Optimality of the Plug-in Estimator for Differential Entropy Estimation under Gaussian Convolutions

Ziv Goldfeld, Kristjan Greenewald, Yury Polyanskiy and Jonathan Weed

MIT, MIT-IBM Watson AI Lab

International Symposium on Information Theory

July 9th, 2019



New Estimation Problem

Setup: Estimate $h(X+Z)$ for d-dimensional, independent X and Z:

New Estimation Problem

Setup: Estimate $h(X+Z)$ for d-dimensional, independent X and Z:

- $X \sim P$, where $P \in \mathcal{F}_d$ is unknown (nonparametric class)

New Estimation Problem

Setup: Estimate $h(X+Z)$ for d-dimensional, independent X and Z:

- $X \sim P$, where $P \in \mathcal{F}_d$ is unknown (nonparametric class)
- $Z \sim \mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$

New Estimation Problem

Setup: Estimate $h(X+Z)$ for d-dimensional, independent X and Z:

- $X \sim P$, where $P \in \mathcal{F}_d$ is unknown (nonparametric class)
- $Z \sim \mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$

Resources: An estimator $\hat{h}$ of $h(X+Z)$ can use

New Estimation Problem

Setup: Estimate $h(X+Z)$ for d-dimensional, independent X and Z:

- $X \sim P$, where $P \in \mathcal{F}_d$ is unknown (nonparametric class)
- $Z \sim \mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$

Resources: An estimator $\hat{h}$ of $h(X+Z)$ can use

1. n i.i.d. samples $X^n \triangleq (X_i)_{i=1}^n$ from P.

New Estimation Problem

Setup: Estimate $h(X+Z)$ for d-dimensional, independent X and Z:

- $X \sim P$, where $P \in \mathcal{F}_d$ is unknown (nonparametric class)
- $Z \sim \mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$

Resources: An estimator $\hat{h}$ of $h(X+Z)$ can use

1. n i.i.d. samples $X^n \triangleq (X_i)_{i=1}^n$ from P.
2. Knowledge of $\mathcal{N}_\sigma$.

New Estimation Problem

Setup: Estimate $h(X+Z)$ for d-dimensional, independent X and Z:

- $X \sim P$, where $P \in \mathcal{F}_d$ is unknown (nonparametric class)
- $Z \sim \mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$

Resources: An estimator $\hat{h}$ of $h(X+Z)$ can use

1. n i.i.d. samples $X^n \triangleq (X_i)_{i=1}^n$ from P.
2. Knowledge of $\mathcal{N}_\sigma$.

Absolute Error Minimax Risk:

$$\mathcal{R}^\star(n, \sigma, \mathcal{F}_d) \triangleq \inf_{\hat{h}} \sup_{P \in \mathcal{F}_d} \mathbb{E}\left| h(P * \mathcal{N}_\sigma) - \hat{h}(X^n, \sigma) \right|$$

New Estimation Problem

Setup: Estimate $h(X+Z)$ for d-dimensional, independent X and Z:

- $X \sim P$, where $P \in \mathcal{F}_d$ is unknown (nonparametric class)
- $Z \sim \mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$

Resources: An estimator $\hat{h}$ of $h(X+Z)$ can use

1. n i.i.d. samples $X^n \triangleq (X_i)_{i=1}^n$ from P.
2. Knowledge of $\mathcal{N}_\sigma$.

Absolute Error Minimax Risk:

$$\mathcal{R}^\star(n, \sigma, \mathcal{F}_d) \triangleq \inf_{\hat{h}} \sup_{P \in \mathcal{F}_d} \mathbb{E}\left| h(P * \mathcal{N}_\sigma) - \hat{h}(X^n, \sigma) \right|$$

⊛ **Sample complexity** $n^\star(\eta, \sigma, \mathcal{F}_d)$: least n needed for η-gap estimation.

Motivation - Information Theory & Deep Learning

Motivation - Information Theory & Deep Learning

- **Information Bottleneck Theory** [Tishby-Zaslavsky'15, Shwartz-Tishby'17]

  Estimate mutual information between layers of a DNN

Motivation - Information Theory & Deep Learning

- **Information Bottleneck Theory** [Tishby-Zaslavsky'15, Shwartz-Tishby'17]

  Estimate mutual information between layers of a DNN

- **Unsupervised Learning** [Hjelm et al'18, Oord-Li-Vinyals'18]

  Mutual information for learning representations (Deep InfoMax, CPC)

Motivation - Information Theory & Deep Learning

- **Information Bottleneck Theory** [Tishby-Zaslavsky'15, Shwartz-Tishby'17]

  Estimate mutual information between layers of a DNN

- **Unsupervised Learning** [Hjelm et al'18, Oord-Li-Vinyals'18]

  Mutual information for learning representations (Deep InfoMax, CPC)

⊛ IT measure **degenerate** over DNNs with fixed parameters

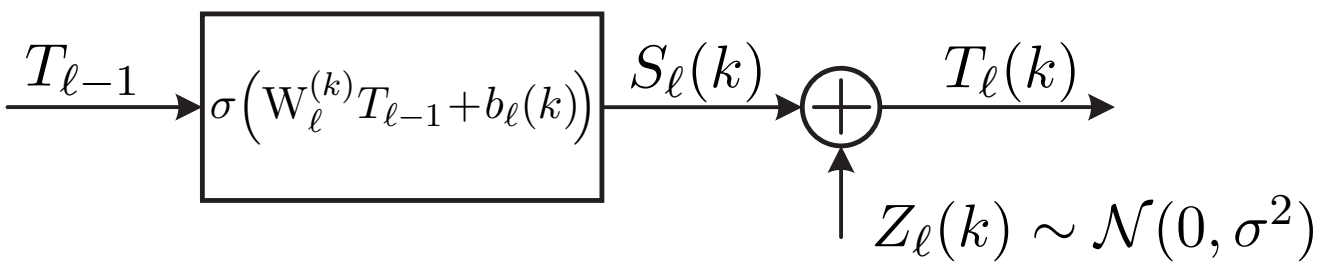
Motivation - Information Theory & Deep Learning

- **Information Bottleneck Theory** [Tishby-Zaslavsky'15, Shwartz-Tishby'17]
  Estimate mutual information between layers of a DNN

- **Unsupervised Learning** [Hjelm et al'18, Oord-Li-Vinyals'18]
  Mutual information for learning representations (Deep InfoMax, CPC)

⊛ IT measure **degenerate** over DNNs with fixed parameters

$\Longrightarrow$ **Study Info. Flow in DNNs:** Stochastic DNNs via noise injection
[Goldfeld-Berg-Greenewald-Melnyk-Nguyen-Kingsbury-Polyanskiy'18]



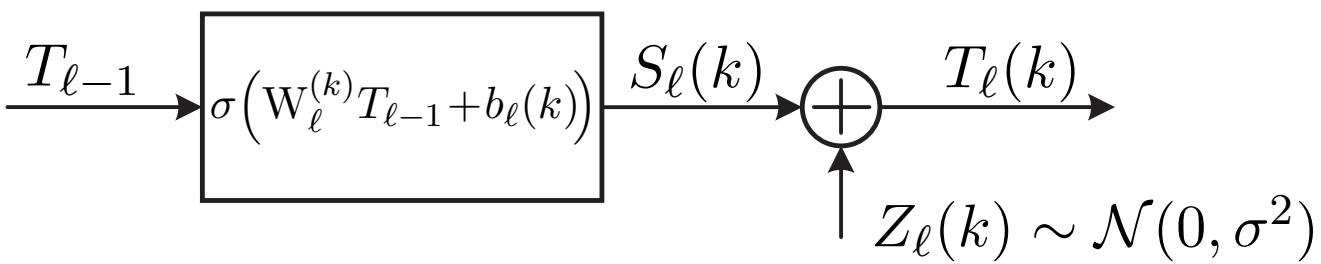
Motivation - Information Theory & Deep Learning

- **Information Bottleneck Theory** [Tishby-Zaslavsky'15, Shwartz-Tishby'17]
  Estimate mutual information between layers of a DNN

- **Unsupervised Learning** [Hjelm et al'18, Oord-Li-Vinyals'18]
  Mutual information for learning representations (Deep InfoMax, CPC)

⊛ IT measure **degenerate** over DNNs with fixed parameters

$\Longrightarrow$ **Study Info. Flow in DNNs:** Stochastic DNNs via noise injection
[Goldfeld-Berg-Greenewald-Melnyk-Nguyen-Kingsbury-Polyanskiy'18]



⊛ Can sample S_ℓ (gen. model) & want to estimate $h(T_\ell) = h(S_\ell + Z_\ell)$

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ *based on* $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ *and knowledge of* $\mathcal{N}_\sigma$.

