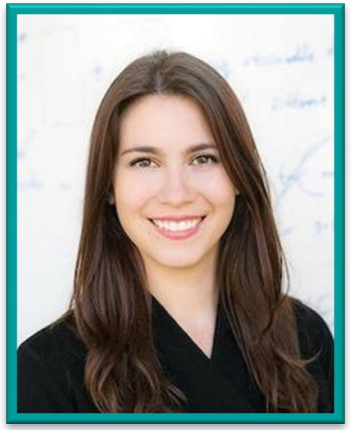
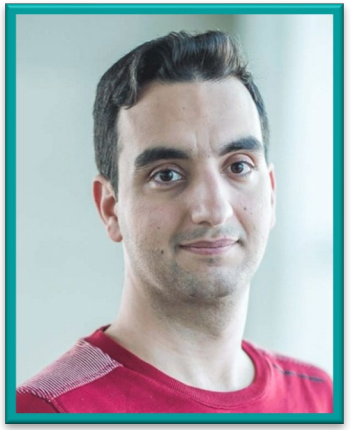
Robust Distribution Estimation via Partial Optimal Transport

Ziv Goldfeld
Cornell University

Sloan Nietert

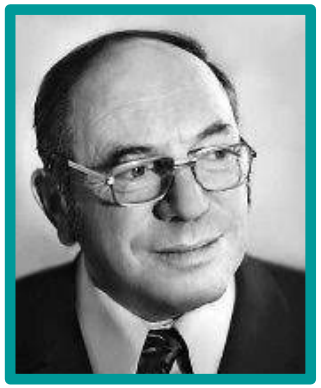


Rachel Cummings



Soroosh Shafiee

Optimal Transport



Kantorovich (1942)

Kantorovich Optimal Transport

$$\mathrm{OT}_c(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \iint c\,d\pi = \sup_{\substack{(f,g)\in L^1(\mu)\times L^1(\nu):\\ f(x)+g(y)\le c(x,y)}} \int f\,d\mu + \int g\,d\nu$$

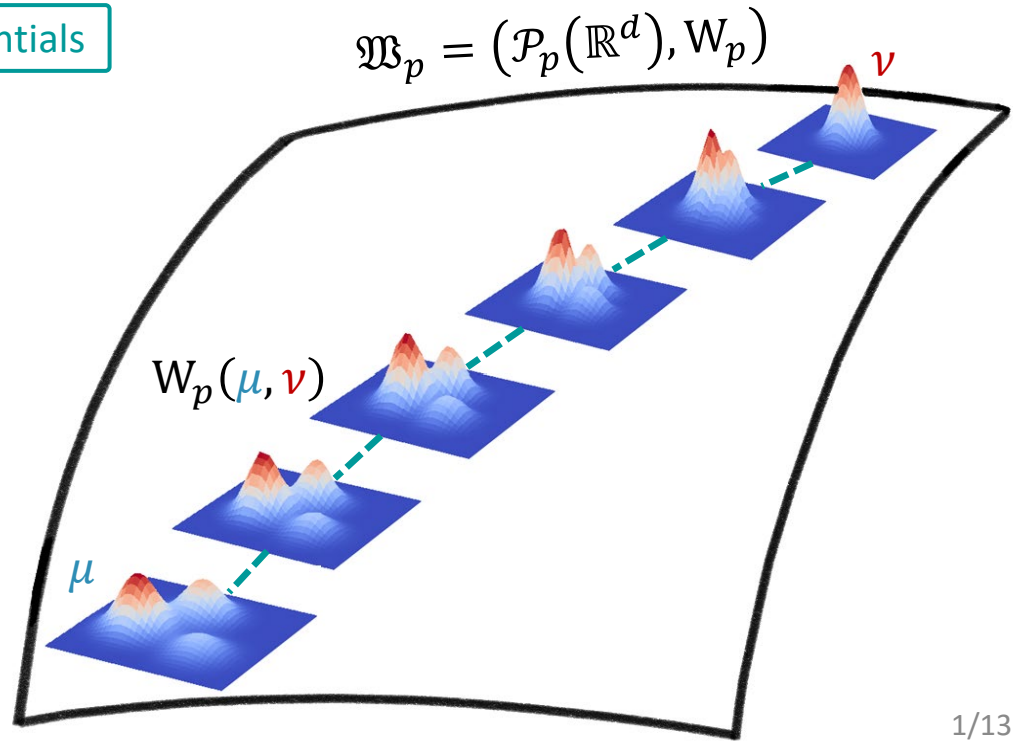
Transport plan: $\pi(x,y) = \mu(x)\pi(y|x)$

Dual potentials

p-Wasserstein metric: $\mathrm{W}_p(\mu,\nu) := \left(\mathrm{OT}_{\|\cdot\|^p}(\mu,\nu)\right)^{1/p}$

