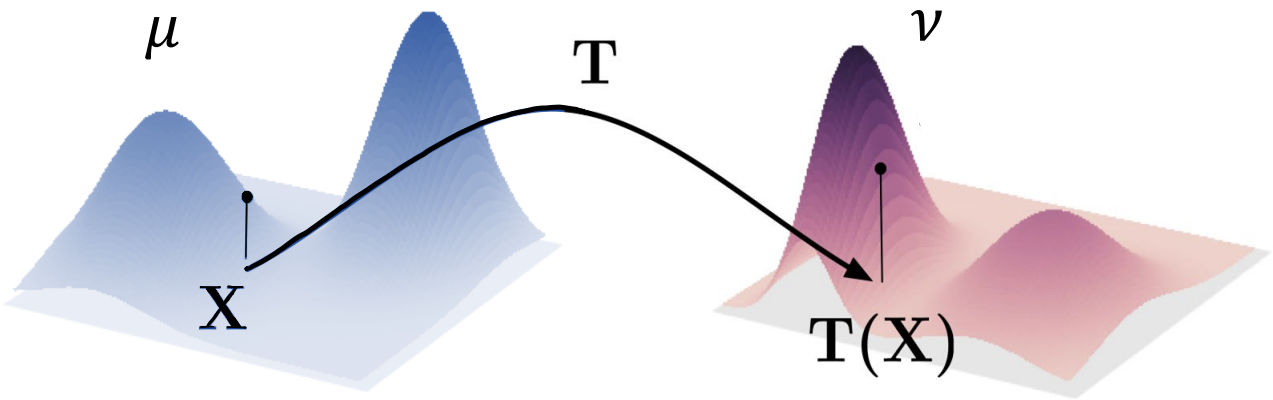
Statistical and Computational Aspects of Sliced Optimal Transport

Ziv Goldfeld
Cornell University

Machine Learning Lunch Seminar, Vanderbilt University
October 17th, 2022

Statistical Divergences for Learning & Inference

Modeling & Analysis: ML tasks boil down to comparing/transforming probability distributions

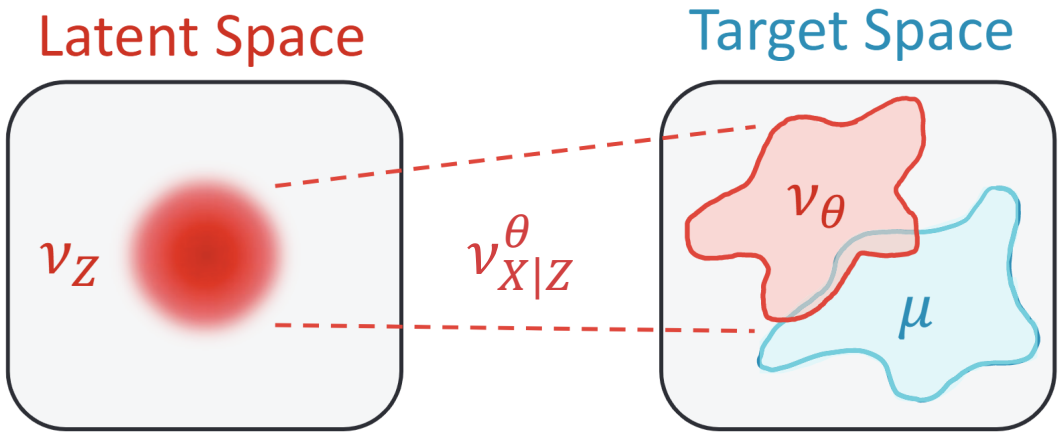


Statistical Divergences for Generative Modeling

Example: Implicit generative modeling aims to learn $\nu_\theta \approx \mu \in \mathcal{P}(\mathbb{R}^d)$

- Map $Z \sim \nu_Z \in \mathcal{P}(\mathbb{R}^p)$, $p \ll d$, to $\mathbb{R}^d$ via parametrized transformation $\nu_{X|Z}^\theta$

$\Longrightarrow$ **Generative model:** $\nu_\theta(\cdot) := \int_{\mathbb{R}^p} \nu_{X|Z}^\theta(\cdot\,|z)\,d\nu_Z(z)$

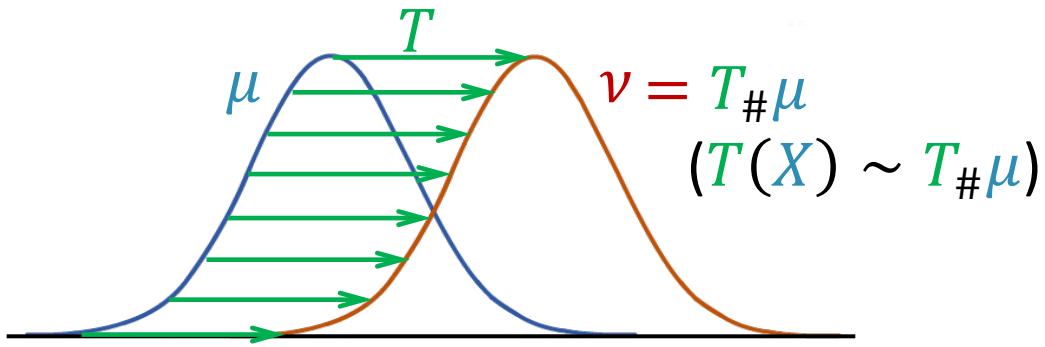


Minimum distance estimation: Solve $\boxed{\theta^* \in \operatorname*{argmin}_\theta \delta(\mu, \nu_\theta)}$

Optimal Transport Distances

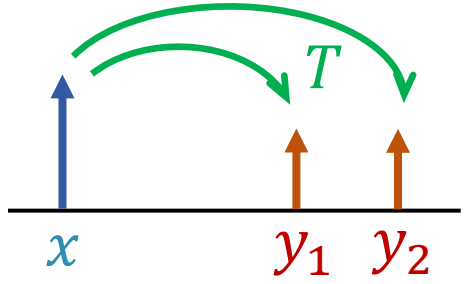
Distributions: $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$

