\boldsymbol{b}$ in the simplex, there exist unique $\boldsymbol{u} \in \mathbb{R}_+^n$ and $\boldsymbol{v} \in \mathbb{R}_+^m$ such that

$$\text{diag}[\boldsymbol{u}]\ \boldsymbol{A}\ \text{diag}[\boldsymbol{v}] \in U_{\boldsymbol{a},\boldsymbol{b}}.$$

› Sinkhorn and Knopp (1967) (extending Sinkhorn (1962, 1964, 1966)) suggested the following algorithm (updating alternatively $D_{\mathtt{A}}$ and $D_{\mathtt{B}}$)

$$\begin{cases} D_{\mathtt{A}}^{(t)} = \text{diag}(\boldsymbol{a}_{\mathtt{A}}/(AD_{\mathtt{B}})^{(t-1)}) \\ D_{\mathtt{B}}^{(t)} = \text{diag}(\boldsymbol{a}_{\mathtt{B}}/(AD_{\mathtt{A}})^{(t)}) \end{cases}$$

(where the division here is element-wise).

# Optimal transport (discrete)

- An alternative way to write the entropic optimization problem is

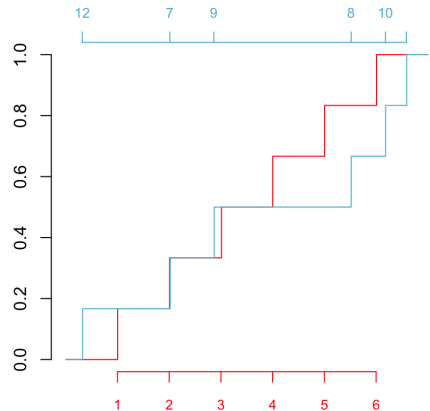
$$P_\gamma^* = \underset{P \in U_{\boldsymbol{a}_A, \boldsymbol{a}_B}}{\operatorname{argmin}} \left\{ \langle P, C \rangle + \gamma \mathrm{KL}(P || \boldsymbol{a}_A \otimes \boldsymbol{a}_B) \right\} \tag{4}$$

Using mutual information here makes it easier to extend to the continuous case...

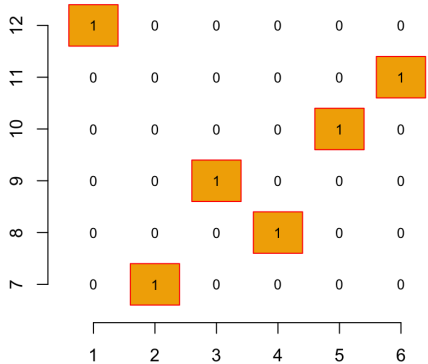
- The extension of Sinkhorn algorithm is the coordinate descent/ascent algorithm.

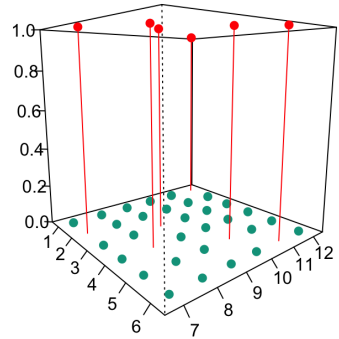
# Optimal transport (discrete)

```
> set.seed(123)
> x = (1:6)/7
> y = runif(9)
> x
[1] 0.14 0.29 0.43 0.57 0.71 0.86
> y[1:6]
[1] 0.29 0.79 0.41 0.88 0.94 0.05
> library(T4transport)
> Wxy = wasserstein(x,y[1:6])
> Wxy$plan
```



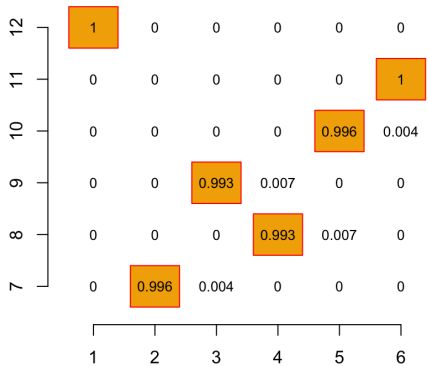
# Optimal transport (discrete)

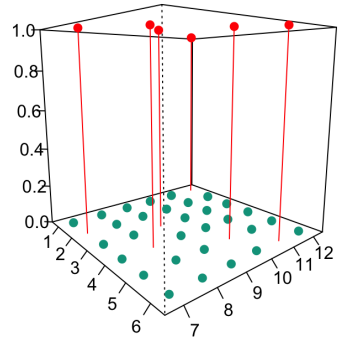



```
> Wxy = wasserstein(x,y[1:6])
> Wxy$plan
```