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Theoretical Guarantees:

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Theoretical Guarantees:

- Most results assume lower bounded density

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Theoretical Guarantees:

- Most results assume lower bounded density $\Longrightarrow$ **Inapplicable**

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Theoretical Guarantees:

- Most results assume lower bounded density $\Longrightarrow$ **Inapplicable**
- **Applicable Here:**

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Theoretical Guarantees:

- Most results assume lower bounded density $\Longrightarrow$ **Inapplicable**
- **Applicable Here:**
  1. **[Han-Jiao-Weissman-Wu'17]:** KDE + Best poly. approximation

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Theoretical Guarantees:

- Most results assume lower bounded density $\Longrightarrow$ **Inapplicable**
- **Applicable Here:**
  1. **[Han-Jiao-Weissman-Wu'17]:** KDE + Best poly. approximation

  $\Longrightarrow$ P subgaussian, $\text{Risk}_{\text{KDE}} \leq O\left(n^{-\frac{2}{2+d}}\right)$ (Analysis: restricted smoothness)*

$\star$ Omitting multiplicative polylogarithmic factors.

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Theoretical Guarantees:

- Most results assume lower bounded density $\Longrightarrow$ **Inapplicable**
- **Applicable Here:**
  1. **[Han-Jiao-Weissman-Wu'17]:** KDE + Best poly. approximation

  $\Longrightarrow$ P subgaussian, $\text{Risk}_{\text{KDE}} \leq O\left(n^{-\frac{2}{2+d}}\right)$ (Analysis: restricted smoothness)*

  2. **[Berrett-Samworth-Yuan'19]:** Weighted kNN (Kozachenko-Leonenko)

Direct Approach - General-Purpose Estimators

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Method: Estimate $h(P * \mathcal{N}_\sigma)$ via i.i.d. (**noisy**) samples from $P * \mathcal{N}_\sigma$

Theoretical Guarantees:

- Most results assume lower bounded density $\Longrightarrow$ **Inapplicable**
- **Applicable Here:**

1. **[Han-Jiao-Weissman-Wu'17]:** KDE + Best poly. approximation

$\Longrightarrow$ P subgaussian, $\text{Risk}_{\text{KDE}} \leq O\left(n^{-\frac{2}{2+d}}\right)$ (Analysis: restricted smoothness)*

2. **[Berrett-Samworth-Yuan'19]:** Weighted kNN (Kozachenko-Leonenko)

$\Longrightarrow$ P compactly supported, $\text{Risk}_{\text{w-kNN}} \leq O\left(1/\sqrt{n}\right)$ (dependence on d?)

Plug-in Estimator

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ *based on* $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ *and knowledge of* $\mathcal{N}_\sigma$.

Plug-in Estimator

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Our Estimator: $\hat{h}(X^n, \sigma) \triangleq h(\hat{P}_{X^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{X^n} \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

Plug-in Estimator

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Our Estimator: $\hat{h}(X^n, \sigma) \triangleq h(\hat{P}_{X^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{X^n} \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

Comment: $\hat{h}$ is **plug-in** estimator for $\mathsf{T}_\sigma(P) \triangleq h(P * \mathcal{N}_\sigma)$

Plug-in Estimator

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Our Estimator: $\hat{h}(X^n, \sigma) \triangleq h(\hat{P}_{X^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{X^n} \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

Comment: $\hat{h}$ is **plug-in** estimator for $\mathsf{T}_\sigma(P) \triangleq h(P * \mathcal{N}_\sigma)$

Nonparametric Classes:

Plug-in Estimator

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Our Estimator: $\hat{h}(X^n, \sigma) \triangleq h(\hat{P}_{X^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{X^n} \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

Comment: $\hat{h}$ is **plug-in** estimator for $\mathsf{T}_\sigma(P) \triangleq h(P * \mathcal{N}_\sigma)$

Nonparametric Classes:

1. **Compact Support:** $\mathcal{F}_d \triangleq \left\{P \middle| \operatorname{supp}(P) \subseteq [-1, 1]^d\right\}$

Plug-in Estimator

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Our Estimator: $\hat{h}(X^n, \sigma) \triangleq h(\hat{P}_{X^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

Comment: $\hat{h}$ is **plug-in** estimator for $\mathsf{T}_\sigma(P) \triangleq h(P * \mathcal{N}_\sigma)$

Nonparametric Classes:

1. **Compact Support:** $\mathcal{F}_d \triangleq \{P \big| \operatorname{supp}(P) \subseteq [-1,1]^d\}$
2. **Subgaussian:** $\mathcal{F}_{d,K}^{(\mathsf{SubG})} \triangleq \{P \big| X \sim P \text{ is } K\text{-SubG}\}$

   where X is K-SubG if $\quad \mathbb{E}e^{\alpha^\mathsf{T}(X-\mathbb{E}X)} \leq e^{\frac{1}{2}K^2\|\alpha\|^2}, \quad \forall \alpha \in \mathbb{R}^d.$

Plug-in Estimator

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Our Estimator: $\hat{h}(X^n, \sigma) \triangleq h(\hat{P}_{X^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

Comment: $\hat{h}$ is **plug-in** estimator for $\mathsf{T}_\sigma(P) \triangleq h(P * \mathcal{N}_\sigma)$

Nonparametric Classes:

1. **Compact Support:** $\mathcal{F}_d \triangleq \{P \big| \operatorname{supp}(P) \subseteq [-1,1]^d\}$
2. **Subgaussian:** $\mathcal{F}_{d,K}^{(\mathsf{SubG})} \triangleq \{P \big| X \sim P \text{ is } K\text{-SubG}\}$

   where X is K-SubG if $\quad \mathbb{E}e^{\alpha^\mathsf{T}(X-\mathbb{E}X)} \leq e^{\frac{1}{2}K^2\|\alpha\|^2}, \quad \forall \alpha \in \mathbb{R}^d.$

⊛ **Relation:** Exists $K' > 0$ such that $\mathcal{F}_d \subseteq \mathcal{F}_{d,K'}^{(\mathsf{SubG})}$

Plug-in Estimator

Differential Entropy Estimation under Gaussian Convolutions

Estimate $h(P * \mathcal{N}_\sigma)$ based on $X^n \overset{iid}{\sim} P \in \mathcal{F}_d$ and knowledge of $\mathcal{N}_\sigma$.

Our Estimator: $\hat{h}(X^n, \sigma) \triangleq h(\hat{P}_{X^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

Comment: $\hat{h}$ is **plug-in** estimator for $\mathsf{T}_\sigma(P) \triangleq h(P * \mathcal{N}_\sigma)$

Nonparametric Classes:

1. **Compact Support:** $\mathcal{F}_d \triangleq \{P \big| \operatorname{supp}(P) \subseteq [-1,1]^d\}$
2. **Subgaussian:** $\mathcal{F}_{d,K}^{(\mathsf{SubG})} \triangleq \{P \big| X \sim P \text{ is } K\text{-SubG}\}$

   where X is K-SubG if $\quad \mathbb{E}e^{\alpha^\mathsf{T}(X-\mathbb{E}X)} \leq e^{\frac{1}{2}K^2\|\alpha\|^2}, \quad \forall \alpha \in \mathbb{R}^d.$

⊛ **Relation:** Exists $K' > 0$ such that $\mathcal{F}_d \subseteq \mathcal{F}_{d,K'}^{(\mathsf{SubG})}$

$\Longrightarrow$ Use $\mathcal{F}_d$ for **Lower Bounds** & $\mathcal{F}_{d,K}^{(\mathsf{SubG})}$ for **Upper Bounds**

Structured Estimator - Convergence Rate

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

For any $\sigma > 0,\ d \geq 1$, *we have*

$$\sup_{P \in \mathcal{F}_{d,K}^{(\mathsf{SubG})}} \mathbb{E}\left|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{X^n} * \mathcal{N}_\sigma)\right| \leq C_{\sigma,d,K} \cdot n^{-\frac{1}{2}}$$

where $C_{\sigma,d,K} = O_{\sigma,K}(c^d)$ *for a constant* c.

Structured Estimator - Convergence Rate

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

For any $\sigma > 0,\ d \geq 1$, *we have*

$$\sup_{P \in \mathcal{F}_{d,K}^{(\mathsf{SubG})}} \mathbb{E}\left|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{X^n} * \mathcal{N}_\sigma)\right| \leq C_{\sigma,d,K} \cdot n^{-\frac{1}{2}}$$

where $C_{\sigma,d,K} = O_{\sigma,K}(c^d)$ *for a constant* c.

Comments:

Structured Estimator - Convergence Rate

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

For any $\sigma > 0,\ d \geq 1$, *we have*

$$\sup_{P \in \mathcal{F}_{d,K}^{(\mathsf{SubG})}} \mathbb{E}\left|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{X^n} * \mathcal{N}_\sigma)\right| \leq C_{\sigma,d,K} \cdot n^{-\frac{1}{2}}$$

where $C_{\sigma,d,K} = O_{\sigma,K}(c^d)$ *for a constant* c.

