Smooth Wasserstein Distance: Metric Structure and Statistical Efficiency

Ziv Goldfeld

Cornell University

CS Theory Seminar

April 27th, 2020

Outline

1. Generative adversarial networks (GANs)
2. Optimal transport and Wasserstein distance
3. Smoothed Wasserstein distance
4. Smoothed GANs: generalization, implementation and initial results
5. Summary

Motivation: Generative Modeling

Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

Motivation: Generative Modeling

Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

- **Goal:** Learn structure (generative model, clusters, dim. reduction)

Motivation: Generative Modeling

Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

- **Goal:** Learn structure (**generative model**, clusters, dim. reduction)

Motivation: Generative Modeling

Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

- **Goal:** Learn structure (**generative model**, clusters, dim. reduction)

(Latent Variable) Generative Modeling: Learn a model $Q_\theta \approx P_X$

Motivation: Generative Modeling

Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

- **Goal:** Learn structure (**generative model**, clusters, dim. reduction)

(Latent Variable) Generative Modeling: Learn a model $Q_\theta \approx P_X$

- Sample latent $Z \sim P_Z \in \mathcal{P}(\mathbb{R}^p)$, $p \ll d$

Motivation: Generative Modeling

Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

- **Goal:** Learn structure (**generative model**, clusters, dim. reduction)

(Latent Variable) Generative Modeling: Learn a model $Q_\theta \approx P_X$

- Sample latent $Z \sim P_Z \in \mathcal{P}(\mathbb{R}^p)$, $p \ll d$
- Expand Z to $\mathbb{R}^d$ space via (random) transformation $Q_{X|Z}^{(\theta)}$

Motivation: Generative Modeling

Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

- **Goal:** Learn structure (**generative model**, clusters, dim. reduction)

(Latent Variable) Generative Modeling: Learn a model $Q_\theta \approx P_X$

- Sample latent $Z \sim P_Z \in \mathcal{P}(\mathbb{R}^p)$, $p \ll d$
- Expand Z to $\mathbb{R}^d$ space via (random) transformation $Q_{X|Z}^{(\theta)}$

$\implies$ **Generative model:** $Q_\theta(\cdot) := \int_{\mathbb{R}^{d_0}} Q_{X|Z}^{(\theta)}(\cdot|z)\, \mathrm{d}P_Z(z)$

Motivation: Generative Modeling

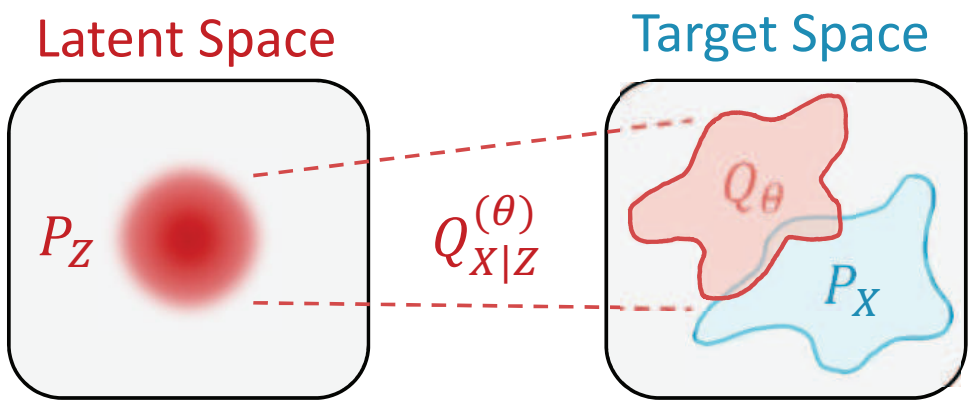
Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

- **Goal:** Learn structure (**generative model**, clusters, dim. reduction)

(Latent Variable) Generative Modeling: Learn a model $Q_\theta \approx P_X$

- Sample latent $Z \sim P_Z \in \mathcal{P}(\mathbb{R}^p)$, $p \ll d$
- Expand Z to $\mathbb{R}^d$ space via (random) transformation $Q_{X|Z}^{(\theta)}$

$\implies$ **Generative model:** $Q_\theta(\cdot) := \int_{\mathbb{R}^{d_0}} Q_{X|Z}^{(\theta)}(\cdot|z)\, \mathrm{d}P_Z(z)$



Motivation: Generative Modeling

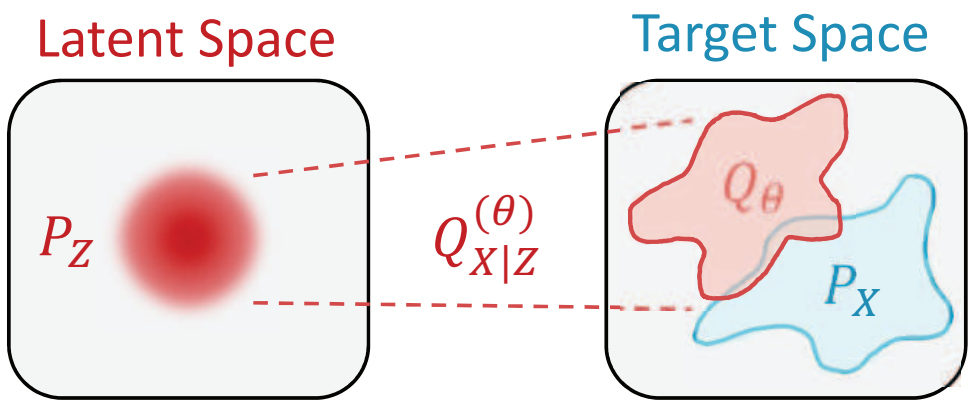
Unsupervised Learning: Get $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X \in \mathcal{P}(\mathbb{R}^d)$ samples

- **Goal:** Learn structure (**generative model**, clusters, dim. reduction)

(Latent Variable) Generative Modeling: Learn a model $Q_\theta \approx P_X$

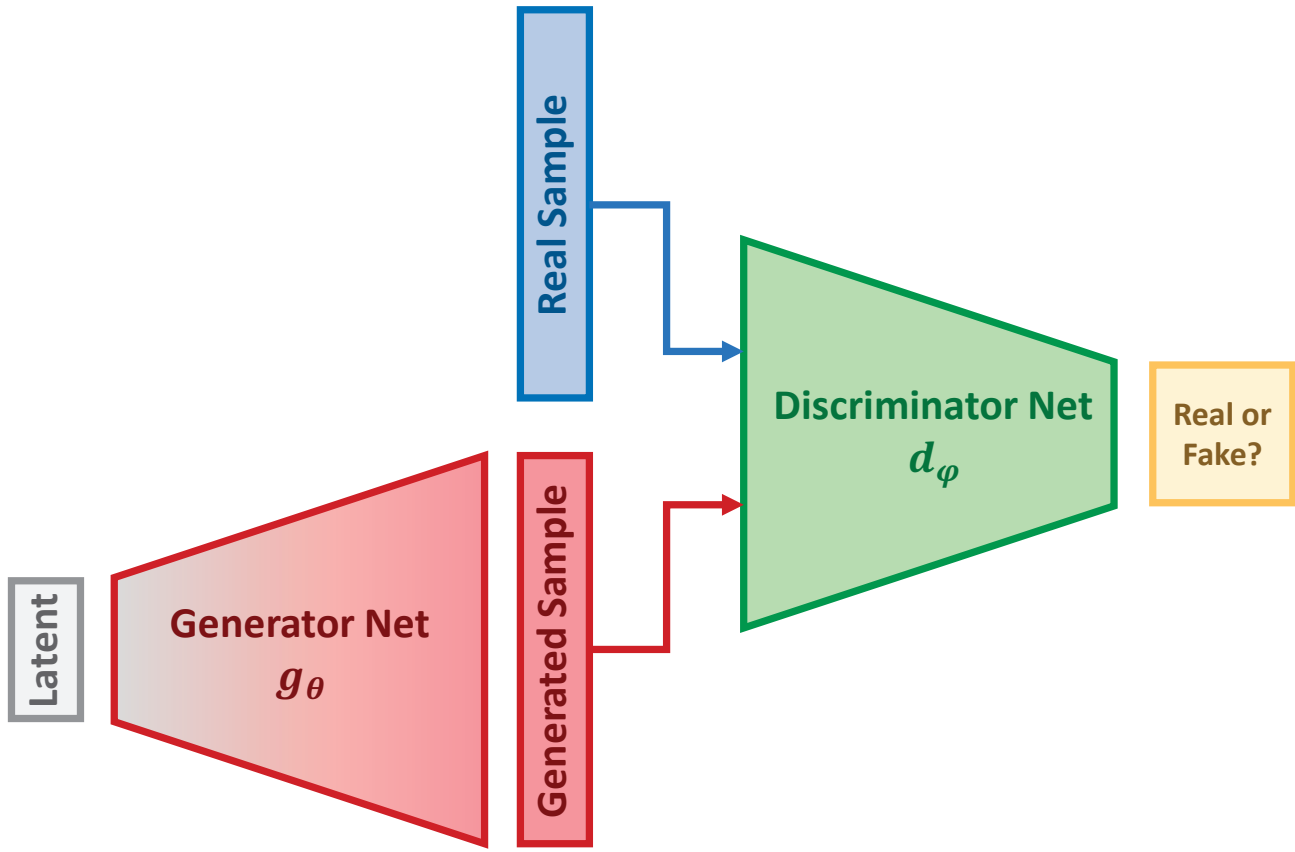
- Sample latent $Z \sim P_Z \in \mathcal{P}(\mathbb{R}^p)$, $p \ll d$
- Expand Z to $\mathbb{R}^d$ space via (random) transformation $Q_{X|Z}^{(\theta)}$

$\Longrightarrow$ **Generative model:** $Q_\theta(\cdot) := \int_{\mathbb{R}^{d_0}} Q_{X|Z}^{(\theta)}(\cdot|z)\,\mathrm{d}P_Z(z)$



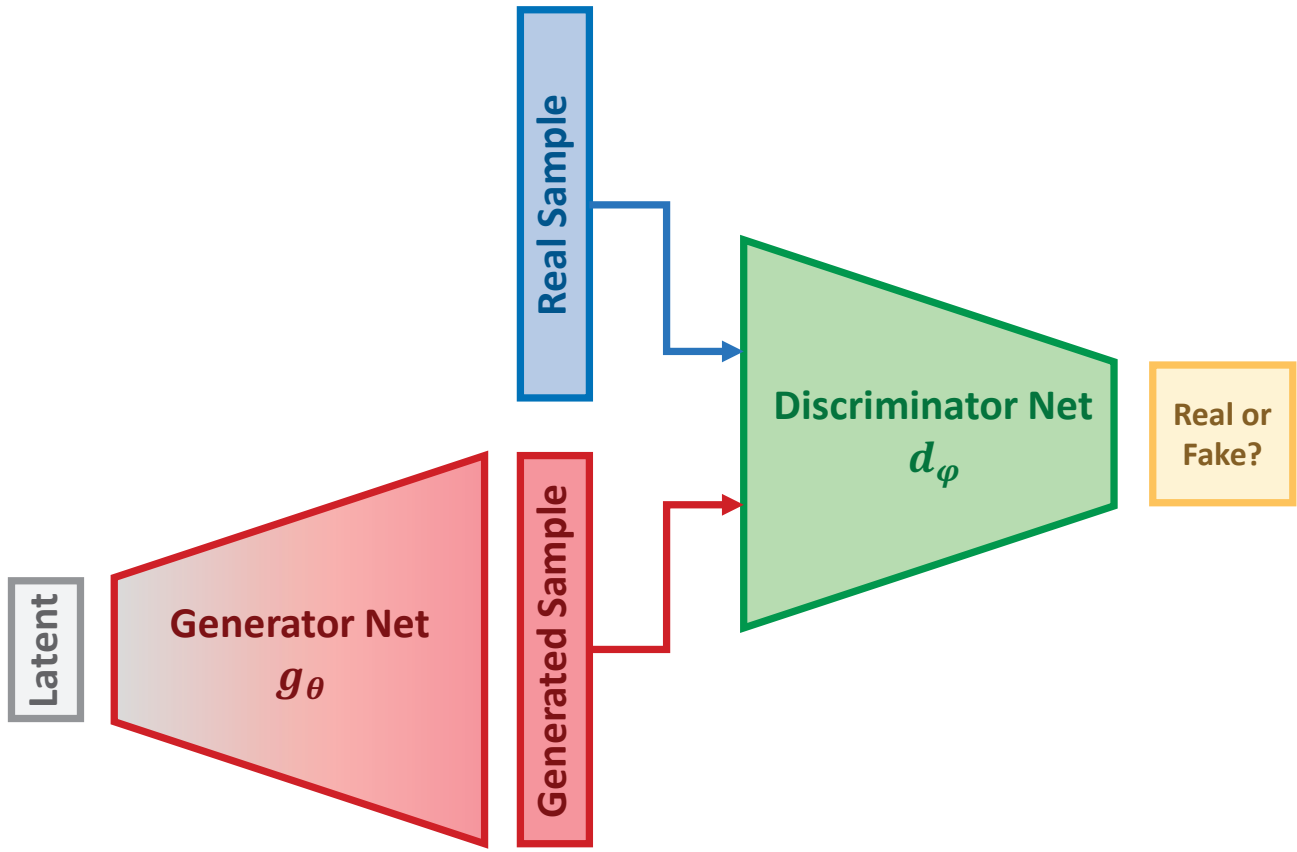
GANs [Goodfellow *et al*'14]: Adversarial training for **implicit** Q_θ

Generative Adversarial Networks



Generative Adversarial Networks

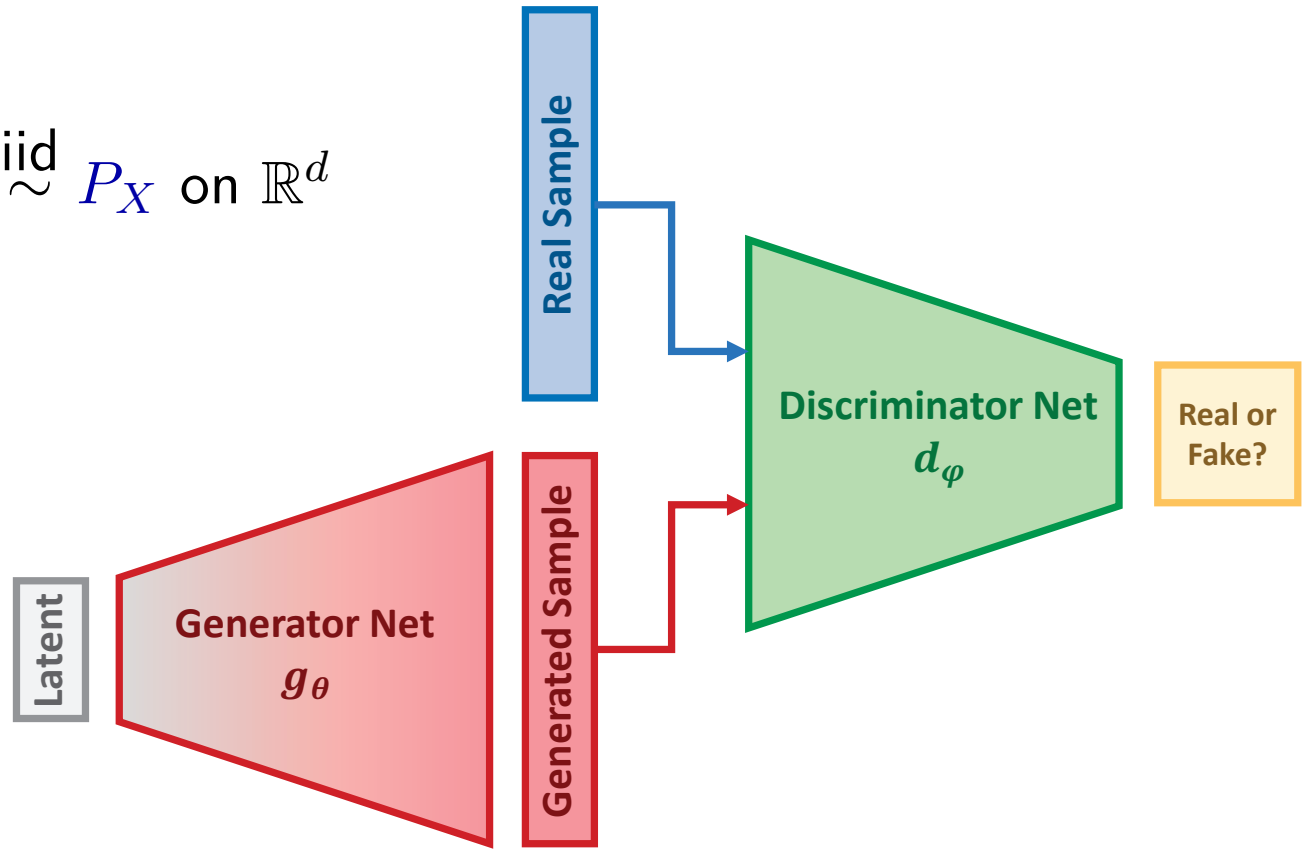
Resources:



Generative Adversarial Networks

Resources:

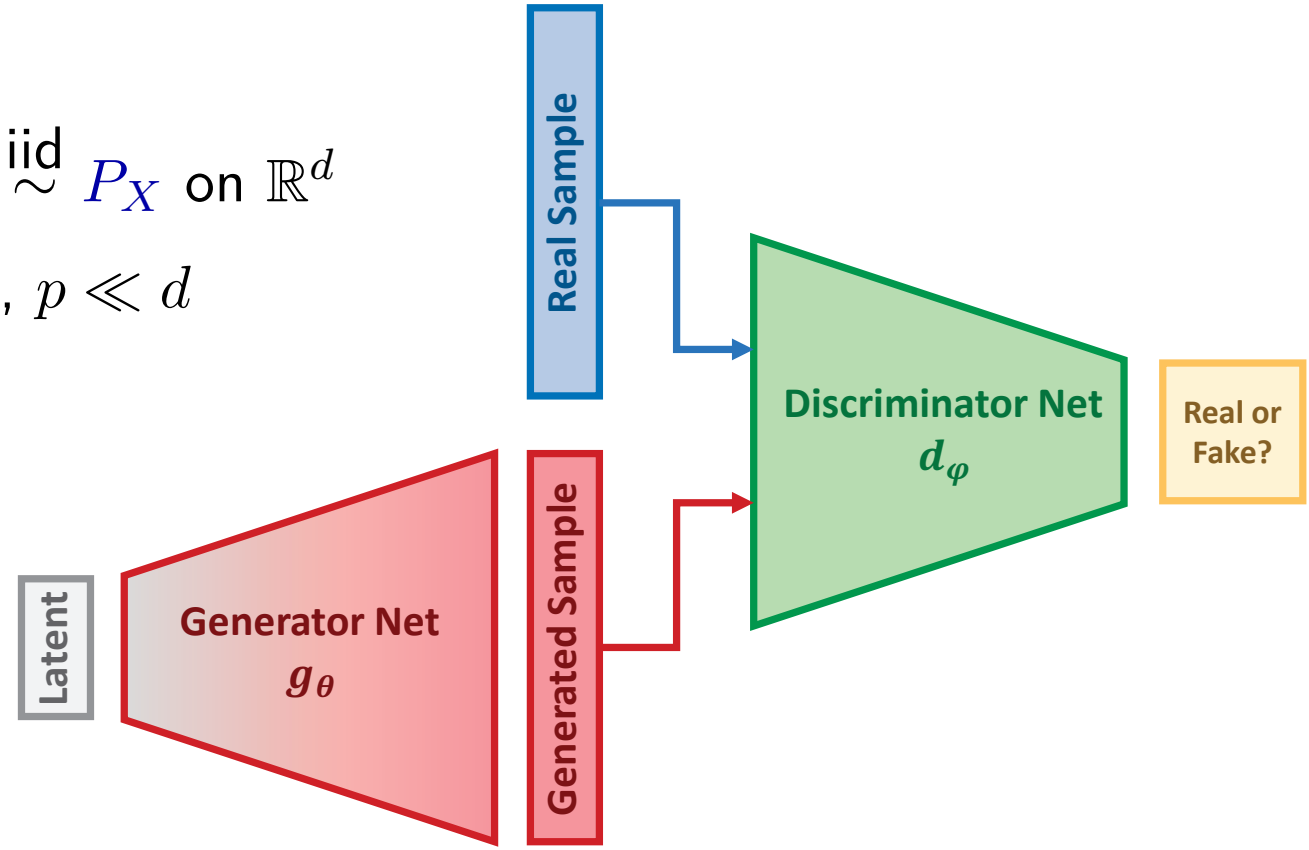
1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$



Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

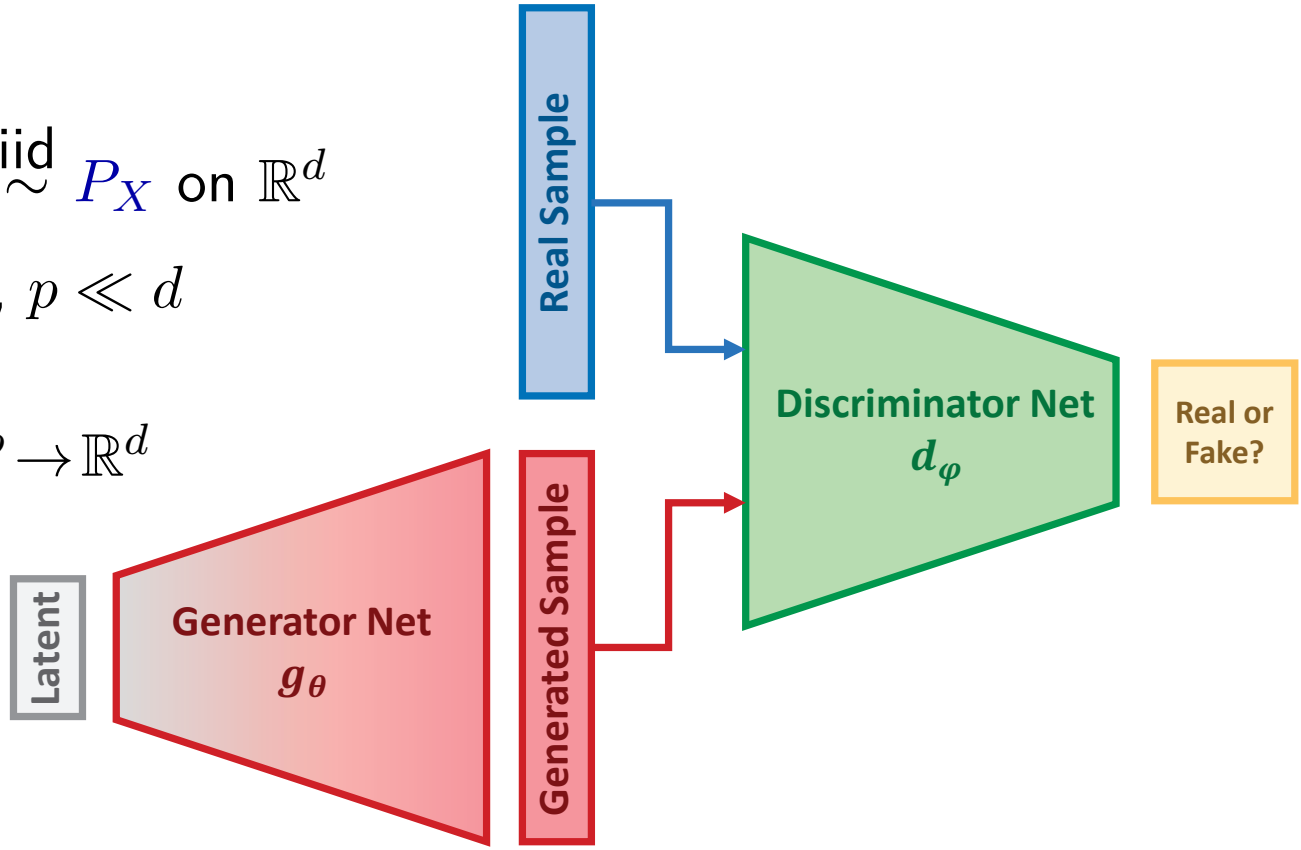


Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$



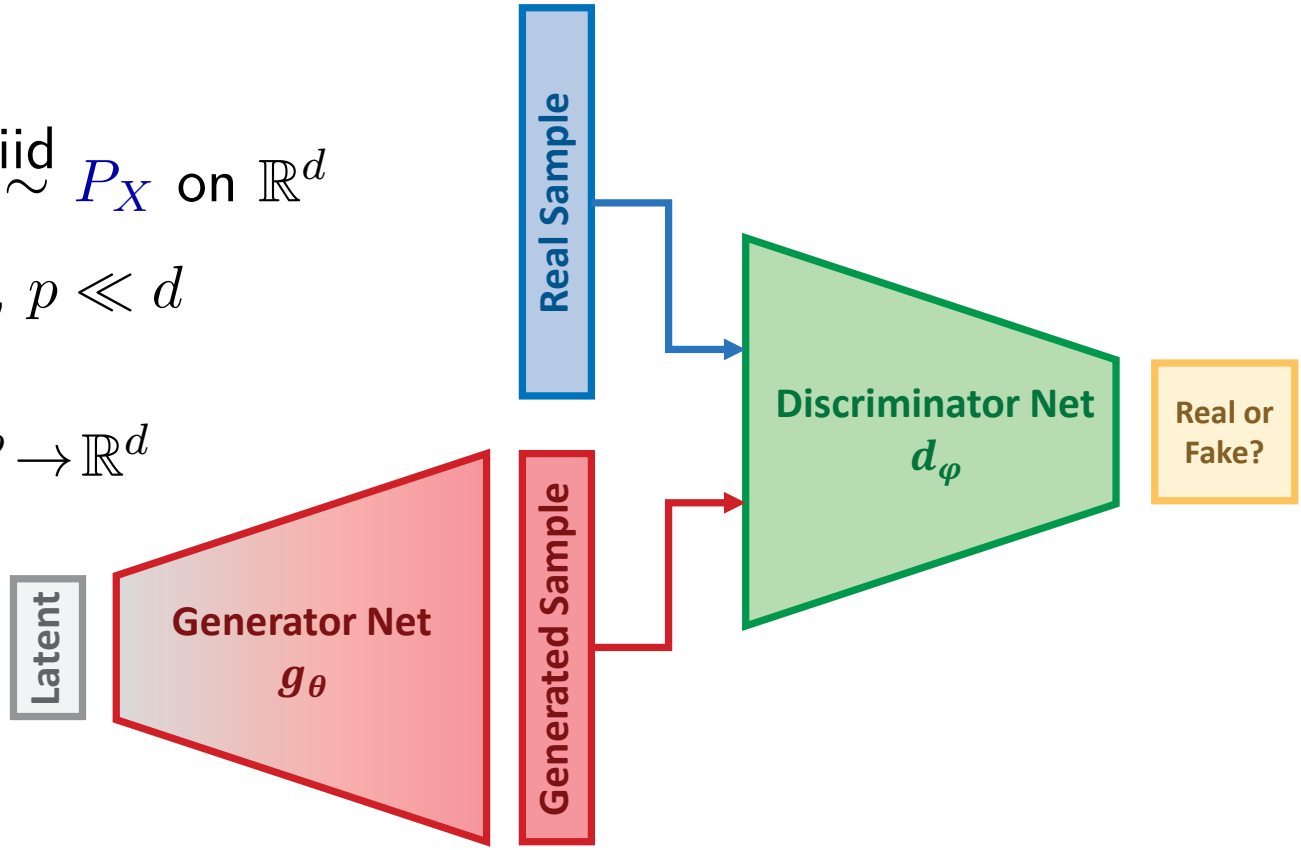
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$



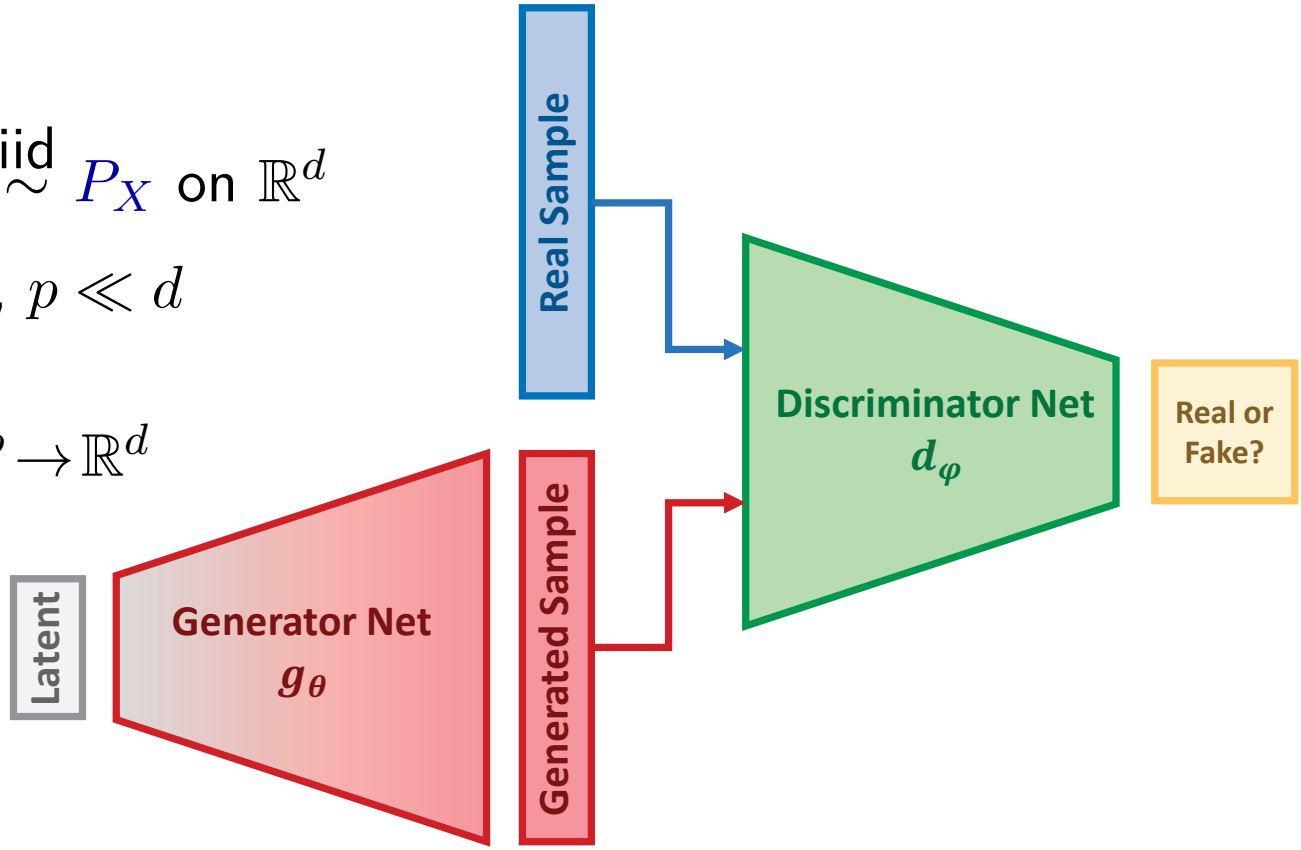
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



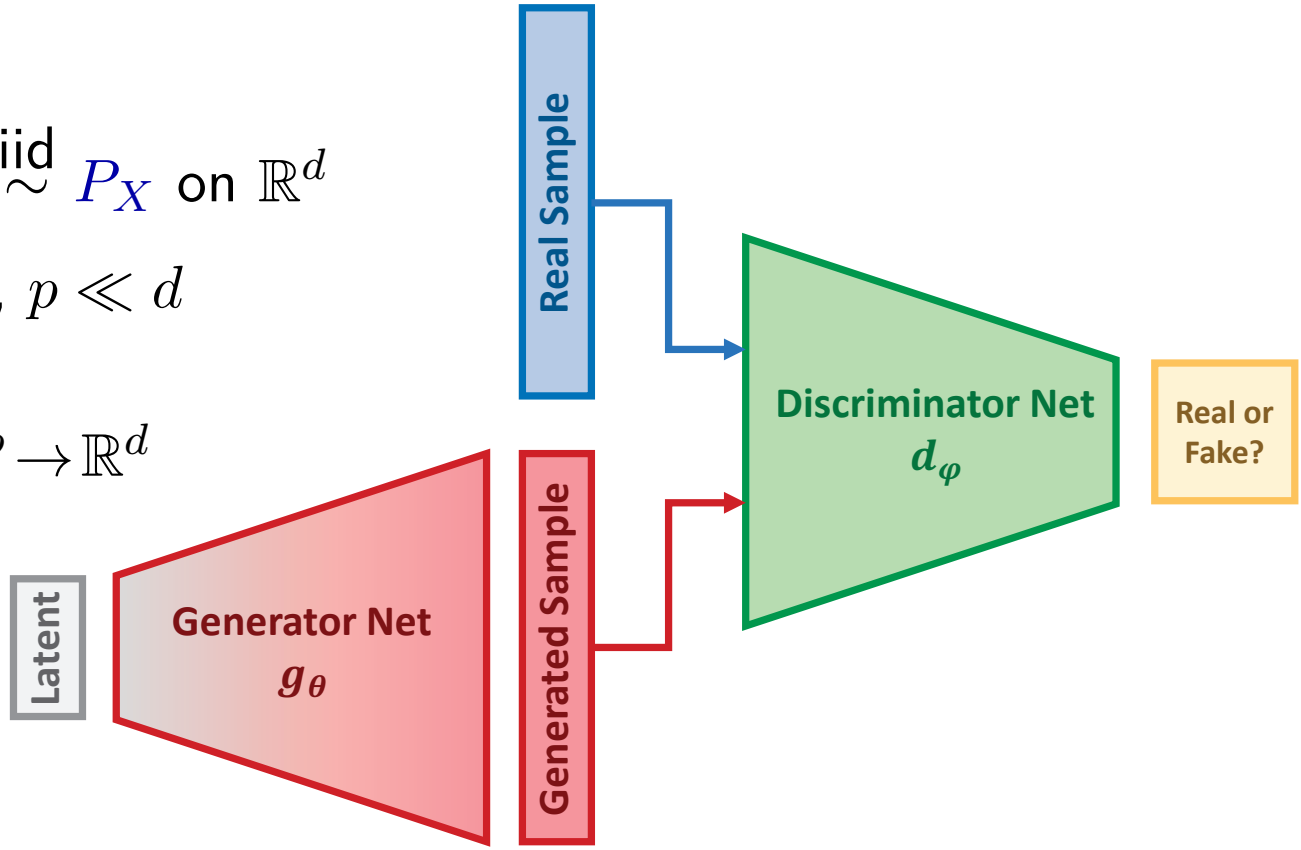
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



