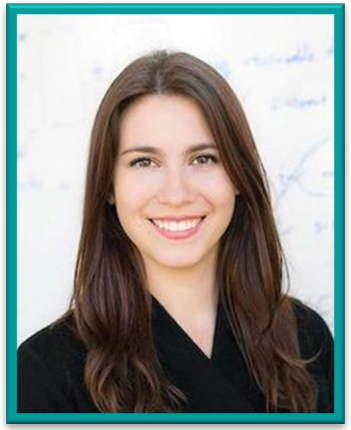
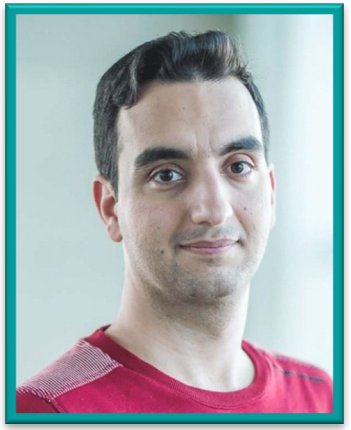
Robust and Geometry-Aware Distribution Estimation via Optimal Transport

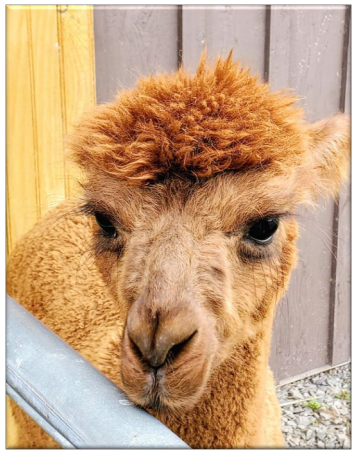
Ziv Goldfeld
Cornell University


Sloan Nietert


Rachel Cummings


Soroosh Shafiee

Machine Learning & Geometry

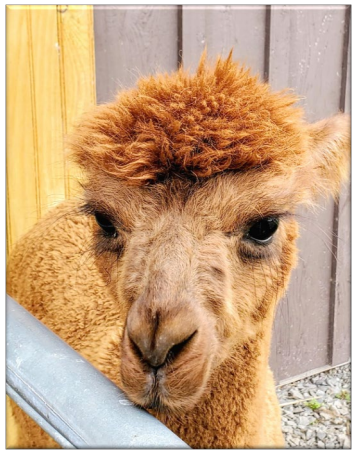


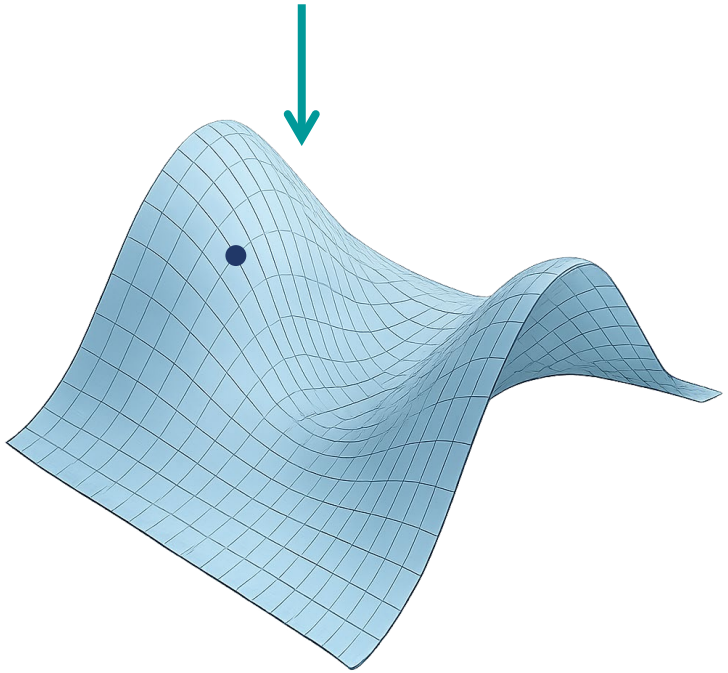
$$\begin{bmatrix} 3 \\ 1.4 \\ 1.5 \\ \vdots \\ 9.2 \end{bmatrix} \in \mathbb{R}^d$$

The quick brown fox jumps over the lazy dog

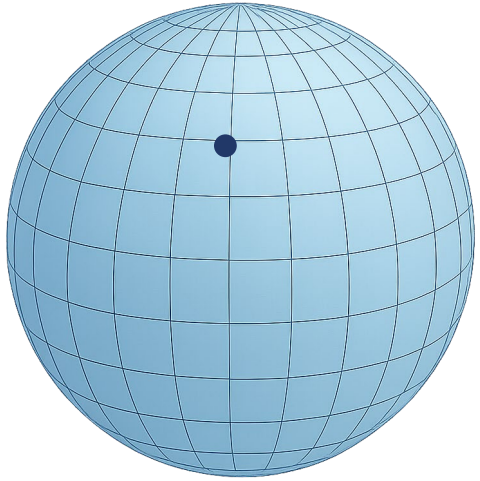
$$\begin{bmatrix} 2.7 \\ 18 \\ 2.8 \\ \vdots \\ 1.8 \end{bmatrix} \in \mathbb{R}^d$$

Machine Learning & Geometry

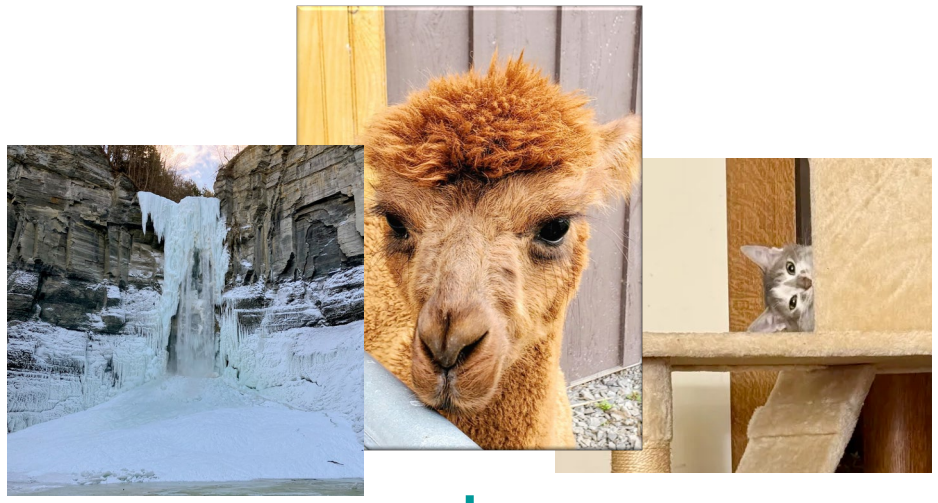


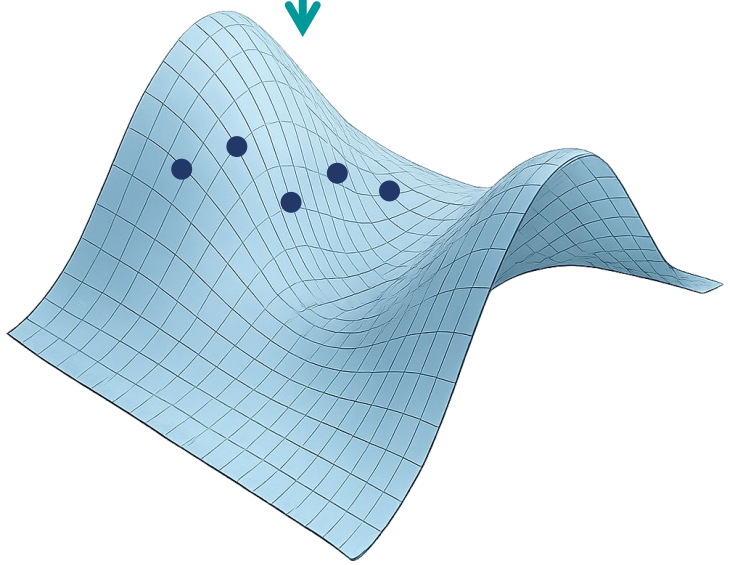


The quick brown fox jumps over the lazy dog

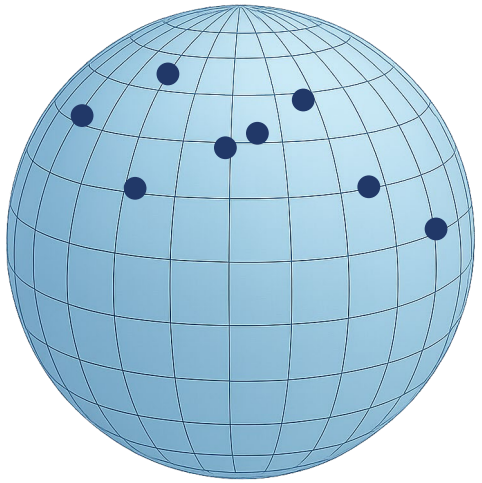


Machine Learning & Geometry



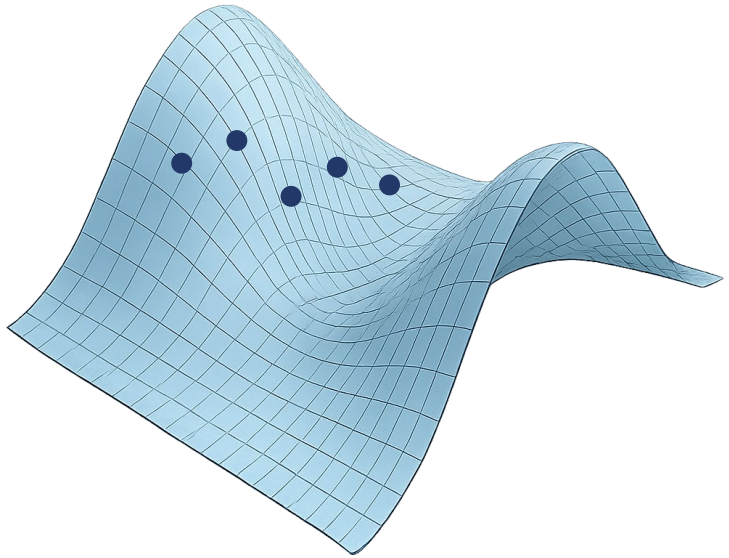


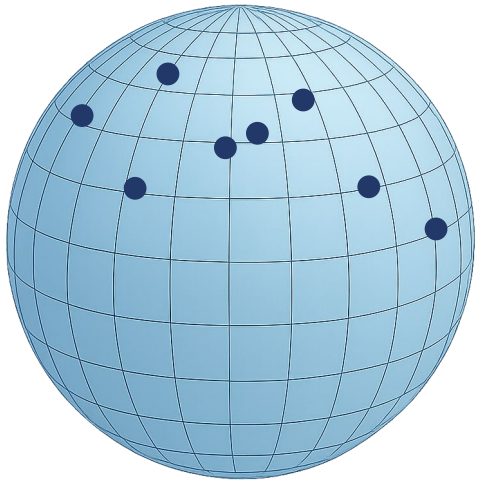
The quick brown fox jumps over the lazy dog