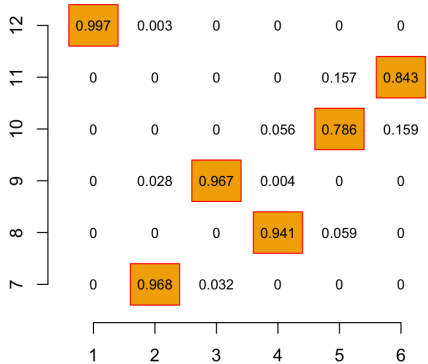
# Optimal transport (discrete)

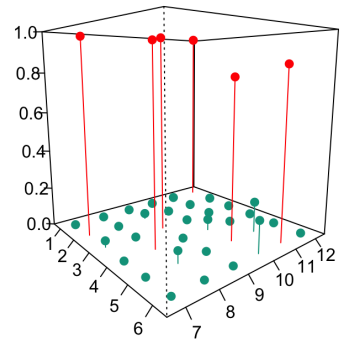



```
> Sxy = sinkhorn(x, y[1:6], p = 2, lambda = 0.001)
> Sxy$plan
```

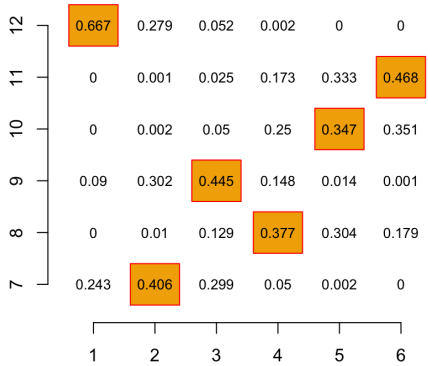
# Optimal transport (discrete)

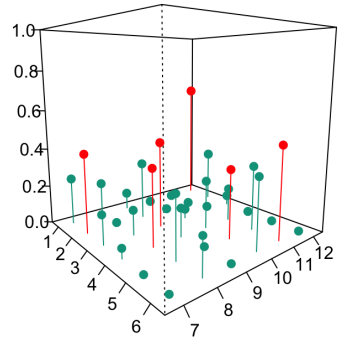



```
> Sxy = sinkhorn(x, y[1:6], p = 2, lambda = 0.005)
> Sxy$plan
```

# Optimal transport (discrete)

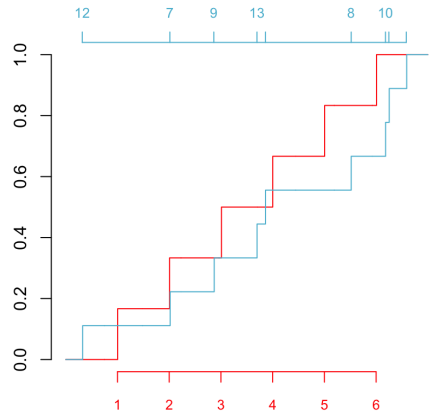


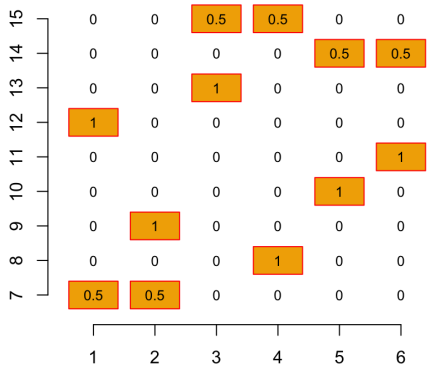

```
> Sxy = sinkhorn(x, y[1:6], p = 2, lambda = 0.05)
> Sxy$plan
```

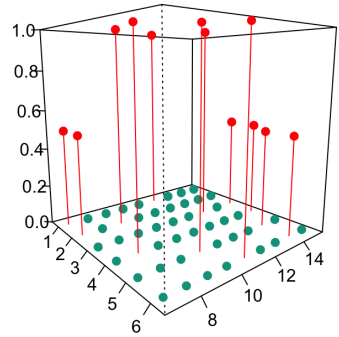
# Optimal transport (discrete)

```
> y
[1] 0.29 0.79 0.41 0.88 0.94 0.05
[7] 0.53 0.89 0.55
> library(T4transport)
> Wxy = wasserstein(x,y)
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]  0.5  0.5  0.0  0.0  0.0  0.0
[2,]  0.0  0.0  0.0  1.0  0.0  0.0
[3,]  0.0  1.0  0.0  0.0  0.0  0.0
[4,]  0.0  0.0  0.0  0.0  1.0  0.0
[5,]  0.0  0.0  0.0  0.0  0.0  1.0
[6,]  1.0  0.0  0.0  0.0  0.0  0.0
[7,]  0.0  0.0  1.0  0.0  0.0  0.0
[8,]  0.0  0.0  0.0  0.0  0.5  0.5
[9,]  0.0  0.0  0.5  0.5  0.0  0.0
```



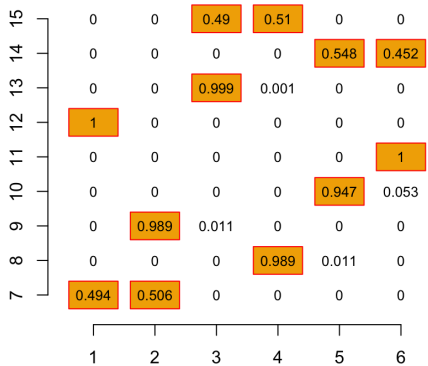
# Optimal transport (discrete)