Discriminator: DNN $d_\phi : \mathbb{R}^d \to \mathbb{R}$

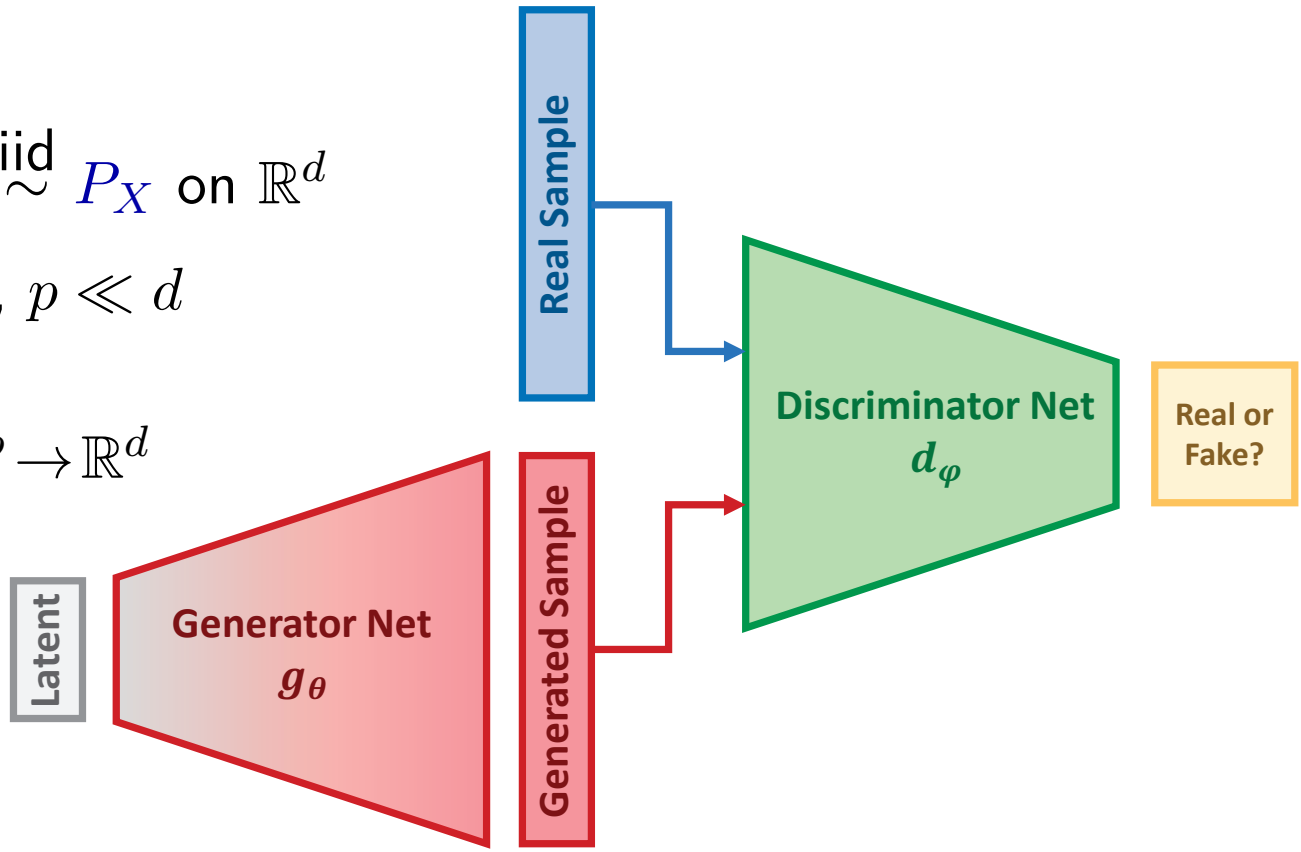
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



Discriminator: DNN $d_\phi : \mathbb{R}^d \to \mathbb{R}$

- Discriminate: $d_\phi(X)$ and $d_\phi(X_\theta)$ are maximally different

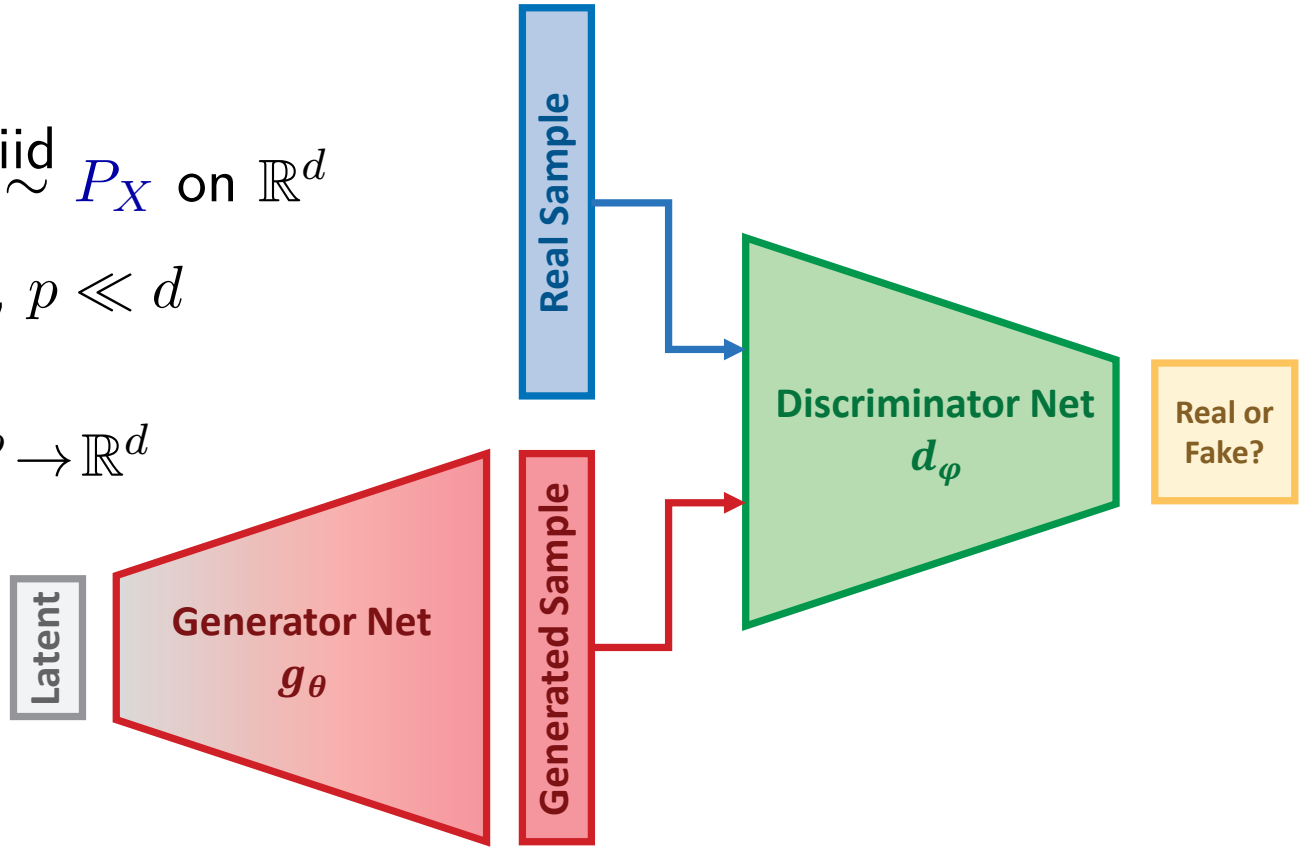
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



Discriminator: DNN $d_\phi : \mathbb{R}^d \to \mathbb{R}$

- Discriminate: $d_\phi(X)$ and $d_\phi(X_\theta)$ are maximally different

$\Longrightarrow$ **Minimax Game:**

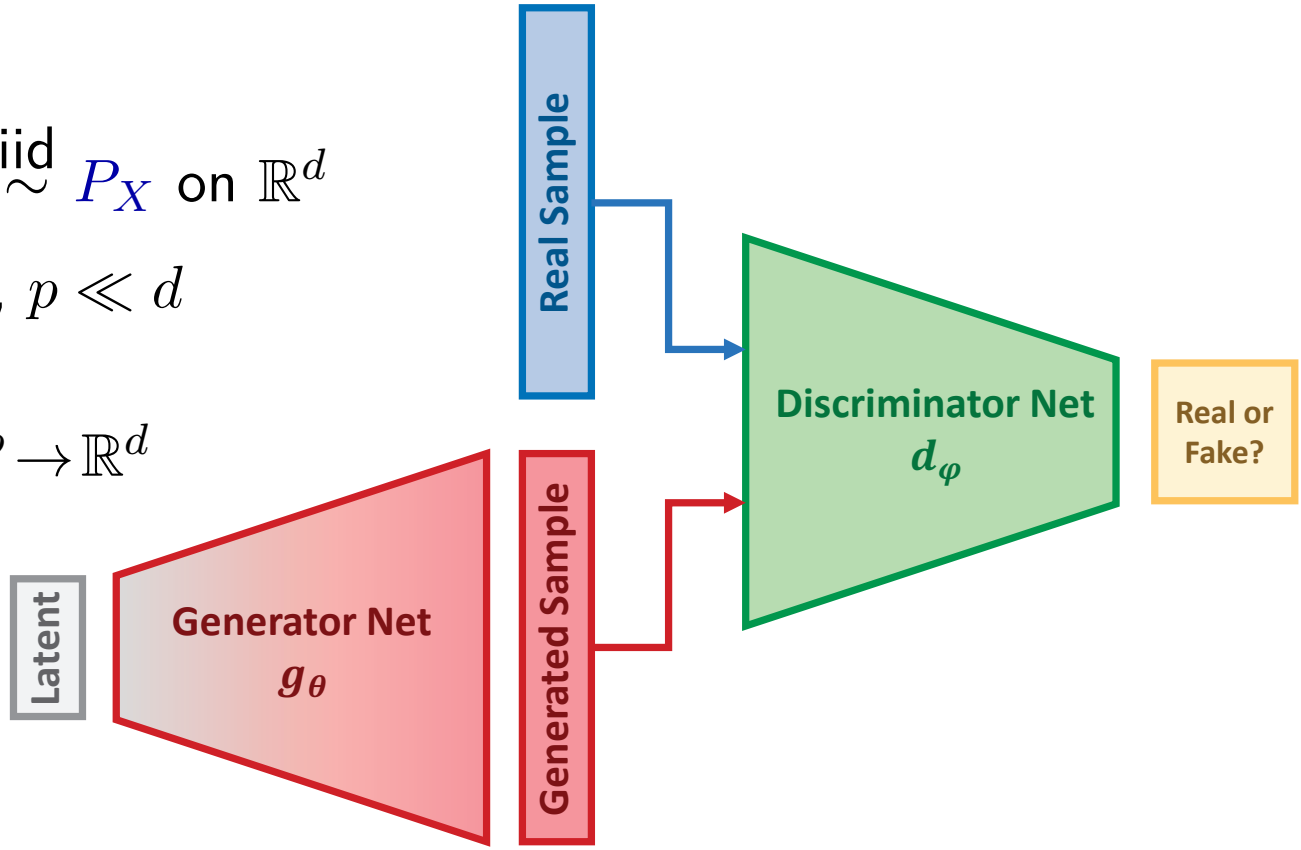
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



Discriminator: DNN $d_\phi : \mathbb{R}^d \to \mathbb{R}$

- Discriminate: $d_\phi(X)$ and $d_\phi(X_\theta)$ are maximally different

$\Longrightarrow$ **Minimax Game:** $\mathbb{E}d_\phi(X)$

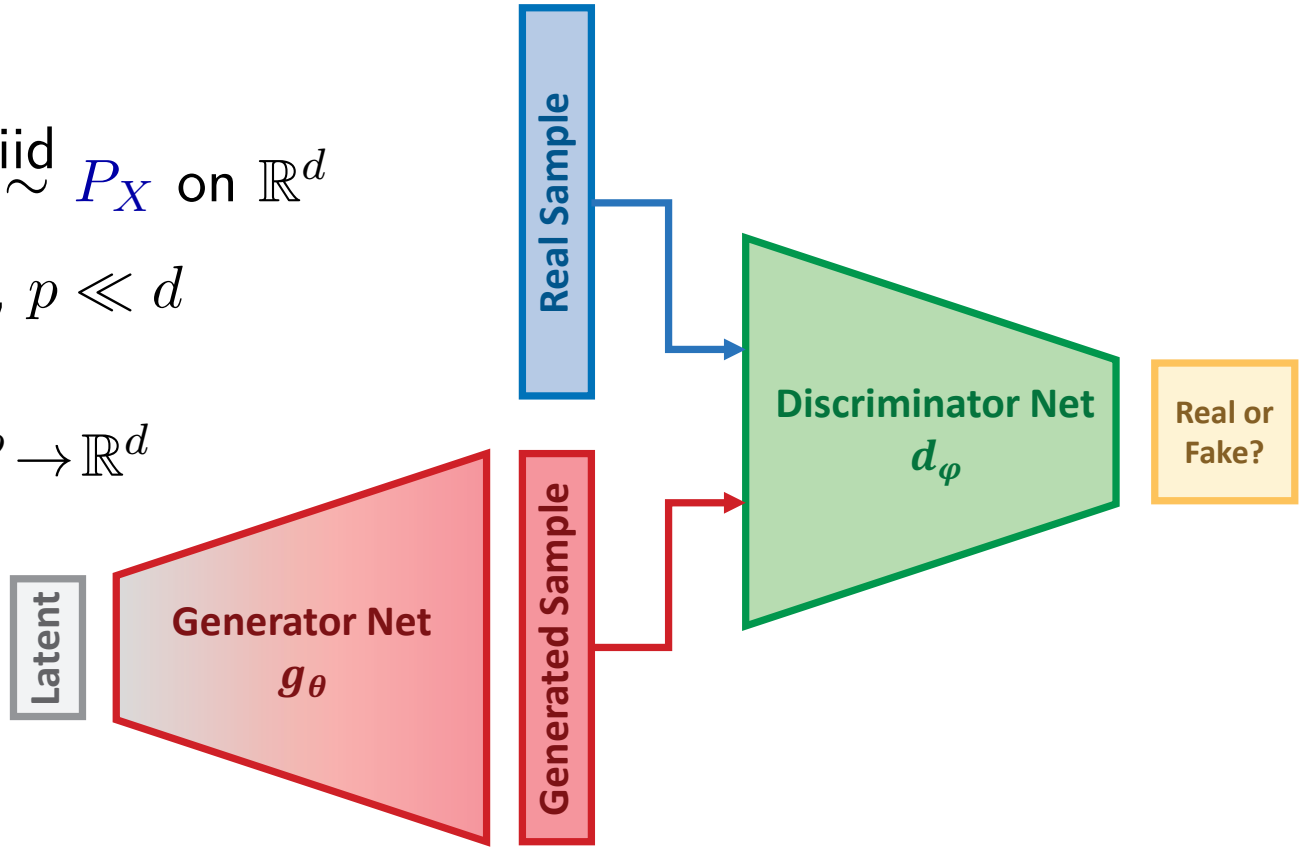
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



Discriminator: DNN $d_\phi : \mathbb{R}^d \to \mathbb{R}$

- Discriminate: $d_\phi(X)$ and $d_\phi(X_\theta)$ are maximally different

$\Longrightarrow$ **Minimax Game:** $\mathbb{E}d_\phi(X) - \mathbb{E}d_\phi(g_\theta(Z))$

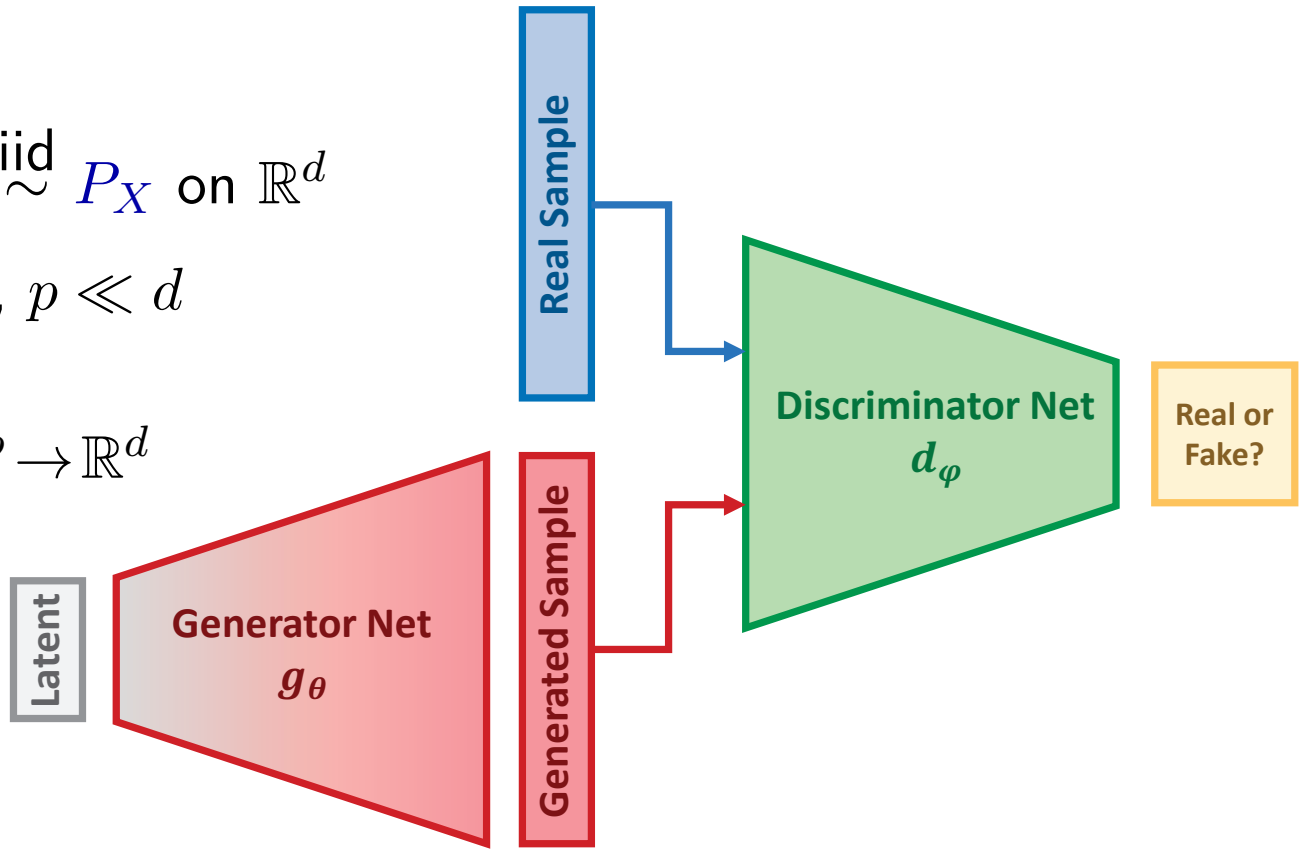
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