- Geodesic curves
- Barycenters
- Gradient flows

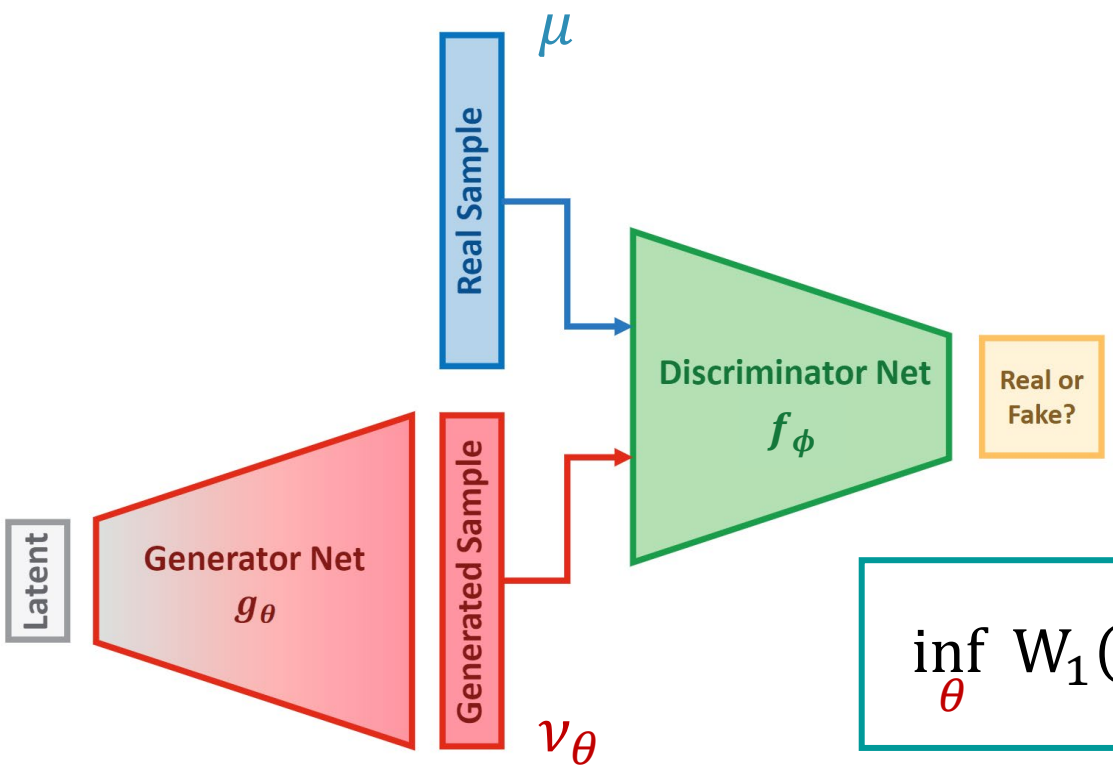
⋮

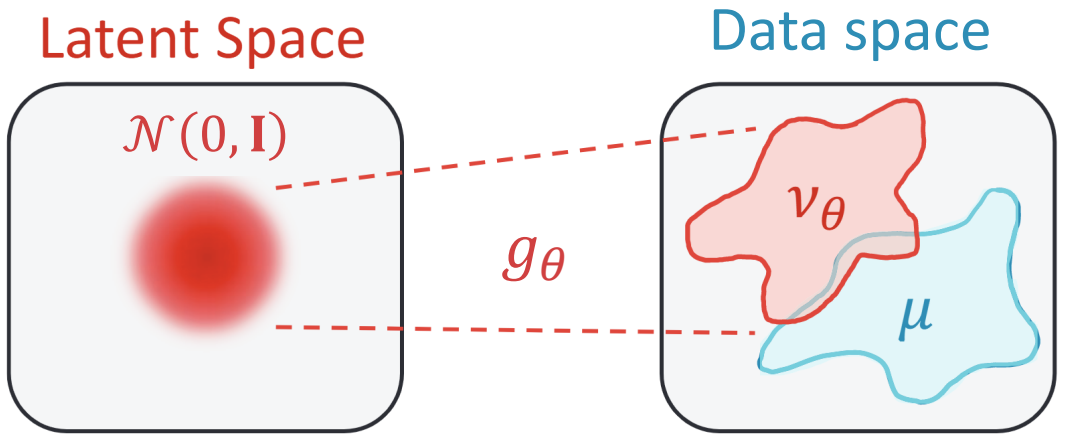


From Wasserstein MDE to Generative Modeling

Minimum distance estimation: Estimate $\mu \in \mathcal{P}_1$ over $\mathcal{G} \subseteq \mathcal{P}_1$ via: $\hat{\mu} \in \operatorname{argmin}_{\nu \in \mathcal{G}} \mathrm{W}_1(\mu, \nu)$

Generative adversarial network: $\mathcal{G} = \{\nu_\theta\}_{\theta \in \Theta}$, $\nu_\theta = (g_\theta)_\sharp \mathcal{N}(0, \mathbf{I})$, w/ g_θ DNN





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

OT duality

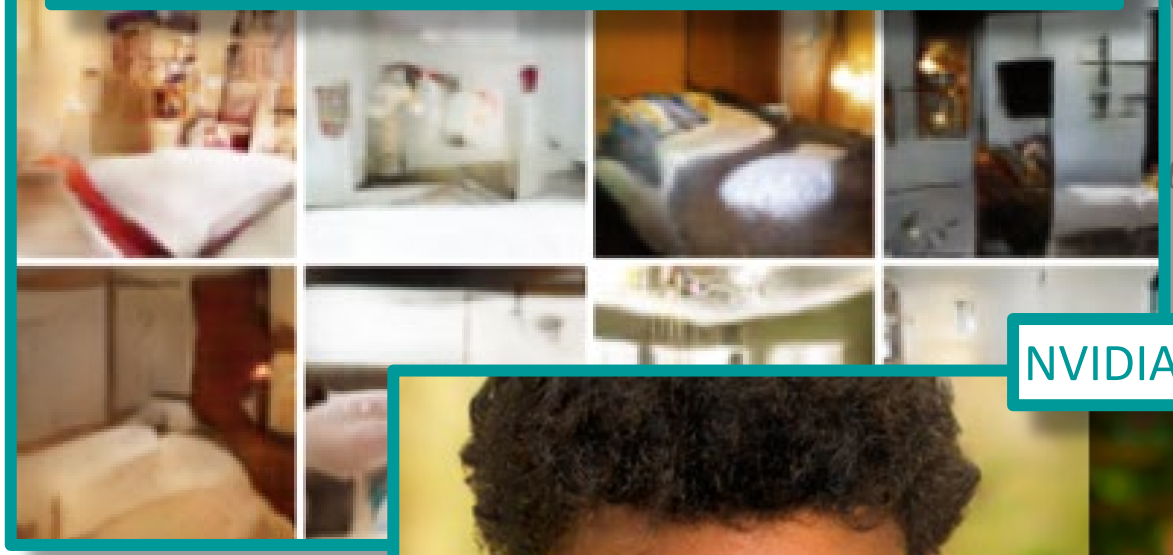
Train w/ alternating optimization

From Wasserstein MDE to Generative Modeling

$\in \mathcal{P}_1$ over $\mathcal{G} \subseteq \mathcal{P}_1$ via: $\hat{\mu} \in \underset{\nu \in \mathcal{G}}{\operatorname{argmin}} \mathrm{W}_1(\mu, \nu)$

