Using optimal transport to mitigate unfair predictions and quantify counterfactual fairness

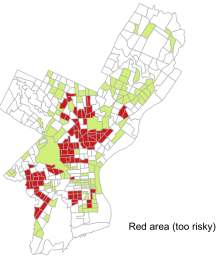
Arthur Charpentier, A. Fernandes-Machado, E. Gallic, F. Hu & P. Ratz

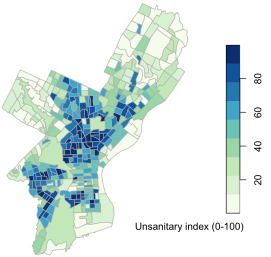
University of Toronto, March 2025

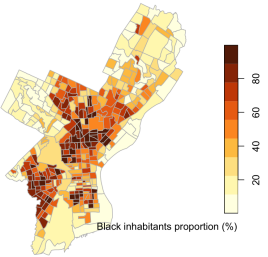
SCOR
FOUNDATION FOR SCIENCE

"Technology is neither good nor bad; nor is it neutral."

(Fictitious maps, inspired by Home Owners' Loan Corporation map, 1937, "redlining")





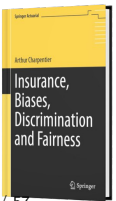


decision $\overline{y}$ | legitimate ratemaking variable x | sensitive attribute a

Direct (intentional) discrimination / indirect (statistical) discrimination / neutrality

"*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitly ask for*," Kearns and Roth (2019)

Charpentier (2024) Insurance: Biases, Discrimination and Fairness.



Agenda

Optimal Transport, the Setting

Consider measures ν_0 and ν_1 on Z_0 Z_1, compact subsets of R^d.

There exists T such that $\nu_1 = T\ \nu_0$, where ν_0 is atomless ($T\ \nu_0(B) = \nu_0\ T^{-1}(B)$).

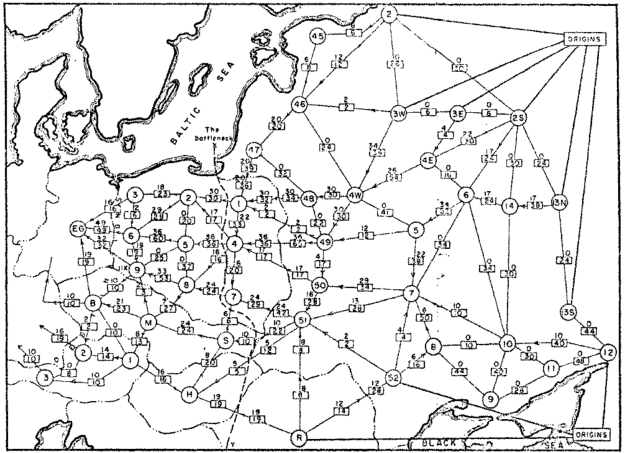
As shown in Villani (2003) and Santambrogio (2015), we can be interested in "optimal" mappings, satisfying Monge problem, from Monge (1781), i.e., solutions of

$$\inf_{T\ \nu_0=\nu_1} \int_{Z_0} c\ \mathbf{z}_0, T(\mathbf{z}_0)\ \nu_0(\mathrm{d}\mathbf{z}_0)\ , \tag{1}$$

for some positive ground cost function $c : Z_0 \times Z_1 \to \mathrm{R}_+$. If $_0$ and $_1$ are not absolutely continuous (with respect to Lebesgue measure), there might not be such deterministic mapping T. This limitation motivates a relaxation of Monge's problem, as considered in Kantorovich (1942),

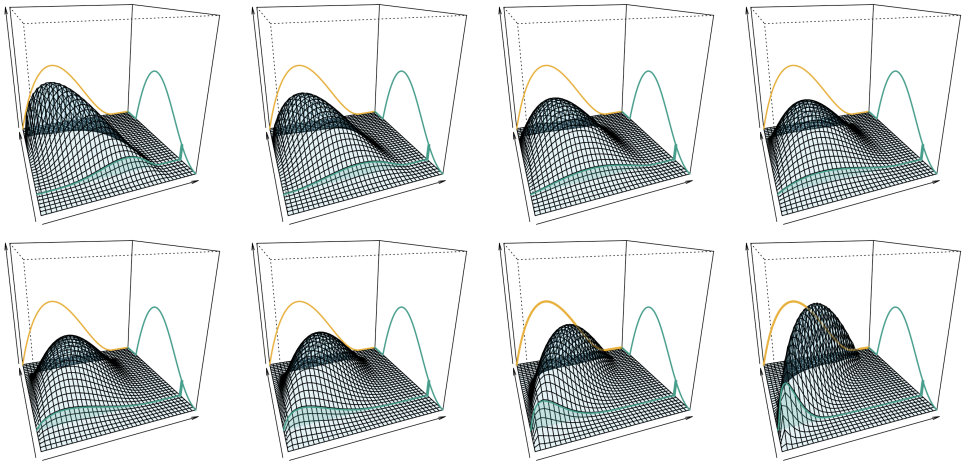
$$\inf_{\pi\ \Pi(\nu_0,\nu_1)} \int_{Z_0\times Z_1} c(\mathbf{z}_0, \mathbf{z}_1)\pi(\mathrm{d}\mathbf{z}_0, \mathrm{d}\mathbf{z}_1)\ , \tag{2}$$

Optimal Transport, the Setting



(flow on networks, via Harris and Ross (1955))

Optimal Transport, the Setting



Optimal Transport, the Gaussian case

With a quadratic cost (and Euclidean distance in R^d) and absolutely continuous measures, the optimal Monge map T is unique, and it is the gradient of a convex function, $T = \phi$, see Brenier (1991).
If $\nu_j \quad N(\mu_j, \mathbf{\Sigma}_j)$,

$$\boldsymbol{z}_1 = T(\boldsymbol{z}_0) = \mu_1 + \mathrm{M}(\boldsymbol{z}_0 - \mu_0),$$

where M is a symmetric positive matrix that satisfies $\mathrm{M}\mathbf{\Sigma}_0\mathrm{M} = \mathbf{\Sigma}_1$, which has a unique solution given by $\mathrm{M} = \mathbf{\Sigma}_0^{-1/2}\mathbf{\Sigma}_0^{1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_0^{1/2}\mathbf{\Sigma}_0^{1/2}\mathbf{\Sigma}_0^{-1/2}$, where $\boldsymbol{M}^{1/2}$ denotes the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008).
In the univariate case

$$z_1 = T(z_0) = \mu_1 + \sigma_1\sigma_0^{-1}(z_0 - \mu_0),$$

that is a linear non-decreasing mapping $\mathrm{R} \to \mathrm{R}$.