Cost: $c: \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$



Transport map: $T: \mathbb{R}^d \to \mathbb{R}^d$ s.t. $T_{\#}\mu = \nu$

OT problem (Monge 1781): $\mathrm{M}_c(\mu, \nu) := \inf_{T:\, T_{\#}\mu = \nu} \int_{\mathbb{R}^d} c\big(x, T(x)\big) d\mu(x)$

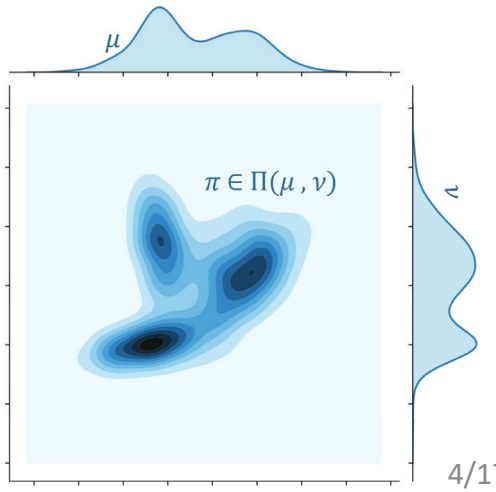


🚫 $\{T: T_{\#}\mu = \nu\}$ may be empty, not closed, non-linear problem, …

Coupling: $\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \colon \pi(\cdot \times \mathbb{R}^d) = \mu,\ \pi(\mathbb{R}^d \times \cdot) = \nu\}$

OT problem (Kantorovich '42)

$$\mathrm{K}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\pi(x, y)$$



The Wasserstein Metric

p-Wasserstein metric

For $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$: $\mathrm{W}_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y) \right)^{1/p}$

Comments: Kantorovich OT with distance cost (or power thereof) $\mathrm{W}_p(\mu, \nu) := \left(\mathrm{K}_{\|\cdot\|^p}(\mu, \nu) \right)^{\frac{1}{p}}$

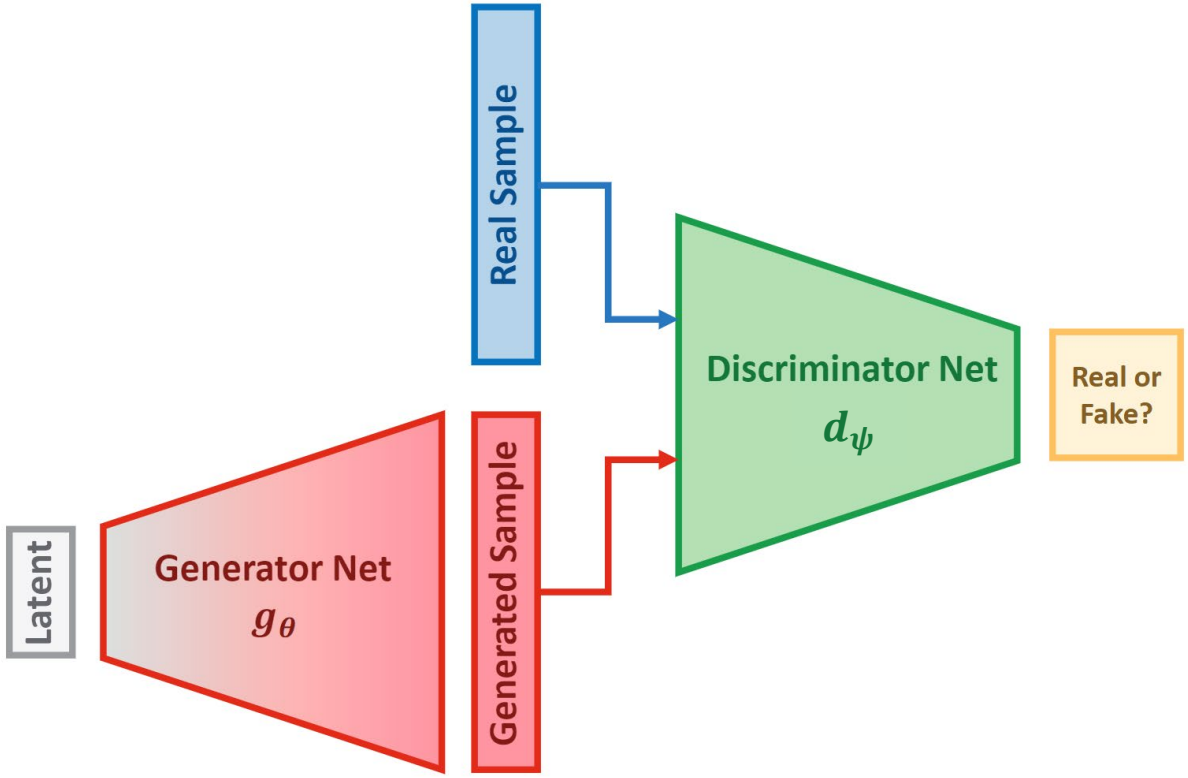
- **Wasserstein space:** $\mathfrak{W}_p = \left(\mathcal{P}_p(\mathbb{R}^d), \mathrm{W}_p \right)$ is a metric space & rich geometric structure
- **Robust to support mismatch:** $\mathrm{W}_p(\mu, \nu) < \infty, \ \forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$
- **Duality:** $\mathrm{W}_1(\mu, \nu) = \sup_{\varphi \in \mathrm{Lip}_1(\mathbb{R}^d)} \int_{\mathbb{R}^d} \varphi d(\mu - \nu)$

From Duality to Generative Modeling

Dual representation: $$\mathrm{W}_1(\mu,\nu) = \sup_{\varphi\in\mathrm{Lip}_1(\mathbb{R}^d)} \mathbb{E}_\mu[\varphi(X)] - \mathbb{E}_\nu[\varphi(Y)]$$

W-GAN [Arjovsky *et al* '17]**:**

- μ (X data sample)
- $\nu = \nu_\theta$ ($Y = g_\theta(Z)$ generated sample)
- $\varphi = d_\psi$ ($\mathrm{Lip}_1(\mathbb{R}^d)$ constraint)



