Smooth Wasserstein Distance: Metric Structure and Statistical Efficiency

Ziv Goldfeld[1], Kristjan Greenewald[2]

[1]Cornell University
[2]MIT-IBM Watson AI Lab

AISTATS 2020

Motivation: Generative Modeling

Generative Modeling:

Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$

Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

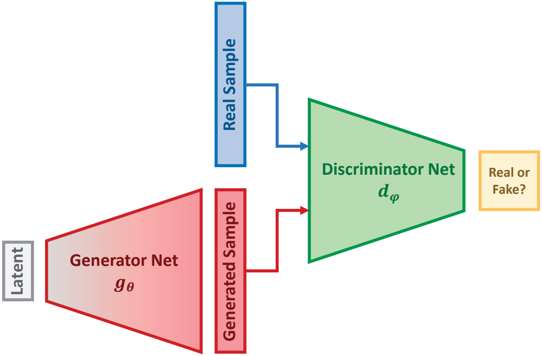
Generative Adversarial Networks: State-of-the-art generative models

Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models



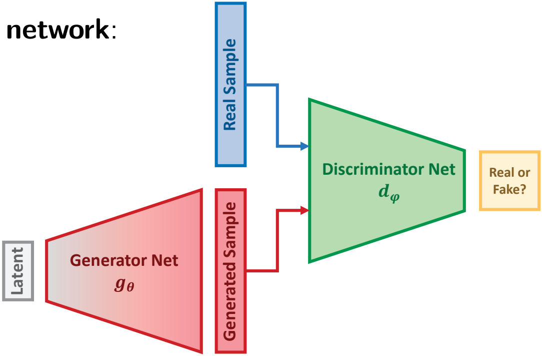
Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models

- Shape noise via **Generator network**:



Motivation: Generative Modeling

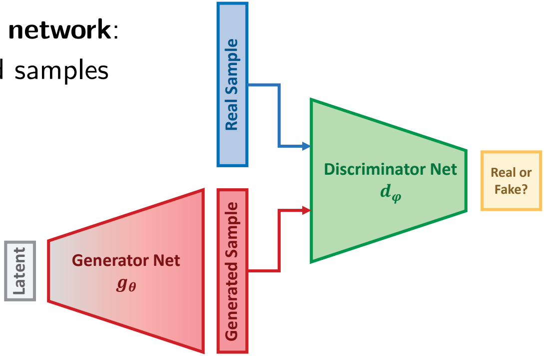
Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models

- Shape noise via **Generator network**:

  $\Longrightarrow$ Produces synthesized samples



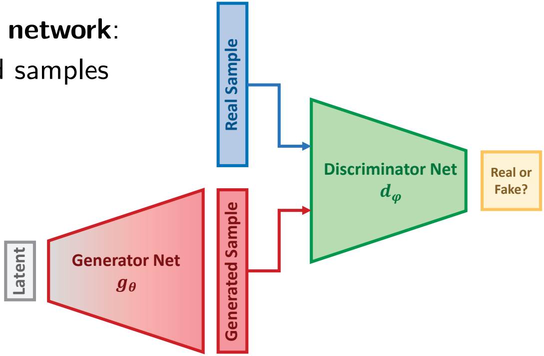
Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models

- Shape noise via **Generator network**:
  $\implies$ Produces synthesized samples
- **Discriminator network**:



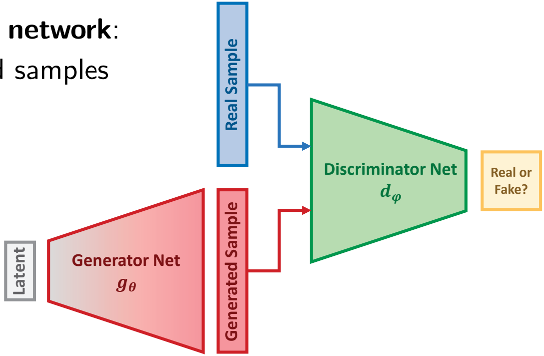
Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models