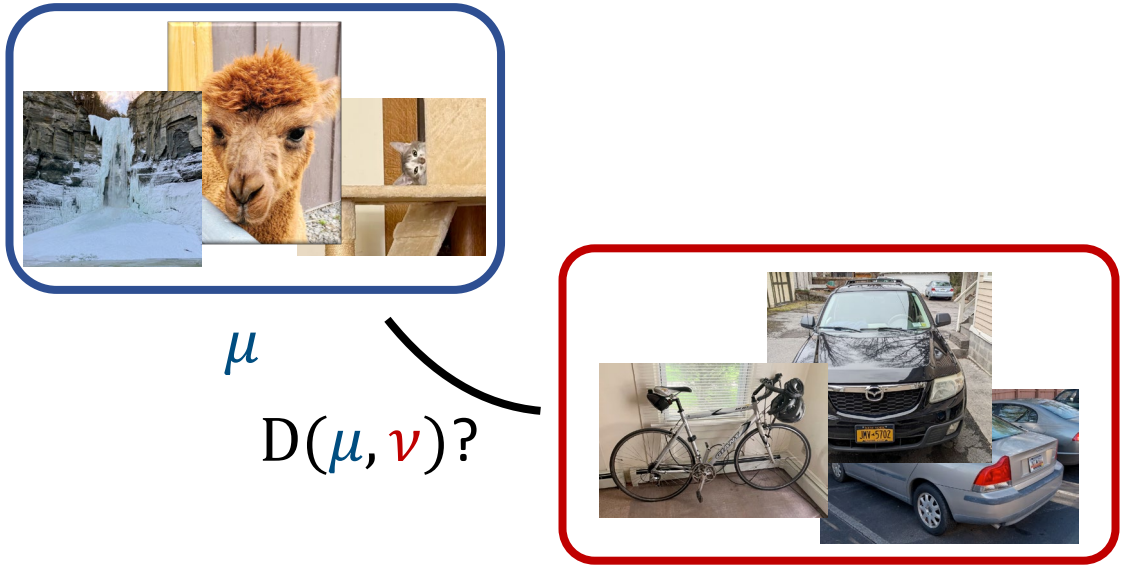
Machine Learning & Geometry

The quick brown fox jumps over the lazy dog



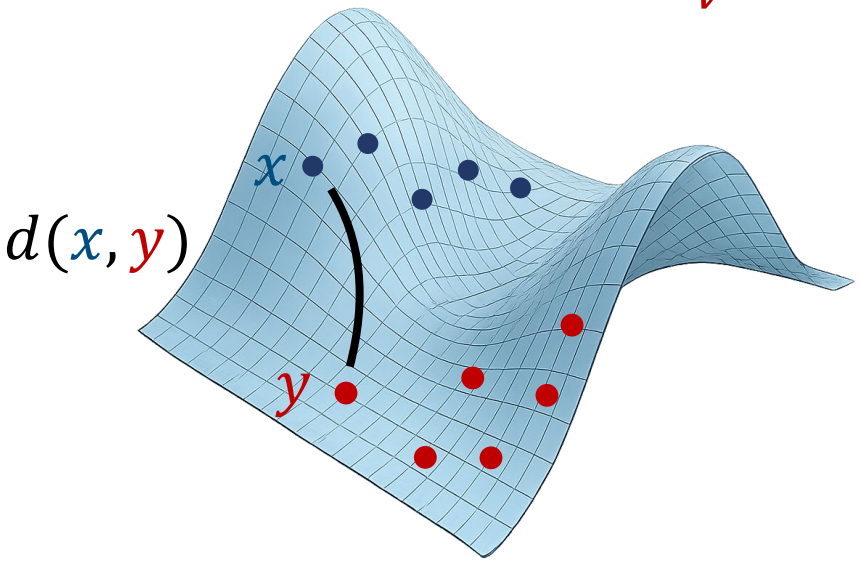


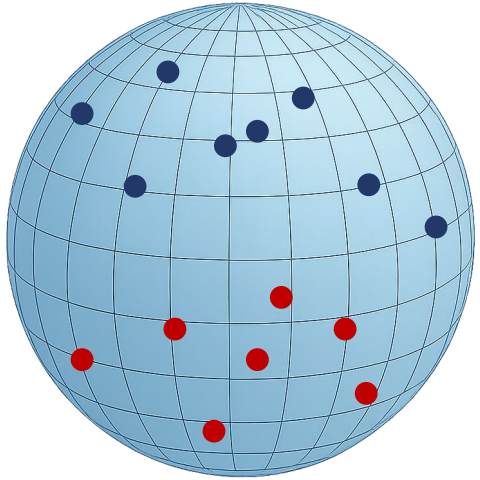
Machine Learning & Geometry



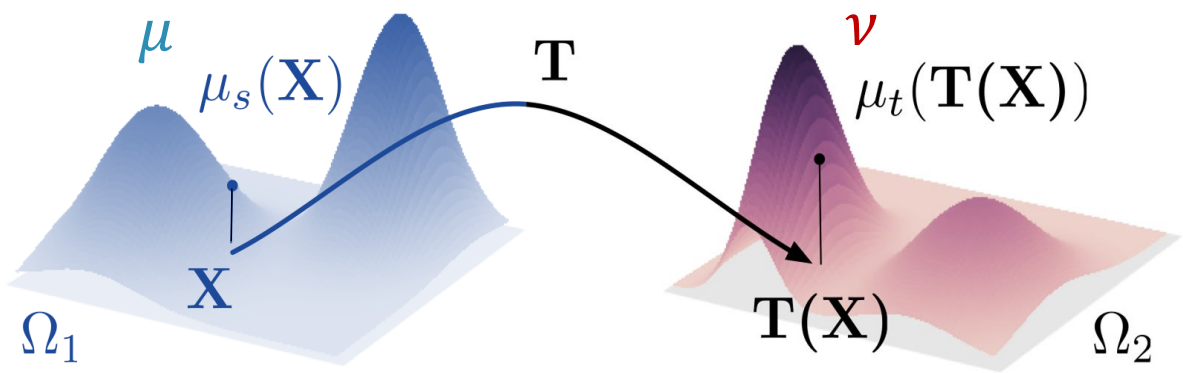
The quick brown fox jumps over the lazy dog

If we are eventually to understand the capability of higher organisms for perceptual recognition…





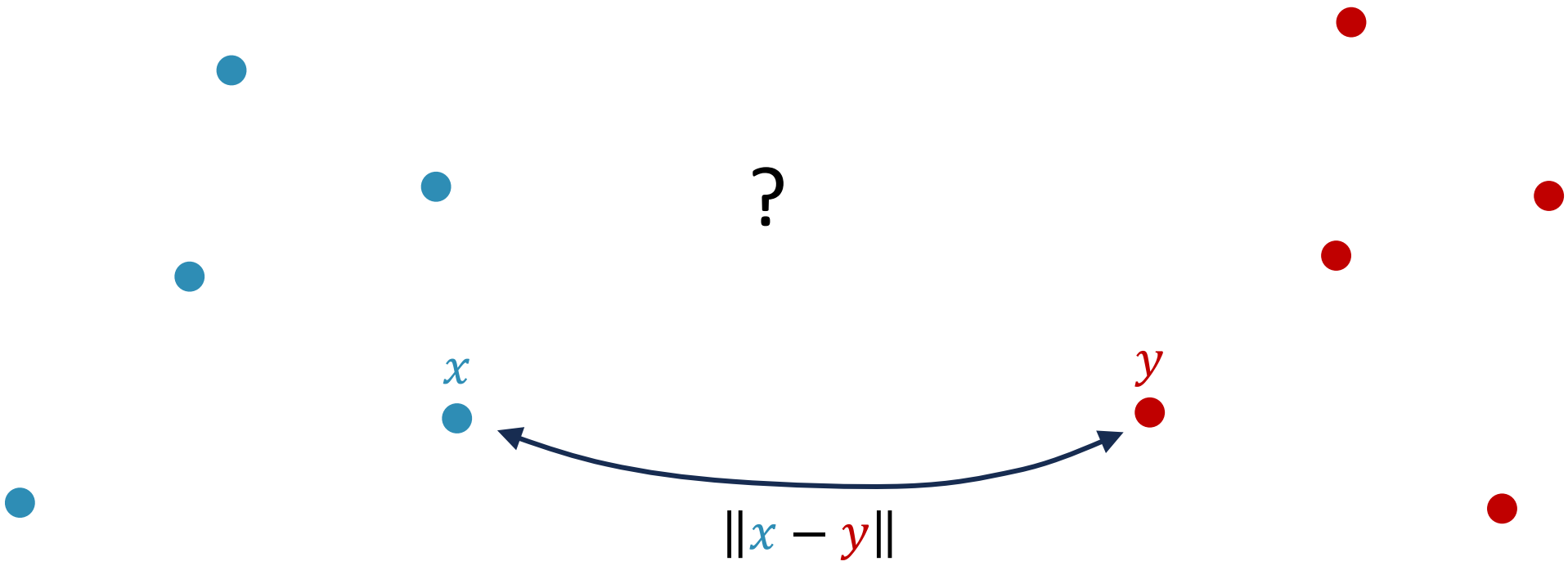
Optimal Transport



Optimal Transport

Given two dataset in $\mathbb{R}^d$, how to compare?

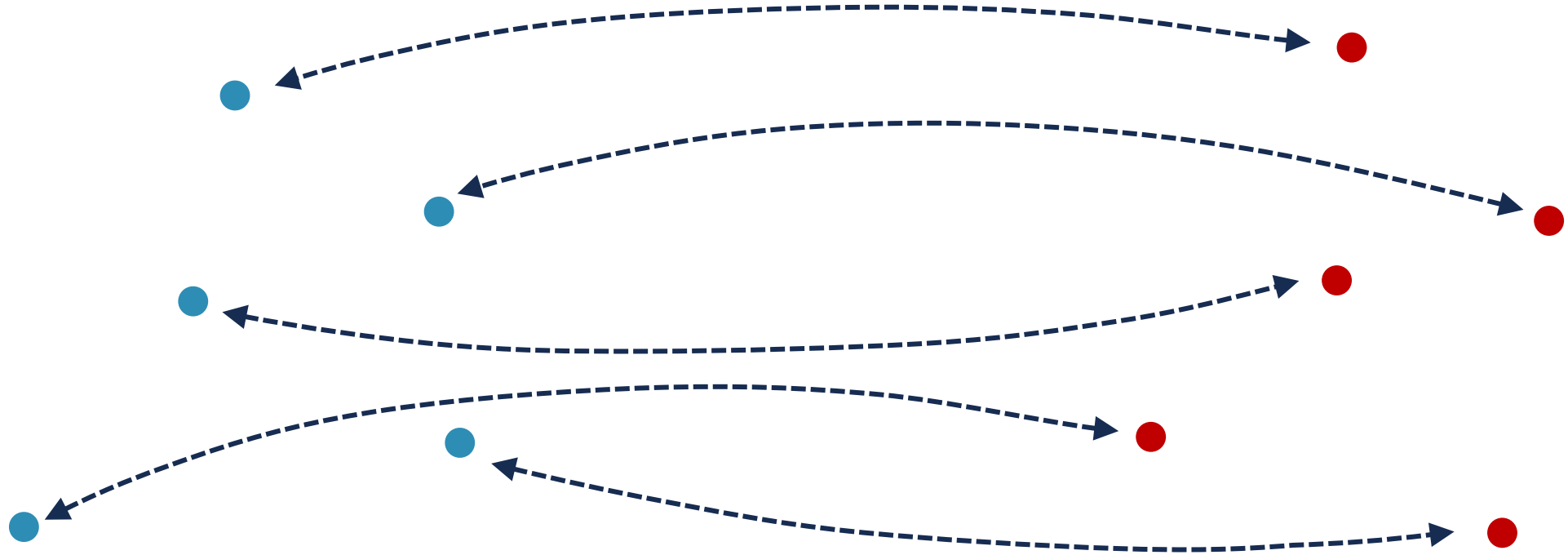
OT approach: Compute min-cost matching with distance-based cost



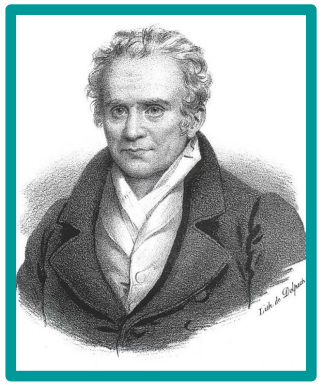
Optimal Transport

Given two dataset in $\mathbb{R}^d$, how to compare?