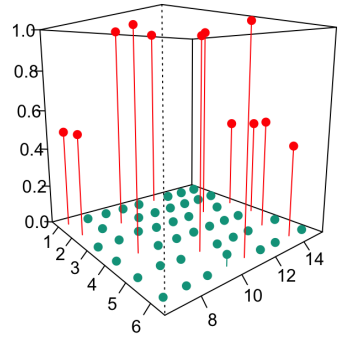



```
> Wxy = wasserstein(x,y)
> Wxy$plan
```

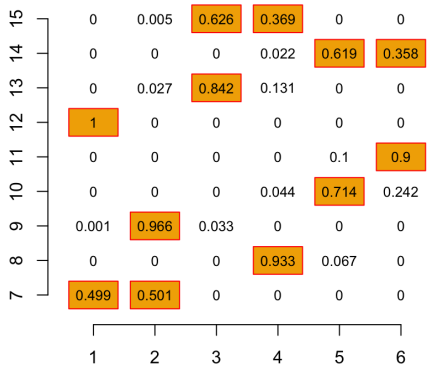
# Optimal transport (discrete)

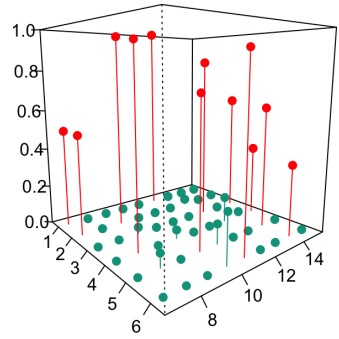



```
> Sxy = sinkhorn(x, y, p = 2, lambda = 0.001)
> Sxy$plan
```

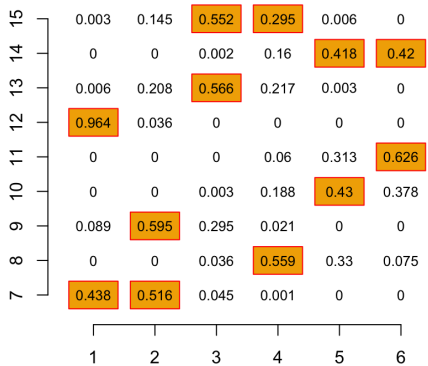
# Optimal transport (discrete)

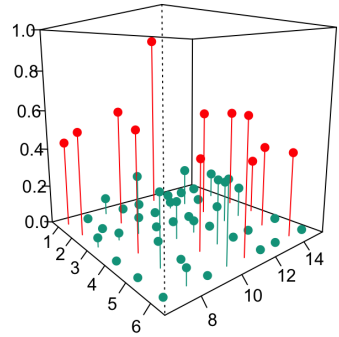



```
> Sxy = sinkhorn(x, y, p = 2, lambda = 0.005)
> Sxy$plan
```

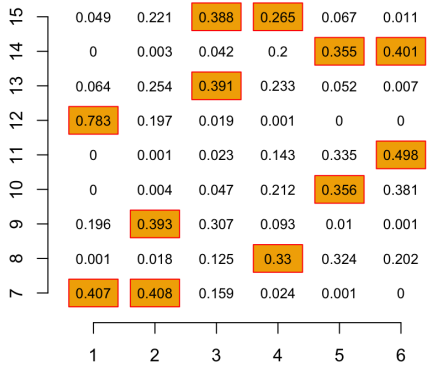
# Optimal transport (discrete)

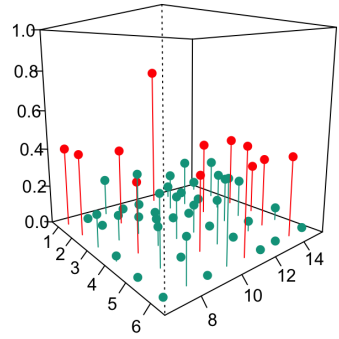



```
> Sxy = sinkhorn(x, y, p = 2, lambda = 0.02)
> Sxy$plan
```

# Optimal transport (discrete)




```
> Sxy = sinkhorn(x, y, p = 2, lambda = 0.05)
> Sxy$plan
```

# Optimal transport (discrete)

# Univariate Optimal Transport

**Proposition 4.26: Hardy–Littlewood–Pólya inequality, Hardy et al. (1952)**

Given $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$ $n$ pairs of ordered real numbers, for every permutation $\sigma$ of $\{1, 2, \cdots, n\}$,

$$\sum_{i=1}^{n} x_i y_{n+1-i} \leq \sum_{i=1}^{n} x_i y_{\sigma(i)} \leq \sum_{i=1}^{n} x_i y_i.$$

› This can be extended, from a product to more general function $\Phi(x_i, y_j)$.