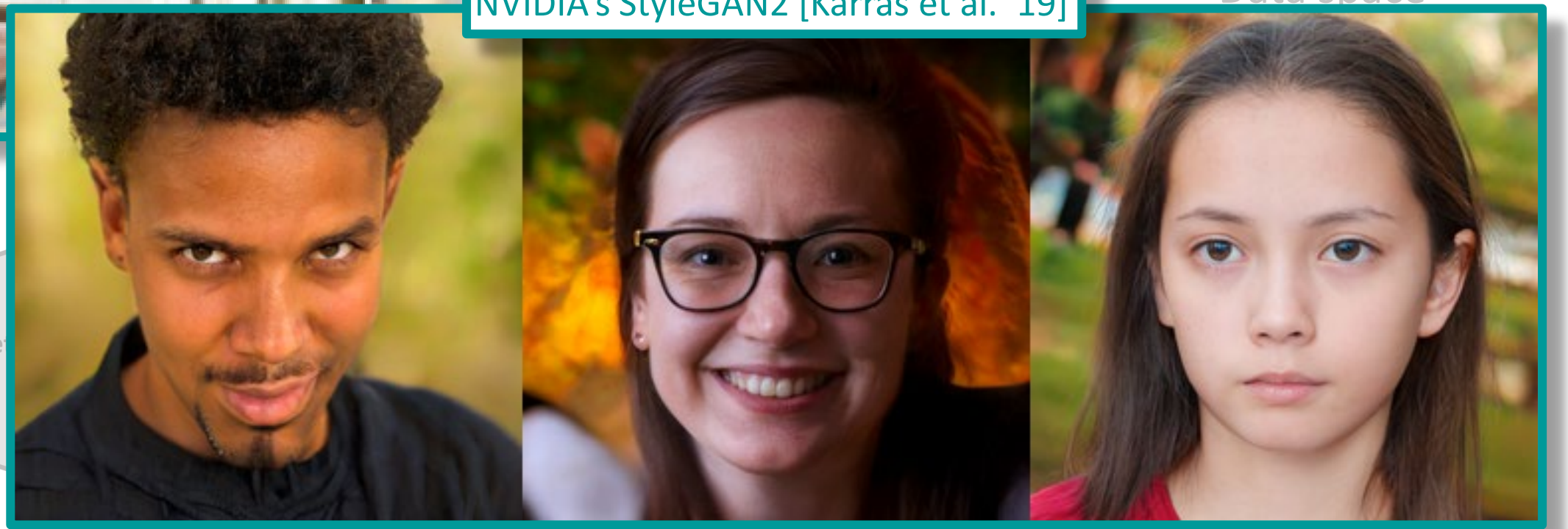
$\theta \in \Theta$, $\nu_\theta = (g_\theta)_\sharp \mathcal{N}(0, \mathbf{I})$, w/ g_θ DNN

Wasserstein GAN [Arjovsky-Chintala-Bottou '17]



NVIDIA's StyleGAN2 [Karras et al. '19]

Data space



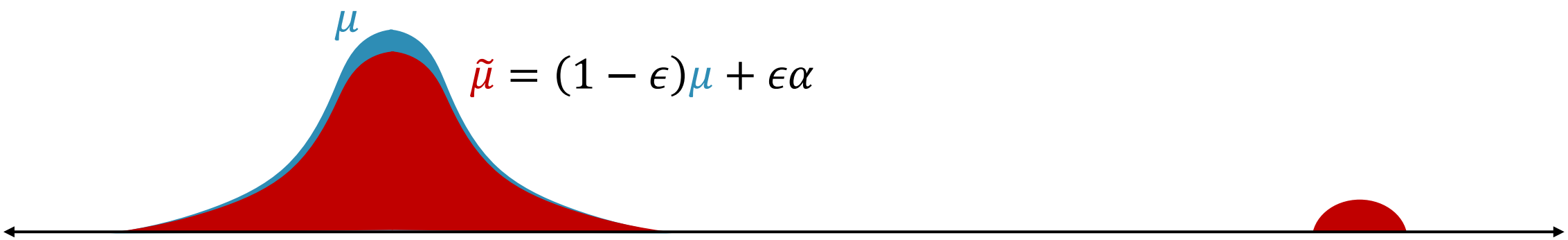
Latent

Generator Ne

g_θ

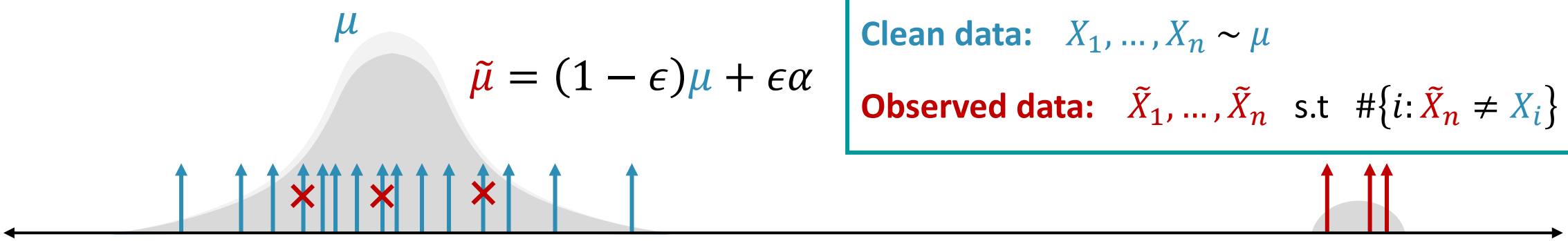
Data Poisoning via Global Corruptions

Real-world data can be messy: Global (possibly adversarial) contamination by outliers



Data Poisoning via Global Corruptions

Real-world data can be messy: Global (possibly adversarial) contamination by outliers