Comments:

- **Explicit Expression:** Enables concrete error bounds

Structured Estimator - Convergence Rate

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

For any $\sigma > 0,\ d \geq 1$, *we have*

$$\sup_{P \in \mathcal{F}_{d,K}^{(\mathsf{SubG})}} \mathbb{E}\left|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{X^n} * \mathcal{N}_\sigma)\right| \leq C_{\sigma,d,K} \cdot n^{-\frac{1}{2}}$$

where $C_{\sigma,d,K} = O_{\sigma,K}(c^d)$ *for a constant* c.

Comments:

- **Explicit Expression:** Enables concrete error bounds
- **Minimax Rate Optimal:** Attains parametric rate $O\left(n^{-\frac{1}{2}}\right)$

Structured Estimator - Convergence Rate

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

For any $\sigma > 0,\ d \geq 1$, we have

$$\sup_{P \in \mathcal{F}_{d,K}^{(\mathsf{SubG})}} \mathbb{E}\left|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{X^n} * \mathcal{N}_\sigma)\right| \leq C_{\sigma,d,K} \cdot n^{-\frac{1}{2}}$$

where $C_{\sigma,d,K} = O_{\sigma,K}(c^d)$ for a constant c.

Comments:

- **Explicit Expression:** Enables concrete error bounds
- **Minimax Rate Optimal:** Attains parametric rate $O\left(n^{-\frac{1}{2}}\right)$

Proof (initial step): Based on [Polyanskiy-Wu'16]

$$\left|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{X^n} * \mathcal{N}_\sigma)\right| \lesssim W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma)$$

Structured Estimator - Convergence Rate

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

For any $\sigma > 0,\ d \geq 1$, *we have*

$$\sup_{P \in \mathcal{F}_{d,K}^{(\mathsf{SubG})}} \mathbb{E}\left|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{X^n} * \mathcal{N}_\sigma)\right| \leq C_{\sigma,d,K} \cdot n^{-\frac{1}{2}}$$

where $C_{\sigma,d,K} = O_{\sigma,K}(c^d)$ *for a constant* c.

Comments:

- **Explicit Expression:** Enables concrete error bounds
- **Minimax Rate Optimal:** Attains parametric rate $O\left(n^{-\frac{1}{2}}\right)$

Proof (initial step): Based on [Polyanskiy-Wu'16]

$$\boxed{\left|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{X^n} * \mathcal{N}_\sigma)\right| \lesssim W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma)}$$

$\Longrightarrow$ Analyze empirical 1-Wasserstein distance under Gaussian convolutions

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P, Q) \triangleq \inf \left(\mathbb{E}\|X - Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P, Q) \triangleq \inf \left(\mathbb{E}\|X - Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q

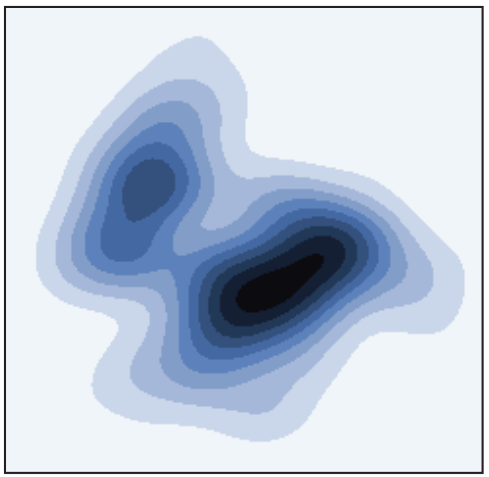
Empirical 1-Wasserstein Distance:

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P, Q) \triangleq \inf \left(\mathbb{E}\|X - Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q



Empirical 1-Wasserstein Distance:

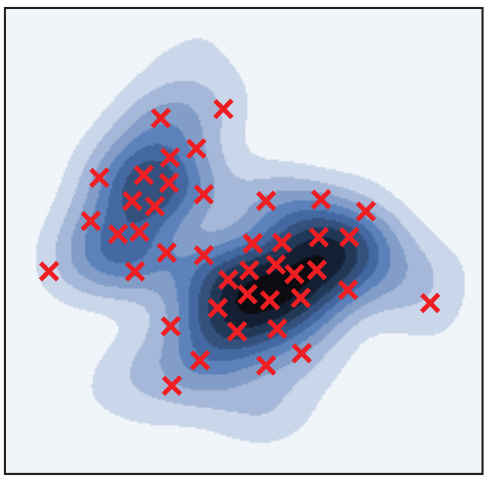
- Distribution P on $\mathbb{R}^d$

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P, Q) \triangleq \inf \left(\mathbb{E}\|X - Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q



Empirical 1-Wasserstein Distance:

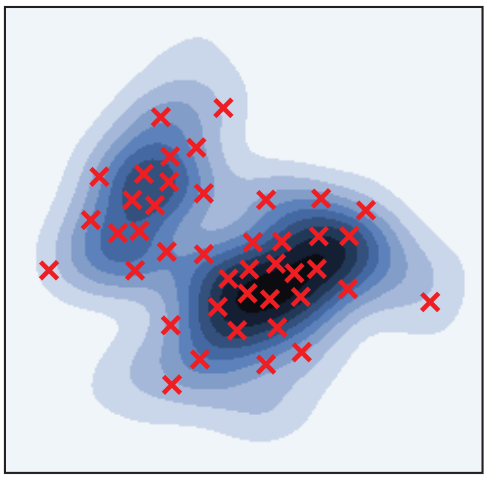
- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P,Q) \triangleq \inf \left(\mathbb{E}\|X-Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q



Empirical 1-Wasserstein Distance:

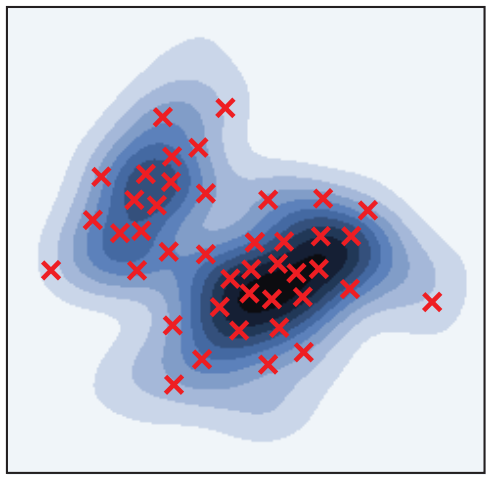
- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$
- Empirical distribution $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P,Q) \triangleq \inf \left(\mathbb{E}\|X-Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q



Empirical 1-Wasserstein Distance:

- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$
- Empirical distribution $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

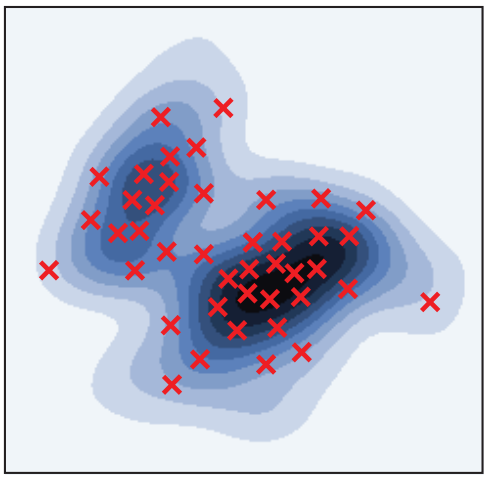
$\implies$ Dependence on (n,d) of $\mathbb{E}W_1(P,\hat{P}_{X^n})$ (for cts. P)

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P,Q) \triangleq \inf \left(\mathbb{E}\|X-Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q



Empirical 1-Wasserstein Distance:

- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$
- Empirical distribution $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

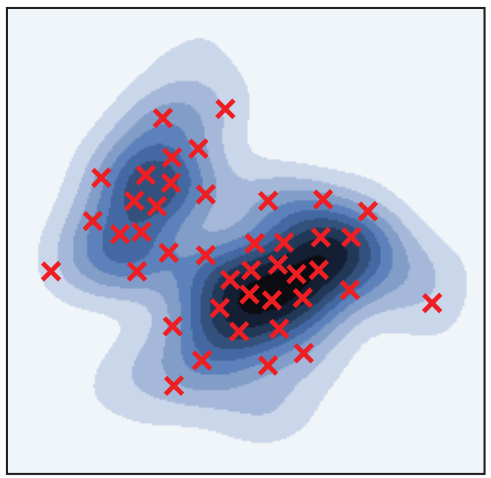
$\implies$ Dependence on (n,d) of $\mathbb{E}W_1\left(P, \hat{P}_{X^n}\right) \asymp n^{-\frac{1}{d}}$ (for cts. P)

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P,Q) \triangleq \inf \left(\mathbb{E}\|X-Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q