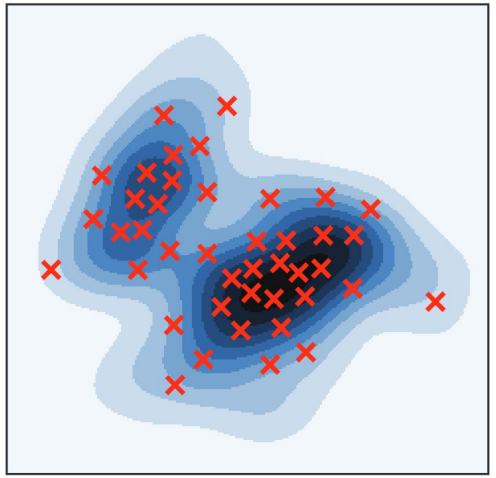
$$\Longrightarrow \quad \boxed{\inf_\theta \mathrm{W}_1(\mu,\nu_\theta) \cong \inf_\theta \sup_{\psi:\, d_\psi\in\mathrm{Lip}_1(\mathbb{R}^d)} \mathbb{E}\big[d_\psi(X)\big] - \mathbb{E}\big[d_\psi\big(g_\theta(Z)\big)\big]}$$

Empirical Estimation in High Dimensions

Sampling: μ, ν are unknown & we get $X_1, \dots, X_n \sim \mu$ and $Y_1, \dots, Y_n \sim \nu$

- **Empirical measures:** $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and $\hat{\nu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$

$\Longrightarrow$ Can we approximate $\mathrm{W}_p(\mu, \nu) \approx \mathrm{W}_p(\hat{\mu}_n, \hat{\nu}_n)$?



Theorem (Dudley '69, Boissard-Le Gouic'14, Niles Weed-Bach'19,…)

For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ and $d > 2p$: $\mathbb{E}\left[\left|\mathrm{W}_p(\mu, \nu) - \mathrm{W}_p(\hat{\mu}_n, \hat{\nu}_n)\right|\right] \asymp n^{-\frac{1}{d}}$

Core issue: For $\mu \ll \mathrm{Leb}(\mathbb{R}^d)$: $\mathrm{W}_p(\hat{\mu}_n, \mu) \gtrsim n^{-\frac{1}{d}}$ a.s.

🚫 **Statistical implication:** Too slow for $d \gg 1$!

🚫 **Computation:** $\mathrm{W}_p(\hat{\mu}_n, \hat{\nu}_n)$ is LP w/ complexity $O(n^3)$ $\Longrightarrow$ infeasible for large n

Regularized OT Distances

Approach: Regularize the OT problem to improve statistical/computational properties

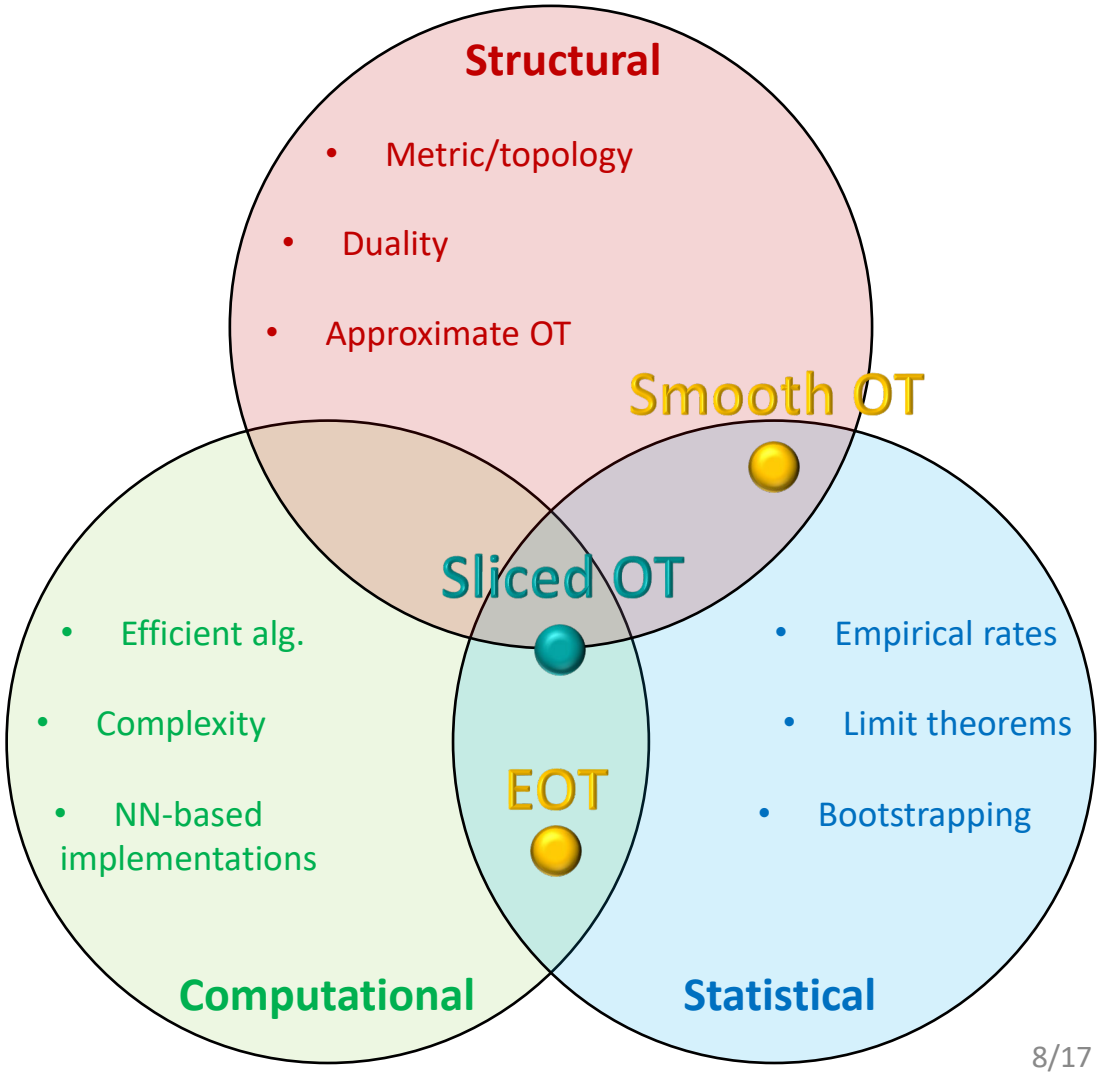
1. **Entropic OT:** Let $\mathrm{D}_{\mathrm{KL}}(\cdot\|\cdot)$ be the KL divergence:

$$\mathrm{E}_c^{\epsilon}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi + \epsilon \mathrm{D}_{\mathrm{KL}}(\pi \parallel \mu \otimes \nu)$$