OT approach: Compute min-cost matching with distance-based cost



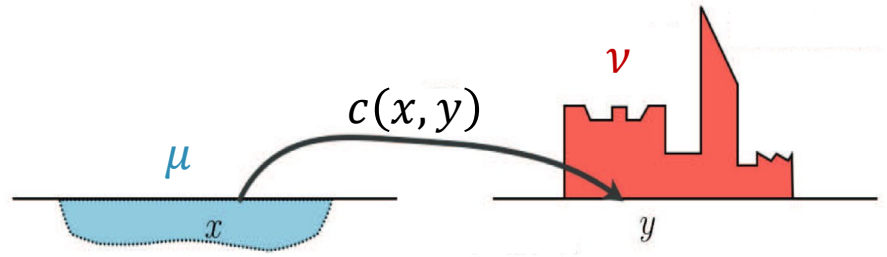
Monge & Kantorovich Formulations

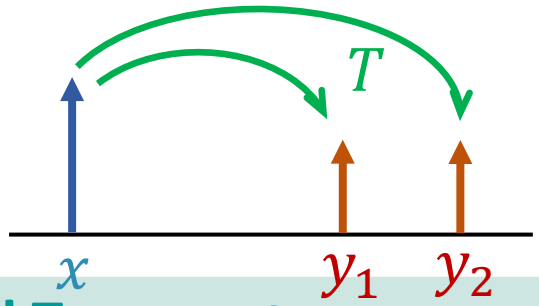


Monge (1781)

$$\mathrm{M}_c(\mu,\nu) := \inf_{\substack{T:\mathcal{X}\to\mathcal{Y} \\ T_{\#}\mu=\nu}} \mathbb{E}_{X\sim\mu}\left[c\left(X,T(X)\right)\right]$$

Transport map:
$\mathrm{Law}\left(T(X)\right) = \nu$



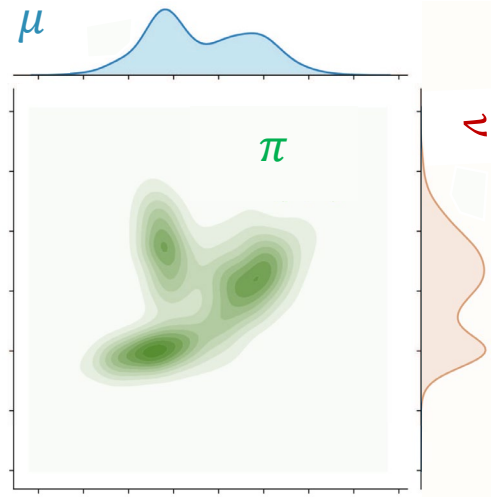


Kantorovich (1942)

Kantorovich Optimal Transport

$$\mathrm{OT}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[c(X,Y)\right]$$

Coupling (transport plan):
$\pi(x,y) = \mu(x)\pi(y|x)$



The Wasserstein Metric

p-Wasserstein Distance ($p \in [1, \infty)$)

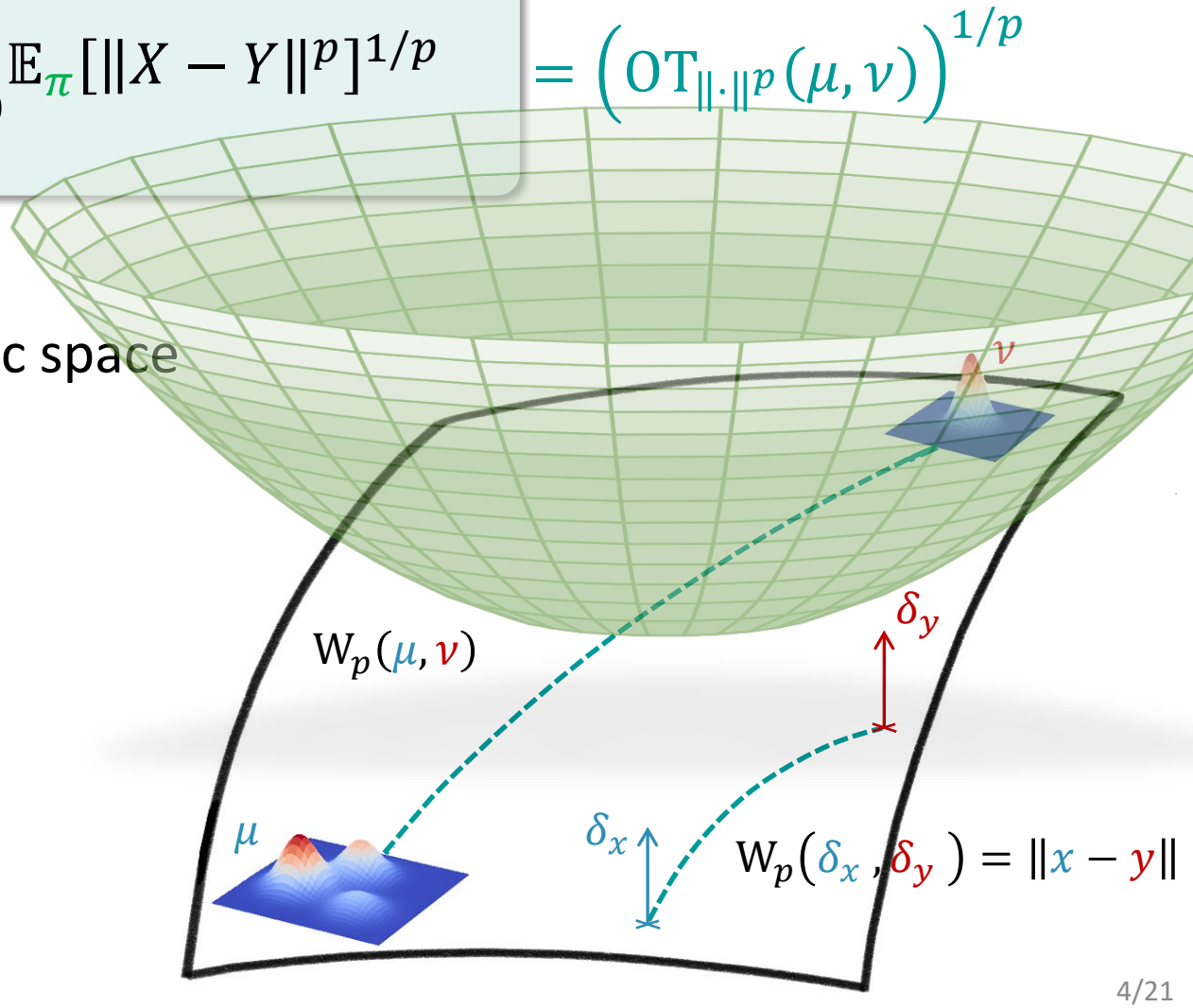
$$\mathrm{W}_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi[\|X - Y\|^p]^{1/p} = \left(\mathrm{OT}_{\|\cdot\|^p}(\mu, \nu)\right)^{1/p}$$

Wasserstein space: $\mathfrak{W}_p = \left(\mathcal{P}_p(\mathbb{R}^d), \mathrm{W}_p\right)$ metric space

Wasserstein geometry:

- Geodesics
- Barycenters
- Projections
- Gradient flows

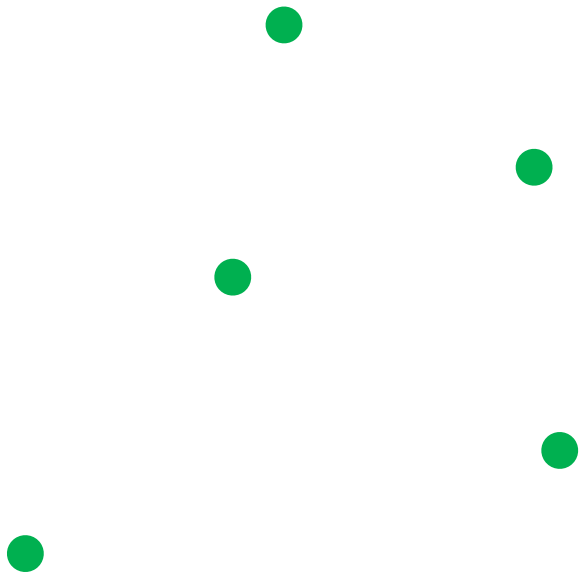
⋮



Corruption Model
&
Partial Optimal Transport

Corruption Model

Clean data: $X_1, \dots, X_n \in \mathbb{R}^d$



Corruption Model

Clean data: $X_1, \dots, X_n \in \mathbb{R}^d$

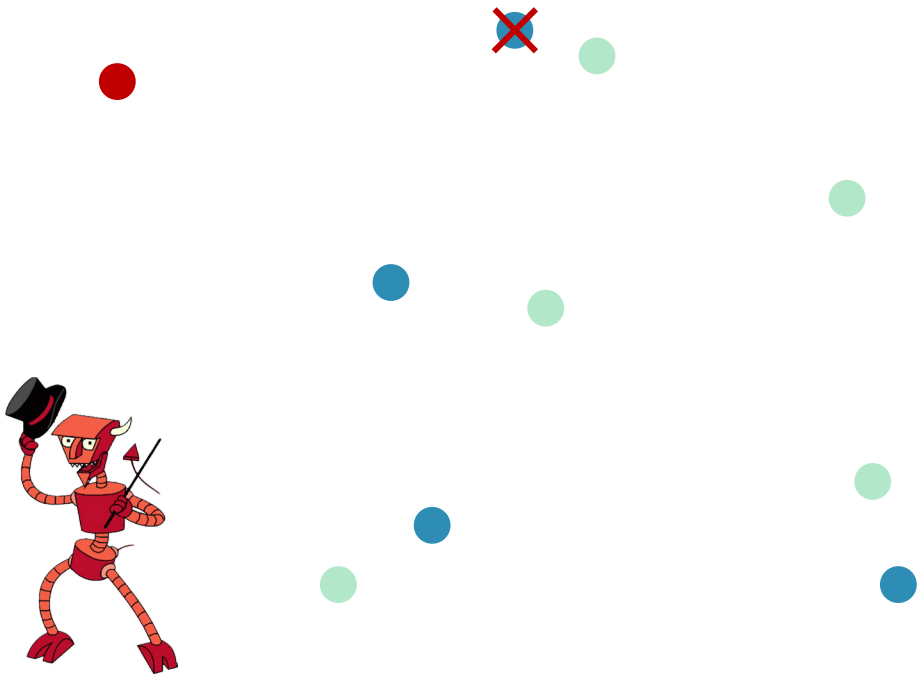
Local perturbation: $X'_1, \dots, X'_n \in \mathbb{R}^d$ s.t. $\frac{1}{n}\sum_{i=1}^{n} \left\| X'_i - X_i \right\| \leq \rho$

Corruption Model

Clean data: $X_1, \dots, X_n \in \mathbb{R}^d$

Local perturbation: $X'_1, \dots, X'_n \in \mathbb{R}^d$ s.t. $\frac{1}{n}\sum_{i=1}^{n} \left\| X'_i - X_i \right\| \leq \rho$

Global perturbation: $\tilde{X}_1, \dots, \tilde{X}_n \in \mathbb{R}^d$ s.t. $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\tilde{X}_i \neq X'_i\}} \leq \epsilon$



Corruption Model

Clean data: $X_1, \dots, X_n \in \mathbb{R}^d$

Local perturbation: $X_1', \dots, X_n' \in \mathbb{R}^d$ s.t. $\frac{1}{n}\sum_{i=1}^n \left\| X_i' - X_i \right\| \le \rho$

Global perturbation: $\tilde{X}_1, \dots, \tilde{X}_n \in \mathbb{R}^d$ s.t. $\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{\tilde{X}_i \neq X_i'\}} \le \epsilon$