Discriminator: DNN $d_\phi : \mathbb{R}^d \to \mathbb{R}$

- Discriminate: $d_\phi(X)$ and $d_\phi(X_\theta)$ are maximally different

$\Longrightarrow$ **Minimax Game:** $\sup_\phi \mathbb{E} d_\phi(X) - \mathbb{E} d_\phi(g_\theta(Z))$

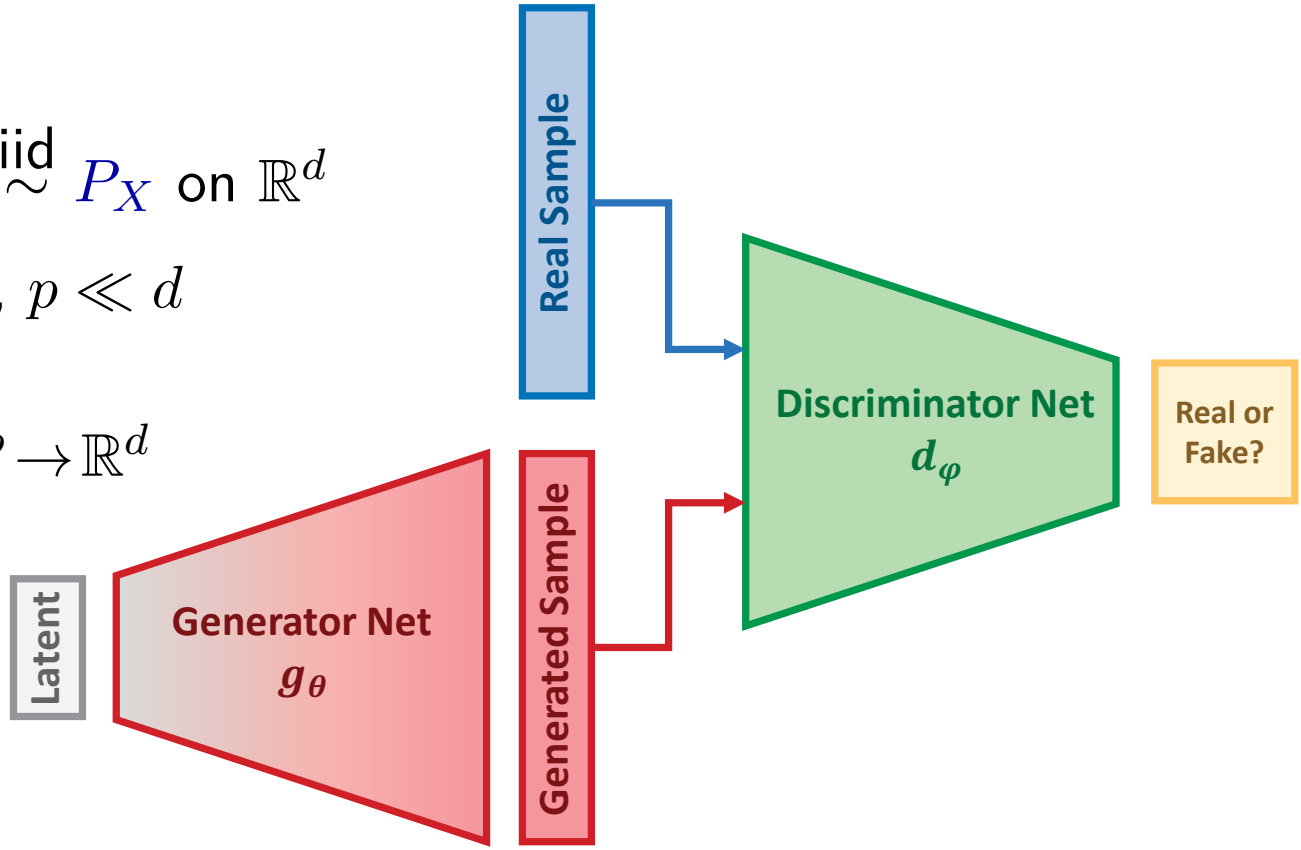
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



Discriminator: DNN $d_\phi : \mathbb{R}^d \to \mathbb{R}$

- Discriminate: $d_\phi(X)$ and $d_\phi(X_\theta)$ are maximally different

$\Longrightarrow$ **Minimax Game:** $\inf_\theta \sup_\phi \mathbb{E} d_\phi(X) - \mathbb{E} d_\phi(g_\theta(Z))$

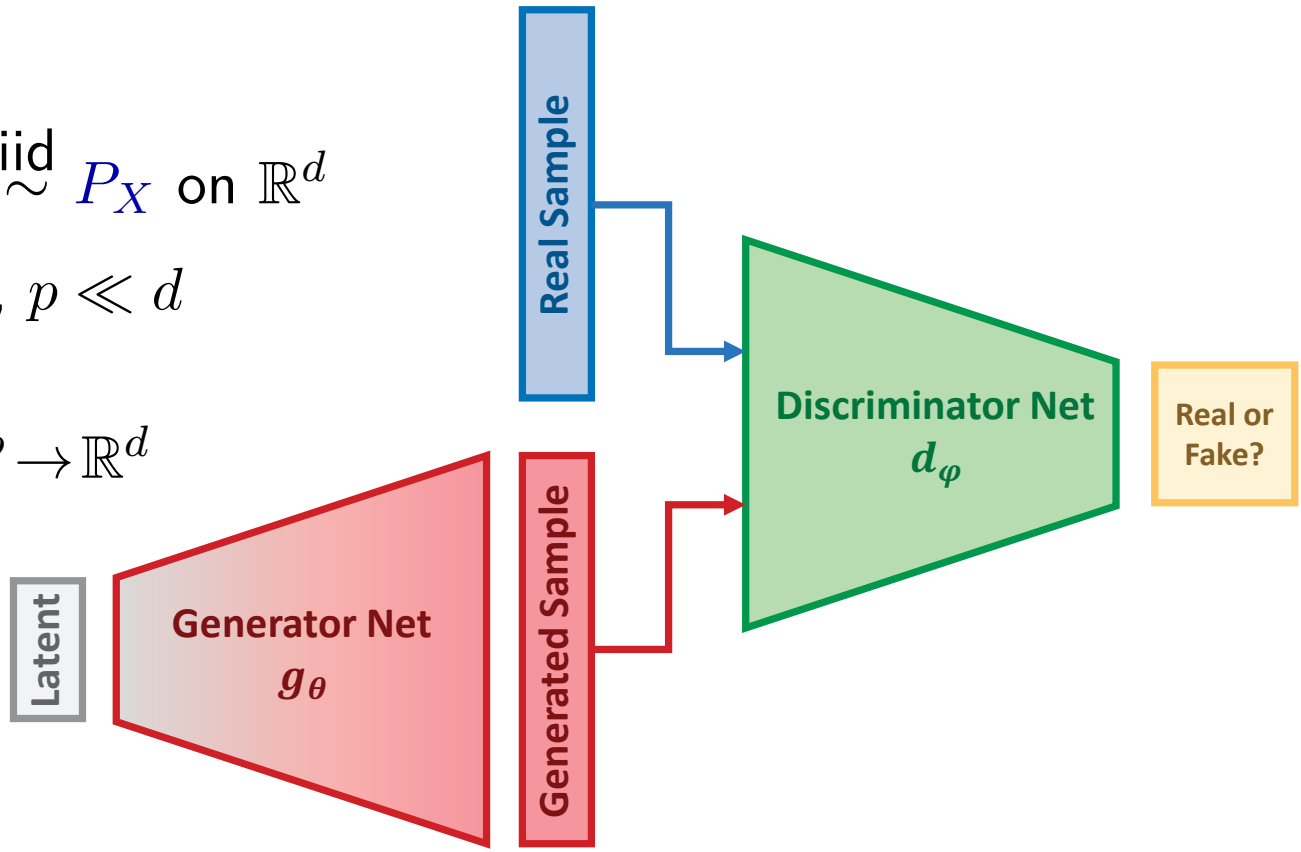
Generative Adversarial Networks

Resources:

1. Data (real) $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_X$ on $\mathbb{R}^d$
2. Latent $Z \sim \mathcal{N}(0, \mathrm{I}_p)$, $p \ll d$

Generator: DNN $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$

- Fake sample:
  $X_\theta = g_\theta(Z)$
  s.t. $X_\theta \sim Q_\theta \approx P_X$



Discriminator: DNN $d_\phi : \mathbb{R}^d \to \mathbb{R}$

- Discriminate: $d_\phi(X)$ and $d_\phi(X_\theta)$ are maximally different

$\Longrightarrow$ **Minimax Game:**

$$\inf_\theta \sup_\phi \mathbb{E} d_\phi(X) - \mathbb{E} d_\phi(g_\theta(Z))$$

Generative Adversarial Networks

NVIDIA's ProGAN 2.0 [Karras *et al*'19]

Principled Approach to Generative Modeling

Statistical Divergences: Measure 'distance' between prob. distributions

Principled Approach to Generative Modeling

Statistical Divergences: Measure 'distance' between prob. distributions

- $\delta : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty)$ s.t. $\delta(P, Q) = 0 \iff P = Q$

Principled Approach to Generative Modeling

Statistical Divergences: Measure 'distance' between prob. distributions

- $\delta : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty)$ s.t. $\delta(P, Q) = 0 \iff P = Q$
- If symmetric & $\delta(P, Q) \leq \delta(P, R) + \delta(R, Q)$ then δ is a metric

Principled Approach to Generative Modeling

Statistical Divergences: Measure 'distance' between prob. distributions

- $\delta : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty)$ s.t. $\delta(P, Q) = 0 \iff P = Q$
- If symmetric & $\delta(P, Q) \leq \delta(P, R) + \delta(R, Q)$ then δ is a metric

Objective Quantification: Pick $\delta(\cdot, \cdot)$ and optimize generator to

$$\inf_{\theta} \delta(P_X, Q_\theta)$$

Principled Approach to Generative Modeling

Statistical Divergences: Measure 'distance' between prob. distributions

- $\delta : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty)$ s.t. $\delta(P, Q) = 0 \iff P = Q$
- If symmetric & $\delta(P, Q) \leq \delta(P, R) + \delta(R, Q)$ then δ is a metric

Objective Quantification: Pick $\delta(\cdot, \cdot)$ and optimize generator to

$$\inf_{\theta} \delta(P_X, Q_\theta)$$

- Both perspectives coincide if δ = 1-Wasserstein distance

Principled Approach to Generative Modeling

Statistical Divergences: Measure 'distance' between prob. distributions

- $\delta : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty)$ s.t. $\delta(P, Q) = 0 \iff P = Q$
- If symmetric & $\delta(P, Q) \leq \delta(P, R) + \delta(R, Q)$ then δ is a metric

Objective Quantification: Pick $\delta(\cdot, \cdot)$ and optimize generator to

$$\inf_{\theta} \delta(P_X, Q_\theta)$$

- Both perspectives coincide if δ = 1-Wasserstein distance
- Wasserstein GAN [Arjovsky *et al*'17]

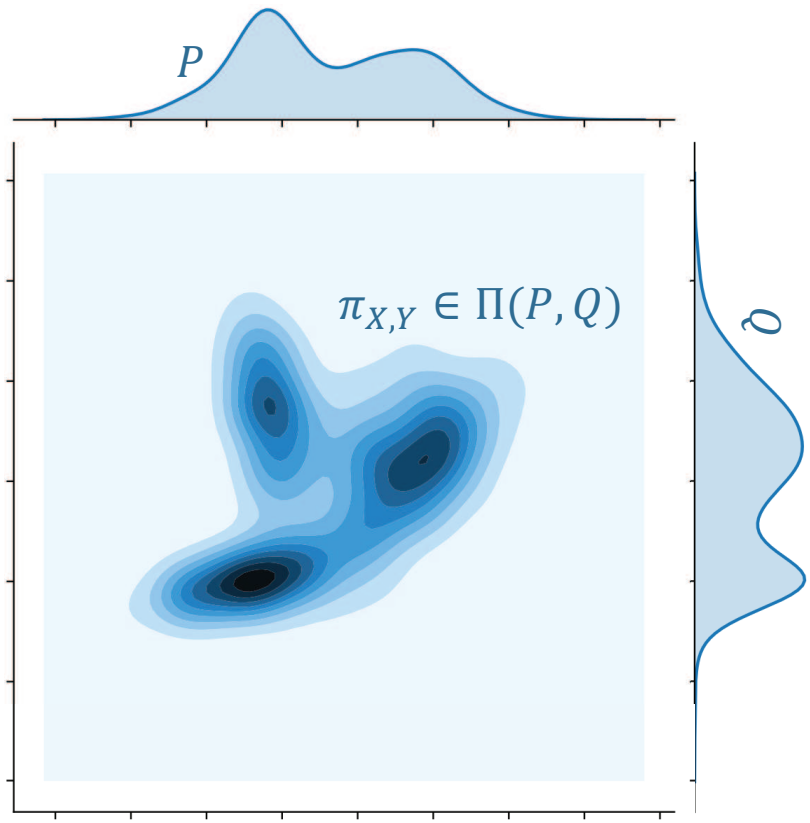
The 1-Wasserstein Distance

Setup: $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (subscript for finite 1st moments)

The 1-Wasserstein Distance

Setup: $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (subscript for finite 1st moments)

- **Coupling:** $\Pi(P,Q) = \left\{ \pi_{X,Y} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, \pi_X = P \;\&\; \pi_Y = Q \right\}$



The 1-Wasserstein Distance

Setup: $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (subscript for finite 1st moments)

- **Coupling:** $\Pi(P, Q) = \left\{ \pi_{X,Y} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, \pi_X = P \;\&\; \pi_Y = Q \right\}$
- **Cost:** $c(x, y) = \|x - y\|$ for transporting x to y

The 1-Wasserstein Distance

Setup: $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (subscript for finite 1st moments)

- **Coupling:** $\Pi(P,Q) = \left\{ \pi_{X,Y} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, \pi_X = P \;\&\; \pi_Y = Q \right\}$
- **Cost:** $c(x,y) = \|x - y\|$ for transporting x to y

Definition (1-Wasserstein)

The 1-Wasserstein distance: $\mathrm{W}_1(P,Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$

The 1-Wasserstein Distance

Setup: $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (subscript for finite 1st moments)

- **Coupling:** $\Pi(P,Q) = \left\{ \pi_{X,Y} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, \pi_X = P \;\&\; \pi_Y = Q \right\}$
- **Cost:** $c(x,y) = \|x - y\|$ for transporting x to y

Definition (1-Wasserstein)

The 1-Wasserstein distance: $\mathrm{W}_1(P,Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$

Comments:

The 1-Wasserstein Distance

Setup: $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (subscript for finite 1st moments)

- **Coupling:** $\Pi(P,Q) = \left\{ \pi_{X,Y} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, \pi_X = P \;\&\; \pi_Y = Q \right\}$
- **Cost:** $c(x,y) = \|x - y\|$ for transporting x to y

Definition (1-Wasserstein)

The 1-Wasserstein distance: $\mathrm{W}_1(P,Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$

Comments:

- **Operational Meaning:** Minimize work of transporting P to Q

The 1-Wasserstein Distance

Setup: $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (subscript for finite 1st moments)

- **Coupling:** $\Pi(P,Q) = \left\{ \pi_{X,Y} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, \pi_X = P \;\&\; \pi_Y = Q \right\}$
- **Cost:** $c(x,y) = \|x - y\|$ for transporting x to y

Definition (1-Wasserstein)

The 1-Wasserstein distance: $\mathrm{W}_1(P,Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$

Comments:

- **Operational Meaning:** Minimize work of transporting P to Q
- **Robustness to Supp. Mismatch:** $\mathrm{W}_1(P,Q) < \infty$, $\forall P, Q \in \mathcal{P}_1(\mathbb{R}^d)$

The 1-Wasserstein Distance

Setup: $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ (subscript for finite 1st moments)

- **Coupling:** $\Pi(P,Q) = \left\{ \pi_{X,Y} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \,\middle|\, \pi_X = P \;\&\; \pi_Y = Q \right\}$
- **Cost:** $c(x,y) = \|x - y\|$ for transporting x to y

Definition (1-Wasserstein)

The 1-Wasserstein distance: $\mathsf{W}_1(P,Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$

Comments:

- **Operational Meaning:** Minimize work of transporting P to Q
- **Robustness to Supp. Mismatch:** $\mathsf{W}_1(P,Q) < \infty$, $\forall P, Q \in \mathcal{P}_1(\mathbb{R}^d)$
- **Metric:** $\left(\mathcal{P}_1(\mathbb{R}^d), \mathsf{W}_1\right)$ is metric space (metrizes weak convergence)

Kantorovich-Rubinstein Duality and GANs

1-Wasserstein: $\mathsf{W}_1(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$

Kantorovich-Rubinstein Duality and GANs

1-Wasserstein: $\mathsf{W}_1(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$

Kantorovich-Rubinstein Duality: Equivalent representation

Kantorovich-Rubinstein Duality and GANs

1-Wasserstein: $\mathsf{W}_1(P,Q) := \inf_{\pi_{X,Y}\in\Pi(P,Q)} \mathbb{E}_\pi \|X-Y\|$

Kantorovich-Rubinstein Duality: Equivalent representation

$$\mathsf{W}_1(P,Q) = \sup_{f\in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)$$