Optimal Transport, the discrete case

Consider two samples in the $\mathbb{R}^d$, $\{\boldsymbol{z}_{0,1}, \cdots, \boldsymbol{x}_{z,n_0}\}$ and $\{\boldsymbol{z}_{1,1}, \cdots, \boldsymbol{z}_{1,n_1}\}$. The discrete version of the Kantorovich problem (Equation 2) is

$$\min_{\mathrm{P}\ U(n_0,n_1)} \sum_{i=1}^{n_0}\sum_{j=1}^{n_1} \mathrm{P}_{i,j}\mathrm{C}_{i,j} = \min_{\mathrm{P}\ U(n_0,n_1)} \{\ \mathrm{P}, \mathrm{C}\ \{ \tag{3}$$

where, as in Brualdi (2006), $U(n_0, n_1)$ is the set of $n_0 \times n_1$ matrices corresponding to the convex transportation polytope

$$U(n_0, n_1) = \quad \mathrm{P} : \mathrm{P}\mathbf{1}_{n_1} = \mathbf{1}_{n_0} \text{ and } \mathrm{P}\ \ \mathbf{1}_{n_0} = \frac{n_0}{n_1}\mathbf{1}_{n_1} \quad ,$$

and where C denotes the $n_0 \times n_1$ cost matrix, $\mathrm{C}_{i,j} = c(\boldsymbol{z}_i, \boldsymbol{z}_j)$, associated with cost c.

Optimal Transport, the discrete case

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

$$\underset{\mathrm{P}\ U(\mathbf{1}_n,\mathbf{1}_n)}{\operatorname{argmin}} \langle \mathrm{P}, \mathrm{C} \rangle$$

over the set of **permutation matrices**.

minimal cost 0.2116

	7	8	9	10	11	12
1	·	·	·	·	1	·
2	·	1	·	·	·	·
3	·	·	1	·	·	·
4	1	·	·	·	·	·
5	·	·	·	1	·	·
6	·	·	·	·	·	1

$1 \leftrightarrow 11$
$2 \leftrightarrow 8$
$3 \leftrightarrow 9$
$4 \leftrightarrow 7$
$5 \leftrightarrow 10$
$6 \leftrightarrow 12$

Optimal Transport, the discrete case

	7	8	9	10	11	12	13	14	15	16
1	0.41	0.55	0.22	0.64	0.04	0.25	0.24	0.77	0.74	0.55
2	0.28	0.24	0.73	0.22	0.64	0.80	0.76	0.76	0.12	0.10
3	0.28	0.47	0.32	0.52	0.16	0.37	0.27	0.68	0.63	0.45
4	0.28	0.62	0.81	0.25	0.64	0.85	0.58	0.32	0.51	0.48
5	0.41	0.37	0.89	0.25	0.81	0.97	0.91	0.81	0.05	0.25
6	0.66	0.76	0.21	0.89	0.22	0.14	0.33	0.96	0.99	0.79

	7	8	9	10	11	12	13	14	15	16	
1	·	·	1/5	·	3/5	·	1/5	·	·	·	$1 \leftrightarrow \{9,11,13\}$
2	·	2/5	·	·	·	·	·	·	·	3/5	$2 \leftrightarrow \{8,16\}$
3	3/5	·	·	·	·	·	2/5	·	·	·	$3 \leftrightarrow \{7,13\}$
4	·	·	·	2/5	·	·	·	3/5	·	·	$4 \leftrightarrow \{10,14\}$
5	·	1/5	·	1/5	·	·	·	·	3/5	·	$5 \leftrightarrow \{8,10,15\}$
6	·	·	2/5	·	·	3/5	·	·	·	·	$6 \leftrightarrow \{9,12\}$

Optimal Transport, the discrete case

	7	8	9	10	11	12
1	0.41	0.55	0.22	0.64	0.04	0.25
2	0.28	0.24	0.73	0.22	0.64	0.80
3	0.28	0.47	0.32	0.52	0.16	0.37
4	0.28	0.62	0.81	0.25	0.64	0.85
5	0.41	0.37	0.89	0.25	0.81	0.97
6	0.66	0.76	0.21	0.89	0.22	0.14

$$\underset{\mathrm{P}\ \in\ U(\mathbf{1}_{n_0},\mathbf{1}_{n_1})}{\operatorname{argmin}} \langle \mathrm{P}, \mathrm{C} \rangle - \gamma \cdot \text{entropy}(\mathrm{P})$$