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

$$\mu' = \frac{1}{n}\sum_{i=1}^n \delta_{X_i'}$$

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{X}_i}$$

$$\mathrm{W}_1(\mu, \mu') \le \rho$$

$$\mathrm{TV}(\mu', \tilde{\mu}) \le \epsilon$$

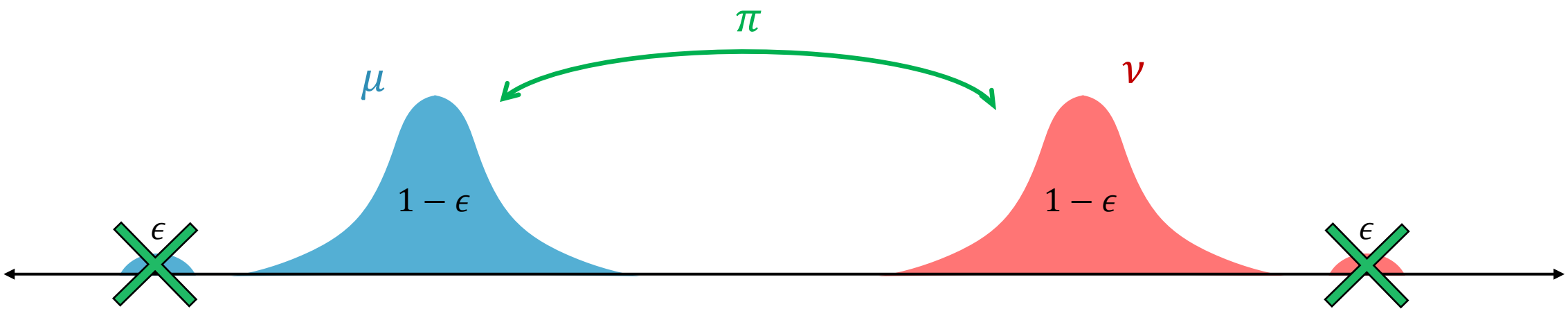
$$\mathrm{W}_1^{\epsilon}(\mu, \tilde{\mu}) \le \rho$$

Partial Optimal Transport

Partial Optimal Transport (Caffarelli-McCann '10)

$$\mathrm{W}_p^{\epsilon}(\mu,\nu) := \inf_{\pi\in\Pi_{\epsilon}(\mu,\nu)} \mathbb{E}_{\pi}[\|X-Y\|^p]^{1/p}$$

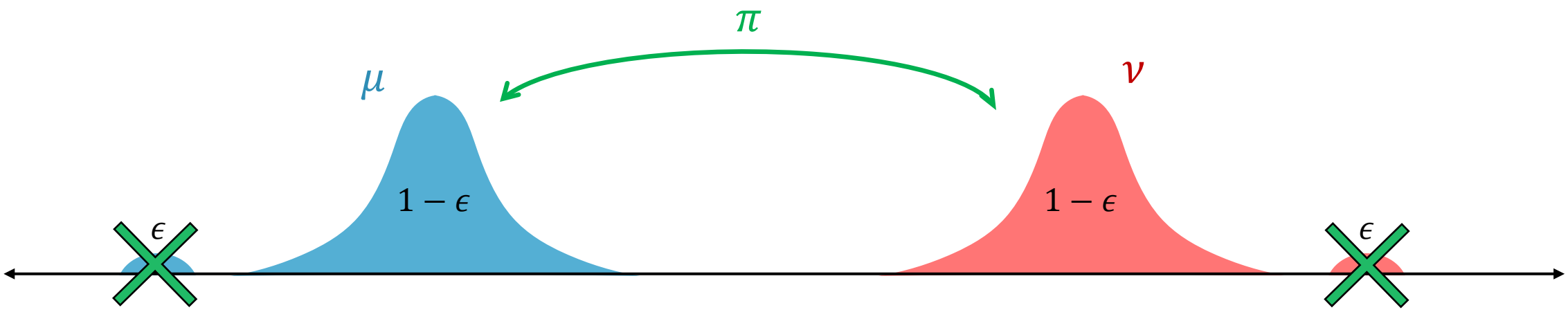
where $\Pi_{\epsilon}(\mu,\nu) := \{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{X}):\ \pi_1 \le \mu, \pi_2 \le \nu, \pi(\mathcal{X}\times\mathcal{X}) = 1-\epsilon\}$



Equivalent Formulations

Theorem (Nietert-Cummings-G '23)

$$\mathrm{W}_p^{\epsilon}(\mu,\nu) := \inf_{\substack{\mu_-,\nu_- \in \mathcal{M}_+(\mathcal{X}):\\ \mu_- \le \mu,\ \nu_- \le \nu \\ \mu_-(\mathcal{X}) = \nu_-(\mathcal{X}) = 1-\epsilon}} \mathrm{W}_p(\mu_-,\nu_-) = \inf_{\substack{\mu_+,\nu_+ \in \mathcal{M}_+(\mathcal{X}):\\ \mu_+ \ge \mu,\ \nu_+ \ge \nu \\ \mu_+(\mathcal{X}) = \nu_+(\mathcal{X}) = 1+\epsilon}} \mathrm{W}_p(\mu_+,\nu_+) = \inf_{\substack{\mu' \in \mathcal{P}(\mathcal{X}):\\ \mathrm{TV}(\mu',\mu) \le \epsilon}} \mathrm{W}_p(\mu',\nu)$$



Approximate Quasi-Metric

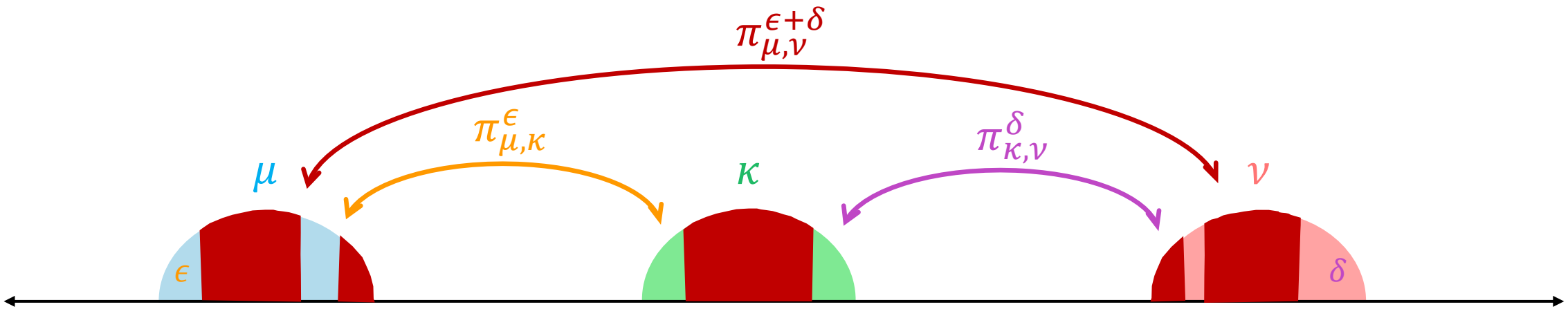
Lemma (Nietert-Cummings-G '23)

For any $\mu, \nu, \kappa \in \mathcal{P}(\mathcal{X})$ and $\epsilon, \delta \in [0,1]$:

a) $\mathrm{W}_p^{\epsilon}(\mu,\nu) \geq 0$ with $\mathrm{W}_p^{\epsilon}(\mu,\nu) = 0$ iff $\mathrm{TV}(\mu,\nu) \leq \epsilon$

b) $\mathrm{W}_p^{\epsilon}(\mu,\nu) = \mathrm{W}_p^{\epsilon}(\nu,\mu)$

c) $\mathrm{W}_p^{\epsilon+\delta}(\mu,\nu) \leq \mathrm{W}_p^{\epsilon}(\mu,\kappa) + \mathrm{W}_p^{\delta}(\kappa,\nu)$



Existing Corruption Models

Global (TV) corruption only ($\epsilon > 0, \rho = 0$):

- Extensive statistics literature dating back to Huber ('69)
- Recent surge of interest from TCS
  [Diakonikolas-Kamath-Kane-Li-Moitra-Stewart '16, '17], …
- Closely connected to heavy-tailed & private estimation
  [Hopkins-Li-Zhang '20], [Hopkins-Kamath-Majid-Narayanan '23], [Asi-Ullman-Zakynthinou '23], …
- **Algorithmic approach:** iterative filtering / spectral reweighting

$$\hat{\mu}_{\mathrm{IF}} = \operatorname*{argmin}_{\nu:\, \mathrm{TV}(\nu, \tilde{\mu}) \le \epsilon} \lambda_{\max}(\Sigma_\nu)$$

Existing Corruption Models

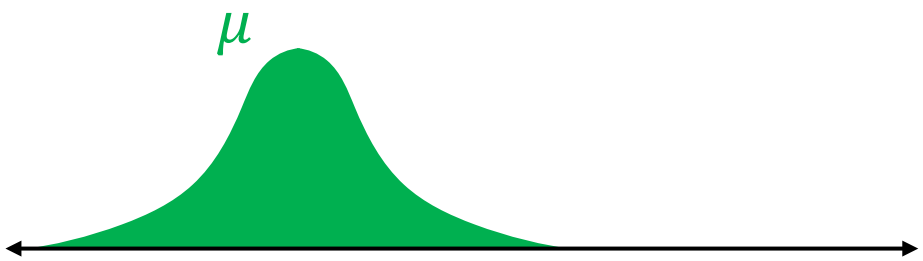
Local (Wasserstein) corruption only ($\rho > 0, \epsilon = 0$):

- Recent interest in ML from adversarial training angle
  [Goodfellow-Shlens-Szegedy '15], [Madry-Makelov-Schmidt-Tsipras-Vladu '18], …
- Explored in OR via distributionally robust stochastic optimization
  [Pflug-Wozabal '07], [Shafiee-Esfahani-Kuhn '15], …
- Examined a bit from robust statistics perspective
  [Liu-Loh '21], [Zhu-Jiao-Steinhardt '22], [Chao-Dobriban '23], …

Robust Estimation

Robust Distribution Estimation

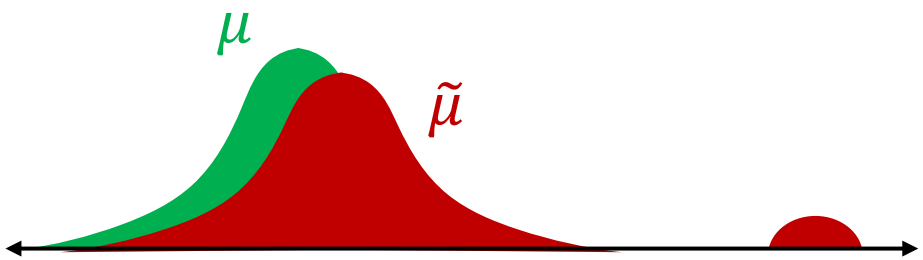
Clean distribution: $\mu \in \mathcal{G}$ (i.e., $\Sigma_\mu \preccurlyeq \mathrm{I}_d$)



Robust Distribution Estimation

Clean distribution: $\mu \in \mathcal{G}_{\mathrm{cov}}$ (i.e., $\Sigma_\mu \preccurlyeq \mathrm{I}_d$)

(ρ, ϵ)**-corruption:** Observe $\tilde{\mu}$ s.t. $\mathrm{W}_1^\epsilon(\mu, \tilde{\mu}) \le \rho$



Robust estimation: Return estimate $\hat{\mu}[\tilde{\mu}] \in \mathcal{P}(\mathcal{X})$ achieving minimax risk

$$R(\rho, \epsilon, \mathcal{G}) \coloneqq \inf_{\hat{\mu}} \sup_{\mu \in \mathcal{G}_{\mathrm{cov}}} \sup_{\substack{\tilde{\mu} \in \mathcal{P}(\mathcal{X}): \\ \mathrm{W}_1^\epsilon(\mu, \tilde{\mu}) \le \epsilon}} \mathrm{W}_1(\hat{\mu}[\tilde{\mu}], \mu)$$

best estimator

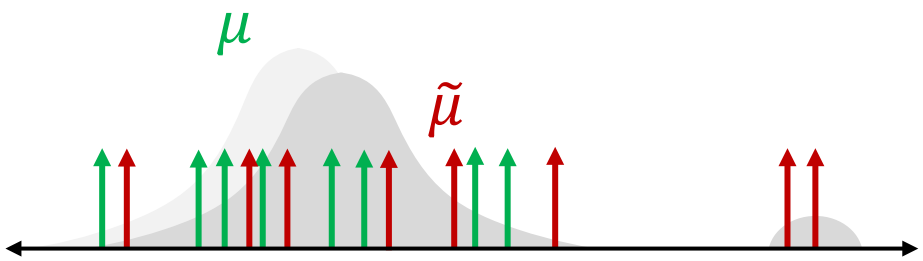
worst-case clean dist.