# Univariate Optimal Transport

**Definition 4.39: Supermodular, Topkis (1998)**

Function $\Phi : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ is supermodular if for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^k$,

$$\Phi(\boldsymbol{z} \wedge \boldsymbol{z}') + \Phi(\boldsymbol{z} \vee \boldsymbol{z}') \geq \Phi(\boldsymbol{z}) + \Phi(\boldsymbol{z}'),$$

where $\boldsymbol{z} \wedge \boldsymbol{z}'$ and $\boldsymbol{z} \vee \boldsymbol{z}'$ denote respectively the maximum and the minimum componentwise. If $-\Phi$ is supermodular, $\Phi$ is said to be submodular.

# Univariate Optimal Transport

**Proposition 4.27: Hardy–Littlewood–Pólya inequality, Hardy et al. (1952)**

Given $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_n$ $n$ pairs of ordered real numbers, and some supermodular function $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, for every permutation $\sigma$ of $\{1, 2, \cdots, n\}$,

$$\sum_{i=1}^{n} \Phi(x_i, y_{n+1-i}) \leq \sum_{i=1}^{n} \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^{n} \Phi(x_i, y_i),$$

while if $\Phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is submodular,

$$\sum_{i=1}^{n} \Phi(x_i, y_i) \leq \sum_{i=1}^{n} \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^{n} \Phi(x_i, y_{n+1-i}).$$

› Functions $\Phi(x, y) = \gamma(x - y)$ for some concave function $\gamma : \mathbb{R} \to \mathbb{R}$, such as $\Phi(x, y) = -|x - y|^k$ with $k \geq 1$, are supermodular.

# Univariate Optimal Transport

```
> permutations = function(n){
+   if(n==1){
+     return(matrix(1))
+   } else {
+   sp = permutations(n-1)
+   p = nrow(sp)
+   A = matrix(nrow=n*p,ncol=n)
+   for(i in 1:n){
+       A[(i-1)*p+1:p,] =
+           cbind(i,sp+(sp>=i))
+         }
+   return(A)
+     }
+ }
```

$\Phi(x, y) = (x - y)^2$, submodular function,

- Consider $x_1 \leq \cdots \leq x_n$

```
> Phi = function(x,y) sum((x-y)^2)
> set.seed(1)
> x = sort(x)
> y = y[1:6]
> vect = permutations(6)
> MY = matrix(vect, ncol=6)
> MPhi = function(i) Phi(x, y[MY[i,]])
> S = Vectorize(MPhi)(1:nrow(MY))
> y[MY[which.min(S),]]
[1] 0.046 0.288 0.409 0.788 0.883 0.940
```

# Univariate Optimal Transport

**Theorem 4.3: Optimal transport for discrete univariate distributions**

Consider $n$ points each group, on $\mathbb{R}$, $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_n\}$, ordered in the senses that $x_1 \leq x_2 \leq \cdots \leq x_n$ and $y_1 \leq y_2 \leq \cdots \leq y_n$, for any $k \geq 1$,

$$W_k = \left( \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

**Theorem 4.4: Optimal transport for continuous univariate distributions**

$$W_k = \left( \int_0^1 |F_x^{-1}(u) - F_y^{-1}(u)|^k du \right)^{1/k}$$

# Univariate Optimal Transport

**Theorem 4.5: Optimal transport for continuous univariate distributions**

Let $\mathbb{P}_{\mathtt{A}}$ and $\mathbb{P}_{\mathtt{B}}$ be two probability measures on $\mathbb{R}$, and suppose that $c(x, y) = h(x-y)$ for some strictly convex function $h$. The there exists a unique $\pi \in \Pi(\mathbb{P}_{\mathtt{A}}, \mathbb{P}_{\mathtt{B}})$ such that

- $\pi$ is optimal to Kantorovich problem (4.35)
- $\pi$ is the comonotone joint distribution with marginals $\mathbb{P}_{\mathtt{A}}$ and $\mathbb{P}_{\mathtt{B}}$.

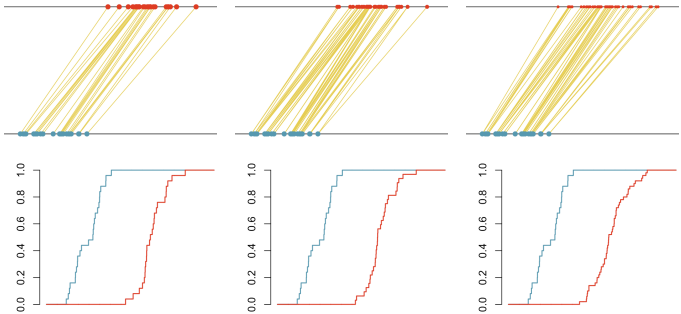
› If $c(x, y) = |x - y|$, the optimal transport solution might be non-unique.