2. **Smooth OT:** For a smoothing kernel $\kappa_\sigma \in \mathcal{C}^\infty$

$$\mathrm{S}_c^{\sigma}(\mu,\nu) := \mathrm{K}_c(\mu * \kappa_\sigma, \nu * \kappa_\sigma)$$

3. **Sliced OT:** Focus of this talk

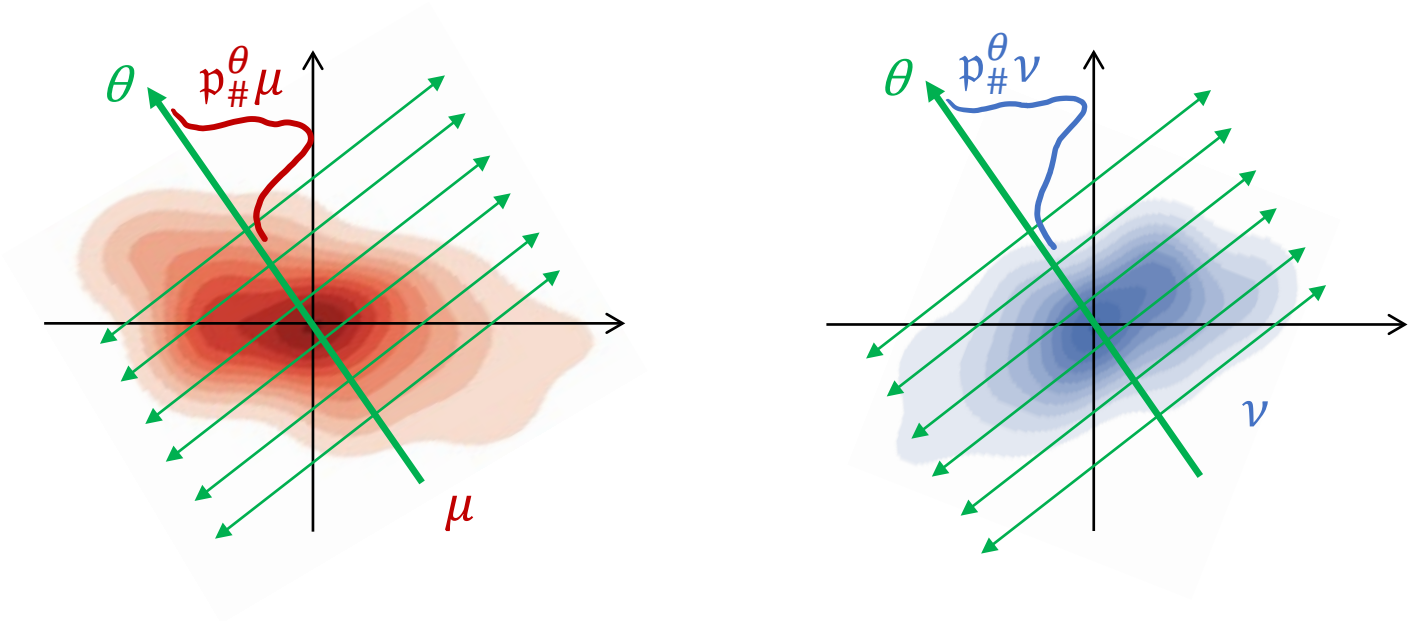


Sliced Wasserstein Distances

Definition (Rabin-Peyré-Delon-Bernot '11)

$$\underline{\mathrm{W}}_p(\mu,\nu) := \left[\int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p\left(\mathfrak{p}^\theta_\#\mu, \mathfrak{p}^\theta_\#\nu\right) d\sigma(\theta)\right]^{1/p} \quad \text{and} \quad \overline{\mathrm{W}}_p(\mu,\nu) := \sup_{\theta\in\mathbb{S}^{d-1}} \mathrm{W}_p\left(\mathfrak{p}^\theta_\#\mu, \mathfrak{p}^\theta_\#\nu\right)$$

where $\mathfrak{p}^\theta(x) = \theta^{\mathrm{T}} x$ and σ is the uniform measure on the unit sphere $\mathbb{S}^{d-1}$



Sliced Wasserstein Distances

Definition (Rabin-Peyré-Delon-Bernot '11)

$$\underline{\mathrm{W}}_p(\mu,\nu) := \left[\int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p\left(\mathfrak{p}_\#^\theta\mu, \mathfrak{p}_\#^\theta\nu\right) d\sigma(\theta)\right]^{1/p} \quad \text{and} \quad \overline{\mathrm{W}}_p(\mu,\nu) := \sup_{\theta\in\mathbb{S}^{d-1}} \mathrm{W}_p\left(\mathfrak{p}_\#^\theta\mu, \mathfrak{p}_\#^\theta\nu\right)$$

where $\mathfrak{p}^\theta(x) = \theta^{\mathrm{T}}x$ and σ is the uniform measure on the unit sphere $\mathbb{S}^{d-1}$

Motivation: OT between 1D distributions $\mu, \nu \in \mathcal{P}(\mathbb{R})$:

- **General:** $\mathrm{W}_p(\mu,\nu) = \left\|F_\mu^{-1} - F_\nu^{-1}\right\|_{L^p([0,1])} \overset{p=1}{=} \left\|F_\mu - F_\nu\right\|_{L^1(\mathbb{R})}$
- **Discrete:** $\mathrm{W}_p^p(\mu,\nu) = \frac{1}{n}\sum_{i=1}^n \left|X_{(i)} - Y_{(i)}\right|^p,$

  where $X_{(1)} \le \cdots \le X_{(n)}$ and $Y_{(1)} \le \cdots \le Y_{(n)}$ are order statistics

$\Longrightarrow$ Computing projected distance $\mathrm{W}_p\left(\mathfrak{p}_\#^\theta\mu, \mathfrak{p}_\#^\theta\nu\right)$ is cheap!