worst-case corruption

Robust Distribution Estimation

Clean distribution: $\mu \in \mathcal{G}_{\mathrm{cov}}$ (i.e., $\Sigma_\mu \preccurlyeq \mathrm{I}_d$)

(ρ, ϵ)**-corruption:** Observe $\tilde{\mu}$ s.t. $\mathrm{W}_1^\epsilon(\mu, \tilde{\mu}) \le \rho$



Robust estimation: Return estimate $\hat{\mu}[\tilde{\mu}] \in \mathcal{P}(\mathcal{X})$ achieving minimax risk

$$R(\rho, \epsilon, \mathcal{G}) \coloneqq \inf_{\hat{\mu}} \sup_{\mu \in \mathcal{G}_{\mathrm{cov}}} \sup_{\substack{\tilde{\mu} \in \mathcal{P}(\mathcal{X}): \\ \mathrm{W}_1^\epsilon(\mu, \tilde{\mu}) \le \epsilon}} \mathrm{W}_1(\hat{\mu}[\tilde{\mu}], \mu)$$

best estimator

worst-case clean dist.

worst-case corruption

Minimax Risk Characterization

Theorem (Nietert-G.-Shafiee '24)

$$R(\rho, \epsilon, \mathcal{G}_{\mathrm{cov}}) \asymp \sqrt{d\epsilon} + \rho$$

and there exists an efficient algorithm that achieves this risk.

Minimax Risk Characterization

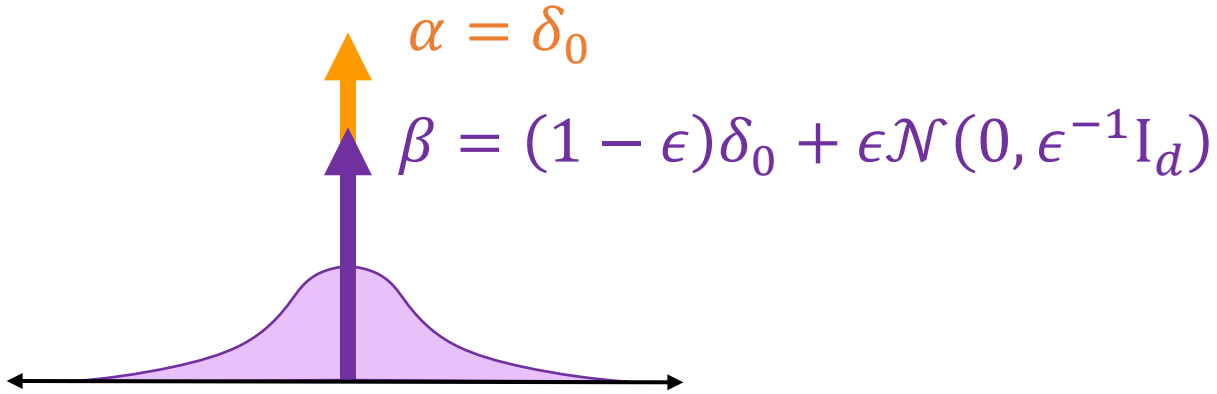
Theorem (Nietert-G.-Shafiee '24)

$$R(\rho, \epsilon, \mathcal{G}_{\mathrm{cov}}) \asymp \sqrt{d\epsilon} + \rho$$

and there exists an efficient algorithm that achieves this risk.

$$\mathrm{TV}(\delta_0, (1-\epsilon)\delta_0 + \epsilon\mathcal{N}(0, \epsilon^{-1}\mathrm{I}_d)\,) = \epsilon$$

$$\mathrm{W}_1(\delta_0, (1-\epsilon)\delta_0 + \epsilon\mathcal{N}(0, \epsilon^{-1}\mathrm{I}_d)\,) = \sqrt{d\epsilon}$$



Minimax Risk Characterization

Theorem (Nietert-G.-Shafiee '24)

$$R(\rho, \epsilon, \mathcal{G}_{\mathrm{cov}}) \asymp \sqrt{d\epsilon} + \rho$$

and there exists an efficient algorithm that achieves this risk.

$$\mathrm{W}_1(\delta_0, \delta_\rho)$$

Minimax Risk Characterization

Theorem (Nietert-G.-Shafiee '24)

$$R(\rho, \epsilon, \mathcal{G}_{\mathrm{cov}}) \asymp \sqrt{d\epsilon} + \rho$$

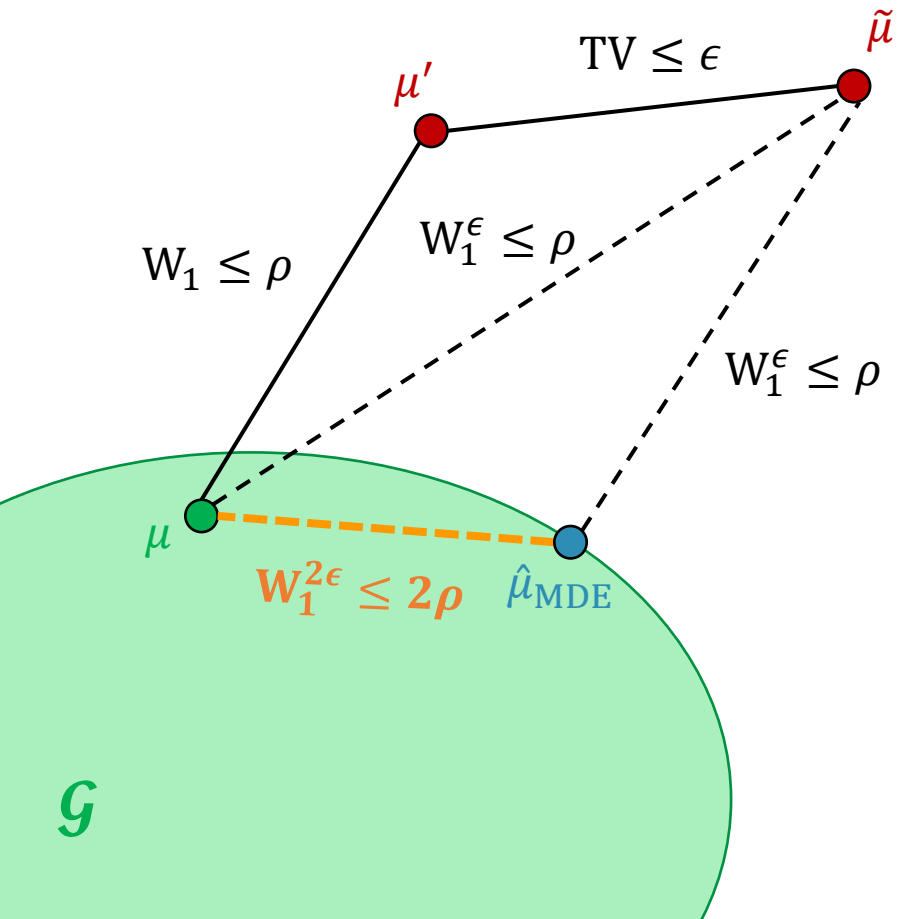
and there exists an efficient algorithm that achieves this risk.

	$\varepsilon = 0$ $\rho > 0$	$\varepsilon > 0$ $\rho = 0$	$\varepsilon > 0$ $\rho > 0$
$\hat{\mu} = \tilde{\mu}$	$\Theta(\rho)$	∞	∞
$\hat{\mu} = \hat{\mu}_{\mathrm{IF}}$	$O(\sqrt{d}\rho)$	$\Theta(\sqrt{d\varepsilon})$	$\Theta(\sqrt{d\varepsilon} + \sqrt{d}\rho)$
our work	$\Theta(\rho)$	$\Theta(\sqrt{d\varepsilon})$	$\Theta(\sqrt{d\varepsilon} + \rho)$

An Ideal Projection Estimator

Minimum distance estimation: Project $\tilde{\mu}$ onto $\mathcal{G}$ wrt W_1^ϵ

$$\hat{\mu}_{\mathrm{MDE}} \in \operatorname*{argmin}_{\nu \in \mathcal{G}} \mathrm{W}_1^\epsilon(\tilde{\mu}, \nu)$$

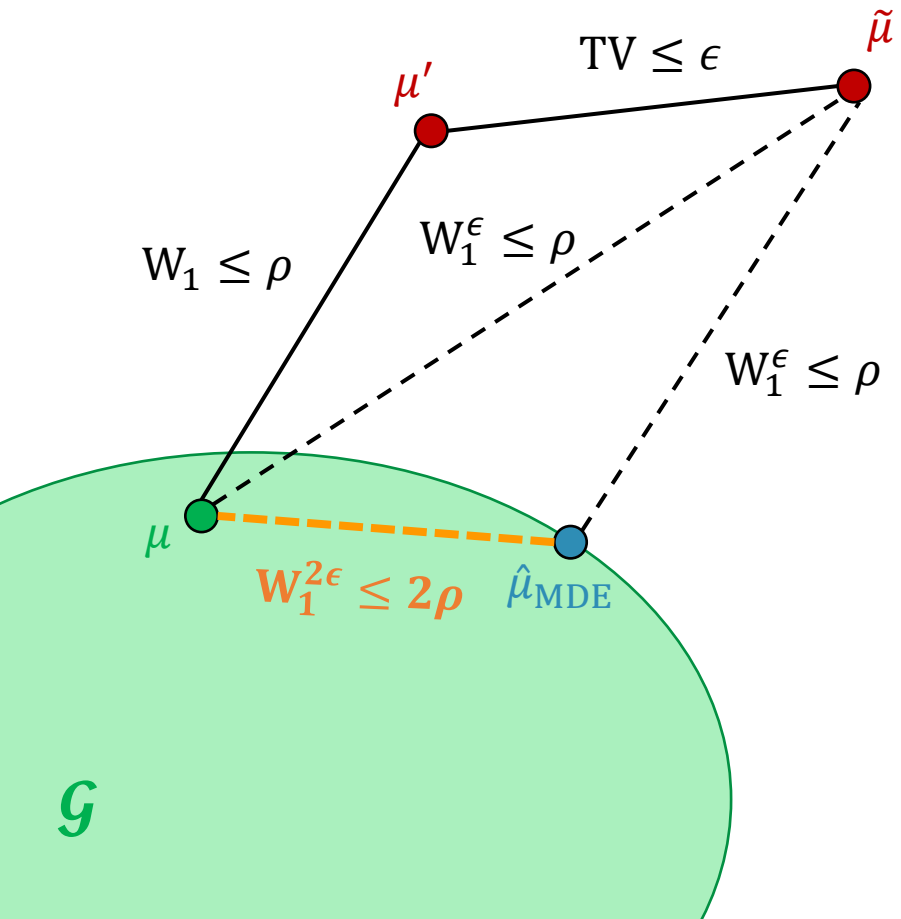


$$\Longrightarrow \quad \mathrm{W}_1(\mu, \hat{\mu}_{\mathrm{MDE}}[\tilde{\mu}]) \le \sup_{\alpha,\beta \in \mathcal{G}:\ \mathrm{W}_1^{2\epsilon}(\alpha,\beta) \le 2\rho} \mathrm{W}_1(\alpha, \beta)$$

$$\lesssim \rho + \underbrace{\mathrm{res}_{\mathrm{W}_1}(\mathcal{G}, \epsilon)}_{\mathcal{G} = \mathcal{G}_{\mathrm{cov}} \Rightarrow \lesssim \rho + \sqrt{d\epsilon}}$$

$$\mathrm{res}_{\mathrm{W}_1}(\mu, \epsilon) = \sup_{\mu' :\ \mu' \le \frac{1}{1-\epsilon}\mu} \mathrm{W}_1(\mu, \mu')$$