**Theorem 4.6: Optimal map for continuous univariate distributions**

The optimal Monge map $T^\star$ such that $T^\star_\# \mathbb{P}_{\mathtt{A}} = \mathbb{P}_{\mathtt{B}}$ is $T^\star = F_{\mathtt{B}}^{-1} \circ F_{\mathtt{A}}$.

# Univariate Optimal Transport

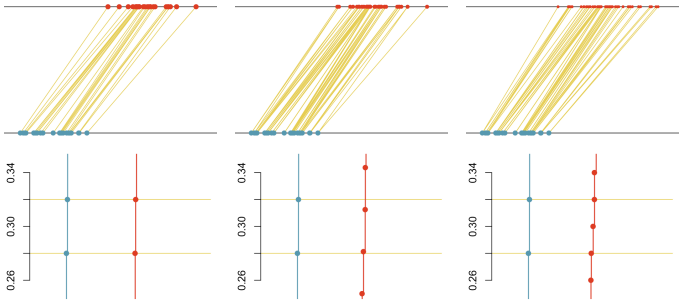
Consider $n_A = 25$ and $n_B = 25$ points in $\mathbb{R}$, $n_B = 32$ and $n_B = 50$




$$\widehat{F}_{n_A}(x) = \frac{1}{n_A}\sum_{i=1}^{n_A} \mathbf{1}(x_i \leq x) \text{ and } \widehat{F}_{n_B}(x) = \frac{1}{n_B}\sum_{i=1}^{n_B} \mathbf{1}(x_i \leq x)$$

# Univariate Optimal Transport

- Consider $n_A = 25$ and $n_B = 25$ points in $\mathbb{R}$, $n_B = 32$ and $n_B = 50$

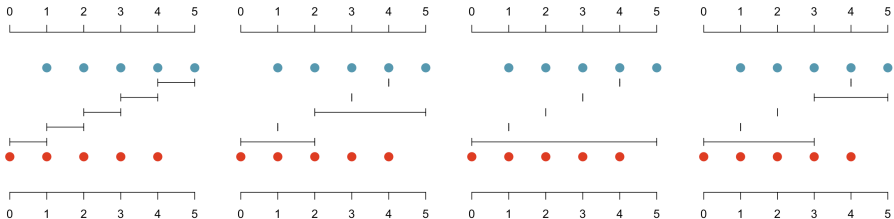



$$\widehat{F}_{n_A}(x) = \frac{1}{n_A}\sum_{i=1}^{n_A} \mathbf{1}(x_i \leq x) \text{ and } \widehat{F}_{n_B}(x) = \frac{1}{n_B}\sum_{i=1}^{n_B} \mathbf{1}(x_i \leq x)$$

# Univariate Optimal Transport

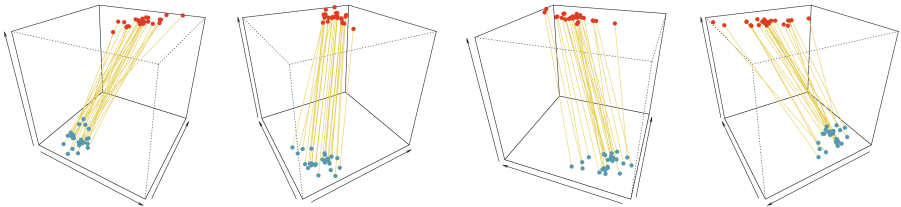
- In the univariate case, if $k = 1$,

$$W_1 = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - y_{\sigma(i)} \right|$$