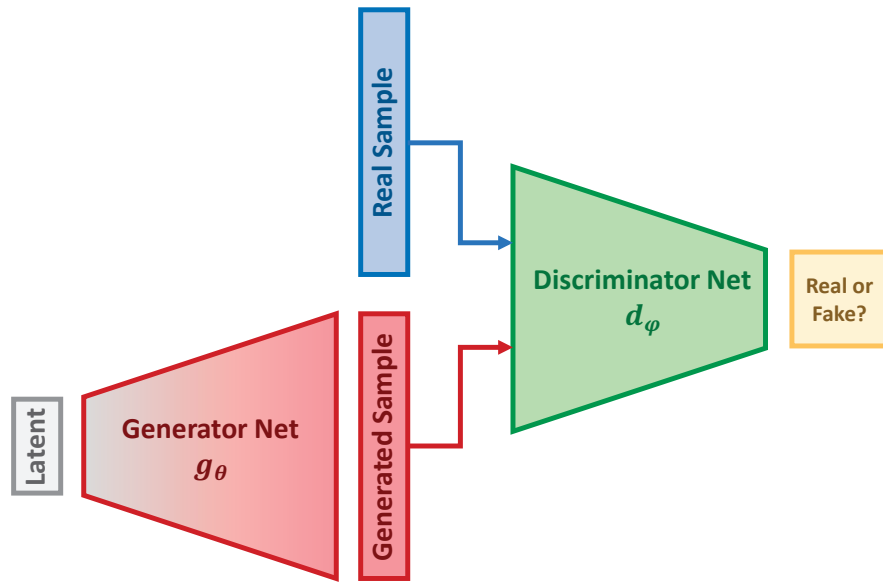
Kantorovich-Rubinstein Duality and GANs

1-Wasserstein: $\mathsf{W}_1(P,Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$

Kantorovich-Rubinstein Duality: Equivalent representation

$$\mathsf{W}_1(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)$$

Back to GANs:



Kantorovich-Rubinstein Duality and GANs

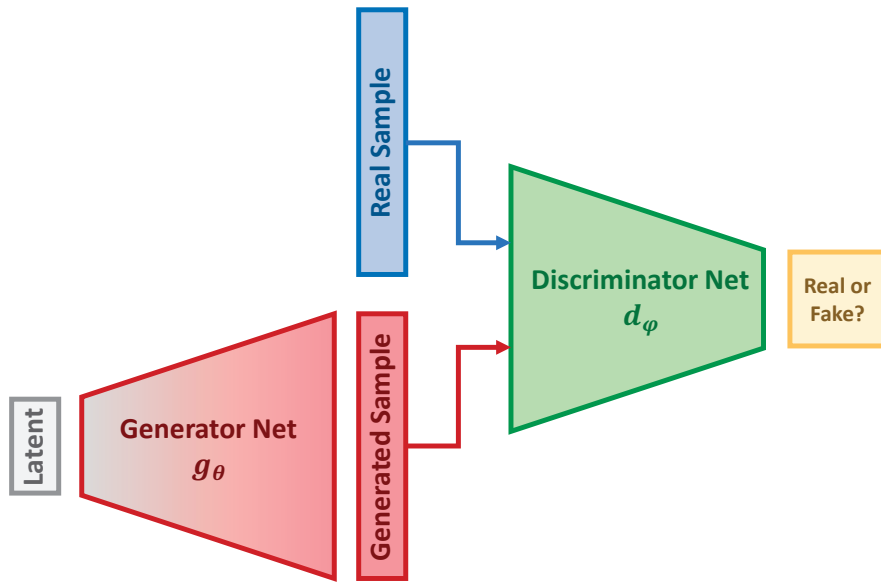
1-Wasserstein: $\mathsf{W}_1(P,Q) := \inf_{\pi_{X,Y}\in\Pi(P,Q)} \mathbb{E}_\pi \|X-Y\|$

Kantorovich-Rubinstein Duality: Equivalent representation

$$\mathsf{W}_1(\boldsymbol{P},Q) = \sup_{f\in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_{\boldsymbol{P}} f(\boldsymbol{X}) - \mathbb{E}_Q f(Y)$$

Back to GANs:

- $P = P_X$ (X (real) data sample)



Kantorovich-Rubinstein Duality and GANs

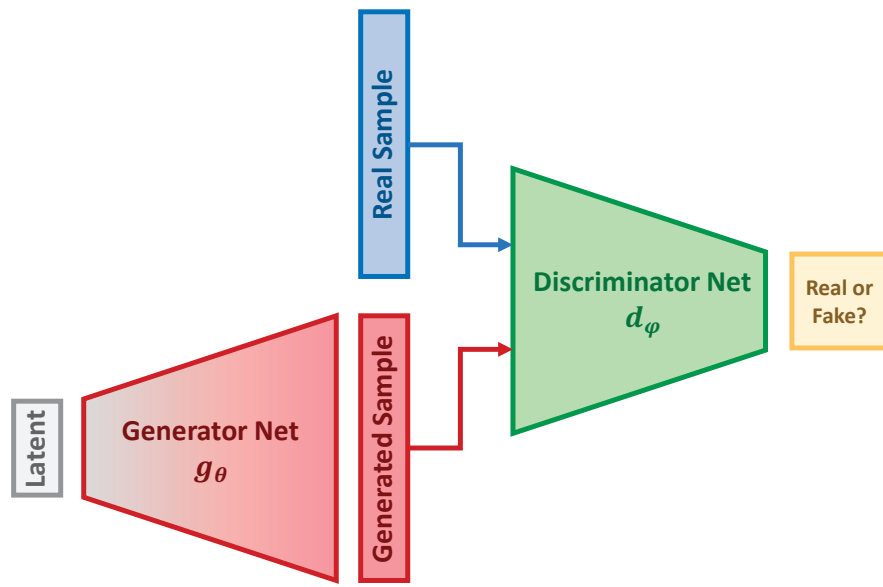
1-Wasserstein: $\mathsf{W}_1(P,Q) := \inf_{\pi_{X,Y}\in\Pi(P,Q)} \mathbb{E}_\pi \|X-Y\|$

Kantorovich-Rubinstein Duality: Equivalent representation

$$\mathsf{W}_1(\boldsymbol{P},\boldsymbol{Q}) = \sup_{f\in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_{\boldsymbol{P}} f(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{Q}} f(\boldsymbol{Y})$$

Back to GANs:

- $P = P_X$ (X (real) data sample)
- $Q = Q_\theta$ ($Y = g_\theta(Z)$ gen. sample)



Kantorovich-Rubinstein Duality and GANs

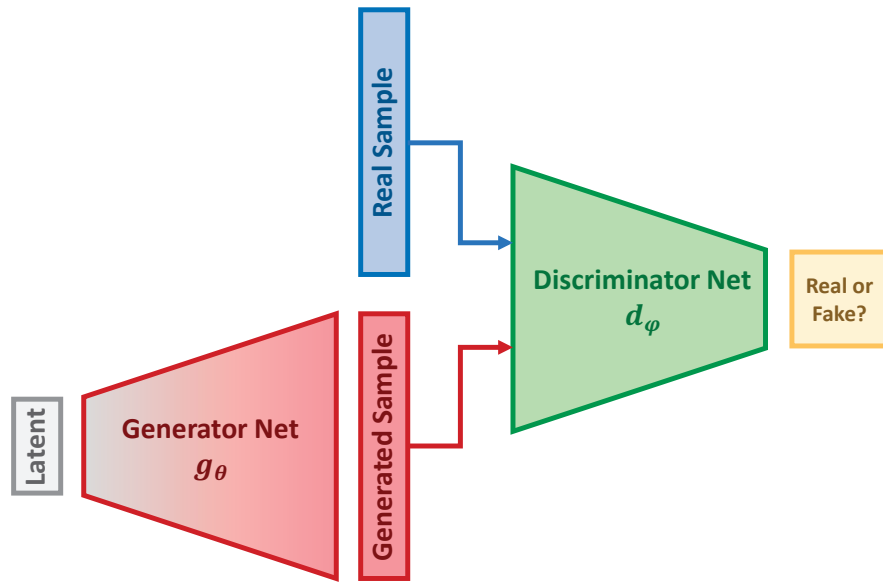
1-Wasserstein: $\mathsf{W}_1(P,Q) := \inf_{\pi_{X,Y}\in\Pi(P,Q)} \mathbb{E}_\pi\|X-Y\|$

Kantorovich-Rubinstein Duality: Equivalent representation

$$\mathsf{W}_1(\boldsymbol{P},\boldsymbol{Q}) = \sup_{\boldsymbol{f}\in\mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_{\boldsymbol{P}}\boldsymbol{f}(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{Q}}\boldsymbol{f}(\boldsymbol{Y})$$

Back to GANs:

- $P = P_X$ (X (real) data sample)
- $Q = Q_\theta$ ($Y = g_\theta(Z)$ gen. sample)
- $f = d_\phi$ (Lip_1 constraint)



Kantorovich-Rubinstein Duality and GANs

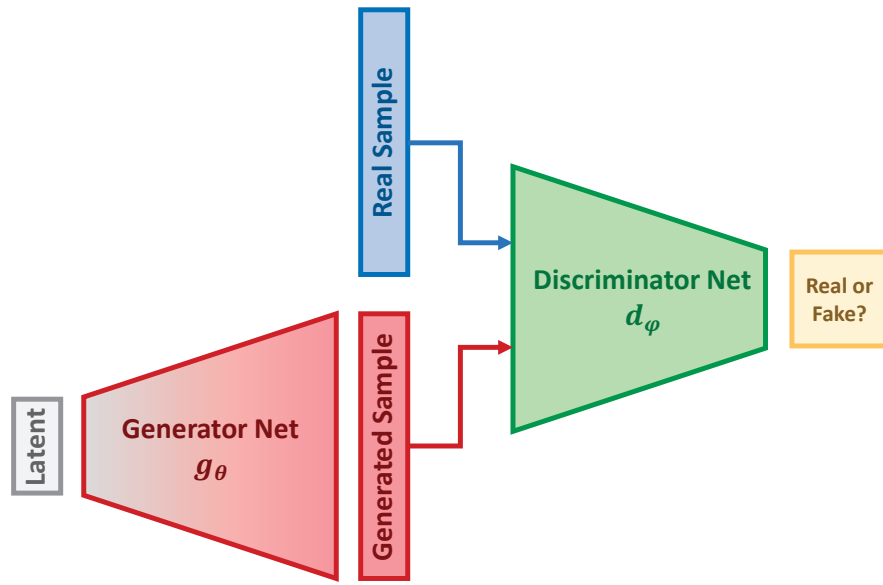
1-Wasserstein: $\mathsf{W}_1(P,Q) := \inf_{\pi_{X,Y}\in\Pi(P,Q)} \mathbb{E}_\pi \|X-Y\|$

Kantorovich-Rubinstein Duality: Equivalent representation

$$\mathsf{W}_1(\boldsymbol{P},\boldsymbol{Q}) = \sup_{\boldsymbol{f}\in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_{\boldsymbol{P}}\boldsymbol{f}(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{Q}}\boldsymbol{f}(\boldsymbol{Y})$$

Back to GANs:

- $P = P_X$ (X (real) data sample)
- $Q = Q_\theta$ ($Y = g_\theta(Z)$ gen. sample)
- $f = d_\phi$ (Lip_1 constraint)



$$\Longrightarrow \quad \boxed{\inf_\theta \mathsf{W}_1(P_X,Q_\theta) \cong \inf_\theta \sup_{\phi:\ d_\phi\in\mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}d_\phi(X) - \mathbb{E}d_\phi\big(g_\theta(Z)\big)}$$

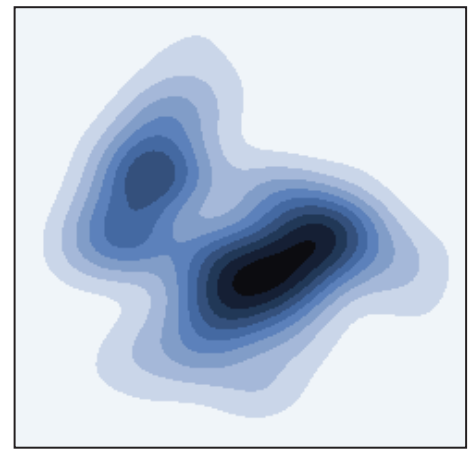
Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

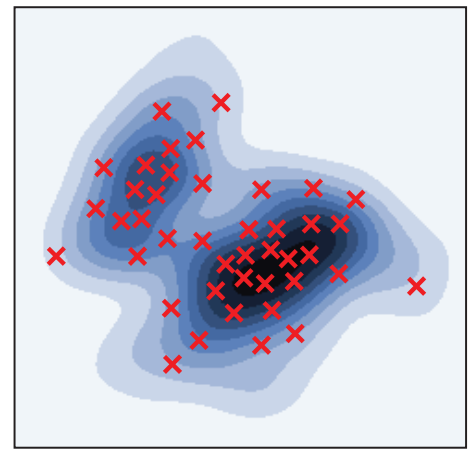
- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$



Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

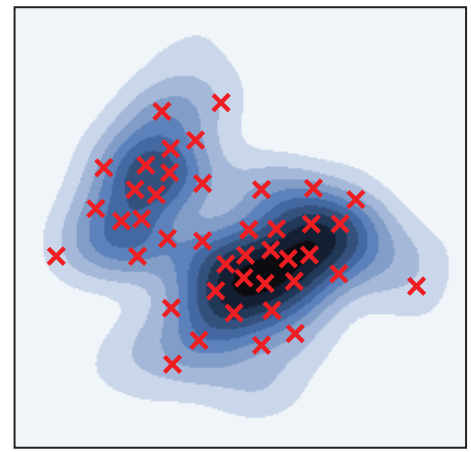
- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$



Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$
- Empirical distribution $P_n \triangleq \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$



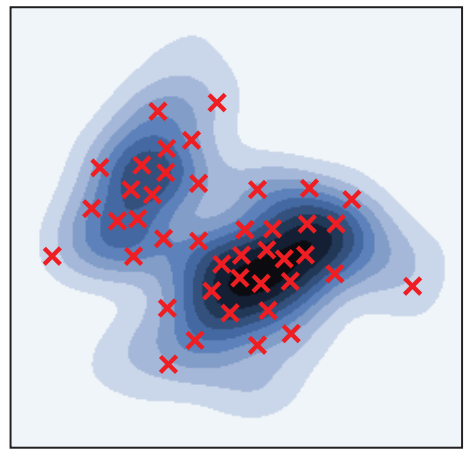
Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$
- Empirical distribution $P_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

$\Longrightarrow$ Inherently we work with $\mathsf{W}_1(P_n, Q_\theta)$

$$\left[\mathsf{W}_1(P_n, Q_\theta) \approx \mathsf{W}_1(P, Q_\theta) \text{ hopefully...}\right]$$



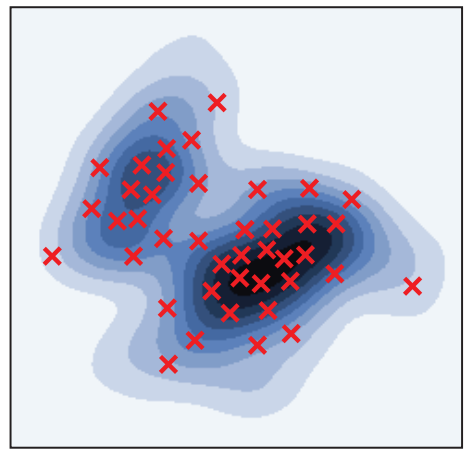
Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$
- Empirical distribution $P_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

$\Longrightarrow$ Inherently we work with $\mathsf{W}_1(P_n, Q_\theta)$

$$\left[\mathsf{W}_1(P_n, Q_\theta) \approx \mathsf{W}_1(P, Q_\theta) \text{ hopefully...}\right]$$



Theorem (Dudley'69)

For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \mathsf{Leb}(\mathbb{R}^d)$: $\mathbb{E}\mathsf{W}_1(P_n, P) \asymp n^{-\frac{1}{d}}$

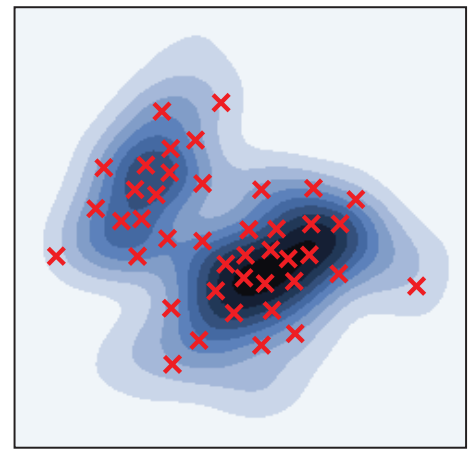
Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$
- Empirical distribution $P_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

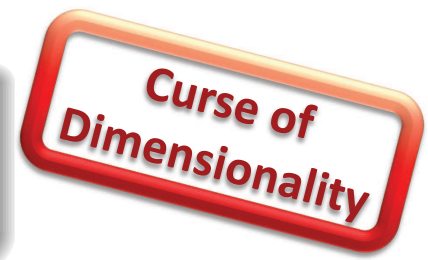
$\Longrightarrow$ Inherently we work with $\mathsf{W}_1(P_n, Q_\theta)$

$$\left[\mathsf{W}_1(P_n, Q_\theta) \approx \mathsf{W}_1(P, Q_\theta) \text{ hopefully...}\right]$$



Theorem (Dudley'69)

For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \mathsf{Leb}(\mathbb{R}^d)$: $\mathbb{E}\mathsf{W}_1(P_n, P) \asymp n^{-\frac{1}{d}}$



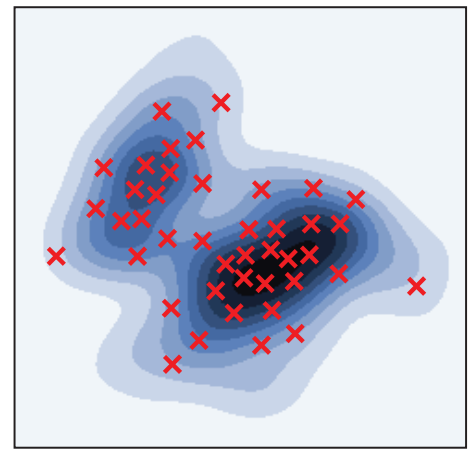
Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$
- Empirical distribution $P_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