An Ideal Projection Estimator



Minimum distance estimation: Project $\tilde{\mu}$ onto $\mathcal{G}$ wrt W_1^ϵ

$$\hat{\mu}_{\mathrm{MDE}} \in \operatorname*{argmin}_{\nu \in \mathcal{G}} \mathrm{W}_1^\epsilon(\tilde{\mu}, \nu)$$

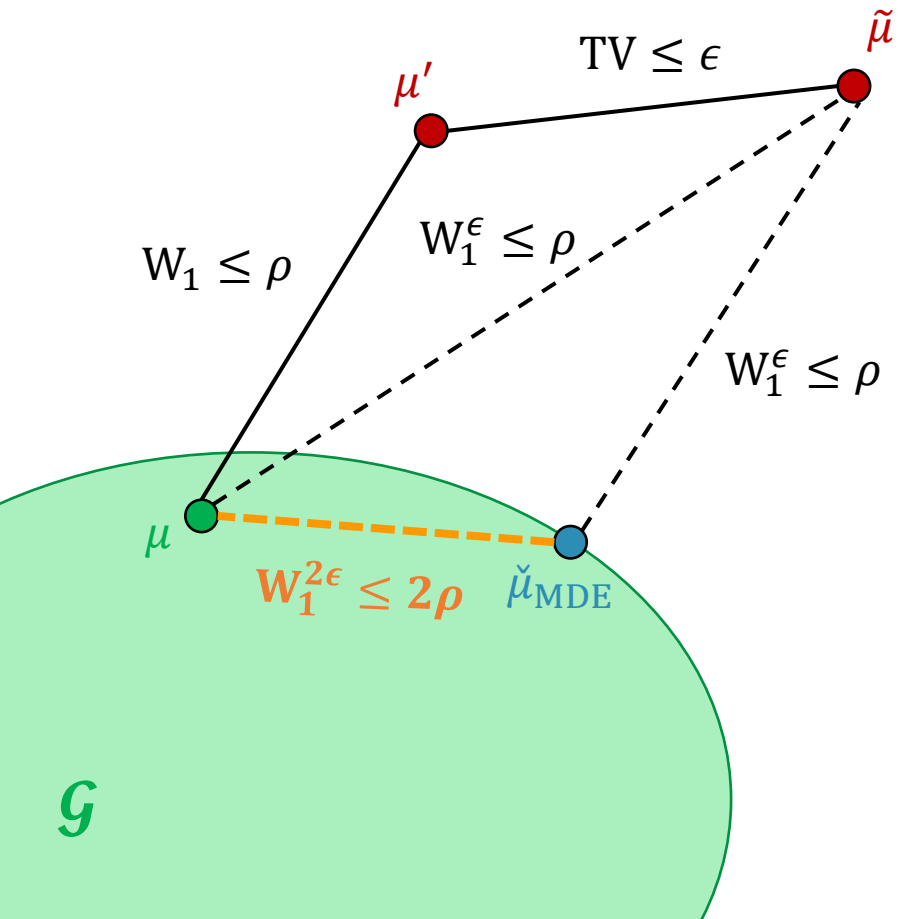
$$\boxed{\mathrm{W}_1(\hat{\mu}_{\mathrm{MDE}}, \tilde{\mu}) \le \rho + \sqrt{d}\boldsymbol{\delta}}$$

Beyond $\mathcal{G}_{cov}$:

- sub-Gaussianity $\quad \boldsymbol{\delta} = \tilde{O}(\epsilon)$
- log-concavity $\quad \boldsymbol{\delta} = \tilde{O}(\epsilon)$
- bdd. qth moments $\quad \boldsymbol{\delta} = \tilde{O}\left(\epsilon^{1-1/q}\right)$

⊘ **Obstacles to efficient alg. w/ formal guarantees**

Another (tractable..?) Ideal Estimator



Use 2-Wasserstein: Project $\tilde{\mu}$ onto $\mathcal{G}$ wrt W_2^ϵ

$$\check{\mu}_{\mathrm{MDE}} \in \operatorname*{argmin}_{\nu \in \mathcal{G}} \mathrm{W}_2^\epsilon(\tilde{\mu}, \nu)$$

- Can show also minimax optimal!

$$\boxed{\mathrm{W}_1(\check{\mu}_{\mathrm{MDE}}, \tilde{\mu}) \le \rho + \sqrt{d\epsilon}}$$

- Key Lemma:

$$\mathrm{W}_1(\mu, \mu') \le \rho \implies \mathrm{W}_2^\epsilon(\mu, \mu') \le \frac{\rho}{\sqrt{\epsilon}}$$

Towards an Efficient Algorithm

Rewrite objective: $\min_{\nu \in \mathcal{G}_{\mathrm{cov}}} \mathrm{W}_2^{\epsilon}(\tilde{\mu}, \nu)$

$$\min_{\nu:\, \mathrm{TV}(\nu,\tilde{\mu}) \le \epsilon} \underbrace{\mathrm{W}_2(\nu, \mathcal{G}_{\mathrm{cov}})}_{} \approx \min_{\nu:\, \mathrm{TV}(\nu,\tilde{\mu}) \lesssim \epsilon} \mathrm{tr}_+(\Sigma_\nu - \mathrm{I}_d)$$

$$= \min_{\eta \in \mathcal{G}_{\mathrm{cov}}} \mathrm{W}_2(\nu, \eta)$$

Tractable approximation: Define $\mathrm{tr}_+(\mathrm{A}) \coloneqq \sum_{i=1}^{d} \max\{\lambda_i(\mathrm{A}), 0\}$

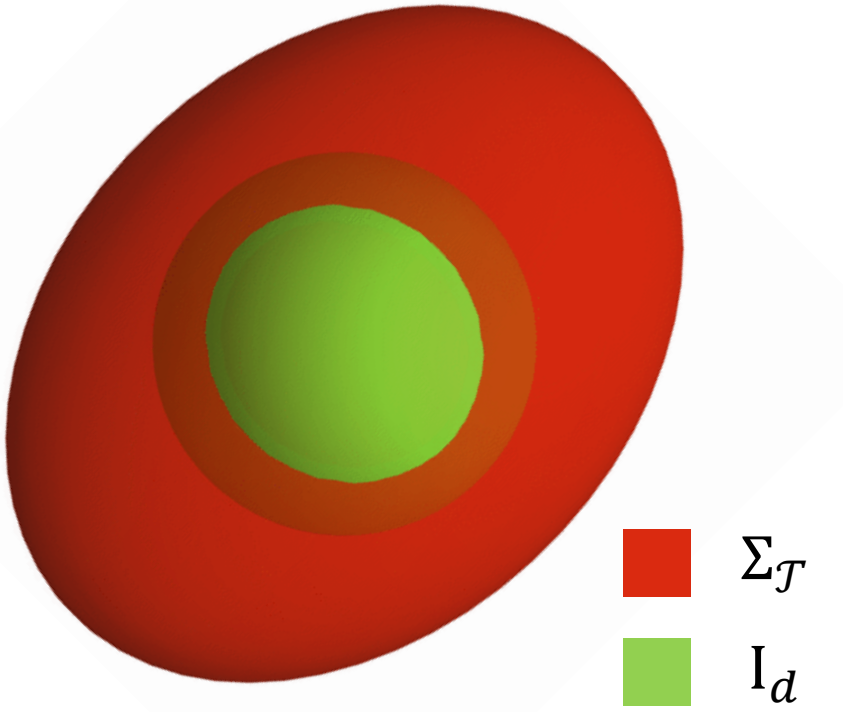
Lemma (Trace norm comparison)

$$\frac{1}{2}\mathrm{tr}_+(\Sigma_\nu - 2\mathrm{I}_d) \le \mathrm{W}_2(\nu, \mathcal{G}_{\mathrm{cov}})^2 \le \mathrm{tr}_+(\Sigma_\nu - \mathrm{I}_d)$$

The W2Project Algorithm

$$\min_{\nu:\, \mathrm{TV}(\nu,\tilde{\mu})\le\epsilon} \mathrm{tr}_+(\Sigma_\nu - \mathrm{I}_d)$$

Initialize: $\mathcal{T} = \mathrm{spt}(\tilde{\mu})$

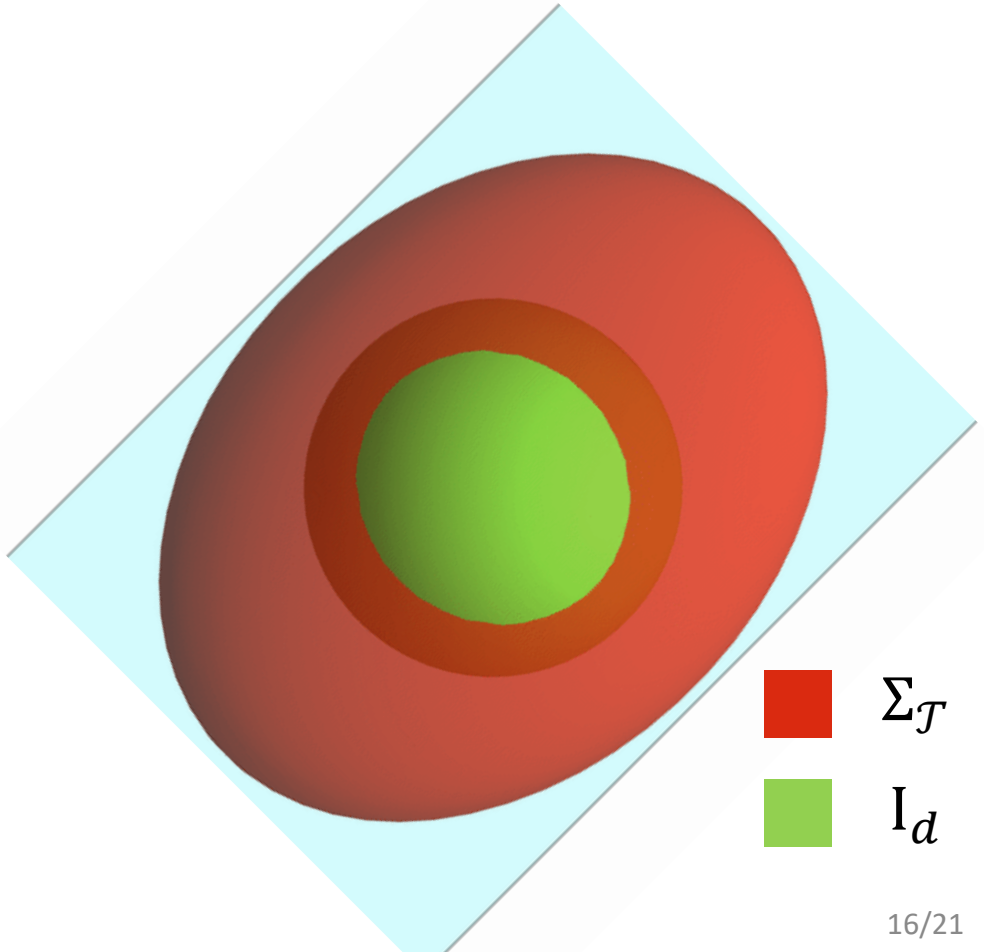


The W2Project Algorithm

$$\min_{\nu:\, \mathrm{TV}(\nu,\tilde{\mu})\le\epsilon} \mathrm{tr}_+(\Sigma_\nu - \mathrm{I}_d)$$

Initialize: $\mathcal{T} = \mathrm{spt}(\tilde{\mu})$

1. $\Pi \leftarrow$ projection onto positive eigenspace of $\Sigma_\nu - \mathrm{I}_d$
2. Set $g(x) = \left\|\Pi\left(x - \mathrm{mean}(\mathcal{T})\right)\right\|$ for $x \in \mathcal{T}$