Empirical 1-Wasserstein Distance:

- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$
- Empirical distribution $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

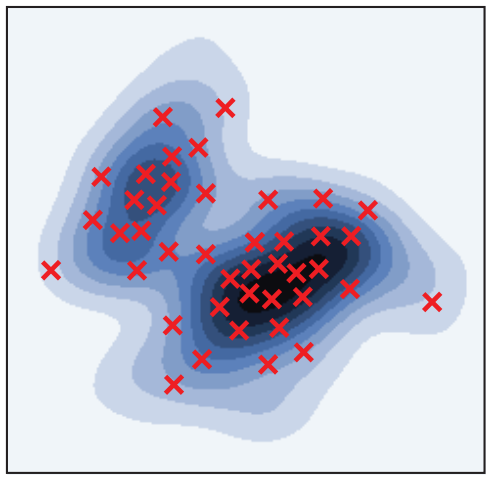
$\implies$ Dependence on (n,d) of $\mathbb{E}W_1\left(P, \hat{P}_{X^n}\right) \asymp n^{-\frac{1}{d}}$ (fo

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P,Q) \triangleq \inf \left(\mathbb{E}\|X-Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q



Empirical 1-Wasserstein Distance:

- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$
- Empirical distribution $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

$\implies$ Dependence on (n,d) of $\mathbb{E}W_1\left(P, \hat{P}_{X^n}\right) \asymp n^{-\frac{1}{d}}$ (fo

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

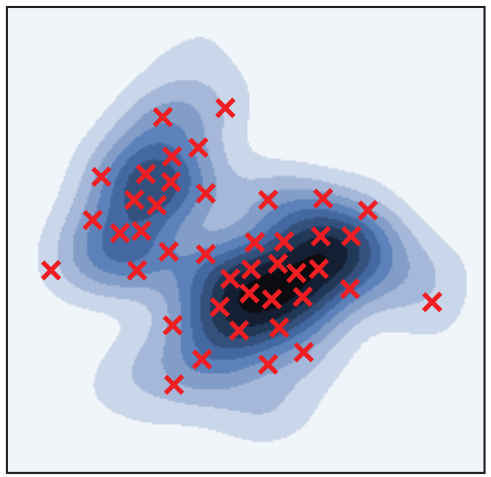
For any d, we have $\mathbb{E}W_1\left(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma\right) \leq O_{\sigma,d}\left(n^{-\frac{1}{2}}\right)$

Gaussian Smoothed Empirical W_1

p-Wasserstein Distance: For two distributions P and Q on $\mathbb{R}^d$ and $p \geq 1$

$$W_p(P,Q) \triangleq \inf \left(\mathbb{E}\|X-Y\|^p\right)^{1/p}$$

infimum over all couplings of P and Q



Empirical 1-Wasserstein Distance:

- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$
- Empirical distribution $\hat{P}_{X^n} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

$\implies$ Dependence on (n,d) of $\mathbb{E}W_1\left(P, \hat{P}_{X^n}\right) \asymp n^{-\frac{1}{d}}$ (fc

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

For any d, we have $\mathbb{E}W_1\left(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma\right) \leq O_{\sigma,d}\left(n^{-\frac{1}{2}}\right) = O_\sigma\left(c^d n^{-\frac{1}{2}}\right)$

Gaussian Smoothed Empirical W_1 - Proof Idea

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

For any d, we have $\mathbb{E}W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma) \leq O_{\sigma,d}(n^{-\frac{1}{2}}) = O_\sigma(c^d n^{-\frac{1}{2}})$

Gaussian Smoothed Empirical W_1 - Proof Idea

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

For any d, we have $\mathbb{E}W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma) \leq O_{\sigma,d}(n^{-\frac{1}{2}}) = O_\sigma(c^d n^{-\frac{1}{2}})$

Main Idea: Consider D_{KL} instead

Gaussian Smoothed Empirical W_1 - Proof Idea

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

For any d, we have $\mathbb{E}W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma) \leq O_{\sigma,d}(n^{-\frac{1}{2}}) = O_\sigma(c^d n^{-\frac{1}{2}})$

Main Idea: Consider D_{KL} instead

- $\mathsf{D}_{KL}(P\|Q) \leq \log\left(1 + \chi^2(P\|Q)\right)$

Gaussian Smoothed Empirical W_1 - Proof Idea

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

For any d, we have $\mathbb{E}W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma) \leq O_{\sigma,d}(n^{-\frac{1}{2}}) = O_\sigma(c^d n^{-\frac{1}{2}})$

Main Idea: Consider D_{KL} instead

- $\mathsf{D}_{KL}(P\|Q) \leq \log\left(1 + \chi^2(P\|Q)\right)$
- Bounding χ^2 with $q =$ PDF of $P * \mathcal{N}_\sigma$:

Gaussian Smoothed Empirical W_1 - Proof Idea

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

For any d, we have $\mathbb{E}W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma) \le O_{\sigma,d}(n^{-\frac{1}{2}}) = O_\sigma(c^d n^{-\frac{1}{2}})$

Main Idea: Consider D_{KL} instead

- $\mathsf{D}_{KL}(P\|Q) \le \log\left(1 + \chi^2(P\|Q)\right)$
- Bounding χ^2 with $q =$ PDF of $P * \mathcal{N}_\sigma$:

$$\mathbb{E}\chi^2\left(\hat{P}_{X^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) = \frac{1}{n}\left(\int_{\mathbb{R}^d} \mathbb{E}\frac{(\varphi(z-X) - q(z))^2}{q(z)}\mathsf{d}z\right) = \frac{1}{n} I_{\chi^2}(X;Y)$$

$$I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y}\|P_X \otimes P_Y), \quad X \sim P,\ Y = X + Z$$

Gaussian Smoothed Empirical W_1 - Proof Idea

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

For any d, we have $\mathbb{E}W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma) \le O_{\sigma,d}(n^{-\frac{1}{2}}) = O_\sigma(c^d n^{-\frac{1}{2}})$

Main Idea: Consider D_{KL} instead

- $\mathsf{D}_{KL}(P\|Q) \le \log\left(1 + \chi^2(P\|Q)\right)$
- Bounding χ^2 with $q =$ PDF of $P * \mathcal{N}_\sigma$:

$$\mathbb{E}\chi^2\left(\hat{P}_{X^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) = \frac{1}{n}\left(\int_{\mathbb{R}^d} \mathbb{E}\frac{(\varphi(z-X) - q(z))^2}{q(z)}\mathrm{d}z\right) = \frac{1}{n} I_{\chi^2}(X;Y)$$

$$I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y}\|P_X \otimes P_Y), \quad X \sim P,\ Y = X + Z$$

$$\Longrightarrow \quad \mathsf{D}_{KL} = O\left(\tfrac{1}{n}\right) \text{ if } I_{\chi^2}(X;Y) < \infty.$$

Gaussian Smoothed Empirical W_1 - Proof Idea

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

For any d, we have $\mathbb{E}W_1(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma) \le O_{\sigma,d}(n^{-\frac{1}{2}}) = O_\sigma(c^d n^{-\frac{1}{2}})$

Main Idea: Consider D_{KL} instead

- $\mathsf{D}_{KL}(P\|Q) \le \log\left(1 + \chi^2(P\|Q)\right)$
- Bounding χ^2 with $q =$ PDF of $P * \mathcal{N}_\sigma$:

$$\mathbb{E}\chi^2\left(\hat{P}_{X^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) = \frac{1}{n}\left(\int_{\mathbb{R}^d} \mathbb{E}\frac{(\varphi(z-X) - q(z))^2}{q(z)}\mathrm{d}z\right) = \frac{1}{n} I_{\chi^2}(X;Y)$$

$$I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y}\|P_X \otimes P_Y), \quad X \sim P,\ Y = X + Z$$

$$\Longrightarrow \quad \mathsf{D}_{KL} = O\left(\tfrac{1}{n}\right) \text{ if } I_{\chi^2}(X;Y) < \infty.$$

Question: Is I_{χ^2} finite, like $\log(1 + \mathrm{SNR})$, for any finite-2nd-moment X?

Gaussian Smoothed Distances & χ^2-Dichotomy (1)

Define: $I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y} \| P_X \otimes P_Y)$, $Y = X + Z$, $X \sim P$, $Z \sim \mathcal{N}_\sigma$.

Gaussian Smoothed Distances & χ^2-Dichotomy (1)

Define: $I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y} \| P_X \otimes P_Y)$, $Y = X + Z$, $X \sim P$, $Z \sim \mathcal{N}_\sigma$.

Question: Decay rate of $\mathbb{E}\delta(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma)$ for different $\delta(\cdot,\cdot)$?