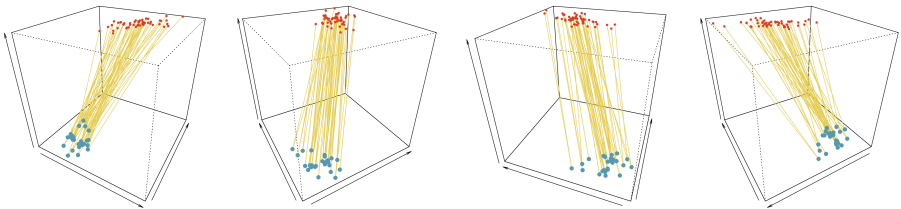


# Multivariate Optimal Transport

- Consider $n$ and $n$ points in $\mathbb{R}^2$

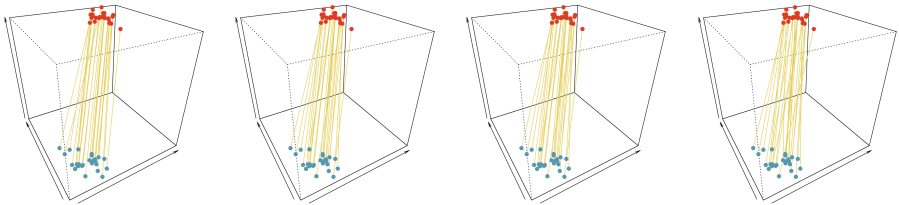


- Consider $n$ and $2n$ points in $\mathbb{R}^2$

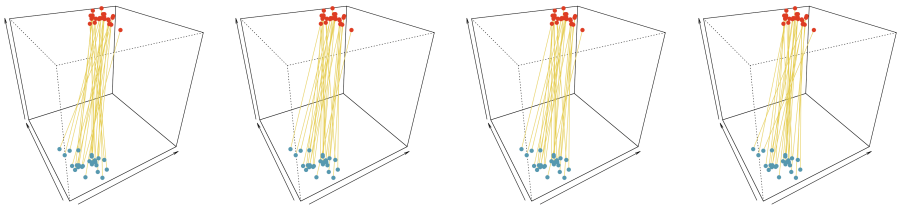


# Multivariate Optimal Transport

- Consider $n$ and $n$ points in $\mathbb{R}^2$, and $p = 1, 2, 3, 4$, $T_{\#}\mathbb{P}_{\mathtt{A}} = \mathbb{P}_{\mathtt{B}}$



- Consider $n$ and $n$ points in $\mathbb{R}^2$, and $p = 1, 2, 3, 4$, $T_{\#}\mathbb{P}_{\mathtt{B}} = \mathbb{P}_{\mathtt{A}}$



# References


Abraham, K. (1986). *Distributing risk: Insurance, legal theory and public policy*. Yale University Press,.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

Amari, S.-i. (2016). *Information geometry and its applications*, volume 194. Springer.

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 249–260.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*, May 23.

Austin, R. (1983). The insurance classification controversy. *University of Pennsylvania Law Review*, 131(3):517–583.

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

# References


Avraham, R., Logue, K. D., and Schwarcz, D. (2013). Understanding insurance antidiscrimination law. *Southern California Law Review*, 87:195.

Bailey, R. A. and Simon, L. J. (1959). An actuarial note on the credibility of experience of a single private passenger car. *Proceedings of the Casualty Actuarial Society*, XLVI:159.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(10).

Barbour, V. (1911). Privateers and pirates of the west indies. *The American Historical Review*, 16(3):529–566.

Barry, L. and Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society*, 7(1):2053951720935143.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Bauschke, H. H., Borwein, J. M., et al. (1997). Legendre functions and the method of random bregman projections. *Journal of convex analysis*, 4(1):27–67.

Bauschke, H. H. and Lewis, A. S. (2000). Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427.

# References


Bellemare, M. G., Dabney, W., and Munos, R. (2017a). A distributional perspective on reinforcement learning. *arXiv:1707.06887*.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. (2017b). The cramer distance as a solution to biased wasserstein gradients. *arXiv:1705.10743*.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404.

Blanpain, N. (2018). L'espérance de vie par niveau de vie-méthode et principaux résultats. *INSEE Document de Travail*, F1801.

Bourdieu, P. (2018). Distinction a social critique of the judgement of taste. In *Inequality Classic Readings in Race, Class, and Gender*, pages 287–318. Routledge.

Box, G. E., Luceño, A., and del Carmen Paniagua-Quinones, M. (2011). *Statistical control by monitoring and adjustment*, volume 700. John Wiley & Sons.

Boyle, J. P. and Dykstra, R. L. (1986). A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in Order Restricted Statistical Inference: Proceedings of the Symposium on Order Restricted Statistical Inference held in Iowa City, Iowa, September 11–13, 1985*, pages 28–47. Springer.

# References


Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Brilmayer, L., Hekeler, R. W., Laycock, D., and Sullivan, T. A. (1979). Sex discrimination in employer-sponsored insurance plans: A legal and demographic analysis. *University of Chicago Law Review*, 47:505.

Brualdi, R. A. (2006). *Combinatorial matrix classes*, volume 13. Cambridge University Press.

Bures, D. (1969). An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite $w^*$-algebras. *Transactions of the American Mathematical Society*, 135:199–212.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Duxbury Advanced Series.

Casey, B., Pezier, J., and Spetzler, C. (1976). *The Role of Risk Classification in Property and Casualty Insurance: A Study of the Risk Assessment Process : Final Report*. Stanford Research Institute.

Censor, Y. and Reich, S. (1998). The dykstra algorithm with bregman projections. *Communications in Applied Analysis*, 2(3):407–420.

# References


Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions, intl. *Journal of Mathematical Models and Methods in Applied Sciences, Issue*, 4.

Chambert-Loir, A. (2023). *Information Theory: Three Theorems by Claude Shannon*, volume 144. Springer Nature.

Charpentier, A. (2023). *Insurance: biases, discrimination and fairness*. Springer Verlag.

Cheney-Lippold, J. (2017). We are data. In *We Are Data*. New York University Press.

Cramér, H. (1928a). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Cramér, H. (1928b). On the composition of elementary errors: second paper: statistical applications. *Scandinavian Actuarial Journal*, 1928(1):141–180.

Crossney, K. B. (2016). Redlining. *https://philadelphiaencyclopedia.org/essays/redlining/*.

Csiszár, I. (1964). Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8:85–108.

Csiszár, I. (1967). On information-type measure of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

# References


Da Silva, N. (2023). *La bataille de la Sécu: une histoire du système de santé*. La fabrique éditions.