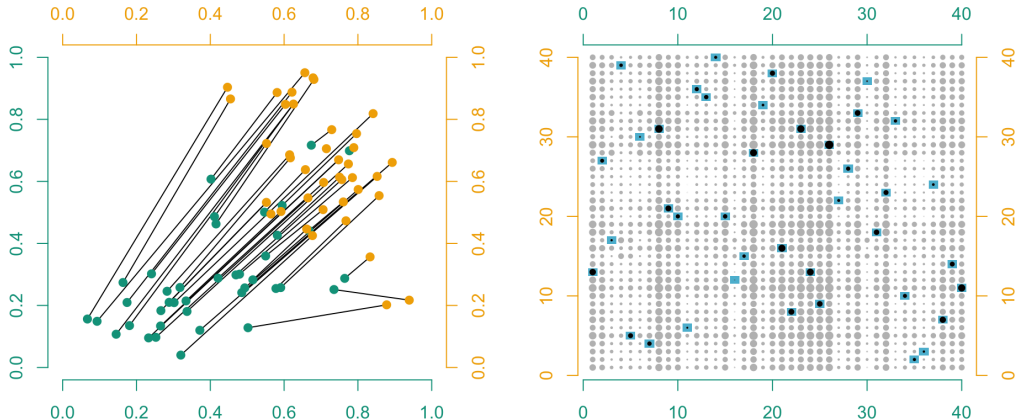
(**regularization** via Sinkhorn algorithm, associated to "Matrix Scaling Problem" in Sinkhorn (1962)) or

$$\underset{\mathrm{P}\ \in\ U(\mathbf{1}_{n_0},\mathbf{1}_{n_1})}{\operatorname{argmin}} \langle \mathrm{P}, \mathrm{C} \rangle + \gamma \cdot \mathrm{d}_{\mathrm{KL}}(\mathrm{P} \| \mathbf{1}_{n_0} \otimes \mathbf{1}_{n_1})$$

minimal cost 0.2189 (> 0.2116)

	7	8	9	10	11	12	
1	·	·	0.35	·	0.59	0.06	$1 \leftrightarrow 11 + \cdots$
2	0.12	0.82	·	0.07	·	·	$2 \leftrightarrow 8 + \cdots$
3	·	·	0.55	·	0.41	0.04	$3 \leftrightarrow 9 + \cdots$
4	0.85	·	·	0.14	·	·	$4 \leftrightarrow 7 + \cdots$
5	0.03	0.18	·	0.79	·	·	$5 \leftrightarrow 10 + \cdots$
6	·	·	0.10	·	·	0.90	$6 \leftrightarrow 12 + \cdots$

Optimal Transport, the discrete case



Discrete matching, $n_0 = n_1 = 40$ individuals in two groups, $\{\boldsymbol{z}_i, s_i = 0\}$ and $\{\boldsymbol{z}_i, s_i = 1\}$, with matrices C ($C_{i,j} = \bullet$) and P (if $P_{i,j} = 1$)

Optimal Transport, the univariate case

Wasserstein distance is defined as the optimal transportation cost, when c is the $_k$ distance, for $k \geq 1$, Wasserstein (1969),

$$W_k(\nu_0, \nu_1)^k := \inf_{T\ \nu_0=\nu_1} \int_{\mathrm{R}} \|z_0 - T(z_0)\|^k \nu_0(\mathrm{d}z_0) = \int_0^1 |Q_1(u) - Q_0(u)|^k \mathrm{d}u$$

i.e., the optimal mapping T (if $k > 1$) is non-decreasing, i.e.,

$$z_1 = T\ (z_0) = Q_1 \circ F_0(z_0) \text{ where } \begin{array}{l} F_j(z) = \mu_j(-\infty, z] \\ Q_j(u) = F_j^{-1}(u) = \inf\{t \quad \mathrm{R} : F_j(t) \geq u\} \end{array}$$

$T\ := Q_1 \circ F_0$ is a non-decreasing mapping $\mathrm{R} \to \mathrm{R}$.

In higher dimension, some multivarite extensions of quantiles can be introduced to keep this construction, see Hallin et al. (2021), Hallin and Konen (2024)

Optimal Transport, the univariate case

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 $\Phi : \mathrm{R} \times \mathrm{R} \to \mathrm{R}$, for every permutation σ of $\{1, 2, \cdots, n\}$,

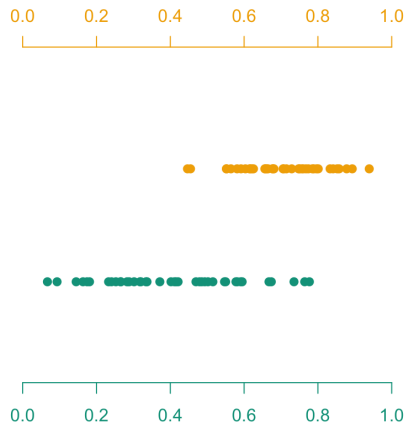
$$\sum_{i=1}^{n} \Phi(x_i, y_{n+1-i}) \leq \sum_{i=1}^{n} \Phi(x_i, y_{\sigma(i)}) \leq \sum_{i=1}^{n} \Phi(x_i, y_i),$$

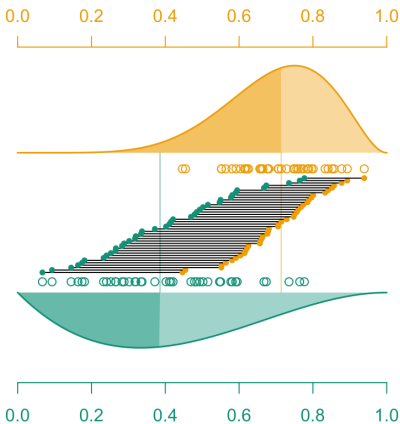
while if $\Phi : \mathrm{R} \times \mathrm{R} \to \mathrm{R}$ is submodular,

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

see Hardy et al. (1952).
See related work on risk measures, Denuit et al. (2006), Galichon (2016), Carlier et al. (2016).

Optimal Transport with Pictures, the univariate case





Univariate predictions for n individuals in two groups, $\{z_i, s_i = 0\}$ and $\{z_i, s_i = 1\}$, or measures ν_0 and ν_1. $T(\cdot) = Q_1 \circ F_0(\cdot)$.

Optimal Transport with Pictures, the univariate case

Univariate predictions for n individuals in two groups, $T^{?}(\cdot) = Q_1 \quad F_0(\cdot)$.

Optimal Transport with Pictures, the univariate case

Multivariate observations for n individuals in two groups, $_0$ and $_1$ (densities).