- Shape noise via **Generator network**:
  $\Longrightarrow$ Produces synthesized samples
- **Discriminator network**:
  $\Longrightarrow$ tells real vs. fake



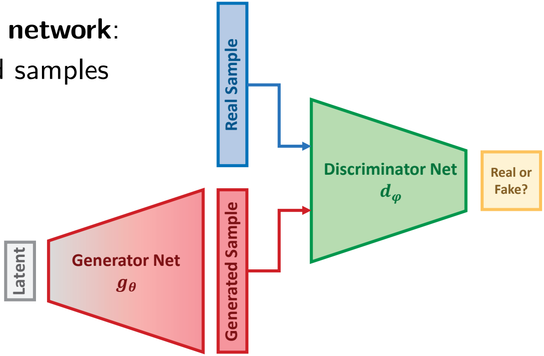
Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models

- Shape noise via **Generator network**:
  $\Longrightarrow$ Produces synthesized samples
- **Discriminator network**:
  $\Longrightarrow$ tells real vs. fake
- Alternating optimization



Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models

- Shape noise via **Generator network**:
  $\Longrightarrow$ Produces synthesized samples
- **Discriminator network**:
  $\Longrightarrow$ tells real vs. fake
- Alternating optimization

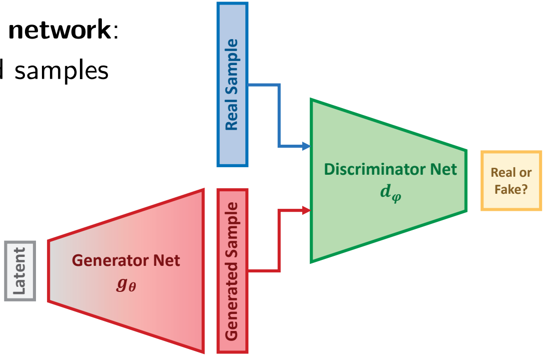
Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models

- Shape noise via **Generator network**:
  $\implies$ Produces synthesized samples
- **Discriminator network**:
  $\implies$ tells real vs. fake
- Alternating optimization



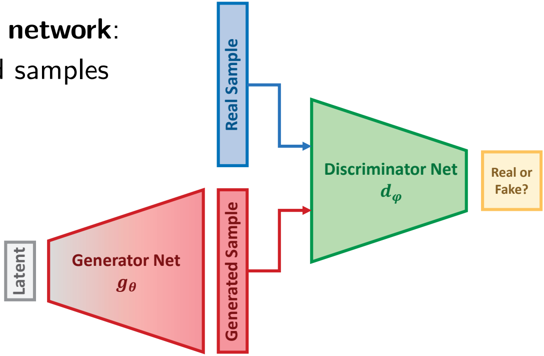
Motivation: Generative Modeling

Generative Modeling:

- **Input:** Unlabeled data $\{x_i\}_{i=1}^n$ i.i.d. from (unknown) $P \in \mathcal{P}(\mathbb{R}^d)$
- **Goal:** Learn underlying structure in data (e.g., $Q_\theta \approx P$)

Generative Adversarial Networks: State-of-the-art generative models

- Shape noise via **Generator network**:
  $\Longrightarrow$ Produces synthesized samples
- **Discriminator network**:
  $\Longrightarrow$ tells real vs. fake
- Alternating optimization



Question:

How to quantify $Q_\theta \approx P$?

Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

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

Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

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

$\Longrightarrow$ **Principled Objective:** $\boxed{\inf_{\theta} \delta(Q_{\theta}, P)}$

Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

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

$\Longrightarrow$ **Principled Objective:** $\boxed{\inf_{\theta} \delta(Q_{\theta}, P)}$

⊛ Coincides with minimax formulation when δ is 1-Wasserstein distance:

Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

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

$\Longrightarrow$ **Principled Objective:** $\boxed{\inf_\theta \delta(Q_\theta, P)}$

⊛ Coincides with minimax formulation when δ is 1-Wasserstein distance:

Definition (1-Wasserstein distance)

For $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$: $\mathsf{W}_1(P, Q) := \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$,

where $\Pi(P, Q)$ is the set of all couplings of P and Q.

Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

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

$\implies$ **Principled Objective:** $\boxed{\inf_\theta \delta(Q_\theta, P)}$

⊛ Coincides with minimax formulation when δ is 1-Wasserstein distance:

Definition (1-Wasserstein distance)

For $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$: $\mathsf{W}_1(P, Q) := \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi \|X - Y\|$,

where $\Pi(P, Q)$ is the set of all couplings of P and Q.

⊛ **Pros:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$ & Robust to supp. mismatch $\mathsf{W}_1(P, Q) < \infty$

Duality & Wasserstein GAN

Kantorovich-Rubinstein Duality:

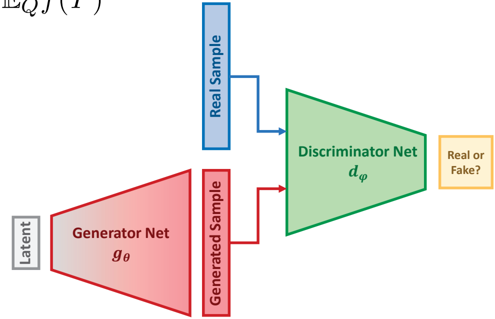
$$\mathsf{W}_1(P, Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)$$

Duality & Wasserstein GAN

Kantorovich-Rubinstein Duality:

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

Correspondence to GANs:



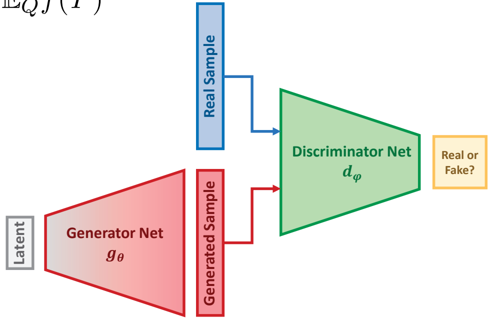
Duality & Wasserstein GAN

Kantorovich-Rubinstein Duality:

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

Correspondence to GANs:

- P = data distribution



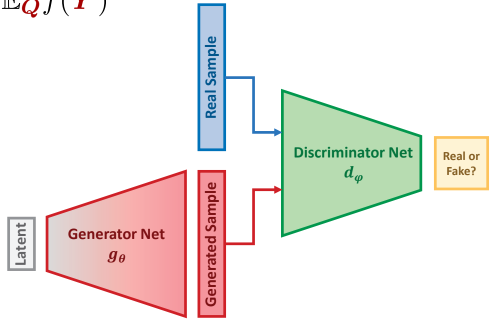
Duality & Wasserstein GAN

Kantorovich-Rubinstein Duality:

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

Correspondence to GANs:

- P = data distribution
- $Q = Q_\theta$ model



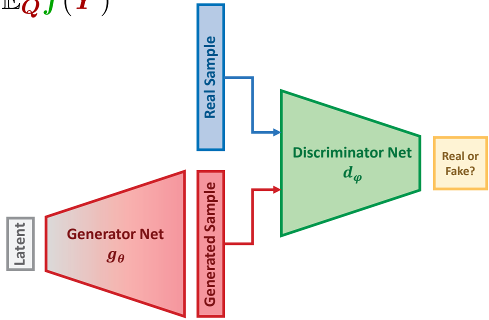
Duality & Wasserstein GAN

Kantorovich-Rubinstein Duality:

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

Correspondence to GANs:

- P = data distribution
- $Q = Q_\theta$ model
- $f = d_\varphi$ disc. (Lip_1 constraint)



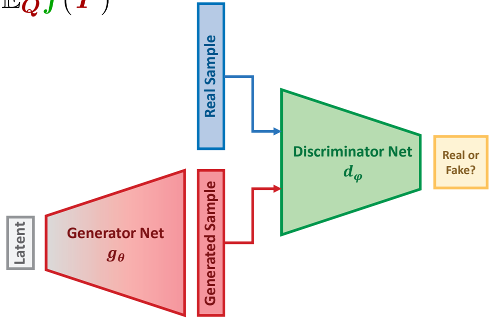
Duality & Wasserstein GAN

Kantorovich-Rubinstein Duality:

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

Correspondence to GANs:

- P = data distribution
- $Q = Q_\theta$ model
- $f = d_\varphi$ disc. (Lip_1 constraint)



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

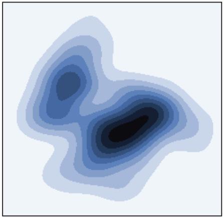
Empirical Approximation in High Dimensions

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

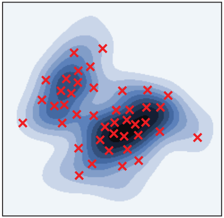
- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$



Empirical Approximation in High Dimensions

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

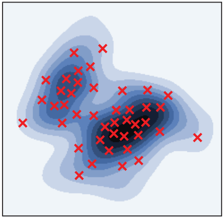
- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$



Empirical Approximation in High Dimensions

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

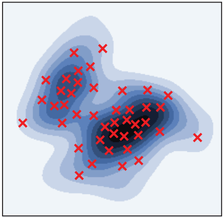
- $\{X_i\}_{i=1}^n$ are i.i.d. samples from $P \in \mathcal{P}_1(\mathbb{R}^d)$
- Empirical distribution $P_n \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$



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



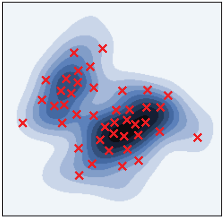
$\Longrightarrow$ Inherently we work with $\mathsf{W}_1(P_n, Q_\theta)$

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

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



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

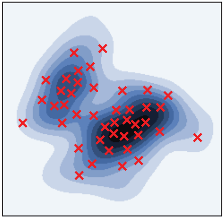
Empirical Approximation in High Dimensions

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 $\mathsf{W}_1(P_n, Q_\theta)$

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



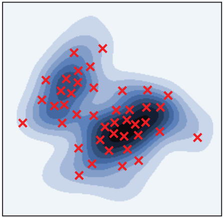
Theorem (Dudley'69)

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

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)

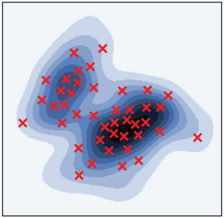
For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \mathsf{Leb}(\mathbb{R}^d)$: $\mathbb{E}\mathsf{W}_1(P_n, P) \asymp n^{-\frac{1}{d}}$

Curse of Dimensionality

⊛ **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 $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \mathrm{Leb}(\mathbb{R}^d)$: $\mathbb{E}\mathsf{W}_1(P_n, P) \asymp n^{-\frac{1}{d}}$

Curse of Dimensionality

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

⊛ **Goal:** Define a new metric that alleviates CoD

Smooth 1-Wasserstein Distance

Definition

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

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

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

Smooth 1-Wasserstein Distance

Definition

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

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

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

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

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

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 $Z_1, Z_2 \sim \mathcal{N}_\sigma$

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

Smooth 1-Wasserstein Distance

Definition

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

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

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 $Z_1, Z_2 \sim \mathcal{N}_\sigma$

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

$\Longrightarrow$ W_1 distance between smoothed distributions

Smooth 1-Wasserstein Distance

Definition

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

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

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 $Z_1, Z_2 \sim \mathcal{N}_\sigma$

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

$\Longrightarrow$ W_1 distance between smoothed distributions

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

$$\mathsf{W}_1^{(\sigma)}(P,Q) = \sup_{f \in \mathsf{Lip}_1(\mathbb{R}^d)} \mathbb{E} f(X + Z_1) - \mathbb{E} f(Y + Z_2)$$

Smooth 1-Wasserstein – Metric Structure

High Level: $\mathsf{W}_1^{(\sigma)}$ 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

$\left(\mathcal{P}_1(\mathbb{R}^d), \mathsf{W}_1^{(\sigma)}\right)$ *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

