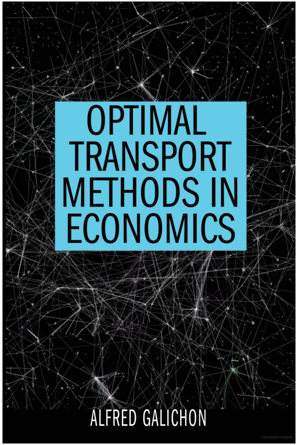
Optimal Transport in Data Science: Theory and Applications

Nian Si

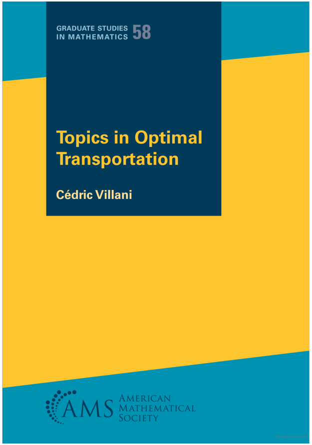
Seminar on Data Science and Applied Mathematics

April 17, 2023

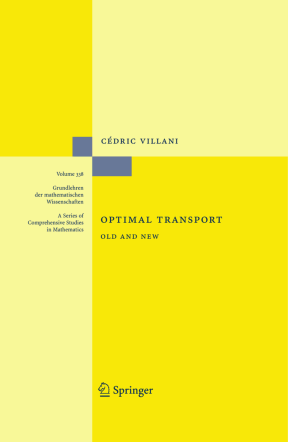
Books


OPTIMAL TRANSPORT METHODS IN ECONOMICS
ALFRED GALICHON

(a)


GRADUATE STUDIES IN MATHEMATICS 58
Topics in Optimal Transportation
Cédric Villani
AMS AMERICAN MATHEMATICAL SOCIETY

(b)


CÉDRIC VILLANI
Volume 338
Grundlehren der mathematischen Wissenschaften
A Series of Comprehensive Studies in Mathematics
OPTIMAL TRANSPORT
OLD AND NEW
Springer

(c)

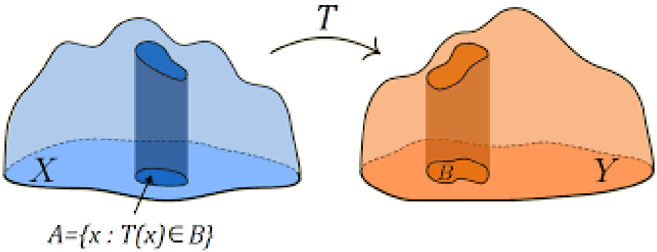
Road Map

1. Optimal Transport Preliminaries
   - Definition
   - Discrete Distributions
   - Continuous Distributions
2. Optimization of Optimal Transport
3. Statistics of Optimal Transport
   - Curse of Dimensionality
   - Projection
   - Smoothness
4. Applications
   - Wasserestein GANs
   - Distributionally Robust Optimization
5. Some New Advances and Open Problems

Monge Map

- Let $P \in \mathcal{P}(S)$ and $Q \in \mathcal{P}(S)$ be two probability distributions defined on a space S; $c : S \times S \to [0, \infty]$ is a cost function.
- Monge problem:

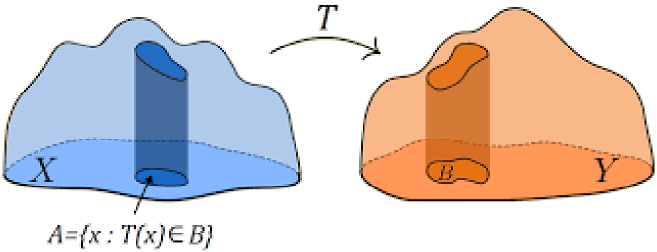
$$\inf_{T(\cdot)} \mathbb{E}_P[c(X, T(X)) | T_\sharp P = Q].$$



Monge Map

- Let $P \in \mathcal{P}(S)$ and $Q \in \mathcal{P}(S)$ be two probability distributions defined on a space S; $c : S \times S \to [0, \infty]$ is a cost function.
- Monge problem:

$$\inf_{T(\cdot)} \mathbb{E}_P[c(X, T(X)) | T_\sharp P = Q].$$



✗ May not always exist: if P supports on one point and Q supports on two points:

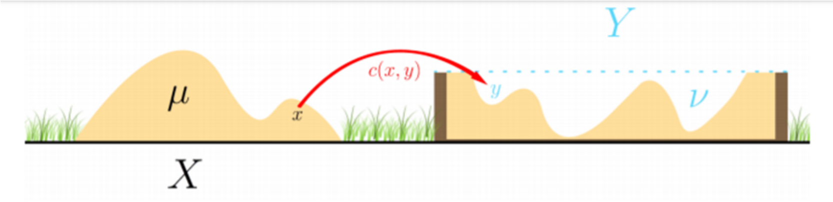
$$P(x_0) = 1 \text{ and } Q(y_0) = Q(y_1) = 1/2.$$

Definition

Definition (Optimal Transport Cost)

Let $P \in \mathcal{P}(S)$ and $Q \in \mathcal{P}(S)$ be two probability distributions defined on a space S; $c : S \times S \to [0, \infty]$ is a cost function. Then, the optimal transport cost is defined as

$$D_c(P, Q) := \inf_{\pi} \Big\{ \mathbb{E}_\pi[c(U, V)] \mid \pi \in \mathcal{P}(S \times S), \\ \pi\,(A \times S) = P(A), \pi\,(S \times B) = Q(B) \text{ for every subsets } A, B \text{ of } S \Big\}$$



Wasserstein Distance, Earth Moving Distance

Let $S = \mathbb{R}^d$ and $c(x, y) = d(x, y)^\rho$ for some metric function $d(\cdot, \cdot)$ and $\rho \geq 1$,

$$W_\rho(P, Q) = D_c(P, Q)^{1/\rho}$$

is a metric on the probability space. We call it the type-ρ Wasserstein distance. In particular, if $\rho = 1$, it is also called the earth moving distance.

Discrete Distributions: Duality

- Let P support on $\{x_1, x_2, \ldots, x_N\}$ and Q support on $\{y_1, y_2, \ldots, y_m\}$. Let $P = \{p_x\}$ and $Q = \{q_y\}$ with $\sum_{x=1}^{N} p_x = \sum_{y=1}^{M} q_y = 1, p_x \geq 0, q_y \geq 0$.
- Optimal transport cost is the optimal value of the linear programming:

$$D_c(P, Q) = \min_{\pi xy \geq 0} \sum_{x=1, y=1}^{N,M} c(x, y)\pi_{xy} \tag{1}$$

$$s.t. \sum_{y=1}^{N} \pi_{xy} = p_x \text{ and } \sum_{x=1}^{N} \pi_{xy} = q_y. \tag{2}$$

Discrete Distributions: Duality

- Let P support on $\{x_1, x_2, \ldots, x_N\}$ and Q support on $\{y_1, y_2, \ldots, y_m\}$. Let $P = \{p_x\}$ and $Q = \{q_y\}$ with $\sum_{x=1}^{N} p_x = \sum_{y=1}^{M} q_y = 1, p_x \geq 0, q_y \geq 0$.
- Optimal transport cost is the optimal value of the linear programming:

$$D_c(P, Q) = \min_{\pi xy \geq 0} \sum_{x=1,y=1}^{N,M} c(x, y)\pi_{xy} \tag{1}$$

$$s.t. \sum_{y=1}^{N} \pi_{xy} = p_x \text{ and } \sum_{x=1}^{N} \pi_{xy} = q_y. \tag{2}$$

- Duality:

$$D_c(P, Q) = \max_{u,v} \sum_{x=1}^{N} p_x u_x + \sum_{y=1}^{M} q_y v_y \tag{3}$$

$$s.t. u_x + v_y \leq c(x, y) \tag{4}$$

Discrete Distributions: Optimal Solutions

- Primal and dual:

$$\min_{\pi_{xy}\geq 0} \sum_{x=1,y=1}^{N,M} c(x,y)\pi_{xy} \text{ s.t. } \sum_{y=1}^{N} \pi_{xy} = p_x \text{ and } \sum_{x=1}^{N} \pi_{xy} = q_y. \tag{P}$$

$$\max_{u,v} \sum_{x=1}^{N} p_x u_x + \sum_{y=1}^{M} q_y v_y \text{ s.t. } u_x + v_y \leq c(x,y) \tag{D}$$