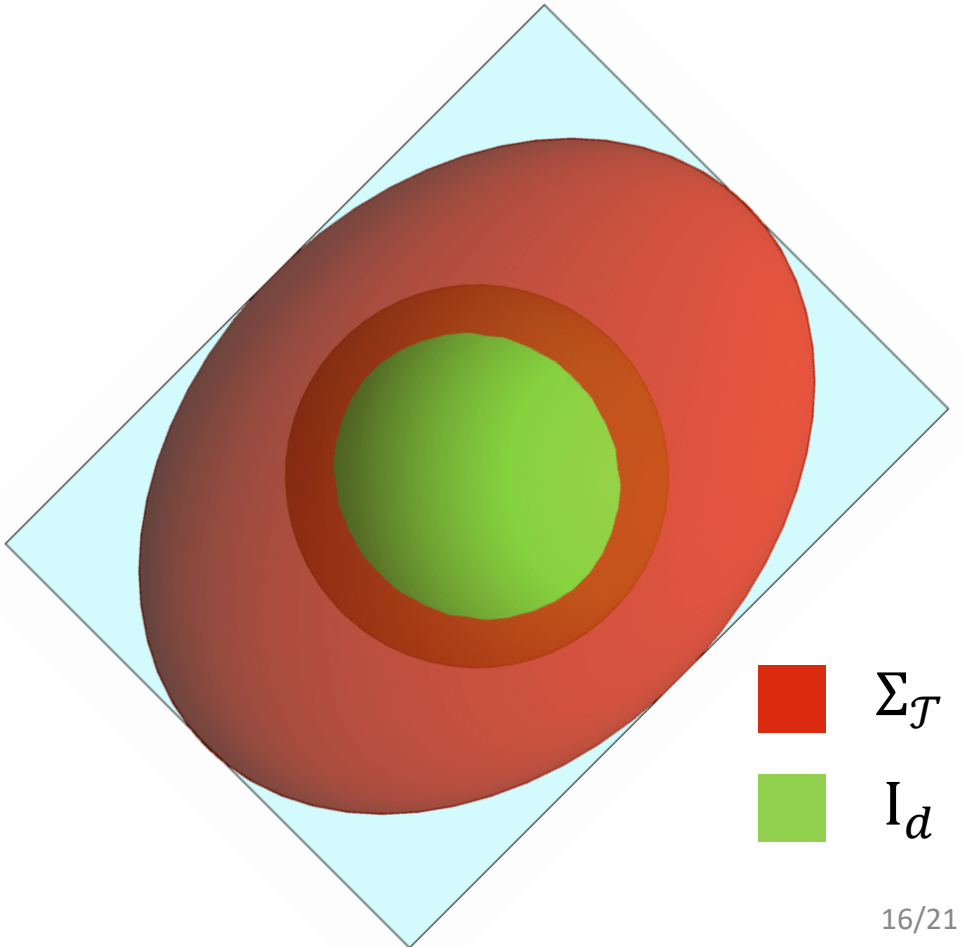


The W2Project Algorithm

$$\min_{\nu:\, \mathrm{TV}(\nu,\tilde{\mu})\le\epsilon} \mathrm{tr}_+(\Sigma_\nu - \mathrm{I}_d)$$

Initialize: $\mathcal{T} = \mathrm{spt}(\tilde{\mu})$

1. $\Pi \leftarrow$ projection onto positive eigenspace of $\Sigma_\nu - \mathrm{I}_d$
2. Set $g(x) = \left\|\Pi\left(x - \mathrm{mean}(\mathcal{T})\right)\right\|$ for $x \in \mathcal{T}$
3. (Randomly) throw out points $x \in \mathcal{T}$ s.t. $g(x)$ is large

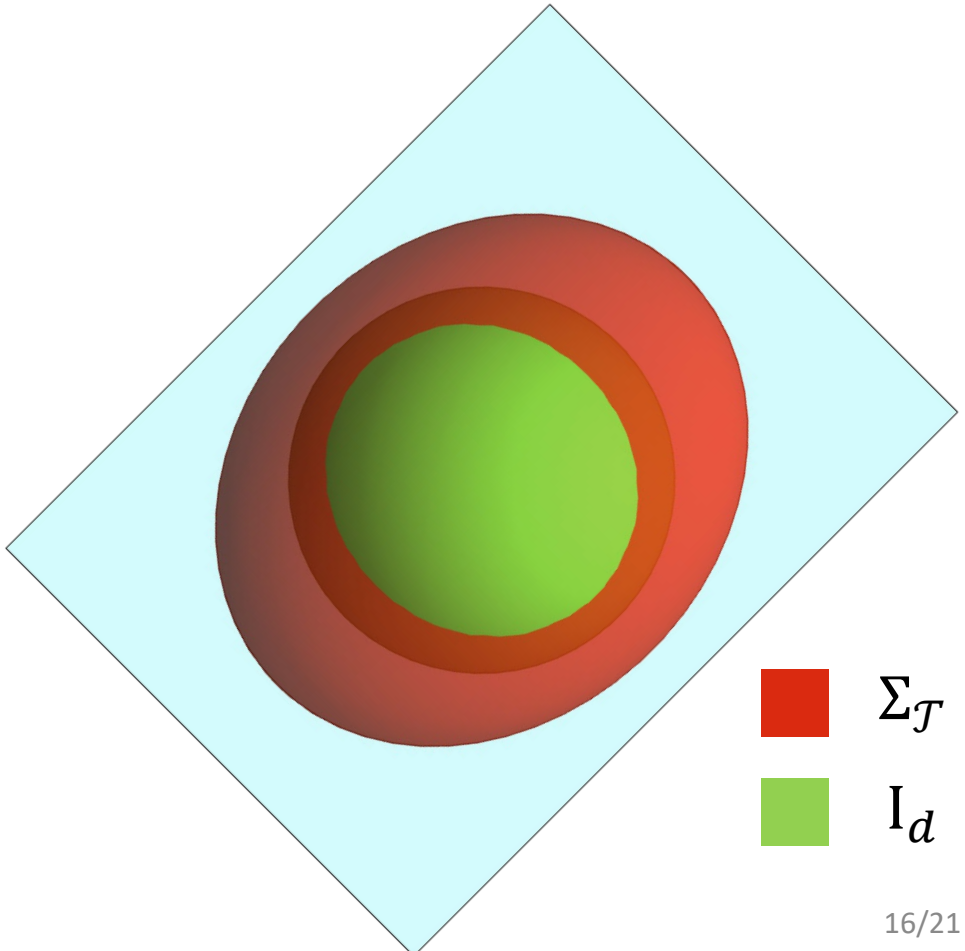


The W2Project Algorithm

$$\min_{\nu:\, \mathrm{TV}(\nu,\tilde{\mu})\le\epsilon} \mathrm{tr}_+(\Sigma_\nu - \mathrm{I}_d)$$

Initialize: $\mathcal{T} = \mathrm{spt}(\tilde{\mu})$

1. $\Pi \leftarrow$ projection onto positive eigenspace of $\Sigma_\nu - \mathrm{I}_d$
2. Set $g(x) = \left\|\Pi\left(x - \mathrm{mean}(\mathcal{T})\right)\right\|$ for $x \in \mathcal{T}$
3. (Randomly) throw out points $x \in \mathcal{T}$ s.t. $g(x)$ is large
4. Repeat from 1



The W2Project Algorithm

$$\min_{\nu:\, \mathrm{TV}(\nu,\tilde{\mu})\le\epsilon} \mathrm{tr}_+(\Sigma_\nu - \mathrm{I}_d)$$

Algorithm 1: W2PROJECT

Input: Contamination levels ε and ρ, uniform discrete measure $\tilde{P}$ supported on set $T \subseteq \mathbb{R}^d$, $|T| = n$

Output: Uniform discrete measure $\hat{P}$

Compute eigen-decomposition $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$, $v_1, \ldots, v_d \in \mathbb{R}^d$ of $\Sigma_T - I_d$

$\Pi \leftarrow \sum_{i:\lambda_i \geq 0} v_i v_i^\top$

if $\mathrm{tr}(\Pi(\Sigma_T - I_d)) < C\rho^2/\varepsilon$ **then return** $\hat{P}_n = \mathrm{Unif}(T)$ `// LHS equals` $\mathrm{tr}(\Sigma_T - I_d)_+$

else

  $g(x) \leftarrow \|\Pi(x - \mu_T)\|_2^2$ for $x \in T$

  Let $L \subseteq T$ be set of $6\varepsilon|T|$ points for which $g(x)$ is largest

  $f(x) \leftarrow g(x)$ for $x \in L$ and $f(x) \leftarrow 0$ otherwise

  Remove each point $x \in T$ from T with probability $f(x)/\max_{x\in T} f(x)$

  Return to Step 1 with new set T

end

Efficient Minimax Optimal Robust Estimator

Theorem (Nietert-G.-Shafiee '24)

Fix any μ with $\Sigma_\mu \preceq \mathrm{I}_d$. Given (ρ, ϵ)-corrupted data $\tilde{\mu}$, the W2Project algorithm produces an estimate $\hat{\mu}$ satisfying:

$$\mathrm{W}_1(\hat{\mu}, \mu) \lesssim \sqrt{d\epsilon} + \rho$$

Efficient Minimax Optimal Robust Estimator

Theorem (Nietert-G.-Shafiee '24)

Fix any μ with $\Sigma_\mu \preceq \mathrm{I}_d$. Given $\boldsymbol{n}$ (ρ, ϵ)-corrupted **samples**, the W2Project algorithm produces an estimate $\hat{\mu}$ satisfying:

$$\mathrm{W}_1(\hat{\mu}, \mu) \lesssim \sqrt{d\epsilon} + \rho + \widetilde{\boldsymbol{O}}\left(\boldsymbol{n}^{-1/d \vee 2}\right)$$

Empirical convergence rate under W_1

with probability at least 0.99.

- Near minimax optimal up to suboptimality of plug-in (logarithmic factors in general)

Dual Formulation

Theorem (Nietert-Cummings-G '23)

$$\mathrm{W}_1^{\epsilon}(\mu,\nu) = \sup_{f\in \mathrm{Lip}_1(\mathcal{X})} \mathbb{E}_{\mu}[f] - \mathbb{E}_{\nu}[f] - 2\epsilon\|f\|_{\infty}$$

Proof sketch:

$$\mathrm{W}_1^{\epsilon}(\mu,\nu) := \inf_{\substack{\mu_+,\nu_+\in\mathcal{M}_+(\mathcal{X}):\\ \mu_+\geq\mu,\ \nu_+\geq\nu \\ \mu_+(\mathcal{X})=\nu_+(\mathcal{X})=1+\epsilon}} \mathrm{W}_1(\mu_+,\nu_+)$$

$$= \inf_{\alpha,\beta\in\mathcal{P}(\mathcal{X})} \sup_{f\in \mathrm{Lip}_1(\mathcal{X})} \int f d\mu - \int f d\nu + \epsilon\left(\int f d\alpha - \int f d\beta\right)$$

Kantorovich duality

$$= \sup_{f\in \mathrm{Lip}_1(\mathcal{X})} \int f d\mu - \int f d\nu + \epsilon\left(\inf_{\alpha\in\mathcal{P}(\mathcal{X})} \int f d\alpha - \sup_{\beta\in\mathcal{P}(\mathcal{X})} \int f d\beta\right)$$

Sion's minimax theorem

$$= \sup_{f\in \mathrm{Lip}_1(\mathcal{X})} \int f d\mu - \int f d\nu - 2\epsilon\|f\|_{\infty}$$

