Smooth Wasserstein Distance: Metric Structure and Statistical Efficiency

Ziv Goldfeld\textsuperscript{1}, Kristjan Greenewald\textsuperscript{2}

\textsuperscript{1}Cornell University
\textsuperscript{2}MIT-IBM Watson AI Lab

AISTATS 2020
Motivation: Generative Modeling

Generative Modeling:

Input:
Unlabeled data \( \{ x_i \} \) \( n \) \( i.i.d. \) 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:

\[ \Rightarrow \]
Produces synthesized samples

Discriminator network:

\[ \Rightarrow \]
tells real vs. fake

Alternating optimization

Question:
How to quantify \( Q_{\theta} \approx P \)?
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**:

![Diagram of Generative Adversarial Networks](image)
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:**
  \[ \Rightarrow \text{Produces synthesized samples} \]

![Diagram of Generative Adversarial Networks](image-url)
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:**
  \( \Rightarrow \) Produces synthesized samples
- **Discriminator network:**

![Diagram of Generative Adversarial Networks]

Real Sample \( \rightarrow \) Discriminator Net \( d_\varphi \)

Real or Fake?

Generated Sample \( \rightarrow \) Generator Net \( g_\theta \)
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:**
  \[ \Longrightarrow \text{ Produces synthesized samples} \]

- **Discriminator network:**
  \[ \Longrightarrow \text{ tells real vs. fake} \]
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:**
  \[ \Rightarrow \text{Produces synthesized samples} \]
- **Discriminator network:**
  \[ \Rightarrow \text{tells real vs. fake} \]
- 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**: 
  \( \Rightarrow \) Produces synthesized samples
- **Discriminator network**: 
  \( \Rightarrow \) 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**: Produces synthesized samples
- **Discriminator network**: 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:**
  \[ \implies \text{Produces synthesized samples} \]
- **Discriminator network:**
  \[ \implies \text{tells real vs. fake} \]
- Alternating optimization

**Question:**

How to quantify \( Q_\theta \approx P \)?
Quantification: Via statistical divergence
Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

\[ \delta : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \rightarrow [0, +\infty) \text{ s.t. } \delta(P, Q) = 0 \iff P = Q \]
Motivation: Generative Modeling (Cont.)

**Quantification:** Via statistical divergence

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

\[ \implies \quad \text{Principled Objective:} \quad \inf_{\theta} \delta(Q_\theta, P) \]
Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

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

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

☆ Coincides with minimax formulation when $\delta$ is 1-Wasserstein distance:
Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

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

\[ \iff \textbf{Principled Objective:} \quad \inf_{\theta} \delta(Q_\theta, P) \]

\( \heartsuit \) Coincides with minimax formulation when \( \delta \) is 1-Wasserstein distance:

**Definition (1-Wasserstein distance)**

For \( P, Q \in \mathcal{P}_1(\mathbb{R}^d) \):

\[ W_1(P, Q) := \inf_{\pi \in \Pi(P, Q)} \mathbb{E}_{\pi} \|X - Y\|, \]

where \( \Pi(P, Q) \) is the set of all couplings of \( P \) and \( Q \).
Motivation: Generative Modeling (Cont.)

Quantification: Via statistical divergence

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

\[ \implies \textbf{Principled Objective:} \quad \inf_{\theta} \delta(Q_{\theta}, P) \]

Pros: Metric on \( \mathcal{P}_1(\mathbb{R}^d) \) & Robust to supp. mismatch \( W_1(P, Q) < \infty \)

Definition (1-Wasserstein distance)

For \( P, Q \in \mathcal{P}_1(\mathbb{R}^d) \):

\[ W_1(P, Q) := \inf_{\pi \in \Pi(P, Q)} \mathbb{E}_{\pi} \|X - Y\|, \]

where \( \Pi(P, Q) \) is the set of all couplings of \( P \) and \( Q \).
Kantorovich-Rubinstein Duality:

\[ W_1(P, Q) = \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \]
Kantorovich-Rubinstein Duality:

\[ W_1(P, Q) = \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \]

Correspondence to GANs:

- **Generator Net** $g_{\theta}$
- **Discriminator Net** $d_{\varphi}$
- **Real Sample**
- **Generated Sample**
- **Real or Fake?**
Kantorovich-Rubinstein Duality:

\[
W_1(P, Q) = \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)
\]

Correspondence to GANs:

- \(P = \) data distribution
Kantorovich-Rubinstein Duality:

\[ W_1(P, Q) = \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \]

Correspondence to GANs:

- \( P = \) data distribution
- \( Q = Q_\theta \) model
Duality & Wasserstein GAN

Kantorovich-Rubinstein Duality:

\[ W_1(P, Q) = \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \]

Correspondence to GANs:

- \( P \) = data distribution
- \( Q = Q_\theta \) model
- \( f = d_\phi \) disc. (\( \text{Lip}_1 \) constraint)
Kantorovich-Rubinstein Duality:

\[ W_1(P, Q) = \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \]

Correspondence to GANs:

- \( P \) = data distribution
- \( Q = Q_\theta \) model
- \( f = d_\varphi \) disc. (\( \text{Lip}_1 \) constraint)

\[ \Rightarrow \inf_{\theta} W_1(P, Q_\theta) \cong \inf_{\theta} \sup_{\varphi: d_\varphi \in \text{Lip}_1(\mathbb{R}^d)} \mathbb{E}d_\varphi(X) - \mathbb{E}d_\varphi(g_\theta(Z)) \]
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 Approx.: In practice we don’t have $P$, only data samples

\[ \{X_i\}_{i=1}^n \text{ are i.i.d. samples from } P \in \mathcal{P}_1(\mathbb{R}^d) \]
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 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 $W_1(P_n, Q_\theta)$

\[
W_1(P_n, Q_\theta) \approx W_1(P, Q_\theta) \text{ hopefully...}
\]
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 $W_1(P_n, Q_\theta)$

$$[W_1(P_n, Q_\theta) \approx W_1(P, Q_\theta) \text{ hopefully...}]$$

Theorem (Dudley’69)

For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \text{Leb}(\mathbb{R}^d)$: $\mathbb{E}W_1(P_n, P) \asymp n^{-\frac{1}{d}}$
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 $P_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

$$\Rightarrow \text{Inherently we work with } W_1(P_n, Q_{\theta})$$

$$\left[ W_1(P_n, Q_{\theta}) \approx W_1(P, Q_{\theta}) \text{ hopefully} \ldots \right]$$

Theorem (Dudley’69)

For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \text{Leb}(\mathbb{R}^d)$: $\mathbb{E}W_1(P_n, P) \asymp n^{-\frac{1}{d}}$
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 $W_1(P_n, Q_\theta)$

\[
W_1(P_n, Q_\theta) \approx W_1(P, Q_\theta) \text{ hopefully...}
\]

Theorem (Dudley’69)

For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \text{Leb}(\mathbb{R}^d)$: $\mathbb{E} W_1(P_n, P) \asymp n^{-\frac{1}{d}}$

Implication: Too slow given dimensionality of real-world data
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 $W_1(P_n, Q_\theta)$

\[
W_1(P_n, Q_\theta) \approx W_1(P, Q_\theta) \text{ hopefully...}
\]

Theorem (Dudley’69)

For $d \geq 3$ and $\mathcal{P}_1(\mathbb{R}^d) \ni P \ll \text{Leb}(\mathbb{R}^d)$: $\mathbb{E}W_1(P_n, P) \asymp n^{-\frac{1}{d}}$

★ 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

$$W_1^{(\sigma)}(P, Q) \triangleq W_1(P \ast N_\sigma, Q \ast N_\sigma),$$

where $N_\sigma \triangleq N(0, \sigma^2 I_d)$ 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
\[
W_1^{(\sigma)}(P, Q) \triangleq W_1(P \ast N_\sigma, Q \ast N_\sigma),
\]
where $N_\sigma \triangleq N(0, \sigma^2 I_d)$ is a $d$-dimensional isotropic Gaussian.