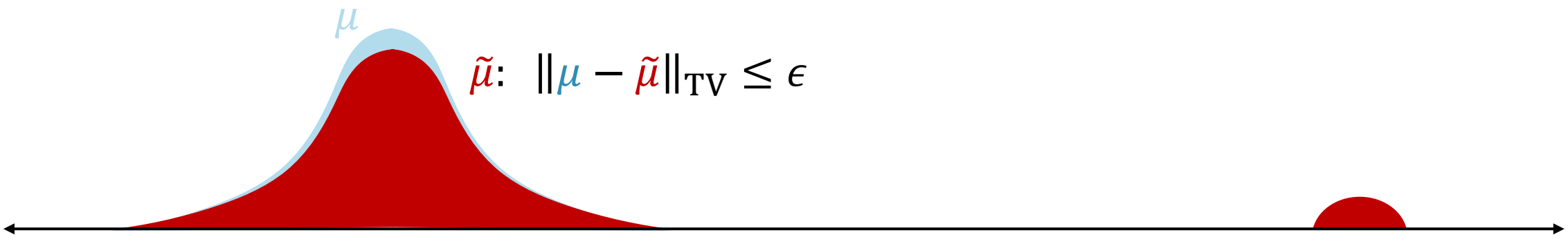
Finite-sample setting

Clean data: $X_1, \dots, X_n \sim \mu$

Observed data: $\tilde{X}_1, \dots, \tilde{X}_n$ s.t $\#\{i: \tilde{X}_n \neq X_i\} \leq n\epsilon$

Data Poisoning via Global Corruptions

Real-world data can be messy: Global (possibly adversarial) contamination by outliers

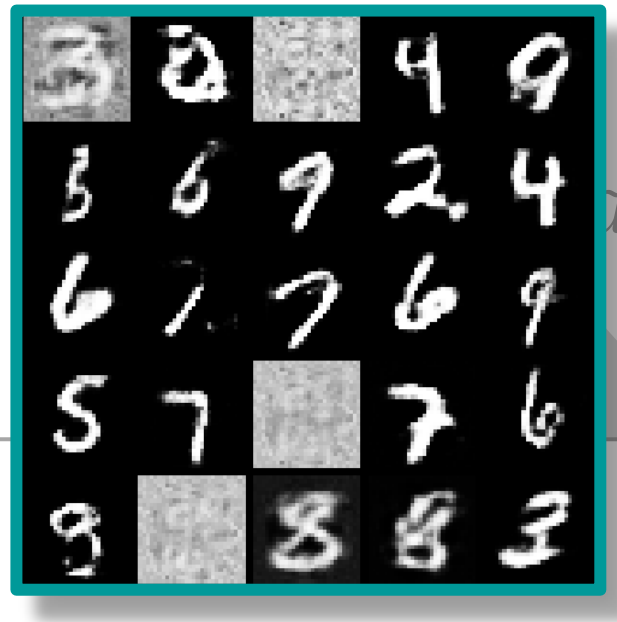


Wasserstein distance is sensitive to outliers: Due to the strict marginal constraints

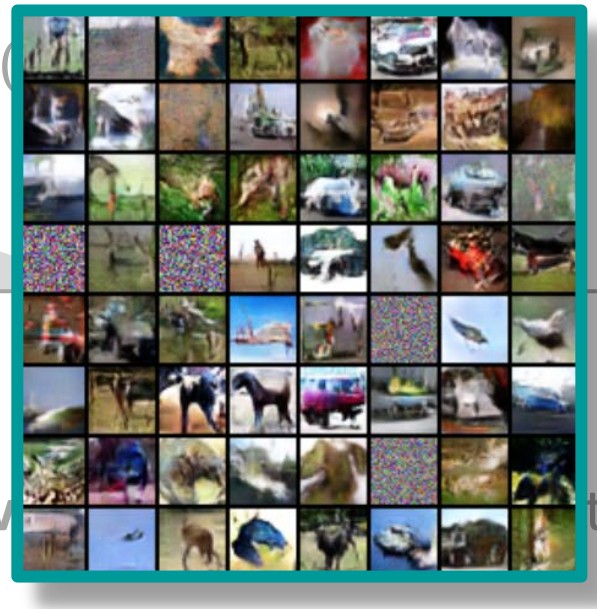
$$\mathrm{W}_p(\mu, (1-\epsilon)\mu + \epsilon\delta_x) \xrightarrow[\|x\|\to\infty]{} \infty$$

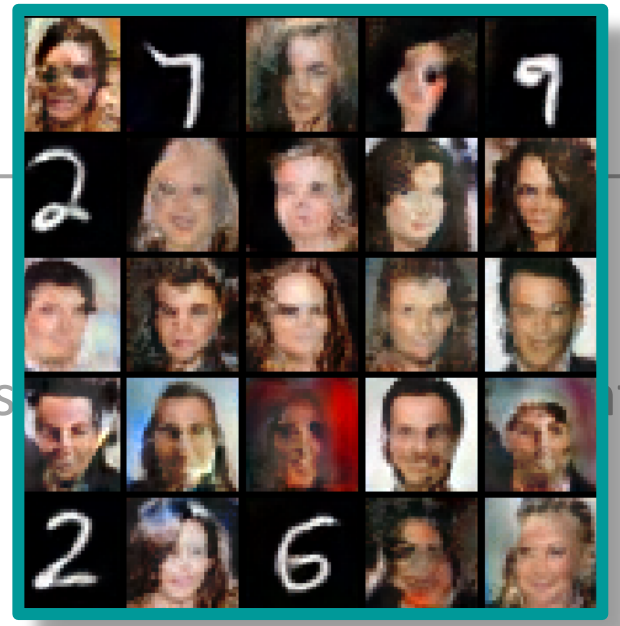
Data Poisoning via Global Corruptions

Real-world data can be messy: Global (possibly adversarial) contamination but outliers



$\tilde{\mu} = ($





⊘ **Wasserstein distance is sensitiv** the s ts

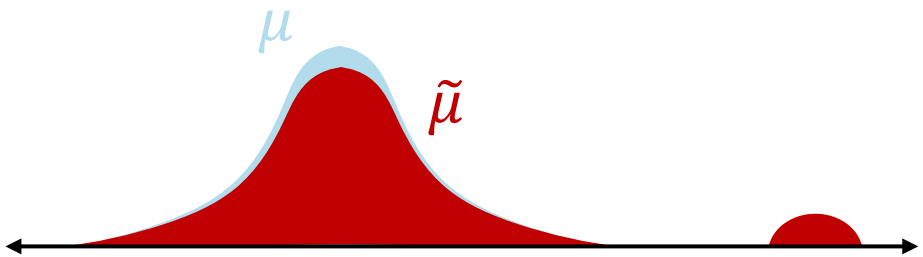
$$\mathrm{W}_p(\mu, (1-\epsilon)\mu + \epsilon\delta_x) \xrightarrow[\|x\|\to\infty]{} \infty$$

Question: How to learn well-performing models from contaminated data?