- The optimal solution satisfies

$$\pi^*_{xy} > 0 \Rightarrow u^*_x + v^*_y = c(x,y)$$
$$u^*_x = \min_{y\in\{1,2,\dots,M\}} c(x,y) - v^*_y \text{ and } v^*_y = \min_{x\in\{1,2,\dots,N\}} c(x,y) - u^*_x.$$

An Economic Interpretation

- Primal and dual:

$$\min_{\pi_{xy}\geq 0} \sum_{x=1,y=1}^{N,M} c(x,y)\pi_{xy} \text{ s.t. } \sum_{y=1}^{N} \pi_{xy} = p_x \text{ and } \sum_{x=1}^{N} \pi_{xy} = q_y. \tag{P}$$

$$\max_{u,v} \sum_{x=1}^{N} p_x u_x + \sum_{y=1}^{M} q_y v_y \text{ s.t. } u_x + v_y \leq c(x,y) \tag{D}$$

- Transfer coal from mines in $\{x_1, x_2, \ldots, x_N\}$ to factories in $\{y_1, y_2, \ldots, y_m\}$:
  - Transportation cost is $c(x,y)$;
  - u_x, v_y are shadow prices: u_x is the price of loading one ton of coal at place x; and v_y is the price of unloading it at destination y.

Pure Assignments and Monge Map

- Consider $N = M$ and $p_x = q_y = 1/N$.
- Then, the optimal solution is a permutation σ: an invertible map from $\{1, 2, \dots, N\}$ onto iteself.

$$\pi^*_{xy} = \frac{1}{N}\mathbb{I}\{y = \sigma(x)\}.$$

- The optimal transport problem is equivalent to the Monge problem: $T(X) = \sigma(X)$.

Continuous Distributions: Duality

- Recall

$$D_c(P,Q) = \inf_{\pi} \Big\{ \mathbb{E}_\pi[c(U,V)] \mid \pi \in \mathcal{P}(S \times S),$$
$$\pi(A \times S) = P(A), \pi(S \times B) = Q(B) \text{ for every subsets } A, B \text{ of } S \Big\}$$

- Duality:

$$D_c(P,Q) = \sup_{\varphi,\psi} \int \varphi \mathrm{d}P + \int \psi \mathrm{d}Q$$
$$s.t. \varphi(x) + \psi(y) \leq c(x,y).$$

- Proof is based on Sion's minimax theorem and compactification.

2-Wasserstein Distance Between Gaussian Distributions

- Cost function $c(x, y) = \|x - y\|_2^2$, $P = \mathcal{N}(\mu_1, \Sigma_1)$ and $Q = \mathcal{N}(\mu_2, \Sigma_2)$.
- Then, the 2-Wasserstein distance between P and Q is

$$W_2^2(P, Q) = \|\mu_1 - \mu_2\|_2^2 + \mathrm{tr}(\Sigma_1) + \mathrm{tr}(\Sigma_2) - 2\mathrm{tr}\left[(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right].$$

- The transportation plan is

$$x \to \mu_2 + A(x - \mu_1),$$

where $A = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}$.

1-Wasserstein Distance: Duality

- $c(x, y) = d(x, y)$;
- Duality:

$$W_1(P, Q) = D_c(P, Q) = \sup_{\varphi} \int \varphi \mathrm{d}P - \int \varphi \mathrm{d}Q$$
$$s.t. \varphi(x) \text{ is 1-Lipschitz with respect to } d(\cdot, \cdot)$$

Total Variation Distance

- Total variation distance is a special case of the Wasserstein distance with $c(x, y) = \mathbb{I}(x \neq y)$;
- $D_c(P, Q) = TV(P, Q)$.

One-Dimensional Case

- $d = 1$, we have

$$W_\rho(P, Q) = \left(\int_0^1 |F_P^{-1}(s) - F_Q^{-1}(s)|^\rho \mathrm{d}s \right)^{1/\rho},$$

  where F_P and F_Q are CDFs of measures P and Q.
- if $\rho = 1$ and $d = 1$, we have

$$W_1(P, Q) = \int_{\mathbb{R}} |F_P(s) - F_Q(s)| \mathrm{d}s,$$

Road Map

Optimization of Optimal Transport: Discrete Case

- Discrete case, linear programming: for simplicity, assume $N = M$

$$\min_{\pi_{xy}\geq 0} \sum_{x=1,y=1}^{N} c(x,y)\pi_{xy} \text{ s.t. } \sum_{y=1}^{N} \pi_{xy} = p_x \text{ and } \sum_{x=1}^{N} \pi_{xy} = q_y. \tag{P}$$

- Linear programming time complexity $O(N^{3.5}\log(1/\epsilon))$.
- Sinkhorn method [Cuturi, 2013] with time complexity $\tilde{O}(N^2/\epsilon^2)$ [Dvurechensky et al., 2018].