De Baere, G. and Goessens, E. (2011). Gender differentiation in insurance contracts after the judgment in case c-236/09, Association Belge des Consommateurs Test-Achats asbl v. conseil des ministres. *Colum. J. Eur. L.*, 18:339.

de La Fontaine, J. (1668). *Fables*. Barbin.

De Pril, N. and Dhaene, J. (1996). Segmentering in verzekeringen. *DTEW Research Report 9648*, pages 1–56.

De Wit, G. and Van Eeghen, J. (1984). Rate making and society's sense of fairness. *ASTIN Bulletin: The Journal of the IAA*, 14(2):151–163.

Dedecker, J. and Merlevède, F. (2007). The empirical distribution function for dependent variables: asymptotic and nonasymptotic results in. *ESAIM: Probability and Statistics*, 11:102–114.

Denuit, M. and Charpentier, A. (2004). *Mathématiques de l'assurance non-vie: Tome I Principes fondamentaux de théorie du risque*. Economica.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional gaussians with the same mean. *arXiv*, 1810.08693.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

# References


Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. (2009). Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9:69–83.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.

Feeley, M. and Simon, J. (1994). Actuarial justice: The emerging new criminal law. *The futures of criminology*, 173:174.

Feeley, M. M. and Simon, J. (1992). The new penology: Notes on the emerging strategy of corrections and its implications. *Criminology*, 30(4):449–474.

Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 17.

Fourcade, M. and Healy, K. (2013). Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society*, 38(8):559–572.

# References


Fox, E. T. (2013). *'Piratical Schemes and Contracts': Pirate Articles and Their Society 1660-1730*. PhD Thesis, University of Exeter.

François, P. (2022). Catégorisation, individualisation. retour sur les scores de crédit. *hal*, 03508245.

Gandy, O. H. (2016). *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*. Routledge.

Gangbo, W. (1999). The monge mass transfer problem and its applications. *Contemporary Mathematics*, 226:79–104.

Garrioch, D. (2011). Mutual aid societies in eighteenth-century paris. *French History & Civilization*, 4.

Ginsburg, M. (1940). Roman military clubs and their social functions. In *Transactions and Proceedings of the American Philological Association*, volume 71, pages 149–156. JSTOR.

Givens, C. R. and Shortt, R. M. (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240.

Glenn, B. J. (2000). The shifting rhetoric of insurance denial. *Law and Society Review*, pages 779–808.

Glenn, B. J. (2003). Postmodernism: the basis of insurance. *Risk Management and Insurance Review*, 6(2):131–143.

# References


Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

Gowri, A. (2014). *The Irony of Insurance: Community and Commodity*. PhD thesis, University of Southern California.

Hacking, I. (1990). *The taming of chance*. Number 17. Cambridge University Press.

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, 27(4):237–243.

Hardy, G. H., Littlewood, J. E., Pólya, G., Pólya, G., et al. (1952). *Inequalities*. Cambridge university press.

Havens, H. V. (1979). Issues and needed improvements in state regulation of the insurance business. *U.S. General Acounting Office*.

# References


He, X. D., Kou, S., and Peng, X. (2022). Risk measures: robustness, elicitability, and backtesting. *Annual Review of Statistics and Its Application*, 9:141–166.

Heimer, C. A. (1985). *Reactive Risk and Rational Action*. University of California Press.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271.

Hoffman, F. L. (1896). *Race traits and tendencies of the American Negro*, volume 11. American Economic Association.

Hoffman, F. L. (1918). *Mortality from respiratory diseases in dusty trades (inorganic dusts)*. Number 231. US Government Printing Office.

Hoffman, F. L. (1931). Cancer and smoking habits. *Annals of surgery*, 93(1):50.

Hubbard, G. N. (1852). *De l'organisation des sociétés de bienfaisance ou de secours mutuels et des bases scientifiques sur lesquelles elles doivent être établies*. Paris, Guillaumin.

Huttegger, S. M. (2013). In defense of reflection. *Philosophy of Science*, 80(3):413–433.

Ismay, P. (2018). *Trust among strangers: friendly societies in modern Britain*. Cambridge University Press.

# References


Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Jordan, C. (1881). Sur la serie de fourier. *Camptes Rendus Hebdomadaires de l'Academie des Sciences*, 92:228–230.

Kantorovich, L. and Rubinstein, G. (1958). On the space of completely additive functions. *Vestnic Leningrad Univ., Ser. Mat. Mekh. i Astron.*, 13(7):52–59. In Russian.

Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Knott, M. and Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49.

Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

# References


Kuhn, H. W. (1956). Variants of the hungarian method for assignment problems. *Naval research logistics quarterly*, 3(4):253–258.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica*, 23-05.

Leeson, P. T. (2009). The calculus of piratical consent: the myth of the myth of social contract. *Public Choice*, 139:443–459.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, K. C.-W. (1996). The private insurance industry's tactics against suspected homosexuals: redlining based on occupation, residence and marital status. *American Journal of Law & Medicine*, 22(4):477–502.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Massey, D. S. (2007). *Categorically unequal: The American stratification system*. Russell Sage Foundation.