Heuristic Neural Implementation

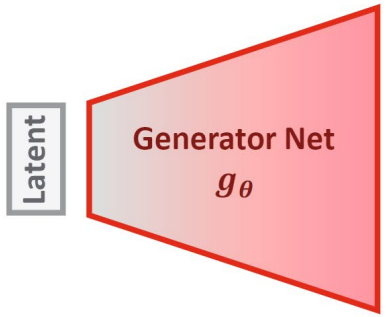
Recall: Minimax optimal (ideal) Wasserstein projection estimator

$$\inf_{\nu} \mathrm{W}_1^{\epsilon}(\tilde{\mu}, \nu)$$

Heuristic Neural Implementation

Recall: Minimax optimal (ideal) Wasserstein projection estimator

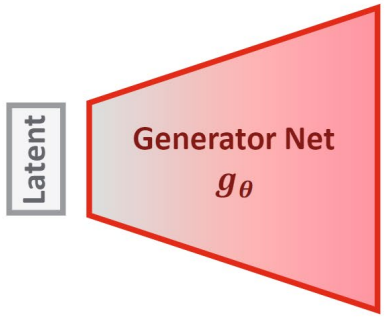
$$\inf_{\theta} \mathrm{W}_1^{\epsilon}(\tilde{\mu}, \nu_\theta)$$



Heuristic Neural Implementation

Recall: Minimax optimal (ideal) Wasserstein projection estimator

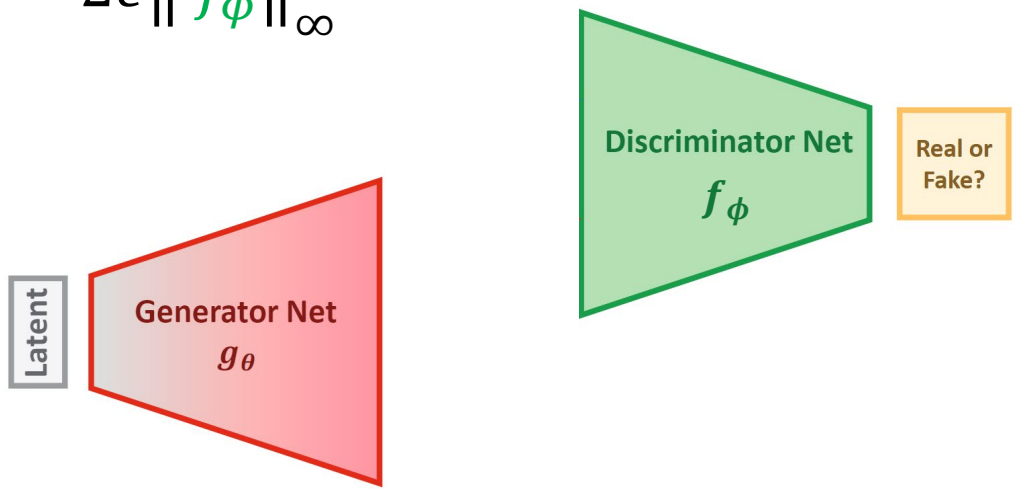
$$\inf_{\theta} \mathrm{W}_1^{\epsilon}(\tilde{\mu}, \nu_\theta) \approx \inf_{\theta} \sup_{f \in \mathrm{Lip}_1} \mathbb{E}_{\mu}[f] - \mathbb{E}_{\mathcal{N}(0,\mathbf{I})}[f \circ g_\theta] - 2\epsilon \|f\|_\infty$$



Heuristic Neural Implementation

Recall: Minimax optimal (ideal) Wasserstein projection estimator

$$\inf_{\theta} \mathrm{W}_1^{\epsilon}(\tilde{\mu}, \nu_\theta) \approx \inf_{\theta} \sup_{\phi:\, f_\phi \in \mathrm{Lip}_1} \mathbb{E}_{\mu}[f_\phi] - \mathbb{E}_{\mathcal{N}(0,\mathbf{I})}[f_\phi \circ g_\theta] - 2\epsilon \| f_\phi \|_\infty$$

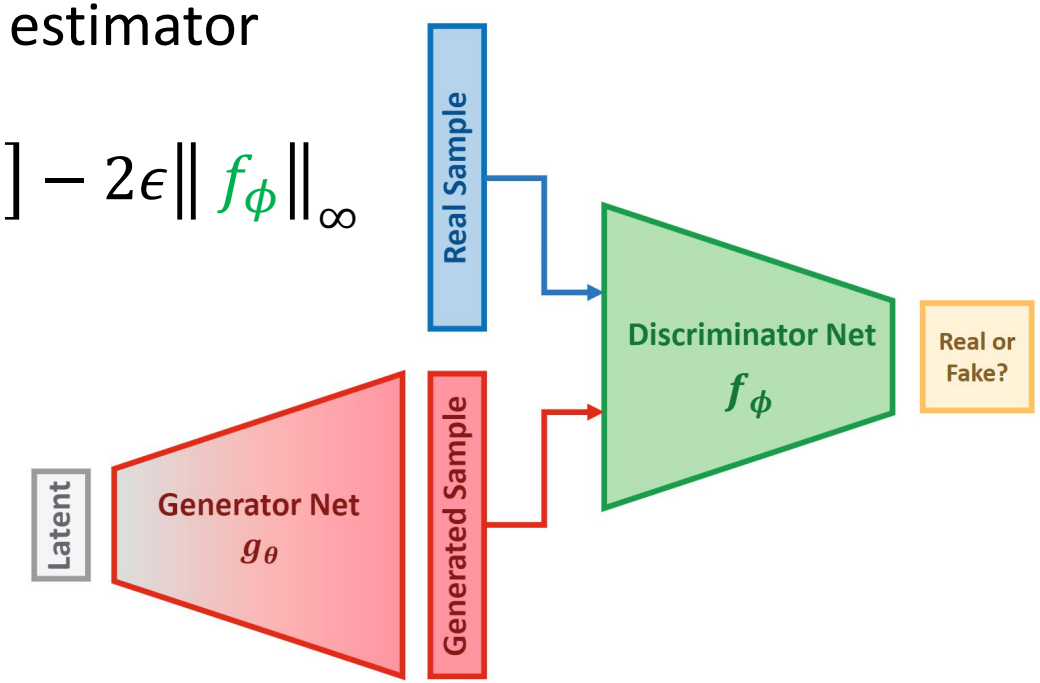


Heuristic Neural Implementation

Recall: Minimax optimal (ideal) Wasserstein projection estimator

$$\inf_{\theta} W_1^{\epsilon}(\tilde{\mu}, \nu_\theta) \approx \inf_{\theta} \sup_{\phi:\, f_\phi \in \mathrm{Lip}_1} \mathbb{E}_{\mu}[f_\phi] - \mathbb{E}_{\mathcal{N}(0,\mathbf{I})}[f_\phi \circ g_\theta] - 2\epsilon \| f_\phi \|_\infty$$

$\Longrightarrow$ Matches W-GAN form up to $\|\cdot\|_\infty$ penalty

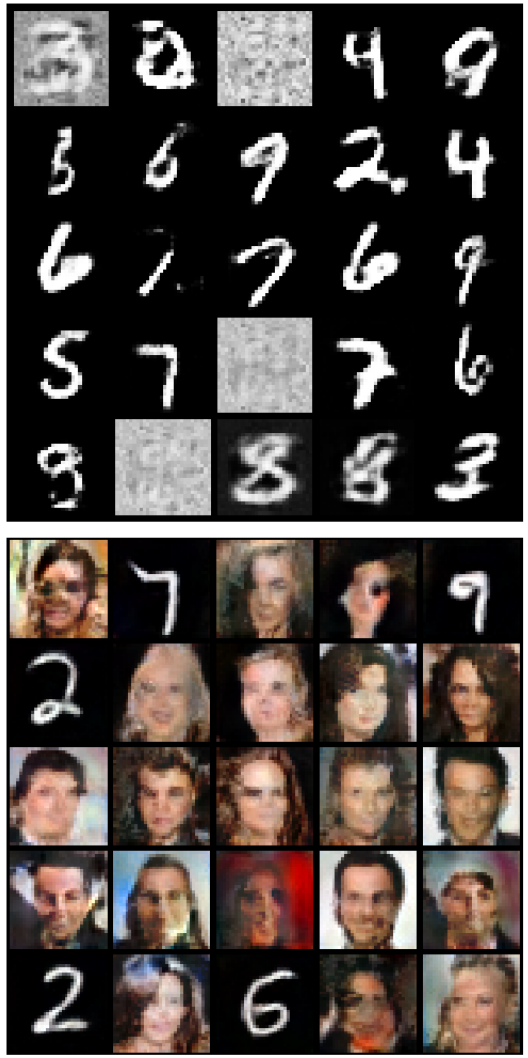


Implementation: 2 lines of code

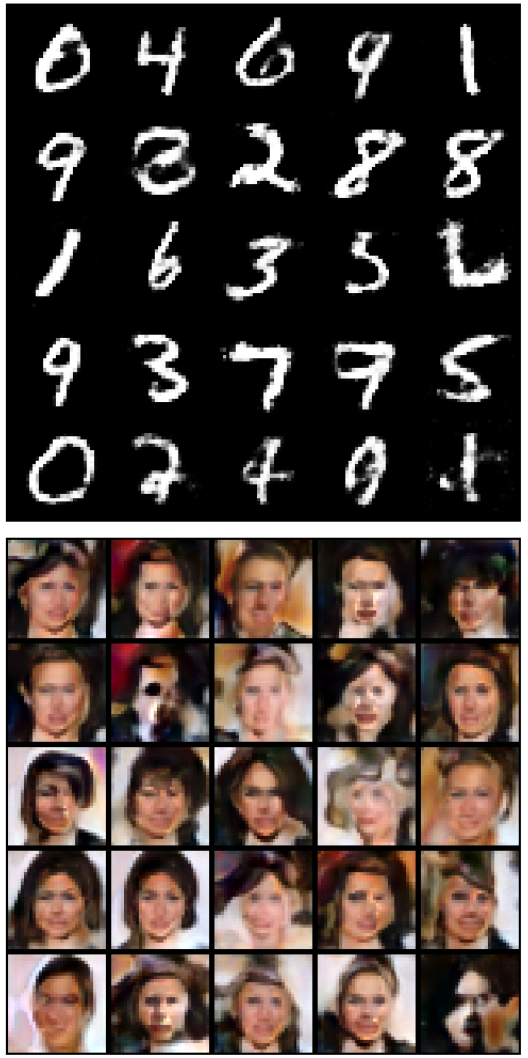
```
score = f_data.mean() - f_generated.mean() # old
score = f_data.mean() - (1-eps)*f_generated.mean() - eps*f_data.max() # new
```

Empirical Results with TV Corruptions

Standard W-GAN



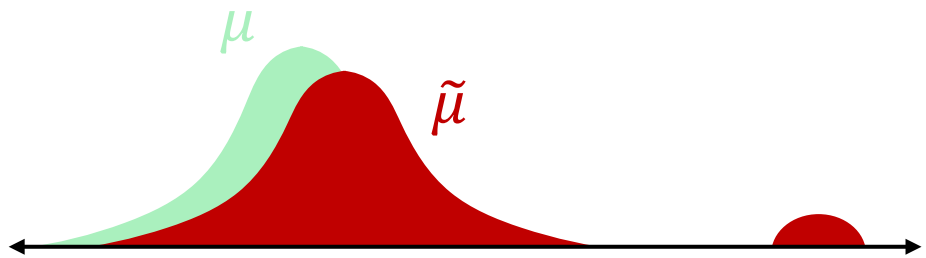
Robust W-GAN



Summary

Robust distribution estimation

- Local & global corruption model
- Under W_1 loss



Wasserstein projection estimator:

- Partial OT is compatible with filtering outliers
- W_1^ϵ or W_2^ϵ POT projection estimator is minimax optimal
- Efficient implementation via W2Project spectral algorithm

Thank you!

Neural implementation: Duality theory for POT + parametrization by neural nets

[A] S. Nietert, R. Cummings, and Z. Goldfeld, "Robust estimation under the Wasserstein distance", ArXiv: 2302.01237

[B] S. Nietert, Z. Goldfeld, and S. Shafieezadeh-Abadeh, "Robust estimation under local and global adversarial corruptions", COLT 2024, ArXiv:2406.06509