Sinkhorn Method

- We optimize the following program:

$$\min_{\pi_{xy}\geq 0} \sum_{x=1,y=1}^{N} c(x,y)\pi_{xy} + \frac{1}{\lambda}\sum_{i,j=1}^{N} \pi_{ij}\log(\pi_{ij}) \text{ s.t. } \sum_{y=1}^{N}\pi_{xy} = p_x \text{ and } \sum_{x=1}^{N}\pi_{xy} = q_y.$$

- The solution admits the form:

$$\pi_{ij}^{\lambda} = u_i \exp(-\lambda c(i,j)) v_j.$$

- By Sinkhorn and Knopp's algorithm [Sinkhorn and Knopp, 1967], we can iteratively update u and v to arrive

$$\pi^{\lambda}\mathbf{1} = P, (\pi^{\lambda})^{\top}\mathbf{1} = Q.$$

.

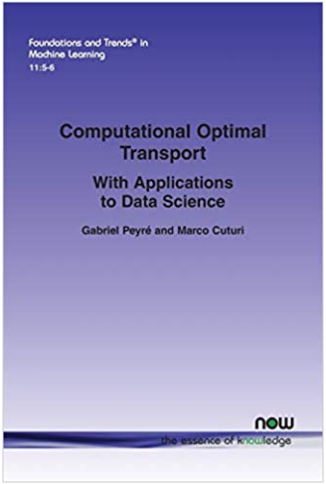
Optimization of Optimal Transport: Semi-Discrete Case

Theorem (Taşkesen et al. [2022])

Computing $W_\rho(P, Q)$ is #P-hard even if $P \sim U[0, 1]^d$ and Q is a two-point distribution.

- The complexity class $\#$P is the set of the counting problems associated with the decision problems in the set NP.
- Consequently, a $\#$P problem is at least as hard as its NP counterpart.

Computational Optimal Transport



Road Map

Curse of Dimensionality

- Let P^* be a measure on $\mathbb{R}^d$ and let P_n be the associated empirical measure, i.e., for i.i.d. sample $X_1, X_2, \ldots, X_n$,

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

- Consistency: $W_\rho(P_n, P^*) \to 0$.
- Curse of Dimensionality: $\mathbb{E}[W_\rho(P_n, P^*)] = O(n^{-1/d})$ [Fournier and Guillin, 2015].
- If P^* supports on an $m-$dimensional manifold of $\mathbb{R}^d$, we have $\mathbb{E}[W_\rho(P_n, P^*)] = O(n^{-1/m})$ [Weed and Bach, 2019].

Curse of Dimensionality

- Let P^* be a measure on $\mathbb{R}^d$ and let P_n be the associated empirical measure, i.e., for i.i.d. sample $X_1, X_2, \ldots, X_n$,

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

- Consistency: $W_\rho(P_n, P^*) \to 0$.
- Curse of Dimensionality: $\mathbb{E}[W_\rho(P_n, P^*)] = O(n^{-1/d})$ [Fournier and Guillin, 2015].
- If P^* supports on an $m-$dimensional manifold of $\mathbb{R}^d$, we have $\mathbb{E}[W_\rho(P_n, P^*)] = O(n^{-1/m})$ [Weed and Bach, 2019].
- CLT [Del Barrio and Loubes, 2019]:

$$\sqrt{n}(W_2^2(P_n, P^*) - \mathbb{E}[W_2^2(P_n, P^*)]) \Rightarrow \mathcal{N}(0, \sigma^2).$$