Gaussian Smoothed Distances & χ^2-Dichotomy (1)

Define: $I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y} \| P_X \otimes P_Y)$, $Y = X + Z$, $X \sim P$, $Z \sim \mathcal{N}_\sigma$.

Question: Decay rate of $\mathbb{E}\delta(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma)$ for different $\delta(\cdot,\cdot)$?

Answer: Dichotomy – all depends on whether $I_{\chi^2}(X;Y)$ is finite (!)

Gaussian Smoothed Distances & χ^2-Dichotomy (1)

Define: $I_{\chi^2}(X;Y) \triangleq \chi^2(P_{X,Y} \| P_X \otimes P_Y)$, $Y = X + Z$, $X \sim P$, $Z \sim \mathcal{N}_\sigma$.

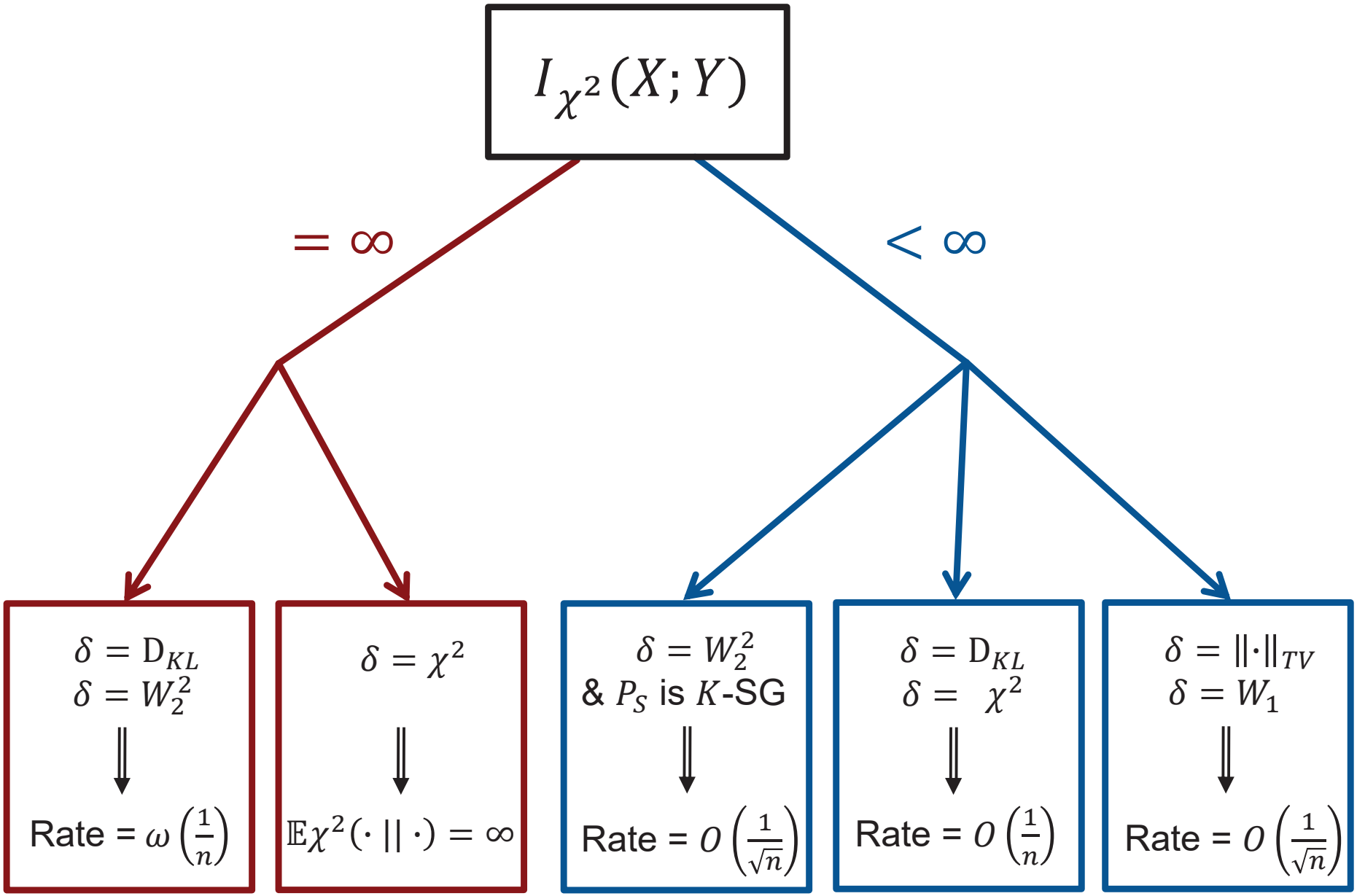
Question: Decay rate of $\mathbb{E}\delta(P * \mathcal{N}_\sigma, \hat{P}_{X^n} * \mathcal{N}_\sigma)$ for different $\delta(\cdot,\cdot)$?

Answer: Dichotomy – all depends on whether $I_{\chi^2}(X;Y)$ is finite (!)

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

1. *If P_X has bounded support, then $I_{\chi^2}(X;Y) < \infty$;*
2. *If P_X be K-subgaussian with $K < \frac{\sigma}{2}$, then $I_{\chi^2}(X;Y) < \infty$;*
3. *If $K > \sqrt{2}\sigma$, then $I_{\chi^2}(X;Y) = \infty$ for some K-subgaussian P.*

Gaussian Smoothed Distances & χ^2-Dichotomy (2)



Summary (for Subgaussian P)

$$\mathbb{E}\left[\delta\left(\hat{P}_{X^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \textbf{???}$$

In All Dimensions: (and different "distances" δ)

Summary (for Subgaussian P)

$$\mathbb{E}\left[\delta\left(\hat{P}_{X^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \textbf{???}$$

In All Dimensions: (and different "distances" δ)

- W_1 and $\|\cdot\|_{\mathsf{TV}}$ are always $O\left(\frac{1}{\sqrt{n}}\right)$

Summary (for Subgaussian P)

$$\mathbb{E}\left[\delta\left(\hat{P}_{X^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \mathbf{???}$$

In All Dimensions: (and different "distances" δ)

- W_1 and $\|\cdot\|_{\mathsf{TV}}$ are always $O\left(\frac{1}{\sqrt{n}}\right)$
- W_2 is $O\left(\frac{1}{\sqrt{n}}\right)$ or $\omega\left(\frac{1}{\sqrt{n}}\right)$. But always $O\left(n^{-\frac{1}{4}}\right)$

Summary (for Subgaussian P)

$$\mathbb{E}\left[\delta\left(\hat{P}_{X^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \mathbf{???}$$

In All Dimensions: (and different "distances" δ)

- W_1 and $\|\cdot\|_{\mathsf{TV}}$ are always $O\left(\frac{1}{\sqrt{n}}\right)$
- W_2 is $O\left(\frac{1}{\sqrt{n}}\right)$ or $\omega\left(\frac{1}{\sqrt{n}}\right)$. But always $O\left(n^{-\frac{1}{4}}\right)$
- D_{KL} is $\frac{1}{n}$ or $\omega\left(\frac{1}{n}\right)$. But always $O\left(\frac{1}{\sqrt{n}}\right)$

Summary (for Subgaussian P)

$$\mathbb{E}\left[\delta\left(\hat{P}_{X^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \mathbf{???}$$

In All Dimensions: (and different "distances" δ)

- W_1 and $\|\cdot\|_{\mathsf{TV}}$ are always $O\left(\frac{1}{\sqrt{n}}\right)$
- W_2 is $O\left(\frac{1}{\sqrt{n}}\right)$ or $\omega\left(\frac{1}{\sqrt{n}}\right)$. But always $O\left(n^{-\frac{1}{4}}\right)$
- D_{KL} is $\frac{1}{n}$ or $\omega\left(\frac{1}{n}\right)$. But always $O\left(\frac{1}{\sqrt{n}}\right)$
- χ^2 is $O\left(\frac{1}{n}\right)$ or $= \infty$

Summary (for Subgaussian P)

$$\mathbb{E}\left[\delta\left(\hat{P}_{X^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp ???$$

In All Dimensions: (and different "distances" δ)

- W_1 and $\|\cdot\|_{\mathsf{TV}}$ are always $O\left(\frac{1}{\sqrt{n}}\right)$
- W_2 is $O\left(\frac{1}{\sqrt{n}}\right)$ or $\omega\left(\frac{1}{\sqrt{n}}\right)$. But always $O\left(n^{-\frac{1}{4}}\right)$
- D_{KL} is $\frac{1}{n}$ or $\omega\left(\frac{1}{n}\right)$. But always $O\left(\frac{1}{\sqrt{n}}\right)$
- χ^2 is $O\left(\frac{1}{n}\right)$ or $= \infty$

Surprise: The dichotomy is fully governed by $I_{\chi^2}(X;Y) \stackrel{?}{=} \infty$