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

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

Smooth 1-Wasserstein – Metric Structure

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

Theorem

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

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

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

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

$P, Q_i, \in \mathcal{P}(\mathbb{R}^d)$, $i = 1, \ldots$ *Then:* $\mathsf{W}_1^{(\sigma)}(Q_i, P) \to 0$ *iff* $\mathsf{W}_1(Q_i, P) \to 0$

Smooth 1-Wasserstein – Metric Structure

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

Theorem

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

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

$P, Q_i, \in \mathcal{P}(\mathbb{R}^d)$, $i = 1, \ldots$ *Then:* $\mathsf{W}_1^{(\sigma)}(Q_i, P) \to 0$ *iff* $\mathsf{W}_1(Q_i, P) \to 0$

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

Smooth 1-Wasserstein – Function of Noise Std

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 Noise Std

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

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

Smooth 1-Wasserstein – Function of Noise Std

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

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 Noise Std

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

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 Noise Std

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

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 Noise Std

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

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 Noise Std

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

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

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

Smooth 1-Wasserstein – Function of Noise Std

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

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

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

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

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get concentration

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get concentration

Theorem

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 concentration

Theorem

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

Under same assumptions: denote $\mathcal{X} \triangleq \mathrm{supp}(\mu)$ and suppose $\mathsf{diam}(\mathcal{X}) < \infty$, *where* $\mathsf{diam}(\mathcal{X}) = \sup_{x \neq y \in \mathcal{X}} \|x - y\|$. *For any $t > 0$ we have*

$$\mathbb{P}_{\mu^{\otimes n}}\left(\left|\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu) - \mathbb{E}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)\right| \geq t\right) \leq 2e^{-\frac{2t^2 n}{\mathsf{diam}(\mathcal{X})^2}}$$

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get concentration

Theorem

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

Under same assumptions: denote $\mathcal{X} \triangleq \operatorname{supp}(\mu)$ *and suppose* $\mathrm{diam}(\mathcal{X}) < \infty$*, where* $\mathrm{diam}(\mathcal{X}) = \sup_{x \neq y \in \mathcal{X}} \|x - y\|$*. For any* $t > 0$ *we have*

$$\mathbb{P}_{\mu^{\otimes n}}\left(\left|\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu) - \mathbb{E}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)\right| \geq t\right) \leq 2e^{-\frac{2t^2 n}{\mathrm{diam}(\mathcal{X})^2}}$$

Comments:

- Achieves $n^{-\frac{1}{2}}$ bias rate vs $n^{-1/d}$ for W_1 - via maximal TV coupling arg

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get concentration

Theorem

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

Under same assumptions: denote $\mathcal{X} \triangleq \mathrm{supp}(\mu)$ and suppose $\mathrm{diam}(\mathcal{X}) < \infty$, *where* $\mathrm{diam}(\mathcal{X}) = \sup_{x \neq y \in \mathcal{X}} \|x - y\|$. *For any $t > 0$ we have*

$$\mathbb{P}_{\mu^{\otimes n}}\left(\left|\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu) - \mathbb{E}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)\right| \geq t\right) \leq 2e^{-\frac{2t^2 n}{\mathrm{diam}(\mathcal{X})^2}}$$

Comments:

- Achieves $n^{-\frac{1}{2}}$ bias rate vs $n^{-1/d}$ for W_1 - via maximal TV coupling arg
- "Variance" bounded at the same asymptotic rate - achieved via McDiarmid's inequality & KR duality