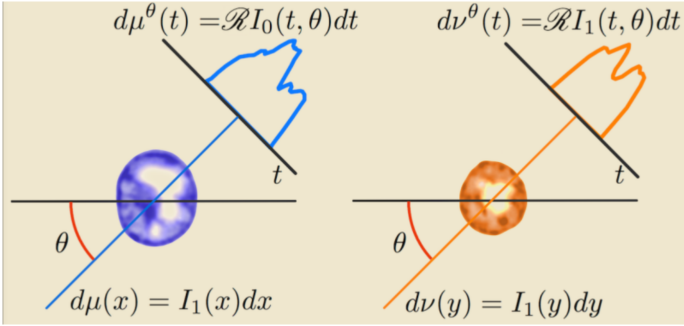
Beating Curse of Dimensionality: Projection

- Sliced Wasserestein distance [Bonneel et al., 2015, Kolouri et al., 2016]:

$$SW_\rho^\rho(P, Q) = \int_{\mathbb{S}^{d-1}} W_\rho^\rho(\theta_\sharp P, \theta_\sharp Q) \mathrm{d}\theta,$$

  where $\theta_\sharp P$ is the push-forward measure:

$$\theta_\sharp P(A) = P(\{x : \theta^\top x \in A\}), \text{ for any Borel set } A \in \mathbb{R}.$$



Beating Curse of Dimensionality: Projection

- Sliced Wasserestein distance [Bonneel et al., 2015, Kolouri et al., 2016]:

$$SW_\rho^\rho(P, Q) = \int_{\mathbb{S}^{d-1}} W_\rho^\rho(\theta_\sharp P, \theta_\sharp Q)\mathrm{d}\theta,$$

  where $\theta_\sharp P$ is the push-forward measure:

$$\theta_\sharp P(A) = P(\{x : \theta^\top x \in A\}), \text{ for any Borel set } A \in \mathbb{R}.$$

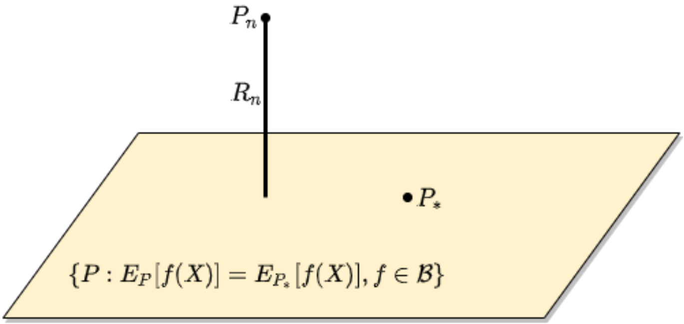
- Convergence rate: $SW_\rho(P_n, P^*) = O(n^{-1/2})$ [Nadjahi et al., 2019].
- Another variance: max-sliced Wasserstein distance [Deshpande et al., 2019]

$$MSW_\rho^\rho(P, Q) = \max_{\theta \in \mathbb{S}^{d-1}} W_\rho^\rho(\theta_\sharp P, \theta_\sharp Q),$$

Beating Curse of Dimensionality: Subspace Projection

- Robust Wasserstein profile function [Si et al., 2020]: consider a function class $\mathcal{B}$

$$R_n(P_*, P_n) := \inf_P \{ D_c(P, P_n) : \mathbb{E}_P[f(X)] = \mathbb{E}_{P_*}[f(X)] \text{ for all } f \in \mathcal{B} \}.$$



Beating Curse of Dimensionality: Subspace Projection

- Robust Wasserstein profile function [Si et al., 2020]: consider a function class $\mathcal{B}$

$$R_n(P_*, P_n) := \inf_P \{D_c\,(P, P_n) : \mathbb{E}_P\,[f\,(X)] = \mathbb{E}_{P_*}\,[f\,(X)] \text{ for all } f \in \mathcal{B}\}.$$

- Duality:

$$R_n(P_*, P_n) = \sup_{f \in \mathcal{LB}} \{\mathbb{E}_{P_*}\,[f\,(X)] - \mathbb{E}_{P_n}\,[f^c\,(X)]\},$$

  where $f^c(x) = \sup_z \{f(z) - c(z, x)\}$ and $\mathcal{LB}$ is a linear space spanned by the function class $\mathcal{B}$:

$$\mathcal{LB} = \left\{ f(\cdot) = \sum_{i=1}^{m} \lambda_i f_i(\cdot) : \{f_i(\cdot)\}_{i=1}^{m} \subset \mathcal{B}, \lambda \in \mathbb{R}^m, \text{ and } m \in \mathbb{Z}_+ \right\}.$$

- $R_n = O(n^{-1/2})$ under some assumptions of $\mathcal{B}$.