For the impatient: direct flight from χ^2 to h

Let $P = \hat{P}_{X^n} * \mathcal{N}$, $Q = P * \mathcal{N}$.

$$h(P) - h(Q) \quad = \quad \left\{ \mathbb{E}_P - \mathbb{E}_Q \left[\log \frac{1}{Q(X)} \right] \right\} - D(P \| Q)$$

For the impatient: direct flight from χ^2 to h

Let $P = \hat{P}_{X^n} * \mathcal{N}$, $Q = P * \mathcal{N}$.

$$\begin{aligned} h(P) - h(Q) &= \left\{ \mathbb{E}_P - \mathbb{E}_Q \left[\log \frac{1}{Q(X)} \right] \right\} - D(P\|Q) \\ &\leq \left\{ \mathbb{E}_P - \mathbb{E}_Q \left[\log \frac{1}{Q(X)} \right] \right\} \end{aligned}$$

For the impatient: direct flight from χ^2 to h

Let $P = \hat{P}_{X^n} * \mathcal{N}$, $Q = P * \mathcal{N}$.

$$\begin{aligned} h(P) - h(Q) &= \left\{ \mathbb{E}_P - \mathbb{E}_Q \left[\log \frac{1}{Q(X)} \right] \right\} - D(P\|Q) \\ &\leq \left\{ \mathbb{E}_P - \mathbb{E}_Q \left[\log \frac{1}{Q(X)} \right] \right\} \\ &\leq \sqrt{\mathrm{Var}_Q \left[\log \frac{1}{Q} \right]} \cdot \sqrt{\chi^2(P\|Q)} \end{aligned}$$

For the impatient: direct flight from χ^2 to h

Let $P = \hat{P}_{X^n} * \mathcal{N}$, $Q = P * \mathcal{N}$.

$$\begin{aligned} h(P) - h(Q) &= \left\{ \mathbb{E}_P - \mathbb{E}_Q \left[\log \frac{1}{Q(X)} \right] \right\} - D(P\|Q) \\ &\leq \left\{ \mathbb{E}_P - \mathbb{E}_Q \left[\log \frac{1}{Q(X)} \right] \right\} \\ &\leq \sqrt{\mathrm{Var}_Q \left[\log \frac{1}{Q} \right]} \cdot \sqrt{\chi^2(P\|Q)} \\ &\lesssim \sqrt{\chi^2(P\|Q)} \end{aligned}$$

Last step: Use the fact that $Q = (\text{something}) * \mathcal{N}_\sigma$.

Is Exponentiality in Dimension Necessary?

Is Exponentiality in Dimension Necessary?

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

For any $\sigma > 0$, *sufficiently large* d *and sufficiently small* $\eta > 0$, *we have* $n^\star(\eta, \sigma, \mathcal{F}_d) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, *where* $\gamma(\sigma) > 0$ *is monotonically decreasing in* σ.

Is Exponentiality in Dimension Necessary?

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

For any $\sigma > 0$, sufficiently large d and sufficiently small $\eta > 0$, we have $n^\star(\eta, \sigma, \mathcal{F}_d) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma) > 0$ is monotonically decreasing in σ.

$\implies O\left(c^d n^{-\frac{1}{2}}\right)$ rate attained by plugin estimator is sharp in n and d

Is Exponentiality in Dimension Necessary?

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

For any $\sigma > 0$, sufficiently large d and sufficiently small $\eta > 0$, we have $n^\star(\eta, \sigma, \mathcal{F}_d) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma) > 0$ is monotonically decreasing in σ.

$\Longrightarrow$ $O\left(c^d n^{-\frac{1}{2}}\right)$ rate attained by plugin estimator is sharp in n and d

Proof (main ideas):

Is Exponentiality in Dimension Necessary?

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

For any $\sigma > 0$, sufficiently large d and sufficiently small $\eta > 0$, we have $n^\star(\eta, \sigma, \mathcal{F}_d) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma) > 0$ is monotonically decreasing in σ.

$\Longrightarrow$ $O\left(c^d n^{-\frac{1}{2}}\right)$ rate attained by plugin estimator is sharp in n and d

Proof (main ideas):

- Relate $h(P * \mathcal{N}_\sigma)$ to Shannon entropy $H(Q)$

  $\mathrm{supp}(Q) =$ peak-constrained AWGN capacity achieving codebook $\mathcal{C}_d$

Is Exponentiality in Dimension Necessary?

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

For any $\sigma > 0$, sufficiently large d and sufficiently small $\eta > 0$, we have $n^\star(\eta, \sigma, \mathcal{F}_d) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma) > 0$ is monotonically decreasing in σ.

$\implies O\left(c^d n^{-\frac{1}{2}}\right)$ rate attained by plugin estimator is sharp in n and d

Proof (main ideas):

- Relate $h(P * \mathcal{N}_\sigma)$ to Shannon entropy $H(Q)$

  $\mathrm{supp}(Q) =$ peak-constrained AWGN capacity achieving codebook $\mathcal{C}_d$

- $H(Q)$ estimation sample complexity $\Omega\left(\frac{|\mathcal{C}_d|}{\eta \log |\mathcal{C}_d|}\right)$

Simulations - Synthetic Experiments

Comparison: General-purpose est. accessing sample of $X + Z \sim P * \mathcal{N}_\sigma$

Simulations - Synthetic Experiments

Comparison: General-purpose est. accessing sample of $X + Z \sim P * \mathcal{N}_\sigma$

1. LOO KDE Estimator from [Kandasamy et al.'15]

Simulations - Synthetic Experiments

Comparison: General-purpose est. accessing sample of $X + Z \sim P * \mathcal{N}_\sigma$

1. LOO KDE Estimator from [Kandasamy et al.'15]
2. Kozachenko-Leonenko (KL) kNN Estimator [Kozachenko-Leonenko'87]

Simulations - Synthetic Experiments

Comparison: General-purpose est. accessing sample of $X + Z \sim P * \mathcal{N}_\sigma$

1. LOO KDE Estimator from [Kandasamy et al.'15]
2. Kozachenko-Leonenko (KL) kNN Estimator [Kozachenko-Leonenko'87]
3. Weighted KL (wKL) Estimator from [Berrett-Samworth-Yuan'19]

Simulations - Synthetic Experiments

Comparison: General-purpose est. accessing sample of $X + Z \sim P * \mathcal{N}_\sigma$

1. LOO KDE Estimator from [Kandasamy et al.'15]
2. Kozachenko-Leonenko (KL) kNN Estimator [Kozachenko-Leonenko'87]
3. Weighted KL (wKL) Estimator from [Berrett-Samworth-Yuan'19]

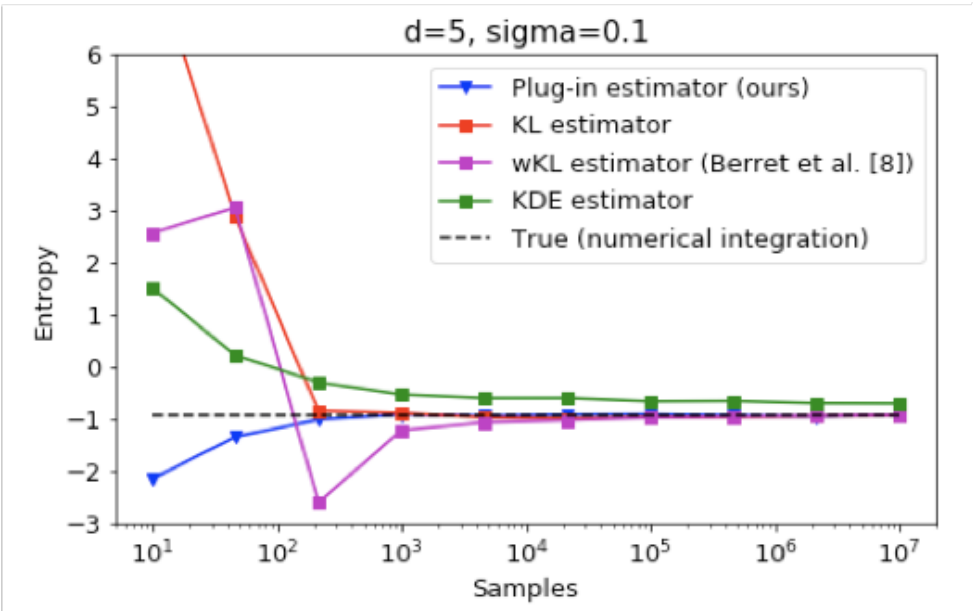
Bdd. Support: $P \propto 2^{-d} \sum_{x \in \{-1,1\}^d} \mathcal{N}(x, \mathrm{I}_d) \cdot \mathbb{1}_{[-1,1]^d}$ (truncated GMM)