Optimal Transport with Pictures, the univariate case

Bivariate observations for n individuals in two groups $\{z_i; s_i = 0\}$ and $\{z_i; s_i = 1\}$.

Optimal Transport with Pictures, the univariate case

Discrete matching, n individuals in two groups $\{z_i; s_i = 0\}$ and $\{z_i; s_i = 1\}$.

Optimal Transport with Pictures, the univariate case

Optimal transport with Gaussian joint distributions, $\mathcal{N}(\mu_0, \sigma_0)$ and $\mathcal{N}(\mu_1, \sigma_1)$.

Machine learning, supervised model

Consider some data (Y, X), and a "model" for Y, and some loss ℓ,

$$m^\star := \underset{m \in \mathcal{M}}{\operatorname{argmin}} \; R(m) \;\; ; \;\; R(m) := \mathbb{E}\left[\ell(Y, m(X))\right] :$$

e.g. if $\ell(y, m) = (y - m)^2$ and $\mathcal{M}$ the set of all measurable functions,

$$:= \underset{m \in \mathcal{M}}{\operatorname{argmin}} \; R(m) \quad \text{is} \quad m(x) := \mathbb{E}[Y \mid X = x] ;$$

corresponding to the "regression function", or "Bayes regressor".

For a subclass of models $\mathcal{G}$, the excess of risk of class $\mathcal{G}$ is

$$E(\mathcal{G}) = \min_{m \in \mathcal{G}} R(m) - \min_{m \in \mathcal{M}} R(m) = \min_{m \in \mathcal{G}} R(m) - R(\quad)$$

Machine learning, supervised model

In fairness problems, we still have Y the quantity of interest, but among covariates,

$$\begin{cases} X = (X_1, \cdots, X_k) & \text{are legitimate autorized variables} \\ A = (A_1, \cdots, A_j) & \text{are proscribed protected attributes} \end{cases}$$

Let $m(x, a)$ denotes a model

- $\mathcal{L}_a(m)$ is the conditional measure of $m(X, A)$ given $A = a$
- $F_{m|a}(u) := \mathbb{P}(m(X, A) \leq u | A = a)$ and,
  with an abuse of notation $F_{m|a}(u) := \mathbb{P}(m(X, A) \leq u | A = a)$
  their cumulative distribution function (CDF);
- $Q_{m|a}(v) := \inf\{u \in \mathbb{R} : F_{m|a}(u) \geq v\}$ and,
  with an abuse of notation $Q_{m|a}(v) := \inf\{u \in \mathbb{R} : F_{m|a_i}(u) \geq v\}$
  their associated quantile functions.

Machine learning, supervised model

$F_{mja=0}(\,)$, $F_{mja=1}(\,)$ and $F_m(\,)$ on the left, $Q_{mja=0}(\,)$, $Q_{mja=1}(\,)$ and $Q_m(\,)$ on the right

Group Fairness

A model m satisfies the independence property if $m(X;A) \perp A$, with respect to the distribution P of the triplet $(X;A;Y)$

For binary classifiers, demographic parity w.r.t. A is defined as

$$\begin{cases} P(\hat{Y}=1 \mid A=a) = P(\hat{Y}=1) \;;\; \forall a \\ P(m(X;A)>t \mid A=a) = P(m(X;A)>t) \;;\; \forall a \end{cases}$$

and a measure of (weak) discrimination would be

$$\begin{cases} \max_{a \in A} P(\hat{Y}=1 \mid A=a) - P(\hat{Y}=1) \\ \max_{a \in A} P(m(X;A)>t \mid A=a) - P(m(X;A)>t) \end{cases}$$

for some threshold t.

Group Fairness & Wasserstein Distance

A model m satisfies the independence property if $m(X;A) \perp A$, with respect to the distribution P of the triplet $(X;A;Y)$

More generally, define a (strong) discrimination measure $U(m) := \max_{a \in A} d(\mu_m, \mu_{m|a})$, for some distance d, to quantify discrimination w.r.t. A.
E.g., total variation:

$$U_{TV|A}(m) := \max_{a \in A} \sup_{I \subset \mathbb{R}} |P(m(X;A) \in I | A = a) - P(m(X;A) \in I)|;$$

E.g., Wasserstein based measure:

$$U_{W_k|A}(m) := \max_{a \in A} W_k(\mu_m, \mu_{m|a})^k = \max_{a \in A} \int_{u \in [0;1]} |Q_m(u) - Q_{m|a}(u)|^k du.$$

Group Fairness & Wasserstein Distance

Possibly multi-attributes measure $U_{W_k A}(m) = U_{W_k A_1}(m) + \quad + U_{W_k A_j}(m)$

Let $\mathcal{M}$ denote the set of all measurable models.
Let G denote a set of models

$$G_{fair} := \{m \in G \text{ s.t. } m(X;A) \perp A\} = \{m \in G \text{ s.t. } U(m) = 0\}$$

price of fairness :

$$E(G_{fair}) := \min_{m \in G_{fair}} R(m) - \min_{m \in \mathcal{M}} R(m) \geq 0$$

We can also consider ε-fairness for class G

$$G_{\varepsilon\text{-fair}} := \{m \in G \text{ s.t. } U(m) \leq \varepsilon\} \text{ and } G_{fair} \subset G_{\varepsilon\text{-fair}} \subset G \subset \mathcal{M}$$

Barycenters

Recall that the barycenter of a set of points z_i in $\mathbb{R}^n$ with weights ω_i is

$$\sum_{i=1}^n \omega_i z_i = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left(\sum_{i=1}^n \omega_i \|z - z_i\|^2 \right)$$

that can be extended to any metric space (E, d), with n-barycenters,

$$\operatorname*{argmin}_{z \in E} \left(\sum_{i=1}^n \omega_i d(z, z_i)^k \right) \quad \text{or} \quad \operatorname*{argmin}_{z \in E} \int d(z, z)^k \, \mathrm{d}\mu(z)$$

for some probability measure μ on (E, d).