Beating Curse of Dimensionality: Smoothness

- If P^* is sufficient smooth, i.e., the density of P^* is in the Besov space $B^s_{p,q}$, then we can construct a wavelet estimator based on data such that $\mathbb{E}W_\rho(\hat{P}^w_n, P^*) = O\left(n^{-\frac{1+s}{d+2s}}\right)$ [Weed and Berthet, 2019].

Beating Curse of Dimensionality: Smoothness

- If P^* is sufficient smooth, i.e., the density of P^* is in the Besov space $B^s_{p,q}$, then we can construct a wavelet estimator based on data such that $\mathbb{E} W_\rho(\hat{P}^w_n, P^*) = O\left(n^{-\frac{1+s}{d+2s}}\right)$ [Weed and Berthet, 2019].
- σ-smooth Wasserstein distance [Nietert et al., 2021]:

$$W_\rho^{(\sigma)}(P, Q) = W_\rho(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma),$$

  where $P * \mathcal{N}_\sigma(A) = \int_{-\infty}^{\infty} P(A - t)\phi_\sigma(t)\mathrm{d}t$ and $\phi_\sigma(t)$ is the PDF of the Gaussian distribution $\mathcal{N}_\sigma$.
  - $\mathbb{E}[W_\rho^{(\sigma)}(P_n, P)] = O(n^{-1/2})$.

More Properties of Smooth Wasserstein Distance

- $W_p^{(\sigma)}$ is continuous and monotonically non-increasing in $\sigma \in [0, +\infty)$;
- $\lim_{\sigma \to 0} W_p^{(\sigma)}(P, Q) = W_p(P, Q)$;
- $\lim_{\sigma \to +\infty} W_p^{(\sigma)}(P, Q) = |\mathbb{E}[X] - \mathbb{E}[Y]|$, for $X \sim P$ and $Y \sim Q$ sub-Gaussian.
- The constants in $\mathbb{E}[W_p^{(\sigma)}(P_n, P)]$ exhibit an exponential dependence on dimension d.

Road Map

Wasserstein Generative Adversarial Networks (GANs) [Arjovsky et al., 2017]

- Goal: learn a generative model $g_\theta(\cdot)$ from data $X_1, X_2, \ldots, X_n$ sampled from a real data distribution P_r. We let P_θ be the distribution induced by the generative model $g_\theta(\cdot)$.

Wasserstein Generative Adversarial Networks (GANs) [Arjovsky et al., 2017]

- Goal: learn a generative model $g_\theta(\cdot)$ from data $X_1, X_2, \ldots, X_n$ sampled from a real data distribution P_r. We let P_θ be the distribution induced by the generative model $g_\theta(\cdot)$.
- Minimize Wasserstein distance:

$$\min_\theta W_1(P_r, P_\theta) = \min_\theta \sup_{\|f\|_L \leq 1} \mathbb{E}_{P_r}[f(x)] - \mathbb{E}_{P_\theta}[f(x)]$$

Wasserstein Generative Adversarial Networks (GANs) [Arjovsky et al., 2017]

- Goal: learn a generative model $g_\theta(\cdot)$ from data $X_1, X_2, \ldots, X_n$ sampled from a real data distribution P_r. We let P_θ be the distribution induced by the generative model $g_\theta(\cdot)$.
- Minimize Wasserstein distance:

$$\min_\theta W_1(P_r, P_\theta) = \min_\theta \sup_{\|f\|_L \le 1} \mathbb{E}_{P_r}[f(x)] - \mathbb{E}_{P_\theta}[f(x)]$$

- Parametric $f(\cdot)$ to be a neural network:

$$\min_\theta W_1(P_r, P_\theta) = \min_\theta \max_w \left\{\mathbb{E}_{P_r}[f_w(x)] - \mathbb{E}_{z\sim p(z)}[f_w(g_\theta(z))]\right\}$$

Wasserstein Generative Adversarial Networks (GANs) [Arjovsky et al., 2017]

- Goal: learn a generative model $g_\theta(\cdot)$ from data $X_1, X_2, \ldots, X_n$ sampled from a real data distribution P_r. We let P_θ be the distribution induced by the generative model $g_\theta(\cdot)$.
- Minimize Wasserstein distance:

$$\min_\theta W_1(P_r, P_\theta) = \min_\theta \sup_{\|f\|_L \leq 1} \mathbb{E}_{P_r}[f(x)] - \mathbb{E}_{P_\theta}[f(x)]$$

- Parametric $f(\cdot)$ to be a neural network:

$$\min_\theta W_1(P_r, P_\theta) = \min_\theta \max_w \left\{\mathbb{E}_{P_r}[f_w(x)] - \mathbb{E}_{z\sim p(z)}[f_w(g_\theta(z))]\right\}$$

- Adversarial.