# References


Mcdonald, S. (2015). Indirect gender discrimination and the 'test-achats ruling': an examination of the uk motor insurance market. In *Royal Economic Society Conf., Manchester*.

Merriam-Webster (2022). *Dictionary*. .

Mowbray, A. (1921). Classification of risks as the basis of insurance rate making with special reference to workmen's compensation. *Proceedings of the Casualty Actuarial Society*.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

Nielsen, F. (2022). The many faces of information geometry. *Notices of the American Mathematical Society*, 69(1):36–45.

Nielsen, F. and Nock, R. (2013). On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13.

Olkin, I. and Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263.

Pardo, L. (2018). *Statistical inference based on divergence measures*. CRC press.

# References


Parlett, B. and Landis, T. (1982). Methods for scaling to doubly stochastic form. *Linear Algebra and its Applications*, 48:53–79.

Parthasarathy, T. (1970). On games over the unit square. *SIAM Journal on Applied Mathematics*, 19(2):473–476.

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Polyanskiy, Y. and Wu, Y. (2022). *Information theory: From coding to learning*. Cambridge University Press.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Proschan, M. A. and Presnell, B. (1998). Expect the unexpected from conditional expectation. *The American Statistician*, 52(3):248–252.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press.

Rhynhart, R. (2020). Mapping the legacy of structural racism in philadelphia. *Philadelphia, Office pf the Controller*.

# References


Rizzo, M. L. and Székely, G. J. (2016). Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1):27–38.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Rothstein, W. G. (2003). *Public health and the risk factor: A history of an uneven medical revolution*, volume 3. Boydell & Brewer.

Rouvroy, A., Berns, T., and Carey-Libbrecht, L. (2013). Algorithmic governmentality and prospects of emancipation. *Réseaux*, 177(1):163–196.

Rudin, W. (1966). *Real and Complex Analysis*. McGraw-hill New York.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*, volume 55. Springer.

Schanze, E. (2013). Injustice by generalization: notes on the Test-Achats decision of the european court of justice. *German Law Journal*, 14(2):423–433.

Schauer, F. (2006). *Profiles, probabilities, and stereotypes*. Harvard University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Simon, J. (1987). The emergence of a risk society-insurance, law, and the state. *Socialist Review*, (95):60–89.

# References


Simon, J. (1988). The ideological effects of actuarial practices. *Law & Society Review*, 22:771.

Sinkhorn, R. (1962). On the factor spaces of the complex doubly stochastic matrices. *Notices of the American Mathematical Society*, 9:334–335.

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879.

Sinkhorn, R. (1966). A relationship between arbitrary positive matrices and stochastic matrices. *Canadian Journal of Mathematics*, 18:303–306.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Squires, G. D. (2003). Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas. *Journal of Urban Affairs*, 25(4):391–410.

Squires, G. D. and Velez, W. (1988). Insurance redlining and the process of discrimination. *The Review of Black Political Economy*, 16(3):63–75.

Stone, D. A. (1993). The struggle for the soul of health insurance. *Journal of Health Politics, Policy and Law*, 18(2):287–317.

# References


Struyck, N. (1912). *Les oeuvres de Nicolas Struyck (1687-1769): qui se rapportent au calcul des chances, à la statistique général, la statistique des décès et aux rentes viagèter*. Société générale néerlandaise d'assurances sur la vie et de rentes viagères.

Székely, G. J. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18.

Takatsu, A. (2008). On wasserstein geometry of the space of gaussian measures. *arXiv*, 0801.2250.

Takatsu, A. and Yokota, T. (2012). Cone structure of l2-wasserstein spaces. *Journal of Topology and Analysis*, 4(02):237–253.

The Zebra (2022). Car insurance rating factors by state. *https://www.thezebra.com/*.

Topkis, D. M. (1998). *Supermodularity and complementarity*. Princeton university press.

Vallender, S. (1974). Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

Van Gerven, G. (1993). Case c-109/91, Gerardus Cornelis Ten Oever v. Stichting bedrijfspensioenfonds voor het glazenwassers-en schoonmaakbedrijf. *EUR-Lex*, 61991CC0109.

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4.

# References


Verboven, K. (2011). Introduction: Professional collegia: Guilds or social clubs? *Ancient Society*, pages 187–195.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer.

von Mises, R. (1939). *Probability, statistics and truth*. Macmillan.

von Neumann, J. (1928). Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320.

Wasserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

Wilkie, D. (1997). Mutuality and solidarity: assessing risks and sharing losses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1357):1039–1044.

Wortham, L. (1986). The economics of insurance classification: The sound of one invisible hand clapping. *Ohio State Law Journal*, 47:835.

Zolotarev, V. M. (1976). Metric distances in spaces of random variables and their distributions. *Mathematics of the USSR-Sbornik*, 30(3):373.