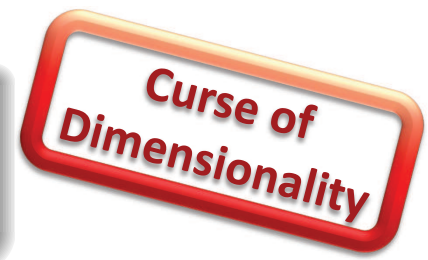
$\Longrightarrow$ Inherently we work with $\mathsf{W}_1(P_n, Q_\theta)$

$$\left[\mathsf{W}_1(P_n, Q_\theta) \approx \mathsf{W}_1(P, Q_\theta) \text{ hopefully...}\right]$$



Theorem (Dudley'69)

For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \mathsf{Leb}(\mathbb{R}^d)$: $\mathbb{E}\mathsf{W}_1(P_n, P) \asymp n^{-\frac{1}{d}}$



⊛ **Implication:** Too slow given dimensionality of real-world data

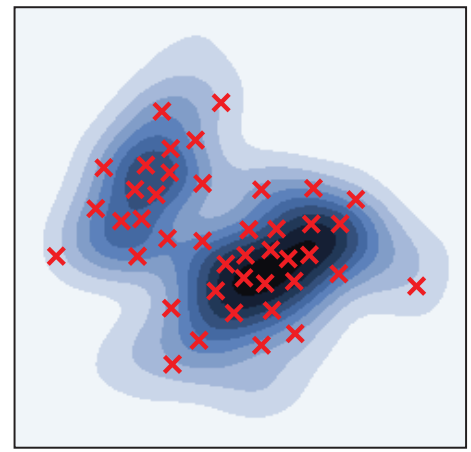
Empirical Approximation in High Dimensions

Empirical Approx.: In practice we don't have P, only data samples

- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$
- Empirical distribution $P_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

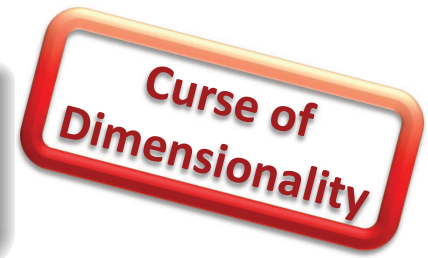
$\Longrightarrow$ Inherently we work with $\mathsf{W}_1(P_n, Q_\theta)$

$$\left[\mathsf{W}_1(P_n, Q_\theta) \approx \mathsf{W}_1(P, Q_\theta) \text{ hopefully...}\right]$$



Theorem (Dudley'69)

For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \mathsf{Leb}(\mathbb{R}^d)$: $\mathbb{E}\mathsf{W}_1(P_n, P) \asymp n^{-\frac{1}{d}}$



⊛ **Implication:** Too slow given dimensionality of real-world data

⊛ **Question:** Can we find a new W_1-like distance that alleviates CoD?

Smooth 1-Wasserstein Distance

Definition (ZG-Greenewald'19)

For $\sigma \geq 0$, the smooth 1-Wasserstein distance between P and Q is

$$\mathsf{W}_1^{(\sigma)}(P, Q) := \mathsf{W}_1(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma),$$

where $\mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ is a d-dimensional isotropic Gaussian.

Smooth 1-Wasserstein Distance

Definition (ZG-Greenewald'19)

For $\sigma \geq 0$, the smooth 1-Wasserstein distance between P and Q is

$$\mathsf{W}_1^{(\sigma)}(P, Q) := \mathsf{W}_1(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma),$$

where $\mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ is a d-dimensional isotropic Gaussian.

Interpretation: $X \sim P$, $Y \sim Q$ and $Z_1, Z_2 \sim \mathcal{N}_\sigma$

Smooth 1-Wasserstein Distance

Definition (ZG-Greenewald'19)

For $\sigma \geq 0$, the smooth 1-Wasserstein distance between P and Q is

$$\mathsf{W}_1^{(\sigma)}(P, Q) := \mathsf{W}_1(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma),$$

where $\mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ is a d-dimensional isotropic Gaussian.

Interpretation: $X \sim P$, $Y \sim Q$ and $Z_1, Z_2 \sim \mathcal{N}_\sigma$

$$X \perp Z_1 \implies X + Z_1 \sim P * \mathcal{N}_\sigma \quad \& \quad Y \perp Z_2 \implies Y + Z_2 \sim Q * \mathcal{N}_\sigma$$

Smooth 1-Wasserstein Distance

Definition (ZG-Greenewald'19)

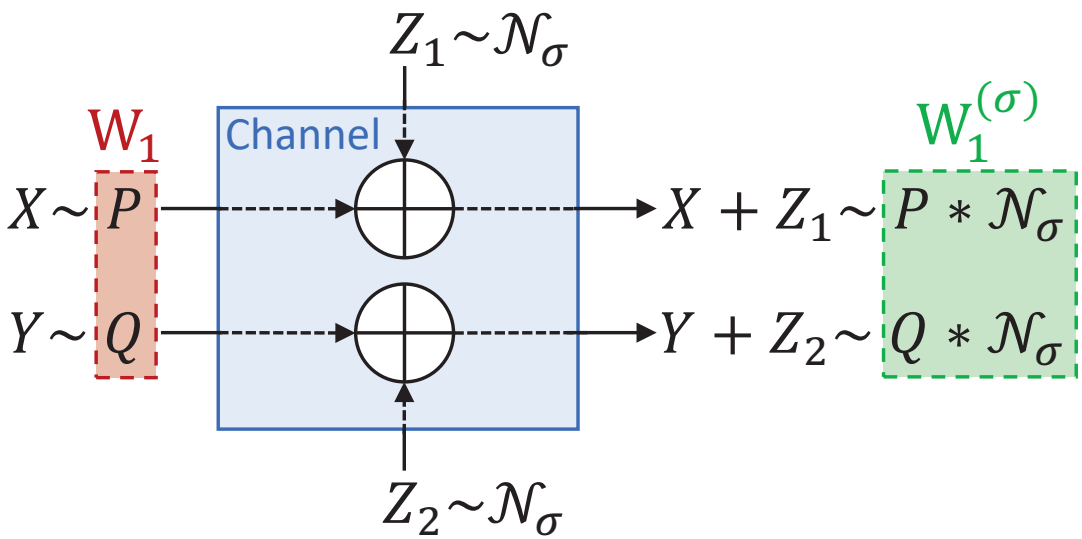
For $\sigma \geq 0$, the smooth 1-Wasserstein distance between P and Q is

$$\mathsf{W}_1^{(\sigma)}(P, Q) := \mathsf{W}_1(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma),$$

where $\mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ is a d-dimensional isotropic Gaussian.

Interpretation: $X \sim P$, $Y \sim Q$ and $Z_1, Z_2 \sim \mathcal{N}_\sigma$

$$X \perp Z_1 \implies X + Z_1 \sim P * \mathcal{N}_\sigma \quad \& \quad Y \perp Z_2 \implies Y + Z_2 \sim Q * \mathcal{N}_\sigma$$



Smooth 1-Wasserstein Distance

Definition (ZG-Greenewald'19)

For $\sigma \geq 0$, the smooth 1-Wasserstein distance between P and Q is

$$\mathsf{W}_1^{(\sigma)}(P, Q) := \mathsf{W}_1(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma),$$

where $\mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ is a d-dimensional isotropic Gaussian.

Interpretation: $X \sim P$, $Y \sim Q$ and $Z_1, Z_2 \sim \mathcal{N}_\sigma$

$$X \perp Z_1 \implies X + Z_1 \sim P * \mathcal{N}_\sigma \quad \& \quad Y \perp Z_2 \implies Y + Z_2 \sim Q * \mathcal{N}_\sigma$$

Retain KR Duality: $\mathsf{W}_1^{(\sigma)}$ is W_1 but between convolved distributions:

$$\mathsf{W}_1^{(\sigma)}(P, Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E} f(X + Z) - \mathbb{E} f(Y + Z)$$

Smooth 1-Wasserstein Distance

Definition (ZG-Greenewald'19)

For $\sigma \geq 0$, the smooth 1-Wasserstein distance between P and Q is

$$\mathsf{W}_1^{(\sigma)}(P, Q) := \mathsf{W}_1(P * \mathcal{N}_\sigma, Q * \mathcal{N}_\sigma),$$

where $\mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ is a d-dimensional isotropic Gaussian.

Interpretation: $X \sim P$, $Y \sim Q$ and $Z_1, Z_2 \sim \mathcal{N}_\sigma$

$$X \perp Z_1 \implies X + Z_1 \sim P * \mathcal{N}_\sigma \quad \& \quad Y \perp Z_2 \implies Y + Z_2 \sim Q * \mathcal{N}_\sigma$$

Retain KR Duality: $\mathsf{W}_1^{(\sigma)}$ is W_1 but between convolved distributions:

$$\begin{aligned} \mathsf{W}_1^{(\sigma)}(P, Q) &= \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E} f(X + Z) - \mathbb{E} f(Y + Z) \\ &= \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}(f * \varphi_\sigma)(X) - \mathbb{E}(f * \varphi_\sigma)(Y) \end{aligned}$$

Smooth 1-Wasserstein – Metric Structure

High Level: $\mathrm{W}_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

Smooth 1-Wasserstein – Metric Structure

High Level: $\mathsf{W}_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

Theorem (ZG-Greenewald'19)

$\left(\mathcal{P}_1(\mathbb{R}^d), \mathsf{W}_1^{(\sigma)}\right)$ *is metric space,* $\forall \sigma \geq 0$ *(and* $\mathsf{W}_1^{(\sigma)}$ *metrizes weak conv.).*

Smooth 1-Wasserstein – Metric Structure

High Level: $\mathsf{W}_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

Theorem (ZG-Greenewald'19)

$\left(\mathcal{P}_1(\mathbb{R}^d), \mathsf{W}_1^{(\sigma)}\right)$ *is metric space,* $\forall \sigma \geq 0$ *(and* $\mathsf{W}_1^{(\sigma)}$ *metrizes weak conv.).*

Key Idea for Pf.: Use characteristic functions $\Phi_P(t) := \mathbb{E}_P\left[e^{itX}\right]$ and:

Smooth 1-Wasserstein – Metric Structure

High Level: $\mathsf{W}_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

Theorem (ZG-Greenewald'19)

$\left(\mathcal{P}_1(\mathbb{R}^d), \mathsf{W}_1^{(\sigma)}\right)$ *is metric space,* $\forall \sigma \geq 0$ *(and* $\mathsf{W}_1^{(\sigma)}$ *metrizes weak conv.).*

Key Idea for Pf.: Use characteristic functions $\Phi_P(t) := \mathbb{E}_P\left[e^{itX}\right]$ and:

$$\Phi_{P*\mathcal{N}_\sigma}(t) = \Phi_P(t)\Phi_{\mathcal{N}_\sigma}(t) \text{ together with } \Phi_{\mathcal{N}_\sigma}(t) = e^{-\frac{\sigma^2\|t\|^2}{2}} \neq 0, \ \forall t.$$

Smooth 1-Wasserstein – Metric Structure

High Level: $\mathsf{W}_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

Theorem (ZG-Greenewald'19)

$\left(\mathcal{P}_1(\mathbb{R}^d), \mathsf{W}_1^{(\sigma)}\right)$ *is metric space,* $\forall \sigma \geq 0$ *(and* $\mathsf{W}_1^{(\sigma)}$ *metrizes weak conv.).*

Key Idea for Pf.: Use characteristic functions $\Phi_P(t) := \mathbb{E}_P\left[e^{itX}\right]$ and:

$$\Phi_{P * \mathcal{N}_\sigma}(t) = \Phi_P(t)\Phi_{\mathcal{N}_\sigma}(t) \text{ together with } \Phi_{\mathcal{N}_\sigma}(t) = e^{-\frac{\sigma^2\|t\|^2}{2}} \neq 0, \ \forall t.$$

Corollary (ZG-Greenewald'19)

Let $P_n, P \in \mathcal{P}_1(\mathbb{R}^d)$, $n \geq 1$. *Then:* $\mathsf{W}_1^{(\sigma)}(P_n, P) \to 0$ *iff* $\mathsf{W}_1(P_n, P) \to 0$

Smooth 1-Wasserstein – Metric Structure

High Level: $\mathsf{W}_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

Theorem (ZG-Greenewald'19)

$\left(\mathcal{P}_1(\mathbb{R}^d), \mathsf{W}_1^{(\sigma)}\right)$ *is metric space,* $\forall \sigma \geq 0$ *(and* $\mathsf{W}_1^{(\sigma)}$ *metrizes weak conv.).*

Key Idea for Pf.: Use characteristic functions $\Phi_P(t) := \mathbb{E}_P\left[e^{itX}\right]$ and:

$$\Phi_{P * \mathcal{N}_\sigma}(t) = \Phi_P(t)\Phi_{\mathcal{N}_\sigma}(t) \text{ together with } \Phi_{\mathcal{N}_\sigma}(t) = e^{-\frac{\sigma^2\|t\|^2}{2}} \neq 0, \ \forall t.$$

Corollary (ZG-Greenewald'19)

Let $P_n, P \in \mathcal{P}_1(\mathbb{R}^d)$, $n \geq 1$. *Then:* $\mathsf{W}_1^{(\sigma)}(P_n, P) \to 0$ *iff* $\mathsf{W}_1(P_n, P) \to 0$

⊛ $\mathsf{W}_1^{(\sigma)}$ and W_1 induce same topology

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:

1. $\mathsf{W}_1^{(\sigma)}(P, Q)$ *is continuous and mono. non-increasing in* $\sigma \in [0, +\infty)$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:

1. $\mathsf{W}_1^{(\sigma)}(P, Q)$ *is continuous and mono. non-increasing in* $\sigma \in [0, +\infty)$
2. $\lim_{\sigma \to 0} \mathsf{W}_1^{(\sigma)}(P, Q) = \mathsf{W}_1(P, Q)$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:

1. $\mathsf{W}_1^{(\sigma)}(P, Q)$ *is continuous and mono. non-increasing in* $\sigma \in [0, +\infty)$
2. $\lim_{\sigma \to 0} \mathsf{W}_1^{(\sigma)}(P, Q) = \mathsf{W}_1(P, Q)$
3. $\lim_{\sigma \to \infty} \mathsf{W}_1^{(\sigma)}(P, Q) \neq 0$, *for some* $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P,Q)$ is well-behaved func. of σ (fixed $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:

1. $\mathsf{W}_1^{(\sigma)}(P,Q)$ *is continuous and mono. non-increasing in* $\sigma \in [0,+\infty)$
2. $\lim_{\sigma\to 0} \mathsf{W}_1^{(\sigma)}(P,Q) = \mathsf{W}_1(P,Q)$
3. $\lim_{\sigma\to\infty} \mathsf{W}_1^{(\sigma)}(P,Q) \neq 0$, *for some* $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$

Pf. Items 1-2: Use dual form to derive stability lemma:

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P,Q)$ is well-behaved func. of σ (fixed $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:

1. $\mathsf{W}_1^{(\sigma)}(P,Q)$ *is continuous and mono. non-increasing in* $\sigma \in [0,+\infty)$
2. $\lim_{\sigma\to 0} \mathsf{W}_1^{(\sigma)}(P,Q) = \mathsf{W}_1(P,Q)$
3. $\lim_{\sigma\to\infty} \mathsf{W}_1^{(\sigma)}(P,Q) \neq 0$, *for some* $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$

Pf. Items 1-2: Use dual form to derive stability lemma:

Lemma (ZG-Greenewald'19)

For $\sigma_1 < \sigma_2$: $\mathsf{W}_1^{(\sigma_2)}(P,Q) \leq \mathsf{W}_1^{(\sigma_1)}(P,Q) \leq \mathsf{W}_1^{(\sigma_2)}(P,Q) + 2d\sqrt{\sigma_2^2 - \sigma_1^2}$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P,Q)$ is well-behaved func. of σ (fixed $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:

1. $\mathsf{W}_1^{(\sigma)}(P,Q)$ *is continuous and mono. non-increasing in* $\sigma \in [0,+\infty)$
2. $\lim_{\sigma\to 0} \mathsf{W}_1^{(\sigma)}(P,Q) = \mathsf{W}_1(P,Q)$
3. $\lim_{\sigma\to\infty} \mathsf{W}_1^{(\sigma)}(P,Q) \neq 0$, *for some* $P,Q \in \mathcal{P}_1(\mathbb{R}^d)$

Pf. Items 1-2: Use dual form to derive stability lemma:

Lemma (ZG-Greenewald'19)

For $\sigma_1 < \sigma_2$: $\mathsf{W}_1^{(\sigma_2)}(P,Q) \le \mathsf{W}_1^{(\sigma_1)}(P,Q) \le \mathsf{W}_1^{(\sigma_2)}(P,Q) + 2d\sqrt{\sigma_2^2 - \sigma_1^2}$

Pf. Item 3: $\mathsf{W}_1^{(\sigma)}(\delta_x,\delta_y) = \mathsf{W}_1(\mathcal{N}(x,\sigma^2\mathrm{I}_d),\mathcal{N}(y,\sigma^2\mathrm{I}_d)) = \|x-y\|$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

- $\{\pi_k\}$, *s.t.* $\pi_k \in \Pi(P * \mathcal{N}_{\sigma_k}, Q * \mathcal{N}_{\sigma_k})$ *is optimal for* $\mathsf{W}_1^{(\sigma_k)}(P, Q)$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

- $\{\pi_k\}$, *s.t.* $\pi_k \in \Pi(P * \mathcal{N}_{\sigma_k}, Q * \mathcal{N}_{\sigma_k})$ *is optimal for* $\mathsf{W}_1^{(\sigma_k)}(P, Q)$

Then $\pi_k \xrightarrow[k\to\infty]{} \pi$ *(weakly) such that:*

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

- $\{\pi_k\}$, *s.t.* $\pi_k \in \Pi(P * \mathcal{N}_{\sigma_k}, Q * \mathcal{N}_{\sigma_k})$ *is optimal for* $\mathsf{W}_1^{(\sigma_k)}(P, Q)$