Wasserstein GANs Results

(d)

(e)

Distributionally Robust Optimization Formulation

Distributionally Robust Optimization (DRO):

$$\inf_{\beta \in \mathbb{R}^d} \underbrace{\sup_{P \in \mathcal{U}} \mathbb{E}_P[\ell(X;\beta)]}_{\text{worst case expectation}},$$

$\mathcal{U}$: distributional uncertainty set.

Distributionally Robust Optimization Formulation

Distributionally Robust Optimization (DRO):

$$\inf_{\beta \in \mathbb{R}^d} \underbrace{\sup_{P \in \mathcal{U}} \mathbb{E}_P[\ell(X;\beta)]}_{\text{worst case expectation}},$$

$\mathcal{U}$: distributional uncertainty set.

Construction of distributional uncertainty set $\mathcal{U}$:

$$\mathcal{U} = \mathcal{U}_\delta(P_n) = \{P \in \mathcal{P}(S) : D_c(P, P_n) \leq \delta\}$$

Why DRO?

- Statistical errors and overfitting;
- Distributional shifts.

Strong Duality for DRO

Theorem (Blanchet and Murthy, 2019; Gao and Kleywegt, 2016; Esfahani and Kuhn, 2018)

Suppose $c(\cdot)$ is a nonnegative lower semicontinuous function satisfying $c(x, y) = 0$ if and only if $x = y$ and $\ell(\cdot)$ is upper semicontinuous. Then,

$$\sup_{P:D_c(P,P_n)\leq\delta} \mathbb{E}_P \left[\ell\left(X;\beta\right)\right] = \inf_{\lambda\geq 0} f(\beta,\lambda),$$

where

$$f(\beta,\lambda) = \lambda\delta + \mathbb{E}_{P_n}[\ell_{rob}\left(X;\beta,\lambda\right)], \textit{ and}$$

$$\ell_{rob}\left(X;\beta,\lambda\right) := \sup_{u\in\mathbb{R}^d} \left\{\ell(u;\beta) - \lambda c(u,X)\right\}.$$

Some DRO Estimators

- **Square-root LASSO** [Belloni, Chernozhukov and Wang 2011]:

$$\ell((x,y);\beta) = \|y - \beta^T x\|_2^2$$

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}(dx,dy)$$

$$c((x,y),(x',y')) = \|x - x'\|_q^2 + \infty \cdot \mathbf{1}\{y \neq y'\}$$

Some DRO Estimators

- **Square-root LASSO** [Belloni, Chernozhukov and Wang 2011]:

$$\ell((x,y);\beta) = \|y-\beta^T x\|_2^2$$

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}(dx,dy)$$

$$c((x,y),(x',y')) = \|x-x'\|_q^2 + \infty\cdot\mathbf{1}\{y\neq y'\}$$

DRO is equivalent to the square-root LASSO [Blanchet, Kang and Murthy, 2016; Gao, Chen and Kleywegt, 2017], $(1/p+1/q=1)$

$$\sup_{P:D_c(P,P_n)\leq\delta} \mathbb{E}_P\left[\ell\left((X,Y);\beta\right)\right] = \left(\sqrt{\mathbb{E}_{P_n}[\ell((X,Y);\beta)]} + \sqrt{\delta}\|\beta\|_p\right)^2.$$

Some DRO Estimators

- **Regularized logistic regression:**

$$\ell((x,y);\beta) = \log(1+\exp(-y\beta^T x))$$

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}(dx,dy)$$

$$c((x,y),(x',y')) = \|x-x'\|_q + \infty \cdot \mathbf{1}\{y \neq y'\}$$

Some DRO Estimators

- **Regularized logistic regression:**

$$\ell((x,y);\beta) = \log(1+\exp(-y\beta^T x))$$

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}(dx,dy)$$

$$c((x,y),(x',y')) = \|x-x'\|_q + \infty \cdot \mathbf{1}\{y \neq y'\}$$

DRO is equivalent to the regularized logistic regression [Blanchet, Kang and Murthy, 2016; Gao, Chen and Kleywegt, 2017; Esfahani and Kuhn, 2015],

$$\sup_{P: D_c(P,P_n)\leq\delta} \mathbb{E}_P\left[\ell\left((X,Y);\beta\right)\right] = \mathbb{E}_{P_n}[\ell((X,Y);\beta)] + \delta\|\beta\|_p.$$