Robust Distribution Estimation under W_p

Clean distribution: $\mu \in \mathcal{G}$ (encoding, e.g., tail bounds)

Contamination model: Observe $\tilde{\mu}$ s.t. $\|\mu - \tilde{\mu}\|_{\mathrm{TV}} \le \epsilon$



Robust estimation: Given ϵ-corruption $\tilde{\mu}$, return est. $\mathrm{T}(\tilde{\mu}) \in \mathcal{P}(\mathcal{X})$ achieving minimax risk

- **Population:** $$R_p(\epsilon, \mathcal{G}) := \inf_{\mathrm{T}} \sup_{\mu \in \mathcal{G}} \sup_{\substack{\tilde{\mu} \in \mathcal{P}(\mathcal{X}): \\ \|\mu - \tilde{\mu}\|_{\mathrm{TV}} \le \epsilon}} \mathrm{W}_p(\mathrm{T}(\tilde{\mu}), \mu)$$

- **Finite-sample:** $$R_{n,p}(\epsilon, \mathcal{G}) := \inf_{\mathrm{T}} \sup_{\mu \in \mathcal{G}} \sup_{\tilde{\mathbb{P}}_n \in \mathcal{A}_n(\mu, \epsilon)} \mathbb{E}_{\tilde{\mathbb{P}}_n}\left[\mathrm{W}_p\left(\mathrm{T}\left(\tilde{X}_1, \ldots, \tilde{X}_n\right), \mu\right)\right]$$

all corrupted data (joint) dist. under TV ϵ-corruption model

⊛ **[Donoho-Liu '88]:** TV MDE is near-minimax optimal in population limit; doesn't work in finite-sample

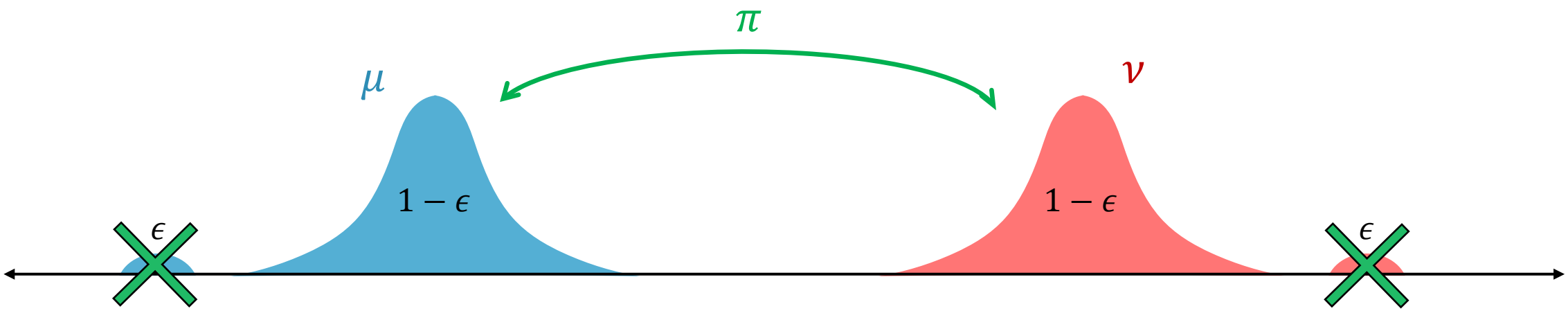
⊛ **[Zhu-Jiao-Steinhardt '22]:** Finite-sample mean & covariance robust estimation; not the distribution itself

Partial OT as an Outlier-Robust Distance

Partial Optimal Transport (Caffarelli-McCann '10)

$$\mathrm{W}_p^{\epsilon}(\mu,\nu) := \left[\inf_{\pi\in\Pi_{\epsilon}(\mu,\nu)} \iint \|x-y\|^p \, d\pi(x,y)\right]^{1/p}$$

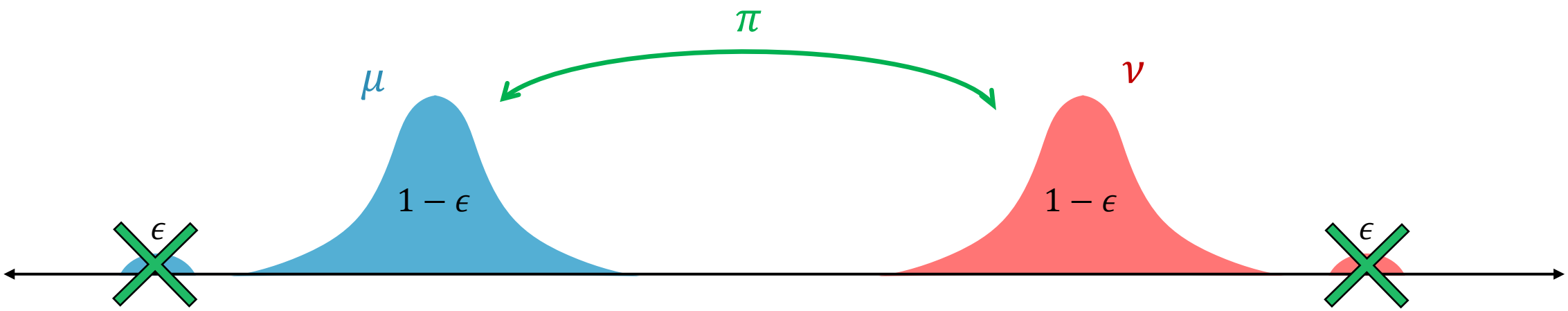
where $\Pi_{\epsilon}(\mu,\nu) := \{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{X}) \colon \pi_1 \le \mu, \pi_2 \le \nu, \pi(\mathcal{X}\times\mathcal{X}) = 1-\epsilon\}$



Equivalent Formulations

Theorem (Nietert-Cummings-G '23)