The existence is related to geodesics. A complete metric space (E, d) is said to be geodesic if $\forall x, y \in E$, $\exists z \in E$ such that

$$d(x, z) = d(y, z) = \frac{1}{2} d(x, y)$$

Barycenters

The p-barycenter of any probability measure on a locally compact geodesic space, with nite moments of orderp, exists.

Given a metric spaceE; d), consider the Wasserstein distance

$$W_k(\ _0;\ _1)^k = \inf_{2(\ _0;\ _1)} \int_E \int_E d(z_0; z_1)^k \ (dz_0; dz_1);$$

for two probability measures on(E; d). Let $P_k(E)$ denote the set of all measures (E; d) for which moments of orderk are nite.
The Wasserstein space$P_k(E); W_k)$ of E is a complete geodesic space if(d) is.

If (E; d) be a separable locally compact geodesic space. Then any set of probability measures$P_k(E); W_k)$ has a barycenter, whenp 1.

Barycenters

Barycenter of univariate distributions, $_0$ and $_1$ and weights $1=2$ $1=2$, and of two Gaussian distributions $N(\ _0;\ _0)$ and $N(\ _1;\ _1)$.

Barycenters

Price of projection and (almost) Wasserstein Barycenter

Write $_A(X) := E\ Y\ X; A$. From Pythagoras' theorem

$$E\ (Y \quad _A(X))^2 + E\ (\ _A(X) \quad m(X; A))^2 = E\ (Y \quad m(X; A))^2$$

$$E\ (\ _A(X) \quad m(X; A))^2 = E\ (Y \quad m(X; A))^2 \quad E\ (Y \quad _A(X))^2$$

therefore

$$\inf_{m2G} E\ \big[\ E\ (\ _A(X) \quad m(X; A))^2\ A\big] = \inf_{m2G} E\ (Y \quad m(X; A))^2 \quad E\ (Y \quad _A(X))^2$$

E(G)

Barycenters

Thus

$$\mathbb{E}\, \| \varphi_A(X) - m(X;A) \|^2 \, A = a \qquad W_2(\mu_a(\cdot), \nu_a(m))^2$$

from the de nition of Wasserstein-2 distance (optimal coupling)

$$W_2(\mu_0, \mu_1)^2 := \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int \|x - y\|^2 \, \pi(dx, dy) = \inf_{(Z_0, Z_1) \in \Pi(\mu_0, \mu_1)} \mathbb{E}\, \|Z_0 - Z_1\|^2$$

i.e.,

$$\mathbb{E}\, \mathbb{E}\, \| \varphi_A(X) - m(X;S) \|^2 \, A \,] \qquad \mathbb{E}\, W_2(\mu_A(\cdot), \nu_A(m))^2$$

and

$$\inf_{m \in G} \mathbb{E}\, (Y - m(X;A))^2 \qquad \mathbb{E}\, (Y - \varphi_A(X))^2 \qquad \inf_{m \in G} \mathbb{E}\, W_2(\mu_A(\cdot), \nu_A(m))^2$$

$$\mathbb{E}(G)$$

Barycenters

If $_A(\ )$ is absolutely continuous (w.r.t. the Lebesgue measure),

$$E(G) = \inf_{m2G} E\ W_2(\ _A(\ );\ _A(m))^2 = \inf_{m2G} \sum_a P[A = a]\ W_2(\ _a(\ );\ _a(m))^2$$

see Gouic et al. (2020). Heuristically, for any there exists $T_a : R \ ! \ R$ s.t.

$$W_2(\ _a(\ );\ _a(m))^2 = E\ jT_a(\ _a(X)) \quad _a(X)j^2$$

Set $h : (x; a) 7! T_a(\ _a(x))$. If $h 2 G$

$$E\ W_2(\ _A(\ );\ _A(m))^2 = E\ jh(X; A) \quad _A(X)j^2 \quad \inf_{g2G} E\ jg(X; A) \quad _A(X)j^2$$

(satis ed if G is \pro le complete", Gouic et al. (2020)), i.e., the inequality is an equality.

Barycenters

Price of fairness and Wasserstein Barycenter

$$E(G_{\text{fair}}) = \inf_{m \in G_{\text{fair}}} \sum_a P[A = a] \; W_2(\nu_{a(\cdot)}, \nu_a(m))^2 = \inf_{\nu \in P} \sum_a P[A = a] \; W_2(\nu_{a(\cdot)}, \nu)^2$$

We recognize on the right the barycenter, with weights $P[A = a]$ and Wasserstein-2 distance, see Chzhen et al. (2020) and Gouic et al. (2020).

Hu et al. (2023b) considered constraints on the set of distributions, i.e.,
$G_{\text{fair}}^P = \{m \in G_{\text{fair}} \text{ s.t. } m \in P\}$

on-going work on constraints on the set of models when G is not "pro le complete"

Mitigating discrimination

See Supreme Court Justice Harry Blackmun stated, in 1978,
"In order to get beyond racism, we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently ," Knowlton (1978), cited in Lippert-Rasmussen (2020)

In the case of a binary sensitive attribute A,

$$\begin{cases} m^?(x; a = 0) = \mathbb{P}[A = 0] \quad m(x; a = 0) + \mathbb{P}[A = 1] \quad Q_{mj1} \quad F_{mj0} \; m(x; a = 0) \\ m^?(x; a = 1) = \mathbb{P}[A = 0] \quad Q_{mj0} \quad F_{mj1} \; m(x; a = 1) \quad + \mathbb{P}[A = 1] \quad m(x; a = 1) \end{cases}$$

or more generally $m^?(x; a) = \sum_{2A} \mathbb{P}[A = \quad] \quad Q_{mj} \quad F_{mja} \; m(x; a)$

See Hu et al. (2023a) and Charpentier et al. (2023b).