Road Map

Martingale Optimal Transport[1]

$$MD_c(P,Q) := \inf_{\pi}\Big\{\mathbb{E}_\pi[c(X,Y)] \mid \pi \in \mathcal{P}(S\times S), \mathbb{E}_\pi[Y|X] = X,$$
$$\pi(A\times S) = P(A), \pi(S\times B) = Q(B) \text{ for every subsets } A, B \text{ of } S\Big\}$$

[1] Guo and Obłój [2019]

Adapted Optimal Transport[2]

Consider two-period case: P is the joint distribution of (X_1, X_2) and Q is the joint distribution of (Y_1, Y_2),

$$
\begin{aligned}
AD_c(P,Q) := \inf_{\pi^1}\Big\{ & \mathbb{E}_{\pi^1}[c(X_1,Y_1) + D_c(P_{X_1}, Q_{Y_1})] \mid \\
& \pi^1(A \times S) = P^1(A), \pi^1(S \times B) = Q^1(B) \text{ for every subsets } A, B\Big\},
\end{aligned}
$$

where P^1, Q^1 are the distributions of X_1 and Y_1 and P_{X_1}, Q_{Y_1} are the distributions of X_2 and Y_2 conditional on X_1 and Y_1.

[2] Backhoff et al. [2022]

Optimization of Optimal Transport

- Can we solve discrete optimal transport in an online fashion for large-scale problems [Mensch and Peyré, 2020]?
- Can we solve semi-discrete or continuous optimal transport approximately under some structural assumptions?

Minimum Wasserstein Distance

- Recall the CLT:

$$\sqrt{n}(W_\rho^\rho(P_n, P^*) - \mathbb{E}[W_\rho^\rho(P_n, P^*)]) \Rightarrow \mathcal{N}(0, \sigma^2).$$

- What can we say about $\hat{\theta}_n$:

$$\hat{\theta}_n = \arg\min_\theta D_c(P_n, P_\theta)$$

Curse of Dimensionality?

Q&A?

References I

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Julio Backhoff, Daniel Bartl, Mathias Beiglböck, and Johannes Wiesel. Estimating processes in adapted wasserstein distance. *The Annals of Applied Probability*, 32(1):529–550, 2022.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Eustasio Del Barrio and Jean-Michel Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

References II

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Gaoyue Guo and Jan Obłój. Computational methods for martingale optimal transport problems. *The Annals of Applied Probability*, 29(6):3311–3347, 2019.

Soheil Kolouri, Serim Park, Matthew Thorpe, Dejan Slepčev, and Gustavo K Rohde. Transport-based analysis, modeling, and learning from signal and data distributions. *arXiv preprint arXiv:1609.04767*, 2016.

Arthur Mensch and Gabriel Peyré. Online sinkhorn: Optimal transport distances from sample streams. *Advances in Neural Information Processing Systems*, 33:1657–1667, 2020.

Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019.

References III

Sloan Nietert, Ziv Goldfeld, and Kengo Kato. Smooth p-wasserstein distance: Structure, empirical approximation, and statistical applications. In *International Conference on Machine Learning*, pages 8172–8183. PMLR, 2021.

Nian Si, Jose Blanchet, Soumyadip Ghosh, and Mark Squillante. Quantifying the empirical wasserstein distance to a set of measures: Beating the curse of dimensionality. *Advances in Neural Information Processing Systems*, 33:21260–21270, 2020.

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

Bahar Taşkesen, Soroosh Shafieezadeh-Abadeh, and Daniel Kuhn. Semi-discrete optimal transport: Hardness, regularization and numerical solution. *Mathematical Programming*, pages 1–74, 2022.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Jonathan Weed and Quentin Berthet. Estimation of smooth densities in wasserstein distance. In *Conference on Learning Theory*, pages 3118–3119. PMLR, 2019.