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



Approximate Quasi-Metric

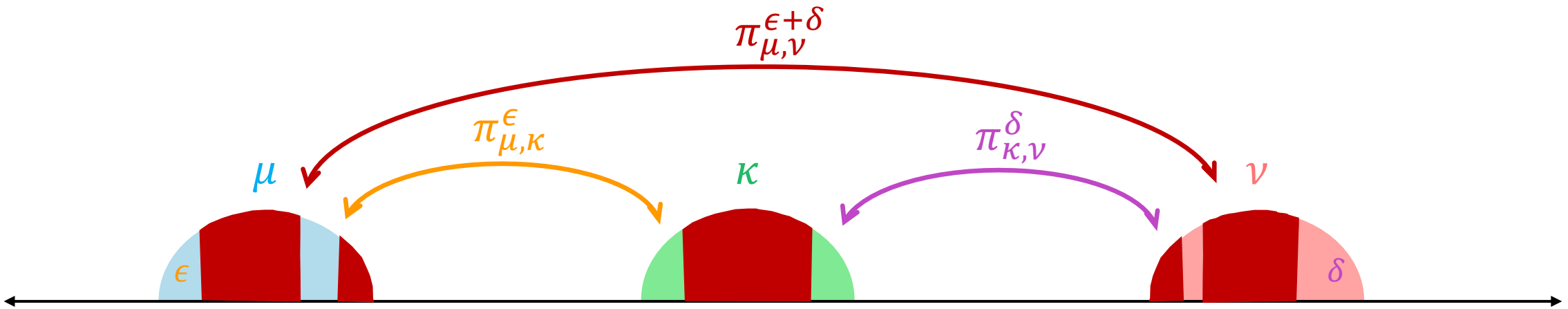
Theorem (Nietert-Cummings-G '23)

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 $\|\mu - \nu\|_{\mathrm{TV}} \leq \epsilon$

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

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



Dual Formulation

Theorem (Nietert-Cummings-G '23)

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

$f^c(y) = \inf_{x\in\mathcal{X}}[c(x,y) - f(x)]$

Proof sketch:

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

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

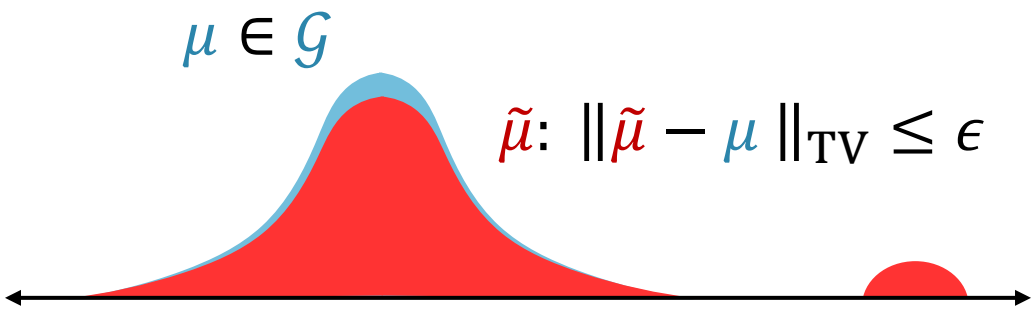
Kantorovich duality

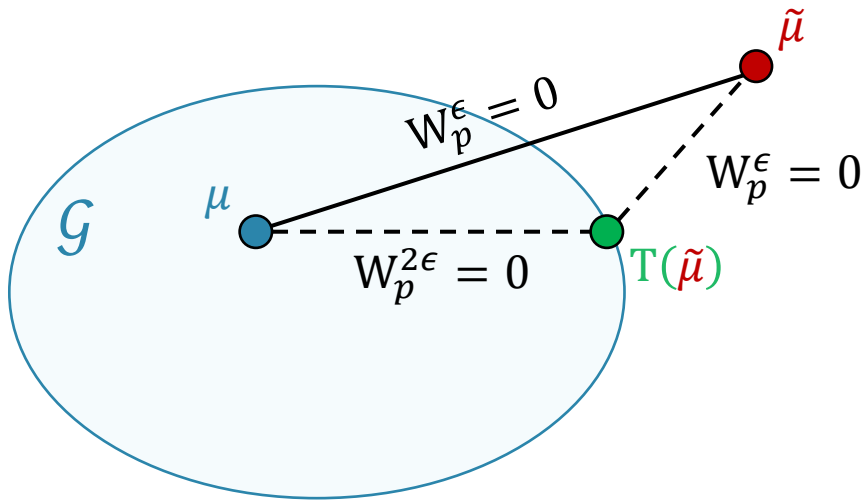
$$= \sup_{f\in\mathrm{Lip}_1(\mathcal{X})} \int f d\mu - \int f d\nu + \epsilon\left(\inf_{\alpha\in\mathcal{P}(\mathcal{X})} \int f d\alpha - \sup_{\beta\in\mathcal{P}(\mathcal{X})} \int f d\beta\right)$$

Sion's minimax theorem

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

Back to Robust Estimation: Population





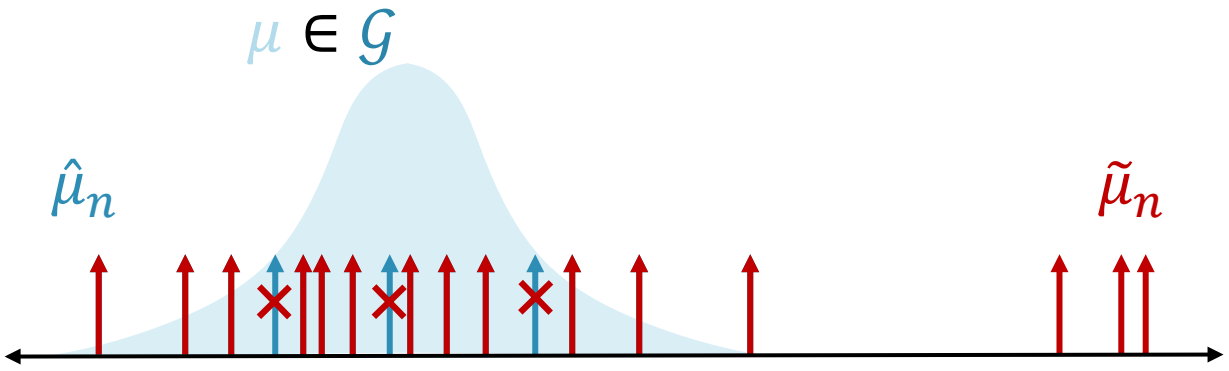
Partial OT MDE: $\mathrm{T}(\tilde{\mu}) \in \underset{\nu \in \mathcal{G}}{\operatorname{argmin}}\, \mathrm{W}_p^\epsilon(\nu, \tilde{\mu})$