**Interpretation:** $X \sim P$, $Y \sim Q$ and $Z_1, Z_2 \sim N_\sigma$
Smooth 1-Wasserstein Distance

**Definition**

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

$$W_1^{(\sigma)}(P, Q) \triangleq W_1(P \ast \mathcal{N}_\sigma, Q \ast \mathcal{N}_\sigma),$$

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

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

- $X \perp Z_1 \implies X + Z_1 \sim P \ast \mathcal{N}_\sigma$
- $Y \perp Z_2 \implies Y + Z_2 \sim Q \ast \mathcal{N}_\sigma$
Smooth 1-Wasserstein Distance

**Definition**

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

$$W_1^{(\sigma)}(P, Q) \triangleq W_1(P \ast \mathcal{N}_\sigma, Q \ast \mathcal{N}_\sigma),$$

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

**Interpretation:**

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

\[
\begin{align*}
X \perp Z_1 & \implies X + Z_1 \sim P \ast \mathcal{N}_\sigma \\
Y \perp Z_2 & \implies Y + Z_2 \sim Q \ast \mathcal{N}_\sigma
\end{align*}
\]

$\implies W_1$ distance between smoothed distributions
Smooth 1-Wasserstein Distance

**Definition**

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

$$W_1^{(\sigma)}(P, Q) \triangleq W_1(P \ast N_\sigma, Q \ast N_\sigma),$$

where $N_\sigma \triangleq N(0, \sigma^2 I_d)$ is a $d$-dimensional isotropic Gaussian.

**Interpretation:** $X \sim P$, $Y \sim Q$ and $Z_1, Z_2 \sim N_\sigma$

- $X \perp Z_1 \implies X + Z_1 \sim P \ast N_\sigma$
- $Y \perp Z_2 \implies Y + Z_2 \sim Q \ast N_\sigma$

$\implies$ $W_1$ distance between smoothed distributions

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

$$W_1^{(\sigma)}(P, Q) = \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \mathbb{E} f(X + Z_1) - \mathbb{E} f(Y + Z_2)$$
**High Level:** $W_{1}^{(\sigma)}$ inherits the metric structure of 1-Wasserstein
Smooth 1-Wasserstein – Metric Structure

**High Level:** $W_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

**Theorem**

$\left(\mathcal{P}_1(\mathbb{R}^d), W_1^{(\sigma)}\right)$ is metric space, $\forall \sigma \geq 0$ (and $W_1^{(\sigma)}$ metrizes weak conv.).
Smooth 1-Wasserstein – Metric Structure

**High Level:** $W_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

**Theorem**

$$\left( \mathcal{P}_1(\mathbb{R}^d), W_1^{(\sigma)} \right) \text{ is metric space, } \forall \sigma \geq 0 \text{ (and } W_1^{(\sigma)} \text{ metrizes weak conv.)}.$$  

**Key Idea for Pf.:** Use Characteristic functions $\Phi_P(t) \triangleq \mathbb{E}_P[e^{itX}]$ and:
**High Level:** $W_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

**Theorem**

$\left( \mathcal{P}_1(\mathbb{R}^d), W_1^{(\sigma)} \right)$ is metric space, $\forall \sigma \geq 0$ (and $W_1^{(\sigma)}$ metrizes weak conv.).

**Key Idea for Pf.:** Use Characteristic functions $\Phi_P(t) \triangleq \mathbb{E}_P[e^{itX}]$ 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:** $W_{1}^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

**Theorem**

$\left( \mathcal{P}_{1}(\mathbb{R}^{d}), W_{1}^{(\sigma)} \right)$ is metric space, $\forall \sigma \geq 0$ (and $W_{1}^{(\sigma)}$ metrizes weak conv.).

**Key Idea for Pf.:** Use Characteristic functions $\Phi_{P}(t) \triangleq \mathbb{E}_{P}[e^{itX}]$ and:

$$\Phi_{P\ast\mathcal{N}_{\sigma}}(t) = \Phi_{P}(t)\Phi_{\mathcal{N}_{\sigma}}(t)$$

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: $W_{1}^{(\sigma)}(Q_{i}, P) \to 0$ iff $W_{1}(Q_{i}, P) \to 0$
High Level: $W_1^{(\sigma)}$ inherits the metric structure of 1-Wasserstein

**Theorem**

\[ \left( \mathcal{P}_1(\mathbb{R}^d), W_1^{(\sigma)} \right) \text{ is metric space, } \forall \sigma \geq 0 \text{ (and } W_1^{(\sigma)} \text{ metrizes weak conv.)}. \]