Then $\pi_k \xrightarrow[k\to\infty]{} \pi$ *(weakly) such that:*

- $\pi \in \Pi(P, Q)$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

- $\{\pi_k\}$, *s.t.* $\pi_k \in \Pi(P * \mathcal{N}_{\sigma_k}, Q * \mathcal{N}_{\sigma_k})$ *is optimal for* $\mathsf{W}_1^{(\sigma_k)}(P, Q)$

Then $\pi_k \xrightarrow[k\to\infty]{} \pi$ *(weakly) such that:*

- $\pi \in \Pi(P, Q)$
- π *is optimal for* $\mathsf{W}_1(P, Q)$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

- $\{\pi_k\}$, *s.t.* $\pi_k \in \Pi(P * \mathcal{N}_{\sigma_k}, Q * \mathcal{N}_{\sigma_k})$ *is optimal for* $\mathsf{W}_1^{(\sigma_k)}(P, Q)$

Then $\pi_k \xrightarrow[k\to\infty]{} \pi$ *(weakly) such that:*

- $\pi \in \Pi(P, Q)$
- π *is optimal for* $\mathsf{W}_1(P, Q)$

Comments:

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

- $\{\pi_k\}$, *s.t.* $\pi_k \in \Pi(P * \mathcal{N}_{\sigma_k}, Q * \mathcal{N}_{\sigma_k})$ *is optimal for* $\mathsf{W}_1^{(\sigma_k)}(P, Q)$

Then $\pi_k \xrightarrow[k\to\infty]{} \pi$ *(weakly) such that:*

- $\pi \in \Pi(P, Q)$
- π *is optimal for* $\mathsf{W}_1(P, Q)$

Comments:

- **In words:** Not only opt. values converge, but also optimizers

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

- $\{\pi_k\}$, *s.t.* $\pi_k \in \Pi(P * \mathcal{N}_{\sigma_k}, Q * \mathcal{N}_{\sigma_k})$ *is optimal for* $\mathsf{W}_1^{(\sigma_k)}(P, Q)$

Then $\pi_k \xrightarrow[k\to\infty]{} \pi$ *(weakly) such that:*

- $\pi \in \Pi(P, Q)$
- π *is optimal for* $\mathsf{W}_1(P, Q)$

Comments:

- **In words:** Not only opt. values converge, but also optimizers
- **Generalization:** Also holds for $\sigma_k \searrow \sigma > 0$

Smooth 1-Wasserstein – Function of σ

High Level: $\mathsf{W}_1^{(\sigma)}(P, Q)$ is well-behaved func. of σ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

Theorem (ZG-Greenewald'19)

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ and $\sigma_k \searrow 0$. Let:

- $\{\pi_k\}$, *s.t.* $\pi_k \in \Pi(P * \mathcal{N}_{\sigma_k}, Q * \mathcal{N}_{\sigma_k})$ *is optimal for* $\mathsf{W}_1^{(\sigma_k)}(P, Q)$

Then $\pi_k \xrightarrow[k\to\infty]{} \pi$ *(weakly) such that:*

- $\pi \in \Pi(P, Q)$
- π *is optimal for* $\mathsf{W}_1(P, Q)$

Comments:

- **In words:** Not only opt. values converge, but also optimizers
- **Generalization:** Also holds for $\sigma_k \searrow \sigma > 0$
- **Pf. Idea:** Γ-convergence & Tightness of $\Pi(P, Q)$

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get limit distribution

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get limit distribution

Theorem (ZG-Greenewald-Polyanskiy-Weed'19)

For any $d \geq 1$, $\sigma > 0$ *and sub-Gaussian* P*:* $\quad \mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}}$

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get limit distribution

Theorem (ZG-Greenewald-Polyanskiy-Weed'19)

For any $d \geq 1$, $\sigma > 0$ *and sub-Gaussian* P*:* $\quad \mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}}$

Theorem (ZG-Kato'20)

Under same assumptions: $\quad \sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P) \xrightarrow{d} \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} G_P^{(\sigma)}(f)$,
where $G_P^{(\sigma)}$ *is a tight centered Gaussian process in* $\ell^\infty(\mathsf{Lip}_1(\mathbb{R}^d))$ *with*
$\mathbb{E}\left[G_P^{(\sigma)}(f) G_P^{(\sigma)}(g)\right] = \mathsf{Cov}_P(f * \varphi_\sigma, g * \varphi_\sigma), \quad \forall f, g, \in \mathsf{Lip}_1(\mathbb{R}^d)$

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get limit distribution

Theorem (ZG-Greenewald-Polyanskiy-Weed'19)

For any $d \geq 1$, $\sigma > 0$ *and sub-Gaussian* P*:* $\quad \mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}}$

Theorem (ZG-Kato'20)

Under same assumptions: $\quad \sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P) \xrightarrow{d} \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} G_P^{(\sigma)}(f)$,
where $G_P^{(\sigma)}$ *is a tight centered Gaussian process in* $\ell^\infty(\mathsf{Lip}_1(\mathbb{R}^d))$ *with*
$\mathbb{E}\left[G_P^{(\sigma)}(f) G_P^{(\sigma)}(g)\right] = \mathsf{Cov}_P(f * \varphi_\sigma, g * \varphi_\sigma), \quad \forall f, g, \in \mathsf{Lip}_1(\mathbb{R}^d)$

Comments:

- Limit distribution shows $n^{-\frac{1}{2}}$ rate is sharp

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get limit distribution

Theorem (ZG-Greenewald-Polyanskiy-Weed'19)

For any $d \geq 1$, $\sigma > 0$ and sub-Gaussian P: $\quad \mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}}$

Theorem (ZG-Kato'20)

Under same assumptions: $\quad \sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P) \xrightarrow{d} \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} G_P^{(\sigma)}(f)$,
where $G_P^{(\sigma)}$ is a tight centered Gaussian process in $\ell^\infty(\mathsf{Lip}_1(\mathbb{R}^d))$ with
$\mathbb{E}\left[G_P^{(\sigma)}(f) G_P^{(\sigma)}(g)\right] = \mathsf{Cov}_P(f * \varphi_\sigma, g * \varphi_\sigma), \quad \forall f, g, \in \mathsf{Lip}_1(\mathbb{R}^d)$

Comments:

- Limit distribution shows $n^{-\frac{1}{2}}$ rate is sharp
- W_1 limit dist. known only for $d = 1$ ($\mathsf{W}_1(P_n, P) = \|F_n - F\|_{L^1(\mathbb{R})}$)

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get limit distribution

Theorem (ZG-Greenewald-Polyanskiy-Weed'19)

For any $d \geq 1$, $\sigma > 0$ *and sub-Gaussian* P*:* $\quad \mathbb{E}\mathsf{W}_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}}$

Theorem (ZG-Kato'20)

Under same assumptions: $\quad \sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P) \xrightarrow{d} \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} G_P^{(\sigma)}(f)$,
where $G_P^{(\sigma)}$ *is a tight centered Gaussian process in* $\ell^\infty(\mathsf{Lip}_1(\mathbb{R}^d))$ *with*
$\mathbb{E}\left[G_P^{(\sigma)}(f) G_P^{(\sigma)}(g)\right] = \mathsf{Cov}_P(f * \varphi_\sigma, g * \varphi_\sigma), \quad \forall f, g, \in \mathsf{Lip}_1(\mathbb{R}^d)$

Comments:

- Limit distribution shows $n^{-\frac{1}{2}}$ rate is sharp
- W_1 limit dist. known only for $d = 1$ ($\mathsf{W}_1(P_n, P) = \|F_n - F\|_{L^1(\mathbb{R})}$)
- $\mathsf{W}_1^{(\sigma)}$ has a stable limit (characterized) in all d!

Limit Distribution Proof Roadmap

1. **KR Duality:** Let $\mathcal{F}_\sigma := \{f * \varphi_\sigma : \ f \in \mathrm{Lip}_1(\mathbb{R}^d)\}$ and write

Limit Distribution Proof Roadmap

1. **KR Duality:** Let $\mathcal{F}_\sigma := \{f * \varphi_\sigma : \ f \in \mathsf{Lip}_1(\mathbb{R}^d)\}$ and write

$$\mathsf{W}_1^{(\sigma)}(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P[f * \varphi_\sigma] - \mathbb{E}_Q[f * \varphi_\sigma] = \sup_{g \in \mathcal{F}_\sigma} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$$

Limit Distribution Proof Roadmap

1. **KR Duality:** Let $\mathcal{F}_\sigma := \{f * \varphi_\sigma : \ f \in \mathsf{Lip}_1(\mathbb{R}^d)\}$ and write

$$\mathsf{W}_1^{(\sigma)}(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P[f * \varphi_\sigma] - \mathbb{E}_Q[f * \varphi_\sigma] = \sup_{g \in \mathcal{F}_\sigma} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$$

$\implies$ **Emp. $\mathsf{W}_1^{(\sigma)}$:**

$$\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P) = \sqrt{n} \sup_{g \in \mathcal{F}_\sigma} \underbrace{\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_P g}_{\triangleq (P_n - P)g}$$

Limit Distribution Proof Roadmap

1. **KR Duality:** Let $\mathcal{F}_\sigma := \{f * \varphi_\sigma : \ f \in \mathsf{Lip}_1(\mathbb{R}^d)\}$ and write

$$\mathsf{W}_1^{(\sigma)}(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P[f * \varphi_\sigma] - \mathbb{E}_Q[f * \varphi_\sigma] = \sup_{g \in \mathcal{F}_\sigma} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$$

$\Longrightarrow$ **Emp. $\mathsf{W}_1^{(\sigma)}$:**

$$\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P) = \sqrt{n} \sup_{g \in \mathcal{F}_\sigma} \underbrace{\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_P g}_{\triangleq (P_n - P)g}$$

✱ **Objective:** $\mathbb{G}_n = \left(\sqrt{n}(P_n - P)g\right)_{g \in \mathcal{F}_\sigma} \overset{w}{\to}$ GP

Limit Distribution Proof Roadmap

1. **KR Duality:** Let $\mathcal{F}_\sigma := \{f * \varphi_\sigma : \ f \in \mathsf{Lip}_1(\mathbb{R}^d)\}$ and write

$$\mathsf{W}_1^{(\sigma)}(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P[f * \varphi_\sigma] - \mathbb{E}_Q[f * \varphi_\sigma] = \sup_{g \in \mathcal{F}_\sigma} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$$

$\implies$ **Emp. $\mathsf{W}_1^{(\sigma)}$:**

$$\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P) = \sqrt{n} \sup_{g \in \mathcal{F}_\sigma} \underbrace{\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_P g}_{\triangleq (P_n - P)g}$$

$\circledast$ **Objective:** $\mathbb{G}_n = \left(\sqrt{n}(P_n - P)g\right)_{g \in \mathcal{F}_\sigma} \overset{w}{\to}$ GP $\iff$ $\mathcal{F}_\sigma$ is $P-$Donsker

Limit Distribution Proof Roadmap

1. **KR Duality:** Let $\mathcal{F}_\sigma := \{f * \varphi_\sigma : \ f \in \mathsf{Lip}_1(\mathbb{R}^d)\}$ and write

$$\mathsf{W}_1^{(\sigma)}(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P[f * \varphi_\sigma] - \mathbb{E}_Q[f * \varphi_\sigma] = \sup_{g \in \mathcal{F}_\sigma} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$$

$\Longrightarrow$ **Emp. $\mathsf{W}_1^{(\sigma)}$:**

$$\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P) = \sqrt{n} \sup_{g \in \mathcal{F}_\sigma} \underbrace{\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_P g}_{\triangleq (P_n - P)g}$$

✱ **Objective:** $\mathbb{G}_n = \left(\sqrt{n}(P_n - P)g\right)_{g \in \mathcal{F}_\sigma} \overset{w}{\to}$ GP $\iff$ $\mathcal{F}_\sigma$ is $P-$Donsker

2. **[van der Vaart'96]:** $K \subseteq \mathbb{R}^d$ bdd & convex: $\mathcal{F}_\sigma\Big|_{\mathcal{K}} \subseteq C_M^N(\mathcal{K}),\ N > \frac{d}{2}$

Limit Distribution Proof Roadmap

1. **KR Duality:** Let $\mathcal{F}_\sigma := \{f * \varphi_\sigma : \ f \in \mathsf{Lip}_1(\mathbb{R}^d)\}$ and write

$$\mathsf{W}_1^{(\sigma)}(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P[f * \varphi_\sigma] - \mathbb{E}_Q[f * \varphi_\sigma] = \sup_{g \in \mathcal{F}_\sigma} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$$

$\Longrightarrow$ **Emp. $\mathsf{W}_1^{(\sigma)}$:**

$$\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P) = \sqrt{n} \sup_{g \in \mathcal{F}_\sigma} \underbrace{\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_P g}_{\triangleq (P_n - P)g}$$

$\circledast$ **Objective:** $\mathbb{G}_n = \left(\sqrt{n}(P_n - P)g\right)_{g \in \mathcal{F}_\sigma} \overset{w}{\to}$ GP $\iff$ $\mathcal{F}_\sigma$ is $P-$Donsker

2. **[van der Vaart'96]:** $K \subseteq \mathbb{R}^d$ bdd & convex: $\mathcal{F}_\sigma\Big|_{\mathcal{K}} \subseteq C_M^N(\mathcal{K}), \ N > \frac{d}{2}$

$\Longrightarrow$ **$\mathcal{F}_\sigma$ is P-Donsker:**

$$\mathbb{G}_n \overset{w}{\to} G_P \text{ (Gaussian process)}$$

Limit Distribution Proof Roadmap

1. **KR Duality:** Let $\mathcal{F}_\sigma := \{f * \varphi_\sigma : \ f \in \mathsf{Lip}_1(\mathbb{R}^d)\}$ and write

$$\mathsf{W}_1^{(\sigma)}(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P[f*\varphi_\sigma] - \mathbb{E}_Q[f*\varphi_\sigma] = \sup_{g \in \mathcal{F}_\sigma} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$$

$\Longrightarrow$ **Emp. $\mathsf{W}_1^{(\sigma)}$:**

$$\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P) = \sqrt{n} \sup_{g \in \mathcal{F}_\sigma} \underbrace{\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}_P g}_{\triangleq (P_n - P)g}$$

✱ **Objective:** $\mathbb{G}_n = \left(\sqrt{n}(P_n - P)g\right)_{g \in \mathcal{F}_\sigma} \overset{w}{\to}$ GP $\iff$ $\mathcal{F}_\sigma$ is $P-$Donsker

2. **[van der Vaart'96]:** $K \subseteq \mathbb{R}^d$ bdd & convex: $\mathcal{F}_\sigma\Big|_{\mathcal{K}} \subseteq C_M^N(\mathcal{K}),\ N > \frac{d}{2}$

$\Longrightarrow$ **$\mathcal{F}_\sigma$ is P-Donsker:**

$$\mathbb{G}_n \overset{w}{\to} G_P \text{ (Gaussian process)}$$

3. **Continuous Mapping Theorem:**

$$\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n,P) \overset{d}{\to} \sup_{g \in \mathcal{F}_\sigma} G_P(g)$$

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \leq \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \leq \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

- **Guarantees:** For $\delta = \mathsf{W}_1^{(\sigma)}$ and $\epsilon = n^{-1/2}$

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \leq \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

- **Guarantees:** For $\delta = \mathsf{W}_1^{(\sigma)}$ and $\epsilon = n^{-1/2}$

  [Zhang *et al*'18]: $\mathsf{W}_1^{(\sigma)}\big(P, Q_{\hat{\theta}_n}\big) - \text{OPT}$

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \leq \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

- **Guarantees:** For $\delta = \mathsf{W}_1^{(\sigma)}$ and $\epsilon = n^{-1/2}$

  [Zhang *et al*'18]: $\mathsf{W}_1^{(\sigma)}\big(P, Q_{\hat{\theta}_n}\big) - \text{OPT} \leq 2\mathsf{W}_1^{(\sigma)}\left(P_n, P\right) + n^{-\frac{1}{2}}$

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \leq \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

- **Guarantees:** For $\delta = \mathsf{W}_1^{(\sigma)}$ and $\epsilon = n^{-1/2}$

  [Zhang *et al*'18]: $\mathsf{W}_1^{(\sigma)}\big(P, Q_{\hat{\theta}_n}\big) - \text{OPT} \leq 2\mathsf{W}_1^{(\sigma)}\left(P_n, P\right) + n^{-\frac{1}{2}} \lesssim n^{-\frac{1}{2}}$

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \leq \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

- **Guarantees:** For $\delta = \mathsf{W}_1^{(\sigma)}$ and $\epsilon = n^{-1/2}$

  [Zhang *et al*'18]: $\mathsf{W}_1^{(\sigma)}\big(P, Q_{\hat{\theta}_n}\big) - \text{OPT} \leq 2\mathsf{W}_1^{(\sigma)}\big(P_n, P\big) + n^{-\frac{1}{2}} \lesssim n^{-\frac{1}{2}}$

  [ZG-Kato'20]: $\mathbb{E}\big[\hat{\theta}_n - \theta^\star\big] \asymp n^{-\frac{1}{2}}$ (in fact, $\sqrt{n}(\hat{\theta}_n - \theta^\star)$ converges in dist.)

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \leq \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

- **Guarantees:** For $\delta = \mathsf{W}_1^{(\sigma)}$ and $\epsilon = n^{-1/2}$

  [Zhang *et al*'18]: $\mathsf{W}_1^{(\sigma)}\big(P, Q_{\hat{\theta}_n}\big) - \text{OPT} \leq 2\mathsf{W}_1^{(\sigma)}(P_n, P) + n^{-\frac{1}{2}} \lesssim n^{-\frac{1}{2}}$

  [ZG-Kato'20]: $\mathbb{E}\big[\hat{\theta}_n - \theta^\star\big] \asymp n^{-\frac{1}{2}}$ (in fact, $\sqrt{n}(\hat{\theta}_n - \theta^\star)$ converges in dist.)