Mitigating discrimination

Fair score, obtained as barycenter of two univariate distributions of score m_{j0} and m_{j1} and weights $\frac{1}{2}\{1=2$ (similar to $n_0 = n_1$).

Mitigating discrimination

"God said: there will be white men, there will be black men, there will be tall men, there will be short men, there will be handsome men and there will be ugly men, and all will be equal; but it won't be easy... And then he added: there will even be some who will be black, short and ugly, and for them, it will be very hard! ," Coluche (1975)

The use of barycenters is interesting when we consider multiple sensitive attributes $A = (A_1; \quad ; A_j)$

Barycenters are associative.

See Hu et al. (2024)
and Fernandes Machado et al. (2025c).

Individual Fairness and Potential Outomes

We have counterfactual fairness if "had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same," Kusner et al. (2017)
If the protected variable is considered as the treatment, individual fairness is close a measuring a treatment effect.

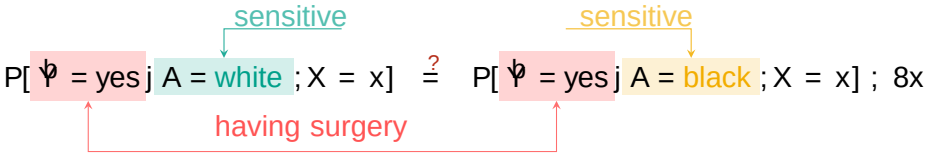
What does "other things being equal" really mean ?

It is possible to suppose that the protected attribute A could affect some explanatory variables X in a non-discriminatory way, Kilbertus et al. (2017) (concept of "resolving variable").

- two groups, distribution of height (cm) with women and men (in the U.S.)
- two groups, distribution of baby weight (g) with non-black and black mothers, delivering babies (in the U.S.)

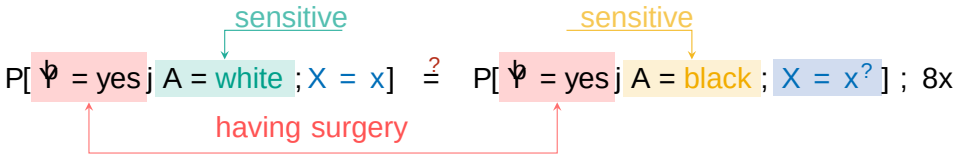
Individual Fairness and Potential Outomes

Individual Fairness and Potential Outomes



e.g., x is the \weight of the baby" (at birth), in g.

Discrimination, with di erent perspectives



e.g., x is the \weight of the baby" (at birth), in g.

Discrimination, with di erent perspectives

See Charpentier et al. (2023a)

(
\ceteris paribus di erence" : $m(A = 1; X = x) \quad m(A = 0; X = x)$
\mutatis mutandis di erence" : $m(A = 1; X = T_{0!1}^{?}(x)) \quad m(A = 0; X = x)$

suggested also in Plo and Meinshausen (2020), Plo et al. (2021) and De Lara et al. (2024). We need to transport $X|A = 0$ to $X|A = 1$ (multivariate transport).

- ok if we assume that $X|A \sim N(\mu_a, \Sigma_a)$: $x_{0!1}^{?}(x) = T_{0!1}^{?}(x) = \mu_1 + M(x - \mu_0)$; where M is a symmetric positive matrix that satis es $M \Sigma_0 M = \Sigma_1$
- ok if we consider individual empirical coupling $x_j^{?} = \hat{T}_{0!1}^{?}(x_i)$ but we cannot get $\hat{T}_{0!1}^{?}(x)$ if x is not an observed individual (in group 0).
- we can consider sequential conditional univariate transport (even if not optimal...)

Transport on a Probabilistic Graphical Model

As explained in Villani (2003); Carlier et al. (2010); Bonnotte (2013), the Knothe-Rosenblatt rearrangement is directly inspired by the Rosenblatt chain rule, from Rosenblatt (1952), and some extensions obtained on general measures by Knothe (1957). The Knothe-Rosenblatt rearrangement is

$$T_{\overline{kr}}(x_1, \ldots, x_d) = \begin{pmatrix} T_1^?(x_1 | x_2, \ldots, x_d) \\ T_2^?(x_2 | x_3, \ldots, x_d) \\ \vdots \\ T_{d-1}^?(x_{d-1} | x_d) \\ T_d^?(x_d) \end{pmatrix} \text{ or } T_{\underline{kr}}(x_1, \ldots, x_d) = \begin{pmatrix} T_1^?(x_1) \\ T_2^?(x_2 | x_1) \\ \vdots \\ T_{d-1}^?(x_{d-1} | x_1, \ldots, x_{d-2}) \\ T_d^?(x_d | x_1, \ldots, x_{d-1}) \end{pmatrix},$$

the "monotone lower triangular map," defined in Bogachev et al. (2005).

Transport on a Probabilistic Graphical Model

Transport on a Probabilistic Graphical Model

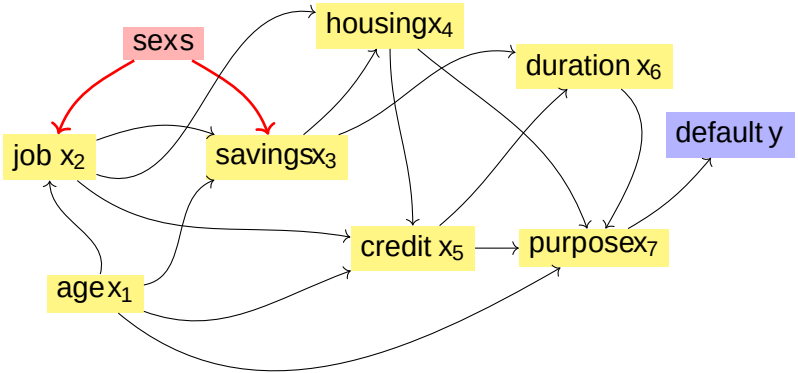
Transport on a Probabilistic Graphical Model

Multivariate observations for individuals in two groups, $_0$ and $_1$ (densities).