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

\[ \Phi_{P*N_\sigma}(t) = \Phi_P(t) \Phi_{N_\sigma}(t) \text{ together with } \Phi_{N_\sigma}(t) = e^{-\frac{\sigma^2\|t\|^2}{2}} \neq 0, \forall t. \]

**Corollary**

Let $P, Q_i, \in \mathcal{P}(\mathbb{R}^d), i = 1, \ldots$. Then: $W_1^{(\sigma)}(Q_i, P) \to 0$ iff $W_1(Q_i, P) \to 0$

$\bigodot W_1^{(\sigma)}$ and $W_1$ induce same topology
High Level: $W_1^{(\sigma)}(P, Q)$ is well-behaved func. of $\sigma$ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)
High Level: $W_1^{(\sigma)}(P, Q)$ is well-behaved func. of $\sigma$ (fixed $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$)

**Theorem**

Fix $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$. The following hold:
High Level: $W_1^{(\sigma)}(P, Q)$ is well-behaved func. of $\sigma$ (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. $W_1^{(\sigma)}(P, Q)$ is continuous and mono. non-increasing in $\sigma \in [0, +\infty)$
High Level: \( W_1^{(\sigma)}(P, Q) \) is well-behaved func. of \( \sigma \) (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. \( W_1^{(\sigma)}(P, Q) \) is continuous and mono. non-increasing in \( \sigma \in [0, +\infty) \)
2. \( \lim_{\sigma \to 0} W_1^{(\sigma)}(P, Q) = W_1(P, Q) \)
Smooth 1-Wasserstein – Function of Noise Std

**High Level:** $W_1^{(\sigma)}(P, Q)$ is well-behaved func. of $\sigma$ (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. $W_1^{(\sigma)}(P, Q)$ is continuous and mono. non-increasing in $\sigma \in [0, +\infty)$
2. $\lim_{\sigma \to 0} W_1^{(\sigma)}(P, Q) = W_1(P, Q)$
3. $\lim_{\sigma \to \infty} W_1^{(\sigma)}(P, Q) \neq 0$, for some $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$
**High Level:** $W_1^{(\sigma)}(P, Q)$ is well-behaved func. of $\sigma$ (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. $W_1^{(\sigma)}(P, Q)$ is continuous and mono. non-increasing in $\sigma \in [0, +\infty)$
2. $\lim_{\sigma \to 0} W_1^{(\sigma)}(P, Q) = W_1(P, Q)$
3. $\lim_{\sigma \to \infty} W_1^{(\sigma)}(P, Q) \neq 0$, 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:** $W_1^{(\sigma)}(P, Q)$ is well-behaved func. of $\sigma$ (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. $W_1^{(\sigma)}(P, Q)$ is continuous and mono. non-increasing in $\sigma \in [0, +\infty)$
2. $\lim_{\sigma \to 0} W_1^{(\sigma)}(P, Q) = W_1(P, Q)$
3. $\lim_{\sigma \to \infty} W_1^{(\sigma)}(P, Q) \neq 0$, for some $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$

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

**Lemma**

For $\sigma_1 < \sigma_2$:

$$W_1^{(\sigma_2)}(P, Q) \leq W_1^{(\sigma_1)}(P, Q) \leq W_1^{(\sigma_2)}(P, Q) + 2d\sqrt{\sigma_2^2 - \sigma_1^2}$$
Smooth 1-Wasserstein – Function of Noise Std

**High Level:** \( W_1^{(\sigma)}(P, Q) \) is well-behaved func. of \( \sigma \) (fixed \( P, Q \in P_1(\mathbb{R}^d) \))

**Theorem**

Fix \( P, Q \in P_1(\mathbb{R}^d) \). The following hold:

1. \( W_1^{(\sigma)}(P, Q) \) is continuous and mono. non-increasing in \( \sigma \in [0, +\infty) \)
2. \( \lim_{\sigma \to 0} W_1^{(\sigma)}(P, Q) = W_1(P, Q) \)
3. \( \lim_{\sigma \to \infty} W_1^{(\sigma)}(P, Q) \neq 0 \), for some \( P, Q \in P_1(\mathbb{R}^d) \)