Computation: How to compute $\mathsf{W}_1^{(\sigma)}(P, Q)$?

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \le \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

- **Guarantees:** For $\delta = \mathsf{W}_1^{(\sigma)}$ and $\epsilon = n^{-1/2}$

  [Zhang *et al*'18]: $\mathsf{W}_1^{(\sigma)}\big(P, Q_{\hat{\theta}_n}\big) - \mathrm{OPT} \le 2\mathsf{W}_1^{(\sigma)}(P_n, P) + n^{-\frac{1}{2}} \lesssim n^{-\frac{1}{2}}$

  [ZG-Kato'20]: $\mathbb{E}\big[\hat{\theta}_n - \theta^\star\big] \asymp n^{-\frac{1}{2}}$ (in fact, $\sqrt{n}(\hat{\theta}_n - \theta^\star)$ converges in dist.)

Computation: How to compute $\mathsf{W}_1^{(\sigma)}(P, Q)$?

- **Naive:** $\sup_{f \in \mathsf{Lip}_1} \mathbb{E}f(X+Z) - \mathbb{E}f(Y+Z)$ & MC sampling for Z

Generative Models: Generalization & Computation

Generalization: Want to solve OPT $:= \inf_\theta \delta\left(P, Q_\theta\right)$ exactly (find $\theta^\star$)

- **In practice:** Can solve $\inf_\theta \delta\left(P_n, Q_\theta\right)$ approximately:

$$\text{Find} \quad \hat{\theta}_n \quad \text{s.t.} \quad \delta\big(P_n, Q_{\hat{\theta}_n}\big) \le \inf_\theta \delta\big(P_n, Q_\theta\big) + \epsilon$$

- **Guarantees:** For $\delta = \mathsf{W}_1^{(\sigma)}$ and $\epsilon = n^{-1/2}$

  [Zhang *et al*'18]: $\mathsf{W}_1^{(\sigma)}\big(P, Q_{\hat{\theta}_n}\big) - \text{OPT} \le 2\mathsf{W}_1^{(\sigma)}\big(P_n, P\big) + n^{-\frac{1}{2}} \lesssim n^{-\frac{1}{2}}$

  [ZG-Kato'20]: $\mathbb{E}\big[\hat{\theta}_n - \theta^\star\big] \asymp n^{-\frac{1}{2}}$ (in fact, $\sqrt{n}(\hat{\theta}_n - \theta^\star)$ converges in dist.)

Computation: How to compute $\mathsf{W}_1^{(\sigma)}(P, Q)$?

- **Naive:** $\sup_{f \in \mathsf{Lip}_1} \mathbb{E}f(X+Z) - \mathbb{E}f(Y+Z)$ & MC sampling for Z
- **Advanced:** $\sup_{f \in \mathsf{Lip}_1} \mathbb{E}(f * \varphi_\sigma)(X) - \mathbb{E}(f * \varphi_\sigma)(Y)$ & implement '$*$'

Smooth Wasserstein GAN: Initial Empirical Results

[Katsman-Kim-ZG-Greenewald'20]

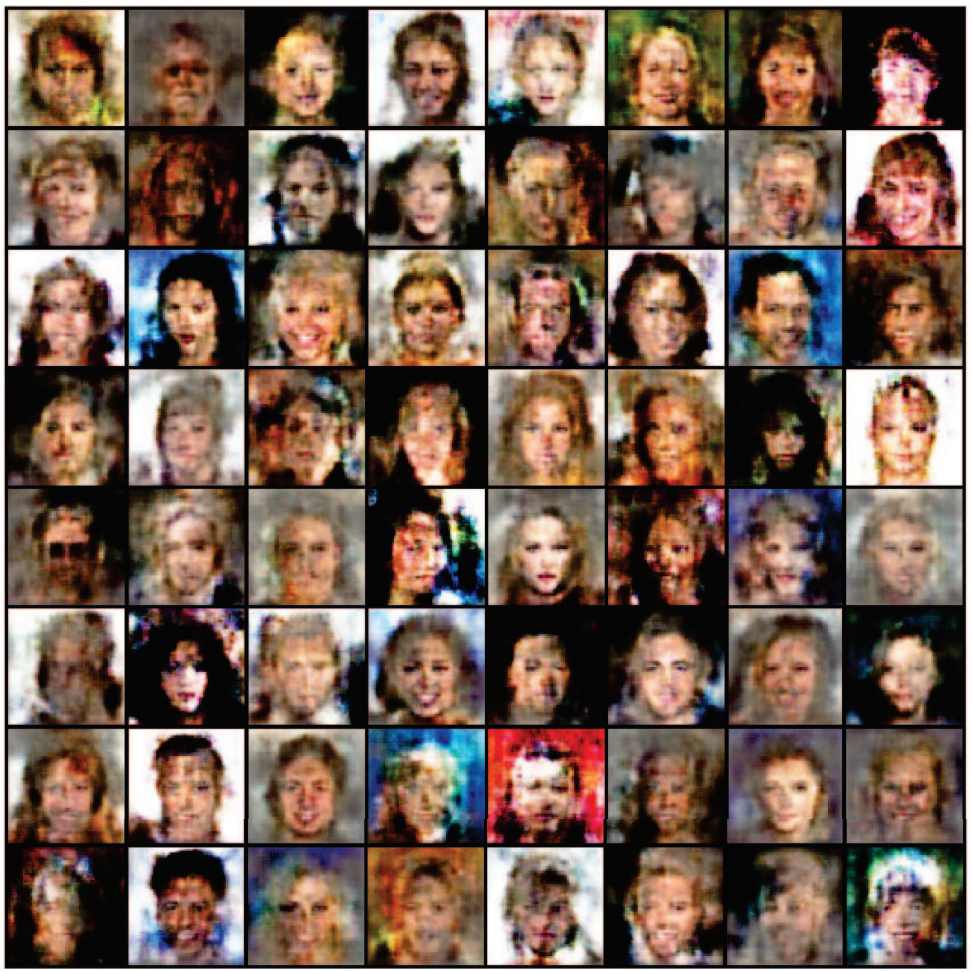
Setup: f parametrized by 2-layer NN & conv. computed in closed form

Smooth Wasserstein GAN: Initial Empirical Results

[Katsman-Kim-ZG-Greenewald'20]

Setup: f parametrized by 2-layer NN & conv. computed in closed form

Baseline W-GAN ($\sigma = 0$)

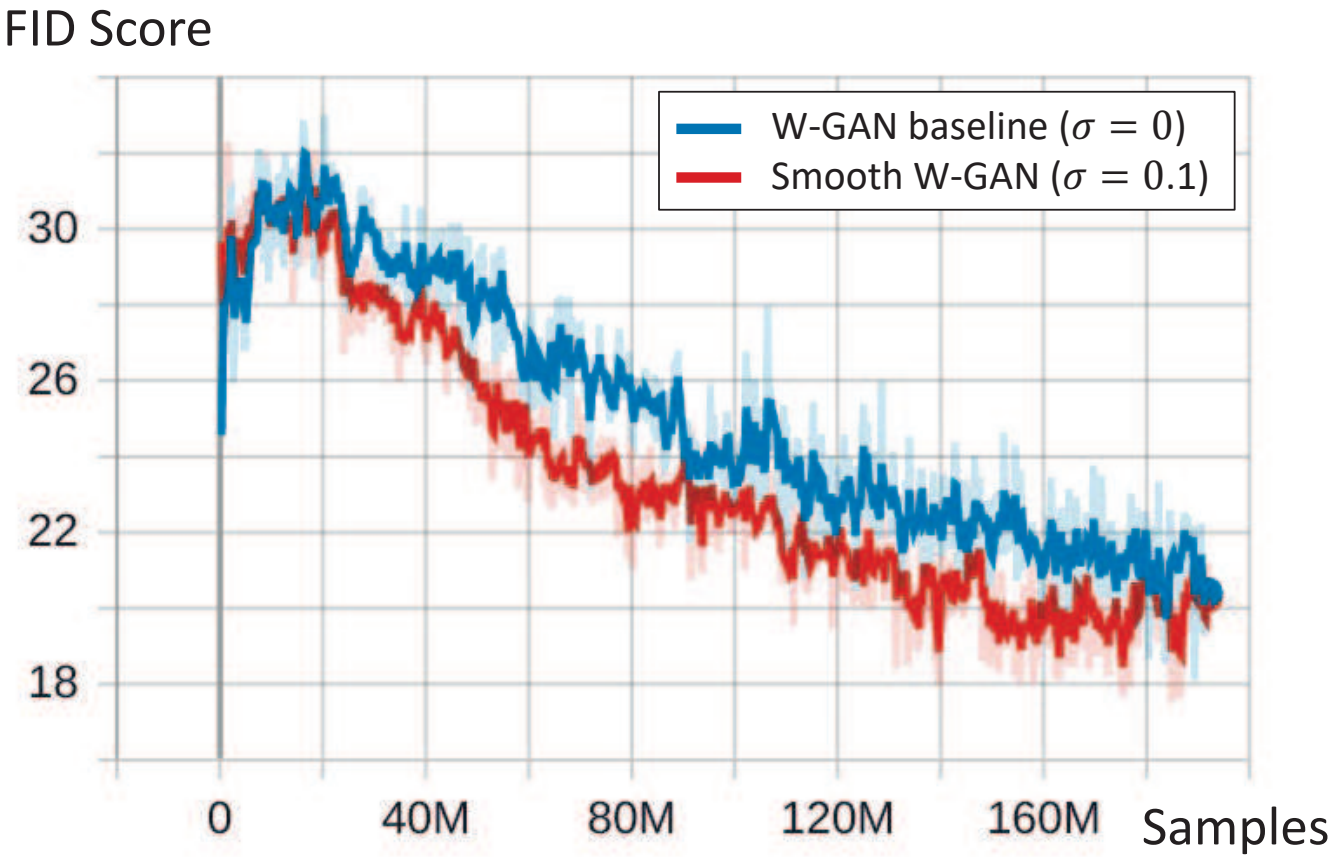


Smooth W-GAN ($\sigma = 0.1$)

Smooth Wasserstein GAN: Initial Empirical Results

[Katsman-Kim-ZG-Greenewald'20]

Setup: f parametrized by 2-layer NN & conv. computed in closed form

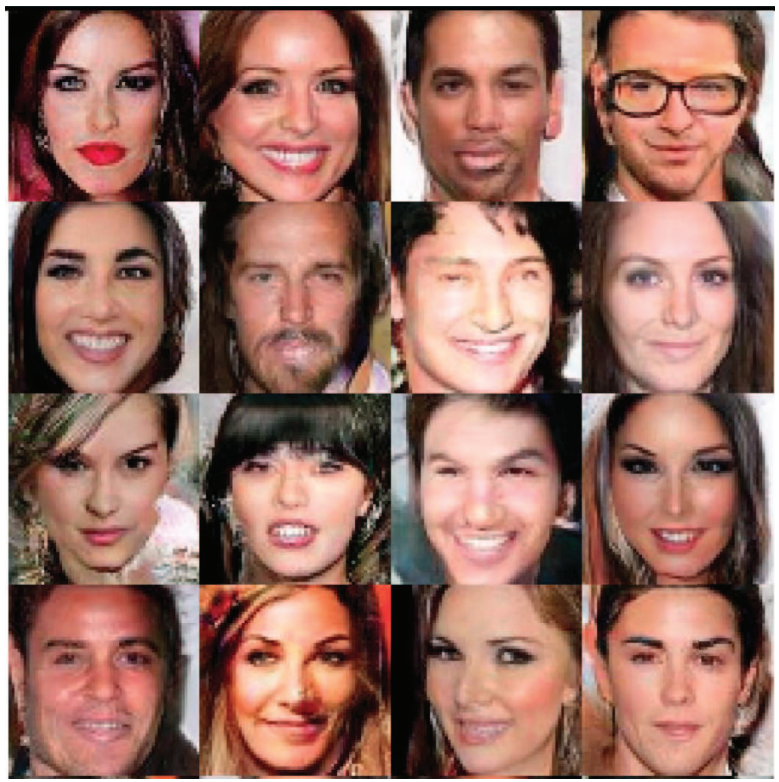


⊛ FID = Fréchet Inception Distance

Smooth Wasserstein GAN: Initial Empirical Results

[Katsman-Kim-ZG-Greenewald'20]

Setup: f parametrized by 2-layer NN & conv. computed in closed form



WGAN-div [Wu et al'2018]: FID $= 16.0$

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality
  - Well-behaved function of smoothing parameter & recovers W_1 in limit

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality
  - Well-behaved function of smoothing parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ empirical convergence in all dimensions

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality
  - Well-behaved function of smoothing parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ empirical convergence in all dimensions
  - Limit distribution for $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P)$ in all dimensions

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality
  - Well-behaved function of smoothing parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ empirical convergence in all dimensions
  - Limit distribution for $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P)$ in all dimensions
- **Smooth W-GAN:** Next-generation systems rooted in theory

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality
  - Well-behaved function of smoothing parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ empirical convergence in all dimensions
  - Limit distribution for $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P)$ in all dimensions
- **Smooth W-GAN:** Next-generation systems rooted in theory
  - Attain competitive performance

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality
  - Well-behaved function of smoothing parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ empirical convergence in all dimensions
  - Limit distribution for $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P)$ in all dimensions
- **Smooth W-GAN:** Next-generation systems rooted in theory
  - Attain competitive performance
  - Consume fewer resources

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality
  - Well-behaved function of smoothing parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ empirical convergence in all dimensions
  - Limit distribution for $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P)$ in all dimensions
- **Smooth W-GAN:** Next-generation systems rooted in theory
  - Attain competitive performance
  - Consume fewer resources
  - Backed up by (useful) generalization guarantees

Summary

- **Classic $\mathbf{W_1}$:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ that is popular in modern ML
  - W-GAN produces SOTA synthetic samples
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth $\mathbf{W_1^{(\sigma)}}$:** Convolve distributions w/ Gaussian kernel
  - Inherits metric structure of W_1 & duality
  - Well-behaved function of smoothing parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ empirical convergence in all dimensions
  - Limit distribution for $\sqrt{n}\mathsf{W}_1^{(\sigma)}(P_n, P)$ in all dimensions
- **Smooth W-GAN:** Next-generation systems rooted in theory
  - Attain competitive performance
  - Consume fewer resources
  - Backed up by (useful) generalization guarantees

Thank you!

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Pros:

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Pros:

- Regularizer favors weakly-dependent coupling

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Pros:

- Regularizer favors weakly-dependent coupling
- $\mathsf{S}_c^{(\epsilon)}(P, Q)$ is strongly convex optimization problem

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Pros:

- Regularizer favors weakly-dependent coupling
- $\mathsf{S}_c^{(\epsilon)}(P, Q)$ is strongly convex optimization problem
- Two-sample statistical efficiency (for certain costs):

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\quad \mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Pros:

- Regularizer favors weakly-dependent coupling
- $\mathsf{S}_c^{(\epsilon)}(P, Q)$ is strongly convex optimization problem
- Two-sample statistical efficiency (for certain costs):

Theorem (Genevay et al'19)

For $\mathcal{C}^\infty$ and Lipschitz cost c, and compactly supported $P, Q \in \mathcal{P}(\mathbb{R}^d)$:

$$\mathbb{E}\left|S_c^{(\epsilon)}(P_n, Q_n) - S_c^{(\epsilon)}(P, Q)\right| \lesssim e^{\frac{c}{\epsilon}} \left(1 + \frac{1}{\epsilon^{\lfloor d/2 \rfloor}}\right) n^{-\frac{1}{2}}$$

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Cons (Specializing to distance cost):

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Cons (Specializing to distance cost):

- $c(x, y) = \|x - y\|$ does not fall under the theorem's framework

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Cons (Specializing to distance cost):

- $c(x, y) = \|x - y\|$ does not fall under the theorem's framework
- $\mathsf{S}_1^{(\epsilon)}$ is not a metric on $\mathcal{P}_1(\mathbb{R}^d)$ for any $\epsilon > 0$ (nor Sinkhorn loss)

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Cons (Specializing to distance cost):

- $c(x, y) = \|x - y\|$ does not fall under the theorem's framework
- $\mathsf{S}_1^{(\epsilon)}$ is not a metric on $\mathcal{P}_1(\mathbb{R}^d)$ for any $\epsilon > 0$ (nor Sinkhorn loss)
- Dual form includes two potentials and regularizer:

$$S_1^{(\epsilon)}(P, Q) = \sup_{u,v \in \mathcal{C}(\mathbb{R}^d)} \mathbb{E}_{P \otimes Q} \left[u(X) + v(Y) - \epsilon e^{\frac{u(X)+v(Y)-\|X-Y\|}{\epsilon}} \right] + \epsilon$$

Entropic Optimal Transport

Definition (Entropic OT)

For $\epsilon > 0$ and $c(x, y)$: $\mathsf{S}_c^{(\epsilon)}(P, Q) := \inf_{\pi_{X,Y} \in \Pi(P,Q)} \mathbb{E}_\pi c(X, Y) + \epsilon I_\pi(X; Y)$

where $I_\pi(X; Y) \triangleq \mathsf{D}_{\mathsf{KL}}(\pi_{X,Y} \| P \otimes Q)$ is the mutual information

Cons (Specializing to distance cost):

- $c(x, y) = \|x - y\|$ does not fall under the theorem's framework
- $\mathsf{S}_1^{(\epsilon)}$ is <u>not</u> a metric on $\mathcal{P}_1(\mathbb{R}^d)$ for any $\epsilon > 0$ (nor Sinkhorn loss)
- Dual form includes two potentials and regularizer:

$$S_1^{(\epsilon)}(P, Q) = \sup_{u,v \in \mathcal{C}(\mathbb{R}^d)} \mathbb{E}_{P \otimes Q} \left[u(X) + v(Y) - \epsilon e^{\frac{u(X)+v(Y)-\|X-Y\|}{\epsilon}} \right] + \epsilon$$

$\Longrightarrow$ No direct correspondence to minimax GAN formulation