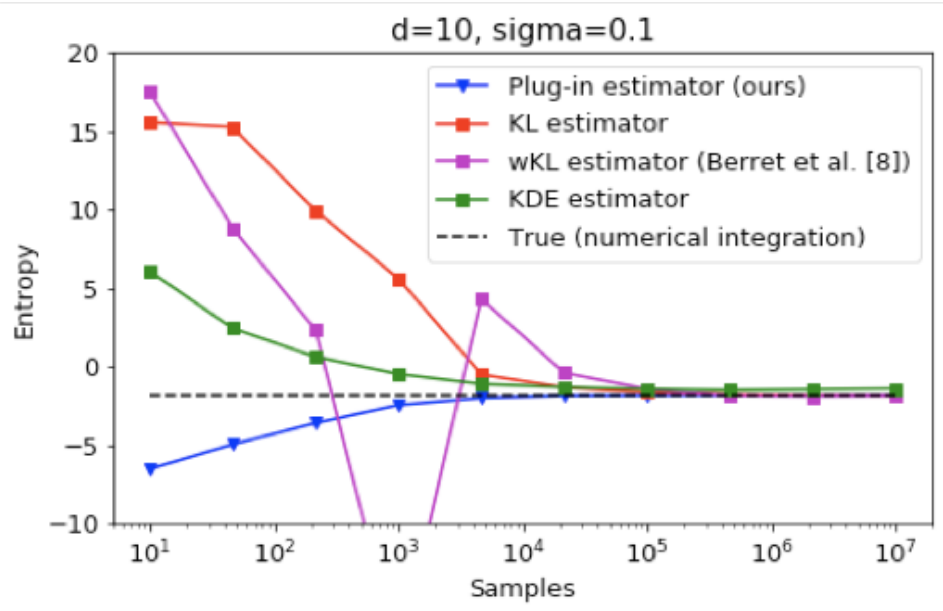
Simulations - Synthetic Experiments

Comparison: General-purpose est. accessing sample of $X + Z \sim P * \mathcal{N}_\sigma$

1. LOO KDE Estimator from [Kandasamy et al.'15]
2. Kozachenko-Leonenko (KL) kNN Estimator [Kozachenko-Leonenko'87]
3. Weighted KL (wKL) Estimator from [Berrett-Samworth-Yuan'19]

Bdd. Support: $P \propto 2^{-d} \sum_{x \in \{-1,1\}^d} \mathcal{N}(x, \mathrm{I}_d) \cdot \mathbb{1}_{[-1,1]^d}$ (truncated GMM)



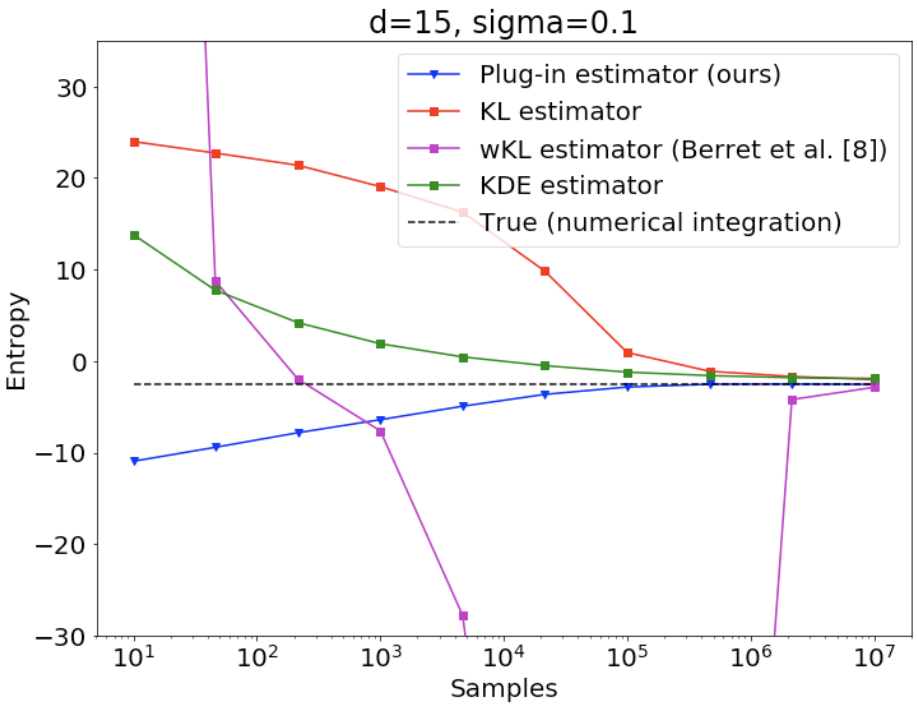


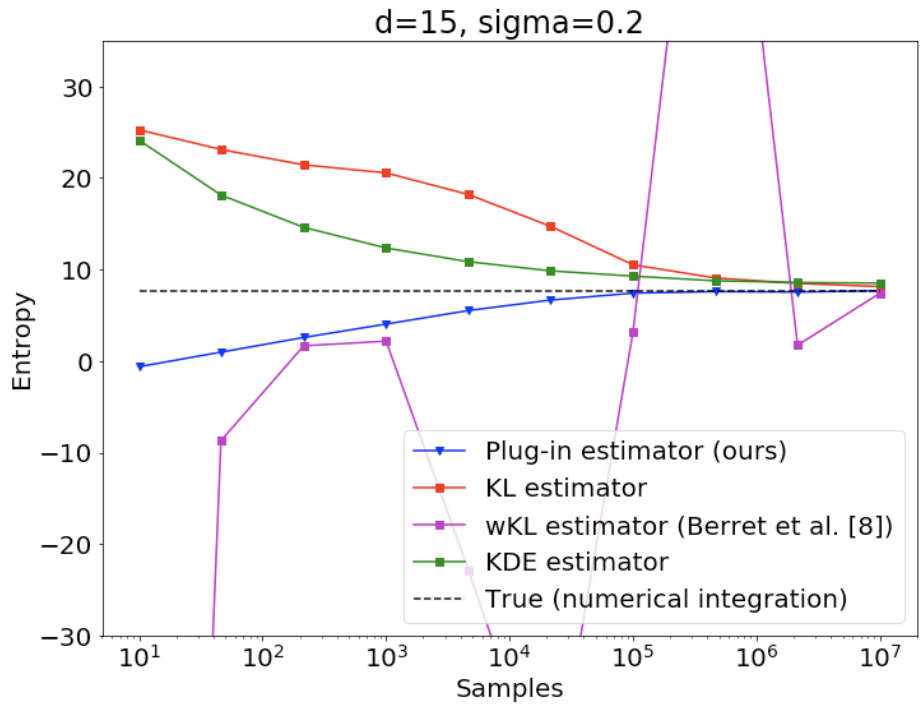
Simulations - Synthetic Experiments

Unbounded Support: P as before but w/o truncation

Simulations - Synthetic Experiments

Unbounded Support: P as before but w/o truncation





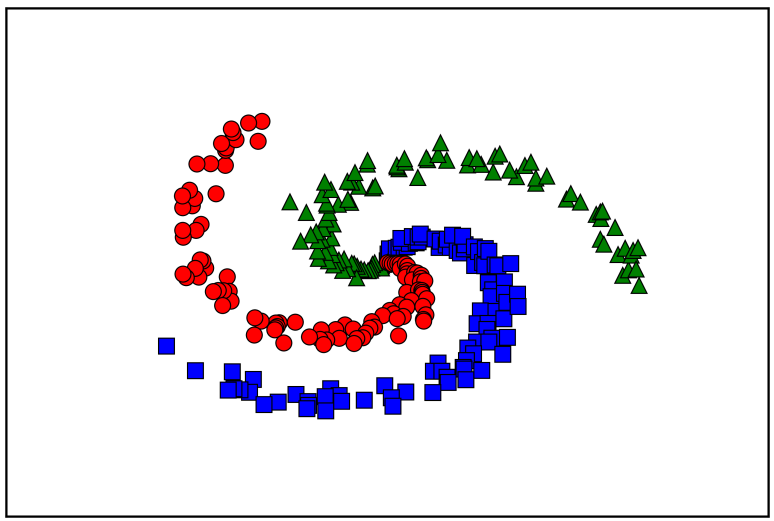
Simulations - Noisy Deep Neural Network Example