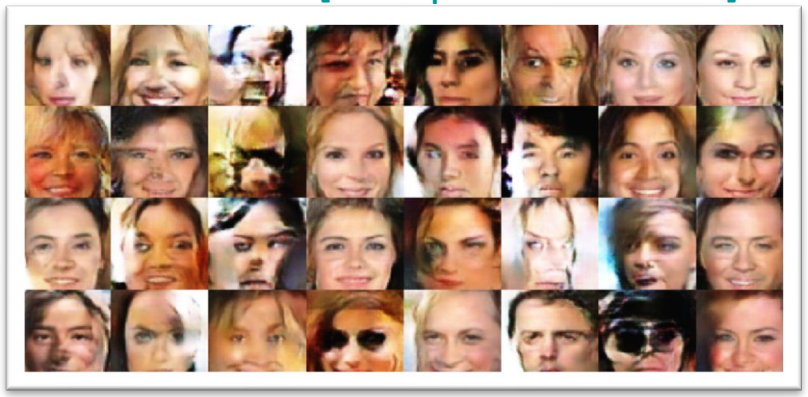
Structural Properties

Metric/topological structure [Nadjahi *et al* '20 , Bayrakta-Guo '21]**:**

$\underline{\mathrm{W}}_p$ and $\overline{\mathrm{W}}_p$ are metrics on $\mathcal{P}_p(\mathbb{R}^d)$ and generate the same topology as W_p

Duality: $$\overline{\mathrm{W}}_1(\mu,\nu) = \sup_{\theta\in\mathbb{S}^{d-1}} \mathrm{W}_1(\mathfrak{p}_\#^\theta\eta, \mathfrak{p}_\#^\theta\nu)$$

Sliced W-GAN [Deshpande *et at* '18]



Approximation: Exponential gap between sliced and classic W_p

Lemma (Bonnotte '13)

There exists $c_{p,d}, C_{p,d} > 0$ s.t. for any $\mu,\nu \in \mathcal{P}(\mathbb{B}_d(0,R))$, we have

$$\underline{\mathrm{W}}_p^p(\mu,\nu) \le c_{p,d}\mathrm{W}_p^p(\mu,\nu) \le C_{p,d}\, R^{p-\frac{1}{d+1}}\, \underline{\mathrm{W}}_p^{\frac{1}{d+1}}(\mu,\nu)$$

$\Longrightarrow$ **Cannot** use sliced distances to approximate classic ones to arbitrary precision

Computational Aspects: Average-Sliced

Recall: $\mathfrak{w}_n^p(\theta) := \mathrm{W}_p^p\left(\mathfrak{p}_\#^\theta \hat{\mu}_n, \mathfrak{p}_\#^\theta \hat{\nu}_n\right)$ computable via sorting & $\underline{\mathrm{W}}_p^p(\hat{\mu}_n, \hat{\mu}_n) = \int_{\mathbb{S}^{d-1}} \mathfrak{w}_n^p(\theta) d\sigma(\theta)$

Average-sliced: MC integration $\underline{\mathrm{W}}_p^p(\mu, \nu) \approx \widehat{\mathrm{W}}_{\mathrm{MC}}^p := \frac{1}{L}\sum_{\ell=1}^L \mathfrak{w}_n^p(\Theta_\ell)$ [Nadjahi *et al* '20]

Theorem (Nietert-Sadhu-G.-Kato '22)

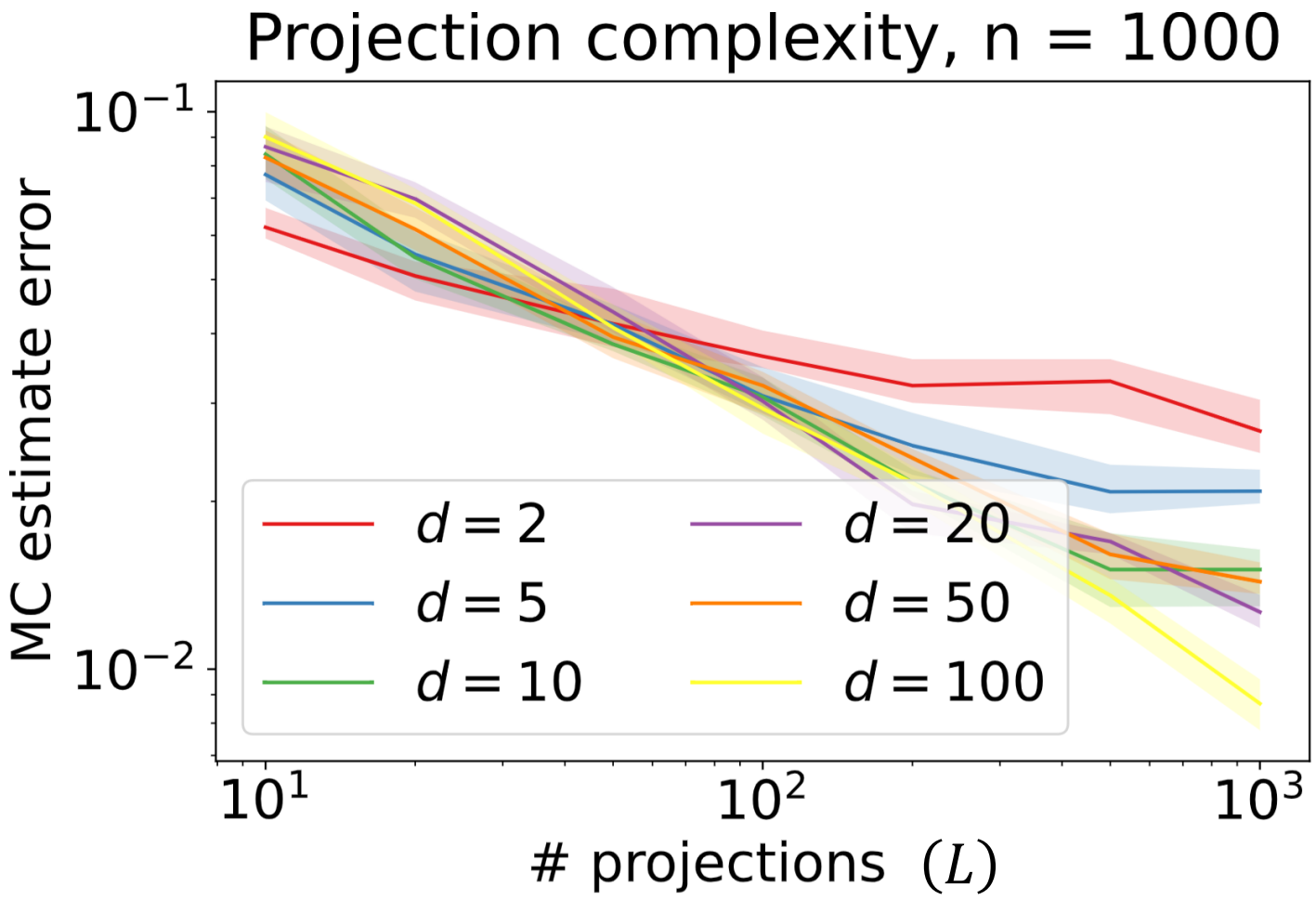
For log-concave $\mu, \nu \in \mathcal{P}\left(\mathbb{R}^d\right)$ with means m_μ, m_ν and covariances Σ_μ, Σ_ν, we have