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

**Lemma**

For \( \sigma_1 < \sigma_2 \): \( W_1^{(\sigma_2)}(P, Q) \leq W_1^{(\sigma_1)}(P, Q) \leq W_1^{(\sigma_2)}(P, Q) + 2d \sqrt{\sigma_2^2 - \sigma_1^2} \)

**Pf. Item 3:** \( W_1^{(\sigma)}(\delta_x, \delta_y) = W_1(\mathcal{N}(x, \sigma^2 I_d), \mathcal{N}(y, \sigma^2 I_d)) = \|x - y\| \)
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 $P$: 

$$\mathbb{E} W_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}}$$
**Smooth 1-Wasserstein – Statistical Efficiency**

**High Level:** Alleviate curse of dimensionality & get concentration

---

**Theorem**

For any $d \geq 1$, $\sigma > 0$ and sub-Gaussian $P$:  
$$\mathbb{E} W_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}}$$

---

**Theorem**

Under same assumptions: denote $\mathcal{X} \triangleq \text{supp}(\mu)$ and suppose  
$\text{diam}(\mathcal{X}) < \infty$, where $\text{diam}(\mathcal{X}) = \sup_{x \neq y \in \mathcal{X}} \|x - y\|$. For any $t > 0$ we have  
$$\mathbb{P}_{\mu^\otimes n}\left( \left| W_1^{(\sigma)}(\hat{\mu}_n, \mu) - \mathbb{E} W_1^{(\sigma)}(\hat{\mu}_n, \mu) \right| \geq t \right) \leq 2e^{-\frac{2t^2n}{\text{diam}(\mathcal{X})^2}}$$
Smooth 1-Wasserstein – Statistical Efficiency

**High Level:** Alleviate curse of dimensionality & get concentration

**Theorem**

For any $d \geq 1$, $\sigma > 0$ and sub-Gaussian $P$:  
$$\mathbb{E} W_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}}$$

**Theorem**

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

$$\mathbb{P}_{\mu \otimes n} \left( \left| W_1^{(\sigma)}(\hat{\mu}_n, \mu) - \mathbb{E} W_1^{(\sigma)}(\hat{\mu}_n, \mu) \right| \geq t \right) \leq 2e^{-\frac{2t^2n}{\text{diam}(\mathcal{X})^2}}$$

**Comments:**

- Achieves $n^{-\frac{1}{2}}$ bias rate vs $n^{-1/d}$ for $W_1$ - via maximal TV coupling arg
**High Level:** Alleviate curse of dimensionality & get concentration

**Theorem**

*For any* \( d \geq 1, \sigma > 0 \) *and sub-Gaussian* \( P \): \[ \mathbb{E} W_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}} \]

**Theorem**

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

\[
\mathbb{P}_{\mu^{\otimes n}} \left( \left| W_1^{(\sigma)}(\hat{\mu}_n, \mu) - \mathbb{E} W_1^{(\sigma)}(\hat{\mu}_n, \mu) \right| \geq t \right) \leq 2e^{-\frac{2t^2n}{\text{diam}(\mathcal{X})^2}}
\]

**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 $P$: \[ \mathbb{E} W_1^{(\sigma)}(P_n, P) \lesssim n^{-\frac{1}{2}} \]

**Theorem**

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

\[
\mathbb{P}_{\mu \otimes n} \left( \left| W_1^{(\sigma)}(\hat{\mu}_n, \mu) - \mathbb{E} W_1^{(\sigma)}(\hat{\mu}_n, \mu) \right| \geq t \right) \leq 2e^{-\frac{2t^2n}{\text{diam}(\mathcal{X})^2}}
\]

**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
Convergence of $W_1^{(\sigma)}(\hat{\mu}_n, \mu)$ as a function of the number of samples $n$ for various values of $\sigma$, shown in log-log space. The measure $\mu$ 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 $d$. 

$d = 5$

$d = 10$

$d = 100$
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

- 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 $n^{-1/2}$ convergence of empirical approximation in all dimensions

Thank you!
- **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
**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

Thank you!
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 $P_1(\mathbb{R}^d)$
  - 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!