Setup: Noisy DNN for spiral dataset classification

Simulations - Noisy Deep Neural Network Example

Setup: Noisy DNN for spiral dataset classification

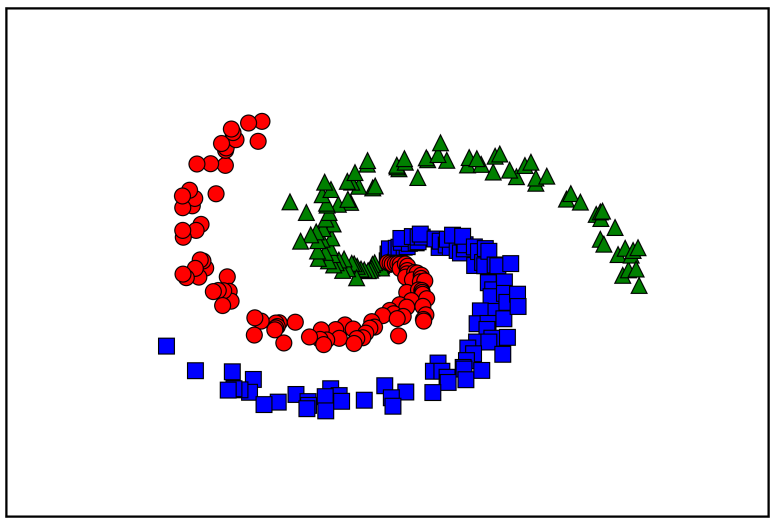
- **Dataset:** 2-dimensional 3-class spiral dataset



Simulations - Noisy Deep Neural Network Example

Setup: Noisy DNN for spiral dataset classification

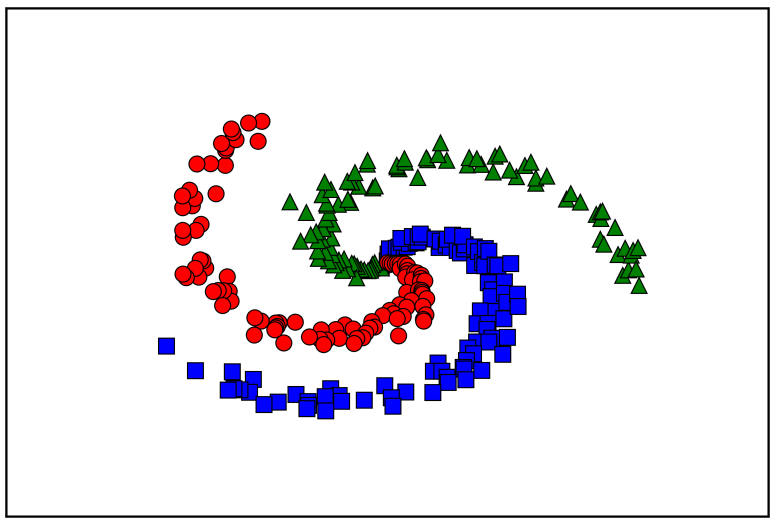
- **Dataset:** 2-dimensional 3-class spiral dataset
- **Network:** 2–8–9–10–3 fully connected noisy ($\sigma = 0.2$) tanh DNN



Simulations - Noisy Deep Neural Network Example

Setup: Noisy DNN for spiral dataset classification

- **Dataset:** 2-dimensional 3-class spiral dataset
- **Network:** 2–8–9–10–3 fully connected noisy ($\sigma = 0.2$) tanh DNN
- **Classification:** Trained to 98% test accuracy

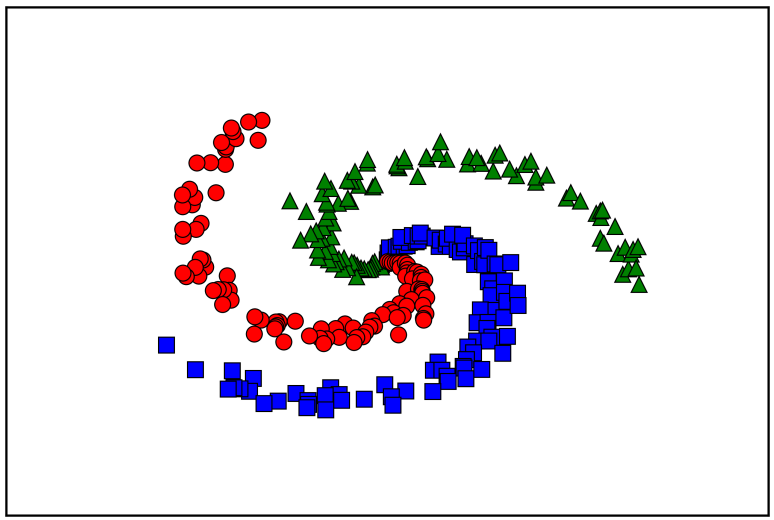


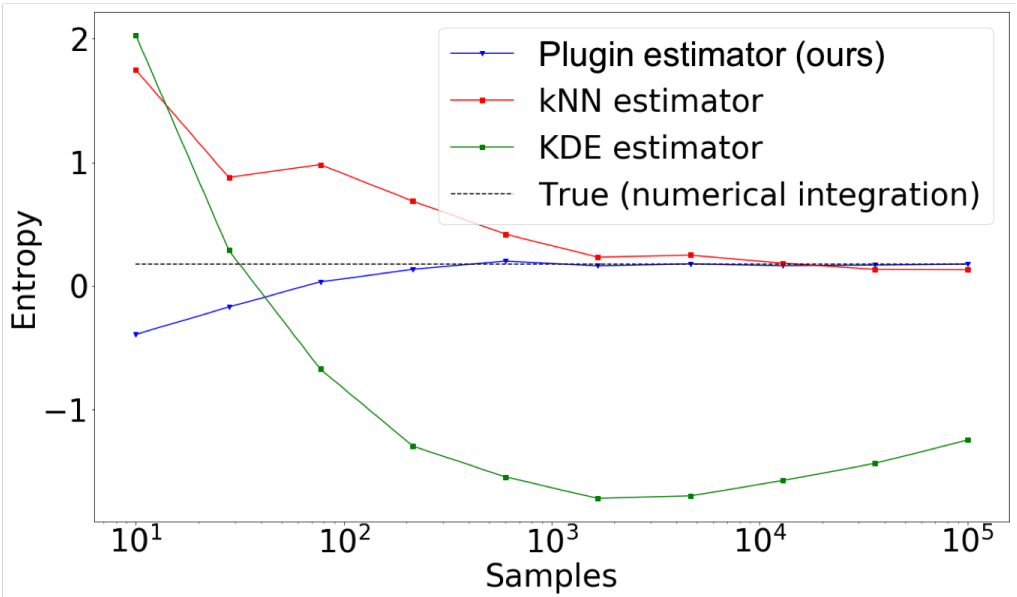
Simulations - Noisy Deep Neural Network Example

Setup: Noisy DNN for spiral dataset classification

- **Dataset:** 2-dimensional 3-class spiral dataset
- **Network:** 2–8–9–10–3 fully connected noisy ($\sigma = 0.2$) tanh DNN
- **Classification:** Trained to 98% test accuracy

⊛ Estimating the entropy of 10-dimensional layer





Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem
- **Intrinsically Difficult:**

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem
- **Intrinsically Difficult:**
  - Sample complexity is exponential in dimension

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem
- **Intrinsically Difficult:**
  - Sample complexity is exponential in dimension
- **Plug-in Estimator:**

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem
- **Intrinsically Difficult:**
  - Sample complexity is exponential in dimension
- **Plug-in Estimator:**
  - Attains parametric estimation rate $O\left(c^d n^{-\frac{1}{2}}\right)$

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem
- **Intrinsically Difficult:**
  - Sample complexity is exponential in dimension
- **Plug-in Estimator:**
  - Attains parametric estimation rate $O\left(c^d n^{-\frac{1}{2}}\right)$
  - Empirically outperforms general-purpose estimation via 'noisy' samples

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem
- **Intrinsically Difficult:**
  - Sample complexity is exponential in dimension
- **Plug-in Estimator:**
  - Attains parametric estimation rate $O\left(c^d n^{-\frac{1}{2}}\right)$
  - Empirically outperforms general-purpose estimation via 'noisy' samples
- **Gaussian Smoothed Empirical Approximation**: χ^2 dichotomy

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem
- **Intrinsically Difficult:**
  - Sample complexity is exponential in dimension
- **Plug-in Estimator:**
  - Attains parametric estimation rate $O\left(c^d n^{-\frac{1}{2}}\right)$
  - Empirically outperforms general-purpose estimation via 'noisy' samples
- **Gaussian Smoothed Empirical Approximation**: χ^2 dichotomy
- **arXiv:1810.05728**: Study MI trends during DNN training (estimation)

Summary

Paper available at arXiv:1905.13576

- **Differential Entropy Estimation under Gaussian Convolutions:**
  - New high-dimensional & nonparametric functional estimation problem
- **Intrinsically Difficult:**
  - Sample complexity is exponential in dimension
- **Plug-in Estimator:**
  - Attains parametric estimation rate $O\left(c^d n^{-\frac{1}{2}}\right)$
  - Empirically outperforms general-purpose estimation via 'noisy' samples
- **Gaussian Smoothed Empirical Approximation**: χ^2 dichotomy
- **arXiv:1810.05728**: Study MI trends during DNN training (estimation)

Thank you!