Transport on a Probabilistic Graphical Model

Sequential transport.

Transport on a Probabilistic Graphical Model



Causal graph in the German Credit dataset from Watson et al. (2021).

Transport on a Probabilistic Graphical Model

The joint distribution of X was written using the chain rule, Rosenblatt (1952)

$$P[x_1, \cdots, x_d] = P[x_1] \prod_{j=2}^{d} P[x_j | x_1, \cdots, x_{j-1}] \text{ or } P[x_d] \prod_{j=1}^{d-1} P[x_j | x_{j+1}, \cdots, x_d]$$

The joint distribution of X satisfies the Markov property

$$P[x_1, \cdots, x_d] = P[x_1] \prod_{j=2}^{d} P[x_j | x_{j-1}]$$

The joint distribution of X satisfies the Markov property w.r.t. PGM:

$$P[x_1, \cdots, x_d] = \prod_{j=1}^{d} P[x_j | \text{parents}(x_j)],$$

where $\text{parents}(x_j)$ are nodes with edges directed towards x_j in PGM.

Transport on a Probabilistic Graphical Model

A classical algorithm for topological sorting is Kahn (1962)'s \Depth First Search" (DFS), and other algorithms are discussed in Section 20.4 in Cormen et al. (2022). For the causal graph on the German Credit dataset, variables are sorted, and

$$T^{?}_{\text{PGM}}(x_1; \quad ; x_7) = \begin{matrix} 0 & & 1 \\ & T^{?}_1(x_1) & \\ \Big\{ & T^{?}_2(x_2 | x_1) & \Big\} \\ & T^{?}_3(x_3 | x_1; x_2) & \\ & T^{?}_4(x_4 | x_2; x_3) & \\ & T^{?}_5(x_5 | x_1; x_2; x_4) & \\ @ & T^{?}_6(x_6 | x_3; x_5) & A \\ & T^{?}_7(x_7 | x_1; x_4; x_5; x_6) & \end{matrix} :$$

See Fernandes Machado et al. (2025b), with an applications on two real datasets.

Optimal Transport on the Simplex (for Categorical Variables)

How to transport X_j if X_j is categorical (e.g. $X_j \in \{A, B, C\}$) ? See Fernandes Machado et al. (2025a)

- use a classifier to get probabilities associated to $\{A, B, C\} \mapsto S_d$,

A	0.69	0.30	0.01
C	0.04	0.31	0.65
A	0.49	0.19	0.32
B	0.01	0.63	0.36
⋮	⋮	⋮	⋮

- transport probabilities (taking into account the geometry of the simplex)

$$S_d = \left\{ x \in (0,1)^d \;\middle|\; x^\top 1 = 1 \right\} \subset \mathbb{R}^d$$

Optimal Transport on the Simplex (for Categorical Variables)

	A	B	C
(0)	43.713%	18.572%	37.715%
(1)	29.831%	48.605%	21.564%
T ( )	29.553%	48.517%	21.930%

	A	B	C
(0)	43.713%	18.572%	37.715%
(1)	29.831%	48.605%	21.564%
T ( )	43.728%	18.553%	37.719%

Optimal Transport on the Simplex (for Categorical Variables)

First approach, "Dirichlet transport " from Baxendale and Wong (2022).

Let $C : \mathbb{R}_+^d \to S_d$ denote the "closure operator",

$$C[x_1, x_2, \cdots, x_d] = \left(\frac{x_1}{\sum_{i=1}^d x_i}, \frac{x_2}{\sum_{i=1}^d x_i}, \cdots, \frac{x_d}{\sum_{i=1}^d x_i} \right),$$

If we define the binary operator $\oplus$ on S_d,

$$x \oplus y = \left(\frac{x_1 y_1}{\sum_{i=1}^d x_i y_i}, \cdots, \frac{x_d y_d}{\sum_{i=1}^d x_i y_i} \right),$$

then $(S_d, \oplus)$ is a commutative group, with identity $e = C(1, 1, \cdots, 1)$, and the inverse of x is

$$x^{-1} = \left(\frac{1/x_1}{\sum_{i=1}^d 1/x_i}, \cdots, \frac{1/x_d}{\sum_{i=1}^d 1/x_i} \right) = C(1/x).$$

Optimal Transport on the Simplex (for Categorical Variables)

Following Pal and Wong (2016, 2018, 2020), consider the optimal transport problem with the following cost function, on $S_d \times S_d$,

$$c(x, y) = \log\left(\frac{1}{d}\sum_{i=1}^{d}\frac{y_i}{x_i}\right) - \frac{1}{d}\sum_{i=1}^{d}\log\frac{y_i}{x_i}, \tag{4}$$

Pal and Wong (2020) proved that there exists an exponentially concave function $\varphi : S_d \to \mathbb{R}$ such that

$$T_\varphi(x) = x \cdot \nabla\varphi(x)^{-1}$$

where $\nabla\varphi : S_d \to S_d$ satisfies

$$\nabla\varphi(x) = \frac{x_1}{1 + \langle e_1 - e_d, x\rangle}, \ldots, \frac{x_d}{1 + \langle e_d - e_d, x\rangle},$$

where $\{e_1, \ldots, e_d\}$ is the standard orthonormal basis of $\mathbb{R}^d$.