$$\mathbb{E}\left[\left|\underline{\mathrm{W}}_p^p(\mu, \nu) - \widehat{\mathrm{W}}_{\mathrm{MC}}^p\right|\right] \lesssim_p \frac{\left\|m_\mu - m_\nu\right\|^p + \left\|\Sigma_\mu\right\|_{\mathrm{op}}^{p/2} + \left\|\Sigma_\nu\right\|_{\mathrm{op}}^{p/2}}{\sqrt{dL}} + \underbrace{\delta_n(\mu, \nu)}_{\text{empirical est. error}}$$

Analysis: 1st term corresponds to $L^{-1/2}\sqrt{\mathrm{var}\left(\mathrm{W}_p^p\left(\mathfrak{p}_\#^\Theta \mu, \mathfrak{p}_\#^\Theta \nu\right)\right)}$, for $\Theta \sim \mathrm{Unif}\left(\mathbb{S}^{d-1}\right)$

1. Show that $\mathrm{W}_p^p\left(\mathfrak{p}_\#^\theta \mu, \mathfrak{p}_\#^\theta \nu\right)$ is a Lipschitz function of $\theta \in \mathbb{S}^{d-1}$
2. Use concentration of Lipschitz functions over the unit sphere to bound variance

Computational Aspects: Average-Sliced (Empirical)



Computational Aspects: Max-Sliced

Max-sliced: Rewrite $\overline{\mathrm{W}}_p^p(\hat{\mu}_n, \hat{\nu}_n) = -\min_{\theta \in \mathbb{B}_d} \max_{\sigma \in S_n} \underbrace{\left(-\frac{1}{n}\sum_{i=1}^n \left|\theta^{\mathrm{T}}\left(X_i - Y_{\sigma(i)}\right)\right|^p\right)}_{:=\rho(\sigma,\theta)}$

Subgradient method: Subgradient of $\widetilde{\mathfrak{w}}_n^p(\theta) := \max_{\sigma \in S_n} \rho(\sigma, \theta)$ is easy to evaluate

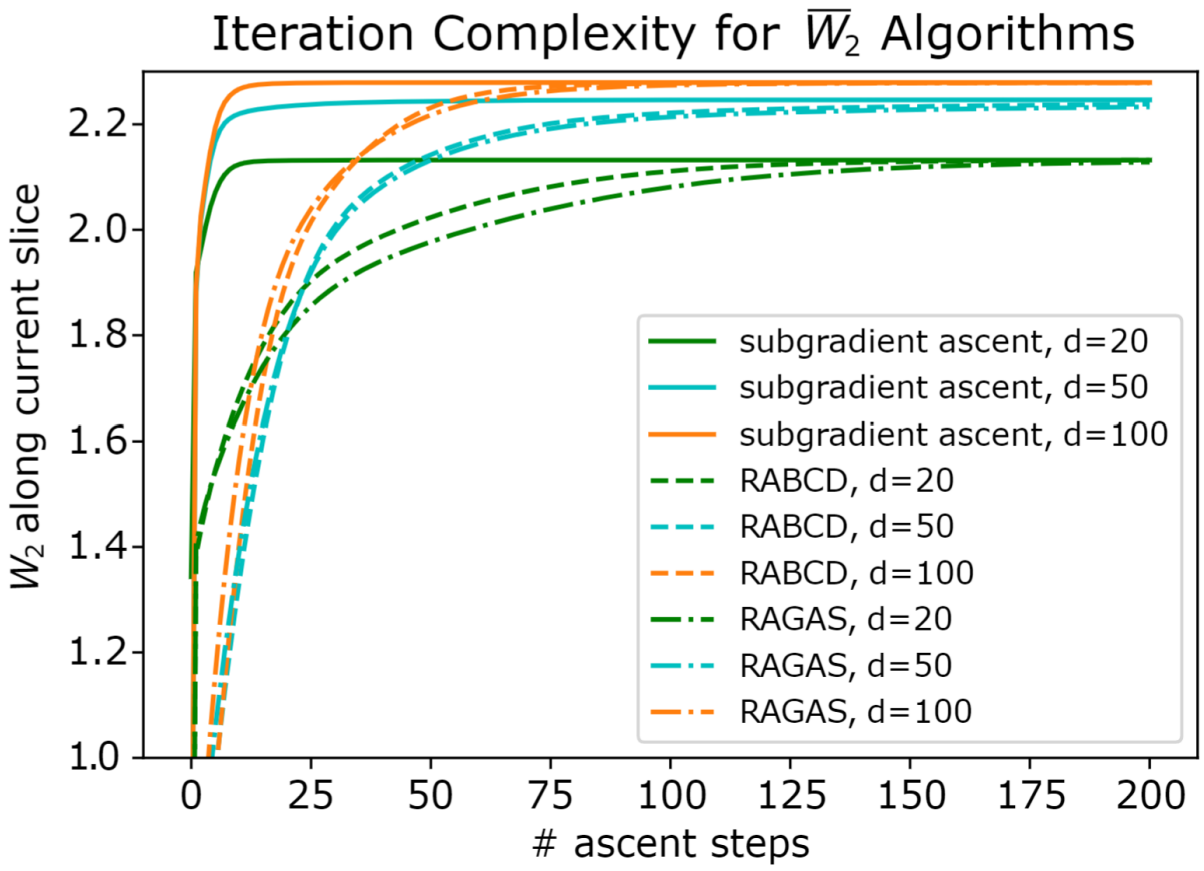
1. Sudifferential of max-function $\partial \widetilde{\mathfrak{w}}_n^p(\theta) = \mathrm{Conv}\left(\left\{\partial_\theta \rho(\sigma^*, \theta) : \sigma^* \in \mathrm{argmax}_{\sigma \in S_n} \rho(\sigma, \theta)\right\}\right)$
2. For fixed θ compute optimal permutation $\sigma^* \in S_n$ via order statistic
3. Evaluate the subgradient vector in $\partial_\theta \rho(\sigma^*, \theta)$