Smooth 1-Wasserstein – Statistical Efficiency

High Level: Alleviate curse of dimensionality & get concentration

Theorem

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

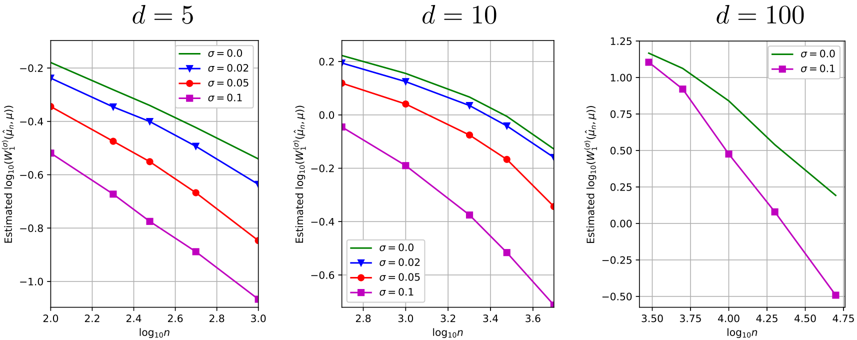
Under same assumptions: denote $\mathcal{X} \triangleq \operatorname{supp}(\mu)$ and suppose $\mathsf{diam}(\mathcal{X}) < \infty$, *where* $\mathsf{diam}(\mathcal{X}) = \sup_{x \neq y \in \mathcal{X}} \|x - y\|$. *For any $t > 0$ we have*

$$\mathbb{P}_{\mu^{\otimes n}}\left(\left|\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu) - \mathbb{E}\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)\right| \geq t\right) \leq 2e^{-\frac{2t^2 n}{\mathsf{diam}(\mathcal{X})^2}}$$

Comments:

- Achieves $n^{-\frac{1}{2}}$ bias rate vs $n^{-1/d}$ for W_1 - via maximal TV coupling arg
- "Variance" bounded at the same asymptotic rate - achieved via McDiarmid's inequality & KR duality
- Paper: more general statements allowing for non-Gaussian convolutions

Synthetic Data Experiments



Convergence of $\mathsf{W}_1^{(\sigma)}(\hat{\mu}_n, \mu)$ as a function of the number of samples n for various values of σ, shown in log-log space. The measure μ is the uniform distribution over $[0, 1]^d$. Note that $\sigma = 0$ corresponds to the vanilla Wasserstein distance, which converges slower than GOT (observe the difference in slopes), especially with larger d.

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)
  - Wasserstein GAN produces outstanding empirical results

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)
  - Wasserstein GAN produces outstanding empirical results
  - Empirical approximation is slow $n^{-\frac{1}{d}}$

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)
  - Wasserstein GAN produces outstanding empirical results
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth 1-Wasserstein:** Convolve distributions w/ Gaussians

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)
  - Wasserstein GAN produces outstanding empirical results
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth 1-Wasserstein:** Convolve distributions w/ Gaussians
  - Inherits metric structure & duality from the Wasserstein distance

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)
  - Wasserstein GAN produces outstanding empirical results
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth 1-Wasserstein:** Convolve distributions w/ Gaussians
  - Inherits metric structure & duality from the Wasserstein distance
  - Well-behaved function of noise parameter & recovers W_1 in limit

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)
  - Wasserstein GAN produces outstanding empirical results
  - Empirical approximation is slow $n^{-\frac{1}{d}}$

- **Smooth 1-Wasserstein:** Convolve distributions w/ Gaussians
  - Inherits metric structure & duality from the Wasserstein distance
  - Well-behaved function of noise parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ convergence of empirical approximation in all dimensions

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)
  - Wasserstein GAN produces outstanding empirical results
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth 1-Wasserstein:** Convolve distributions w/ Gaussians
  - Inherits metric structure & duality from the Wasserstein distance
  - Well-behaved function of noise parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ convergence of empirical approximation in all dimensions

Recap

- **Classic 1-Wasserstein:** Metric on $\mathcal{P}_1(\mathbb{R}^d)$
  - Popular in machine learning (esp. generative modeling)
  - Wasserstein GAN produces outstanding empirical results
  - Empirical approximation is slow $n^{-\frac{1}{d}}$
- **Smooth 1-Wasserstein:** Convolve distributions w/ Gaussians
  - Inherits metric structure & duality from the Wasserstein distance
  - Well-behaved function of noise parameter & recovers W_1 in limit
  - Fast $n^{-\frac{1}{2}}$ convergence of empirical approximation in all dimensions

Thank you!