Optimal Transport on the Simplex (for Categorical Variables)

Second approach, "transport on transformed space"

Define a transport mapping based on some isomorphism $m : S_d \to E$ and then define the inverse mapping $m^{-1} : E \to S_d$, where E is some Euclidean space, classically $\mathbb{R}^{d-1}$. This is the dual transport problem in Pal and Wong (2018). E.g.,

$$\text{alr}(x) = \left[\log \frac{x_1}{x_d}; \quad ; \log \frac{x_{d-1}}{x_d} \right]$$

with inverse, for any $z \in \mathbb{R}^{d-1}$,

$$\text{alr}^{-1}(z) = C(\exp(z_1); \quad ; \exp(z_{d-1}); 1) = C \exp([z; 0])$$

Optimal Transport on the Simplex (for Categorical Variables)

A more popular one is

$$\text{clr}(x) = \left[\log \frac{x_1}{\bar{x}_g}; \quad ; \log \frac{x_D}{\bar{x}_g} \right];$$

where $\bar{x}_g$ denotes the geometric mean of x. Its inverse is

$$\text{clr}^{-1}(z) = C(\exp(z_1); \quad ; \exp(z_d)) = C \exp(z) \; ; \; z \in \mathbb{R}^d:$$

that is the softmax function

Optimal Transport on the Simplex (for Categorical Variables)

with two models, random forest (left) and gradient boosting (right)

Optimal Transport on the Simplex (for Categorical Variables)

	random forest			boosting		
	cars	equipmt.	other	cars	equipmt.	other
categorical (M)	30.323%	53.226%	16.452%	30.323%	53.226%	16.452%
categorical (F)	35.217%	44.638%	20.145%	35.217%	44.638%	20.145%
composition (M)	32.055%	49.882%	18.063%	31.843%	50.712%	17.445%
composition (F)	34.977%	45.409%	19.614%	34.714%	45.341%	19.945%
T ( )	32.046%	49.883%	18.070%	31.871%	50.675%	17.455%

See Fernandes Machado et al. (2025a) for more details.

References

Baxendale, P. and Wong, T.-K. L. (2022). Random concave functions. The Annals of Applied Probability, 32(2):812–852.

Bogachev, V. I., Kolesnikov, A. V., and Medvedev, K. V. (2005). Triangular transformations of measures. Sbornik: Mathematics, 196(3):309.

Bonnotte, N. (2013). From Knothe's rearrangement to Brenier's optimal transport map. SIAM Journal on Mathematical Analysis, 45(1):64–87.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. Communications on pure and applied mathematics, 44(4):375–417.

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

Carlier, G., Chernozhukov, V., and Galichon, A. (2016). Vector quantile regression: an optimal transport approach. The Annals of Statistics, 44.

Carlier, G., Galichon, A., and Santambrogio, F. (2010). From Knothe's transport to Brenier's map and a continuation method for optimal transport. SIAM Journal on Mathematical Analysis, 41(6):2554–2576.

Charpentier, A. (2024). Insurance: biases, discrimination and fairness. Springer Verlag.

References

Charpentier, A., Flachaire, E., and Gallic, E. (2023a). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, Optimal Transport Statistics for Economics and Related Topics. Springer Verlag.

Charpentier, A., Hu, F., and Ratz, P. (2023b). Mitigating discrimination in insurance with wasserstein barycenters. bias. In 3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020). Fair regression with wasserstein barycenters. In Advances in Neural Information Processing Systems

Coluche (1975). Le blouson noir. https://www.youtube.com/watch?v=07xkVU_dztc.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2022). Introduction to algorithms. MIT press.

De Lara, L., González-Sanz, A., Asher, N., Risser, L., and Loubes, J.-M. (2024). Transport-based counterfactual models. Journal of Machine Learning Research, 25(136):1–59.

Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. (2006). Actuarial theory for dependent risks: measures, orders and models. John Wiley & Sons.

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2025a). Optimal transport on categorical data for counterfactuals using compositional data and dirichlet transport. arXiv, 2501.15549.

References

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2025b). Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. 39th Annual AAAI Conference on Artificial Intelligence.

Fernandes Machado, A., Grondin, S., Ratz, P., Charpentier, A., and Hu, F. (2025c). Equipy: Sequential fairness using optimal transport in Python. soon on ArXiv.

Galichon, A. (2016). Optimal transport methods in economics. Princeton University Press.

Gouic, T. L., Loubes, J.-M., and Rigollet, P. (2020). Projection to fairness in statistical learning. arXiv, 2005.11720.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. The Annals of Statistics, 49(2):1139{1165.

Hallin, M. and Konen, D. (2024). Multivariate quantiles: Geometric and measure-transportation-based contours.

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

Harris, T. and Ross, F. (1955). Fundamentals of a method for evaluating rail net capacities. Technical report.

Higham, N. J. (2008). Functions of matrices: theory and computation. SIAM.

References

Hu, F., Ratz, P., and Charpentier, A. (2023a). Fairness in multi-task learning via wasserstein barycenters. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases – ECML PKDD*.

Hu, F., Ratz, P., and Charpentier, A. (2023b). Parametric fairness with statistical guarantees. *arXiv*, 2310.20508.

Hu, F., Ratz, P., and Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. *Annual AAAI Conference on Artificial Intelligence*.

Kahn, A. B. (1962). Topological sorting of large networks. *Communications of the ACM*, 5(11):558–562.

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.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

References

Knothe, H. (1957). Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52.

Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

Lippert-Rasmussen, K. (2020). *Making sense of a rmative action*. Oxford University Press.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Pal, S. and Wong, T.-K. L. (2016). The geometry of relative arbitrage. *Mathematics and Financial Economics*, 10:263–293.

Pal, S. and Wong, T.-K. L. (2018). Exponentially concave functions and a new information geometry. *The Annals of probability*, 46(2):1070–1113.

Pal, S. and Wong, T.-K. L. (2020). Multiplicative schrödinger problem and the dirichlet transport. *Probability Theory and Related Fields*, 178(1):613–654.

Plečko, D., Bennett, N., and Meinshausen, N. (2021). fairadapt: Causal reasoning for fair data pre-processing.

References

Plečko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

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

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Society.

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

Watson, D. S., Gultchin, L., Taly, A., and Floridi, L. (2021). Local explanations via necessity and su□ciency: Unifying theory and practice. *Uncertainty in Artificial Intelligence*, pages 1382–1392.