$\Longrightarrow$ EM-like alg. **+** projected subgradient method to compute $\overline{\mathrm{W}}_p^p(\hat{\mu}_n, \hat{\nu}_n)$ [Nadjahi *et al* '20]

Theorem (Nietert-Sadhu-G.-Kato '22)

$\widetilde{\mathfrak{w}}_n^2(\theta)$ is weakly-convex in $\theta \in \mathbb{B}_d$.

Computational Aspects: Max-Sliced (Empirical)



Statistical Analysis: Empirical Convergence Rates

- [Nadjahi *et al* '20]: Rates for average-sliced distances follow 1D rates of base distance
- [Niles Weed-Rigollet '19, Lin *et al* '21]: $O_d\left(n^{-1/(2p)}\right)$, under $T_q(\sigma^2)$, Poincaré, bdd moments

Theorem (Nietert-Sadhu-G.-Kato '22)

For log-concave $\mu \in \mathcal{P}\left(\mathbb{R}^d\right)$ with covariance Σ s.t. $\mathrm{rank}(\Sigma) = k$, we have

$$\mathbb{E}\left[\underline{\mathrm{W}}_p(\hat{\mu}_n, \mu)\right] \lesssim \frac{\left(\|\Sigma\|_{\mathrm{op}}^{1/2} (\log n)^{\mathbb{I}\{p=2\}}\right)}{n^{1/(2\vee p)}}$$

$$\mathbb{E}\left[\overline{\mathrm{W}}_p(\hat{\mu}_n, \mu)\right] \lesssim \frac{\|\Sigma\|_{\mathrm{op}}^{1/2} k \log n}{n^{1/p}} + \frac{\|\Sigma\|_{\mathrm{op}}^{1/2} \sqrt{k \log n}}{n^{1/(2\vee p)}} + \frac{\left(\|\Sigma\|_{\mathrm{op}}^{1/2} (\log n)^{\mathbb{I}\{p=2\}}\right)}{n^{1/(2\vee p)}}$$

Analysis (for $\underline{\mathrm{W}}_p$): If $\mu \in \mathcal{P}_p\left(\mathbb{R}^d\right)$ is log-concave then so is $\mathfrak{p}_{\#}^{\theta}\mu$ & has bounded Cheeger constant

$\Longrightarrow \sup_{\theta} \mathbb{E}\left[\mathrm{W}_p\left(\mathfrak{p}_{\#}^{\theta}\hat{\mu}_n, \mathfrak{p}_{\#}^{\theta}\mu\right)\right] \lesssim$ RHS via 1D W_p rates for bdd Cheeger const [Bobkov-Ledoux '16]

Statistical Analysis: Limit Distribution (Method)

Limit theorems: Key for valid inference but remained an open question

Approach: Relies on extended functional delta method in normed spaces [Römisch '04]

Theorem (G.-Kato-Rioux-Sadhu '22)

Setup: $\mathcal{F}$ class of Borel functions with finite envelope F

$\ell^\infty(\mathcal{F})$ space of bdd functionals from $\mathcal{F}$ to $\mathbb{R}$ equipped w/ $\|\mu\|_{\infty,\mathcal{F}} := \sup_{f\in\mathcal{F}} |\underbrace{\mu(f)}_{\mathbb{E}_\mu[f]}|$

Φ is a functional from (a subset of) $\ell^\infty(\mathcal{F})$ to $\mathbb{R}$

Assume:

1. The empirical process weakly converges $\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} G_\mu$ in $\ell^\infty(\mathcal{F})$
2. Φ is a Lipschitz functional wrt $\|\cdot\|_{\infty,\mathcal{F}}$, i.e., $|\Phi(\nu) - \Phi(\nu')| \leq C\|\nu - \nu'\|_{\infty,\mathcal{F}}$
3. Φ is Gâteaux differentiable at μ w/ derivative $\Phi'_\mu(\rho) = \lim_{t\to 0^+} \frac{1}{t}\big(\Phi(\mu + t\rho) - \Phi(\mu)\big)$

Then: $\sqrt{n}\big(\Phi(\hat{\mu}_n) - \Phi(\mu)\big) \xrightarrow{d} \Phi'_\mu\big(G_\mu\big)$

Statistical Analysis: Limit Distribution (Derivation)

Goal: Limit thm for $\sqrt{n}\left(\underline{\mathrm{W}}_p(\hat{\mu}_n,\nu)-\underline{\mathrm{W}}_p(\mu,\nu)\right)$ for fixed $\nu\in\mathcal{P}\left(\mathbb{R}^d\right)$ with $\mu\neq\nu$ and $p>1$

Derivation: Assume compact supports and take $\Phi(\eta)\coloneqq\underline{\mathrm{W}}_p^p(\eta,\nu)$