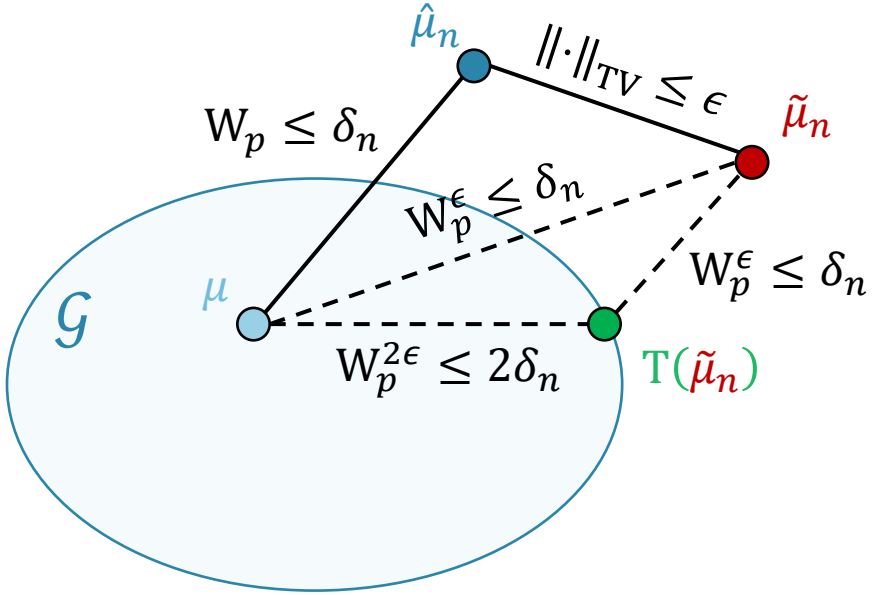
$$\Longrightarrow \mathrm{W}_p\big(\mu, \mathrm{T}(\tilde{\mu})\big) \le \sup_{\substack{\alpha,\beta \in \mathcal{G}: \\ \mathrm{W}_p^{2\epsilon}(\alpha,\beta)=0}} \mathrm{W}_p(\alpha,\beta) \simeq R_p(\epsilon, \mathcal{G})$$

Theorem (Nietert-Cummings-G. '23)

$$\simeq \begin{cases} \sigma \epsilon^{\frac{1}{p}-\frac{1}{q}}, & \mathcal{G} = \mathcal{G}_{\mathrm{mom}}(q,\sigma) \\ \sigma\sqrt{d}\,\epsilon^{\frac{1}{p}-\frac{1}{2}}, & \mathcal{G} = \mathcal{G}_{\mathrm{cov}}(\sigma) \\ \sigma\sqrt{d + p + \log(1/\epsilon)}\,\epsilon^{\frac{1}{p}}, & \mathcal{G} = \mathcal{G}_{\mathrm{subG}}(\sigma) \end{cases}$$

Back to Robust Estimation: Finite-Sample





Partial OT MDE: $T(\tilde{\mu}_n) \in \operatorname{argmin}_{\nu \in \mathcal{G}} W_p^\epsilon(\nu, \tilde{\mu}_n)$

$$\Longrightarrow W_p\big(\mu, T(\tilde{\mu}_n)\big) \le \sup_{\substack{\alpha,\beta \in \mathcal{G}: \\ W_p^{2\epsilon}(\alpha,\beta) \le 2\delta_n}} W_p(\alpha,\beta) \lesssim R_p(\epsilon, \mathcal{G}) + \mathbb{E}\big[W_p(\mu, \hat{\mu}_n)\big]$$

W_p resilience

Theorem (Nietert-Cummings-G. '23)

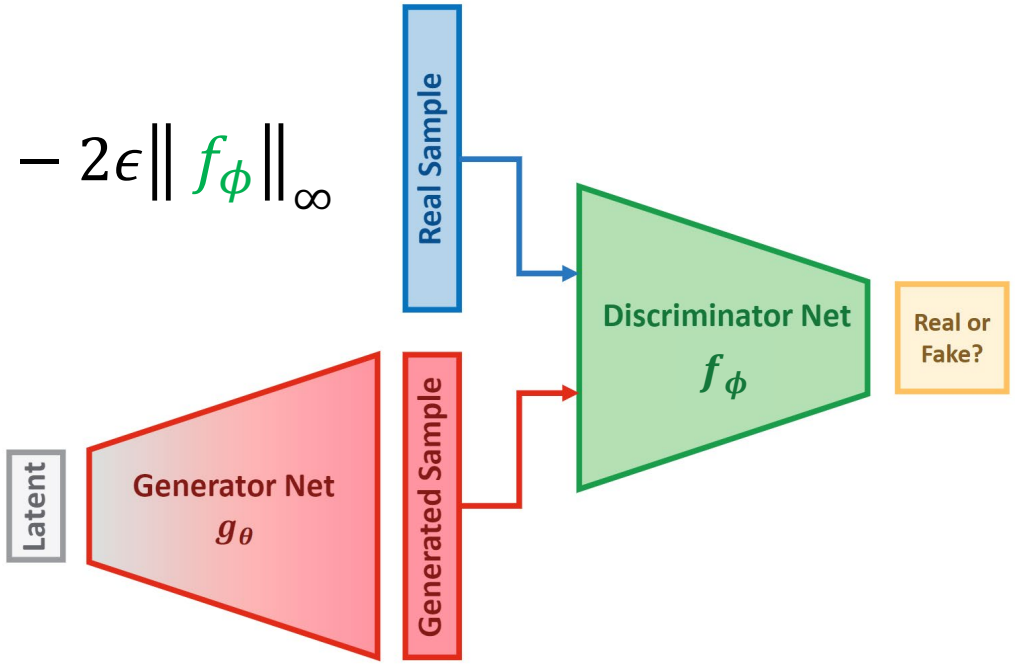
For the $\mathcal{G}$ families before: $R_{p,n}(\epsilon, \mathcal{G}) \simeq R_p(\epsilon, \mathcal{G}) + n^{-1/d}$, achieved by $T(\tilde{\mu}_n)$.