1. $\underline{\mathrm{W}}_p^p$ is Lipschitz in $\overline{\mathrm{W}}_1$ & $\overline{\mathrm{W}}_1(\eta,\nu)=\|\eta-\nu\|_{\infty,\mathcal{F}}$ w/ $\mathcal{F}=\left\{\varphi\circ\mathfrak{p}^\theta:\varphi\in\mathrm{Lip}_1(\mathbb{R}),\theta\in\mathbb{S}^{d-1}\right\}$
   - **Recall:** $\overline{\mathrm{W}}_1(\eta,\nu)=\sup\limits_{\substack{\varphi\in\mathrm{Lip}_1(\mathbb{R})\\ \theta\in\mathbb{S}^{d-1}}}(\eta-\nu)\left(\varphi\circ\mathfrak{p}^\theta\right)=\|\eta-\nu\|_{\infty,\mathcal{F}}$
2. $\mathcal{F}$ is a μ-Donsker class (via Donskerness of $\mathrm{Lip}_1(\mathbb{R})$) $\Longleftrightarrow$ $\sqrt{n}(\hat{\mu}_n-\mu)\xrightarrow{d}G_\mu$ in $\ell^\infty(\mathcal{F})$
3. Evaluate the Gâteaux derivative $\frac{d}{dt^+}\underline{\mathrm{W}}_p^p(\mu+t\rho,\nu)=\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}^d}\varphi_\theta\left(\theta^{\mathrm{T}}x\right)d\rho(x)d\sigma(\theta)$

$\Longrightarrow$ Apply general theorem

Statistical Analysis: Limit Distribution (Result)

Theorem (G.-Kato-Rioux-Sadhu '22)

Fix $p > 1$. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be compactly supported s.t. μ is a.c. and $\mathrm{spt}(\mu)$ is convex.

For $\theta \in \mathbb{S}^{d-1}$, let φ^θ be an OT potential from $\mathfrak{p}^\theta_\# \mu$ to $\mathfrak{p}^\theta_\# \nu$ (unique up to additive const.). Then:

$$\sqrt{n}\left(\underline{\mathrm{W}}_p^p(\hat{\mu}_n, \nu) - \underline{\mathrm{W}}_p^p(\mu, \nu)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_p^2\right)$$

where $\sigma_p^2 = \int \int \mathrm{Cov}\left(\varphi^\theta \circ \mathfrak{p}^\theta, \varphi^\psi \circ \mathfrak{p}^\psi\right) d\sigma(\theta) d\sigma(\psi)$.

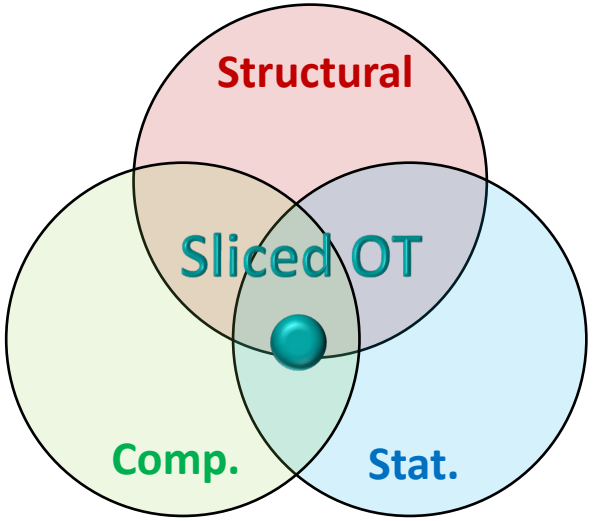
Comments: All results extend to $\overline{\mathrm{W}}_p$; separate (more general) derivations for $p = 1$

1. Linearity of the Gâteaux derivative directly implies consistency of the bootstrap
2. pth power can be removed via standard delta method for the map $x \mapsto x^{1/p}$
3. Similar approach yields distributional limits for entropic/smooth OT

Summary

Classic $\mathbf{W}_p$: Metric on $\mathcal{P}_p(\mathbb{R}^d)$ with rich & meaningful structure

- Many applications in ML and beyond
- Hard to compute & statistical curse of dimensionality $n^{-1/d}$



Sliced $\mathbf{W}_p$: Project distributions to low dimension & average/maximize

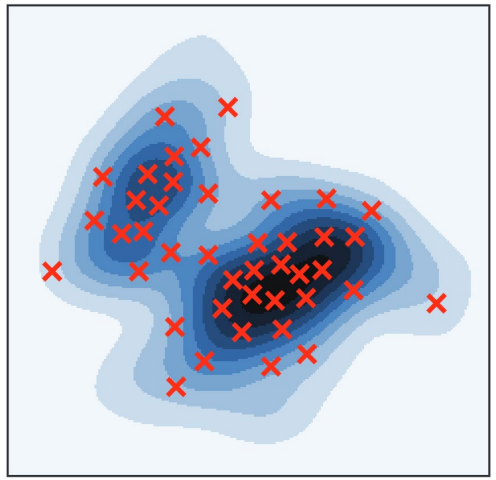
- Inherits structure of W_p (metric, topology, duality, but not proxy)
- Easy to compute & formal guarantees [A]
- Fast empirical convergence [A] & rich limit distribution theory [B]

[A] Nietert, Sadhu, Goldfeld, Kato, "Statistical, robustness, and computational guarantees for sliced W_p",ArXiv:2210.09160

[B] Goldfeld, Kato, Rioux, Sadhu, "Statistical inference with regularized optimal transport", ArXiv:2205.04283

Thank you!

Generative Modeling: Generalization



Goal: Solve $\mathrm{OPT} = \inf_{\theta} \mathrm{W}_1(\mu, \nu_\theta)$ exactly (find θ^*)

Estimation: We don't have μ and only get samples $X_1, \dots, X_n$

- **Empirical distribution:** $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

$\Longrightarrow$ Inherently we work with $\mathrm{W}_1(\hat{\mu}_n, \nu_\theta)$

Optimization: Can solve $\inf_{\theta} \mathrm{W}_1(\hat{\mu}_n, \nu_\theta)$ approximately

$$\text{Find } \hat{\theta}_n \text{ s.t. } \mathrm{W}_1\left(\hat{\mu}_n, \nu_{\hat{\theta}_n}\right) \leq \inf_{\theta} \mathrm{W}_1(\hat{\mu}_n, \nu_\theta) + \epsilon$$

Generalization: $\mathrm{W}_1\left(\mu, \nu_{\hat{\theta}_n}\right) - \mathrm{OPT} \leq 2\mathrm{W}_1(\hat{\mu}_n, \mu) + \epsilon$

$\Longrightarrow$ **Boils down to empirical approximation question under $\mathbf{W_1}$**