Robust Generative Modeling

Partial OT MDE:

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

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



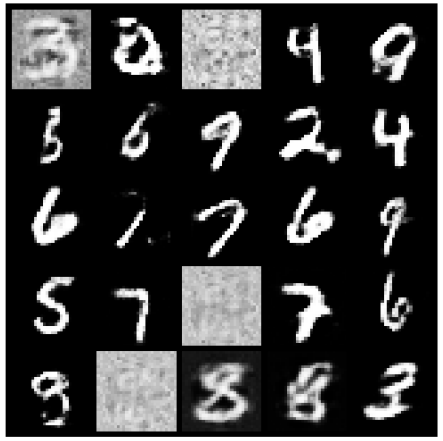
Implementation: 2 lines of code

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

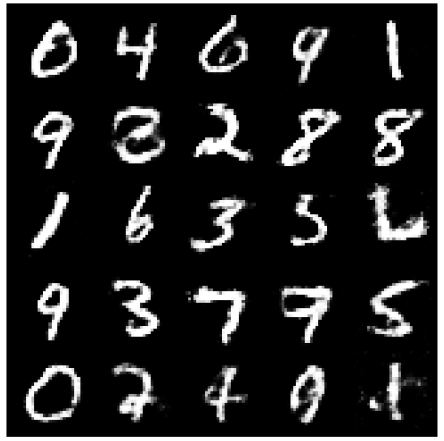
Robust Generative Modeling: Empirical Results

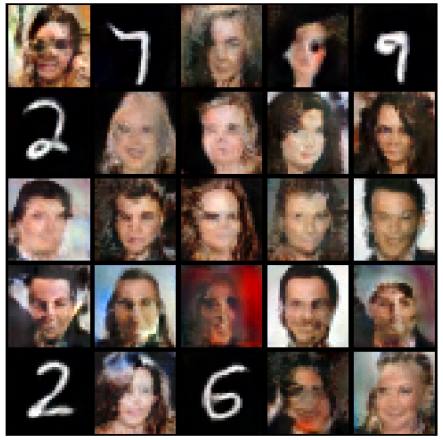
Effect of Robustification

Standard GAN



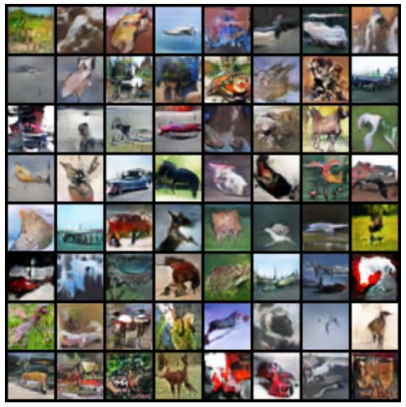
Robust GAN



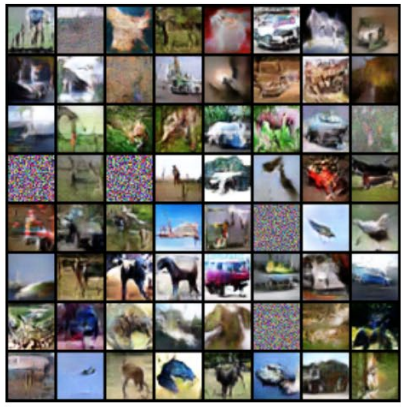


Comparisons

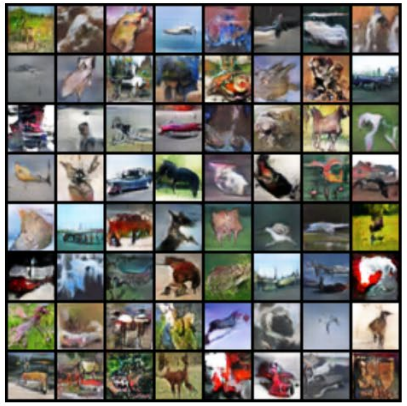
Ours



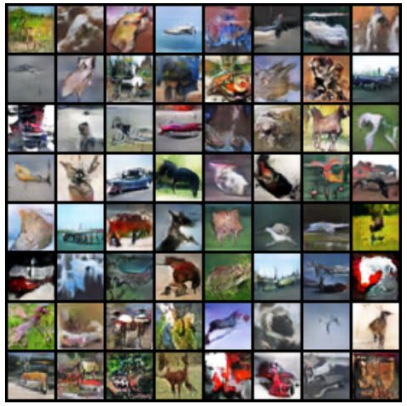
[Balaji et al. ’20]



[Staerman et al. ’21]



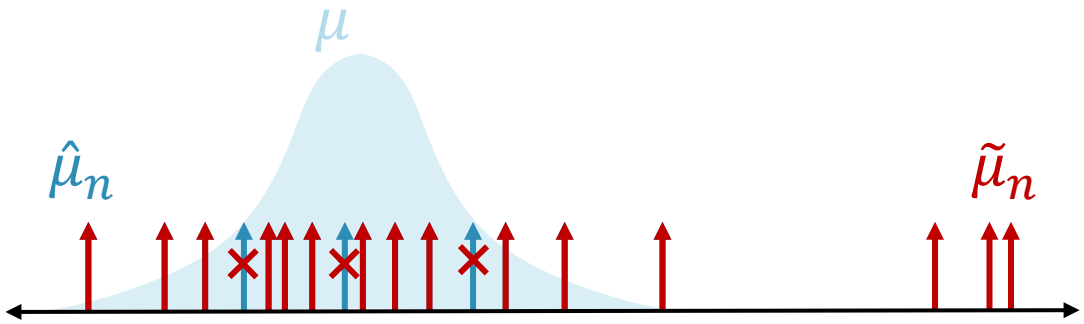
Control



Summary

Limitations of OT: Sensitivity to outliers

- Popular for applications
- Methods break down due to data poisoning



Partial OT: Outlier-robust OT framework

- Structural properties, including dual form
- Partial OT MDE is minimax optimal robust dist. estimator
- Enables learning well-performing models from contaminated data

Thank you!

Additional results: Polytime estimator, outlier-robust map estimation & stochastic optimization

[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