Convergence of Smoothed Empirical Measures under Wasserstein Distance

Yury Polyanskiy
Joint work Zeyu Jia, Adam Block, and Sasha Rakhlin

Massachusetts Institute of Technology

November 30, 2021

Smoothed Empirical Measures

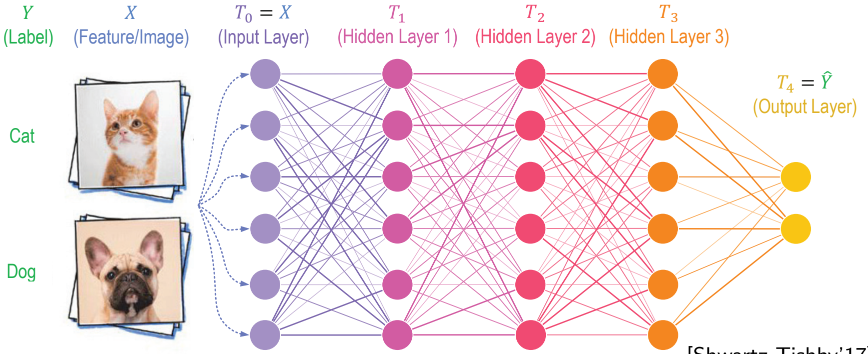
- **Empirical Measures:** Given distribution $\mathbb{P}$, the empirical measure of $\mathbb{P}$ is $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, where $X_i \sim \mathbb{P}$;
- **Smoothed Empirical Measures:** For given σ, the smoothed empirical measure is the convolution of empirical measure and $\mathcal{N}(0, \sigma^2)$:

$$\mathbb{P}_n * \mathcal{N}(0, \sigma^2).$$

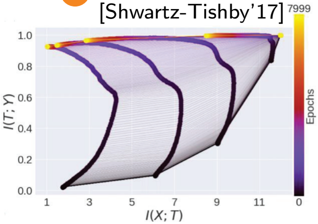
Smoothed Empirical Measures

- **Empirical Measures:** Given distribution $\mathbb{P}$, the empirical measure of $\mathbb{P}$ is $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, where $X_i \sim \mathbb{P}$;
- **Smoothed Empirical Measures:** For given σ, the smoothed empirical measure is the convolution of empirical measure and $\mathcal{N}(0, \sigma^2)$:
$$\mathbb{P}_n * \mathcal{N}(0, \sigma^2).$$
- Why?

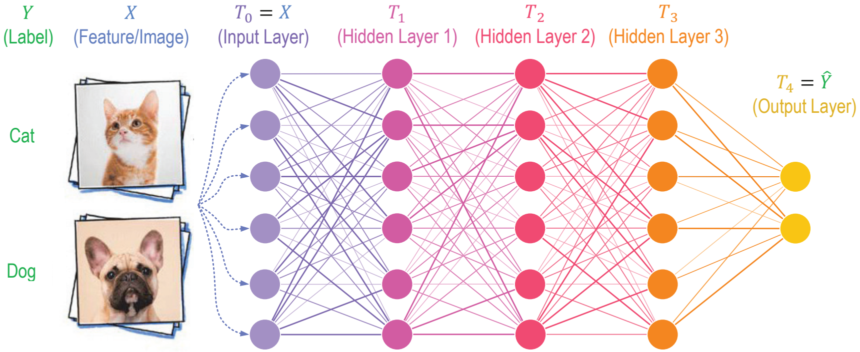
Feedforward DNN: Each layer $T_\ell = f_\ell(T_{\ell-1})$



[Shwartz-Tishby'17]



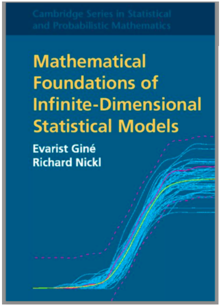
Feedforward DNN: Each layer $T_\ell = f_\ell(T_{\ell-1})$



How to talk about $I(Y; T_\ell)$ & $I(X; T_\ell)$?

1. *Formally:* these are (almost) indep of DNN weights if X is discrete
2. *Practically:* Should not bother about info at 10^{-6} scale...
3. *Our solution:* **add noise** to neuron outputs

Textbook idea



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

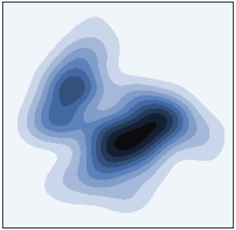
Empirical 1-Wasserstein Distance:

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:

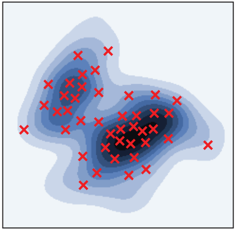
- Distribution P on $\mathbb{R}^d$

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:

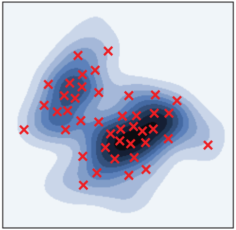
- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$

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:

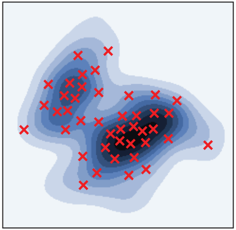
- Distribution P on $\mathbb{R}^d \implies$ i.i.d. Samples $(X_i)_{i=1}^n$
- Empirical distribution $\mathbb{P}_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

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 $\mathbb{P}_n \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

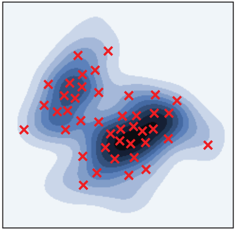
$\implies$ Dependence on (n, d) of $\mathbb{E}W_1(P, \mathbb{P}_n)$

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 $\mathbb{P}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

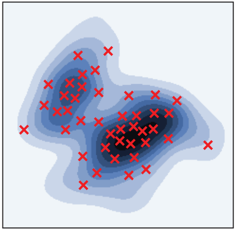
$\implies$ Dependence on (n, d) of $\mathbb{E} W_1\left(P, \mathbb{P}_n\right) \asymp n^{-\frac{1}{d}}$

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 $(X_i)_{i=1}^n$
- Empirical distribution $\mathbb{P}_n \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$

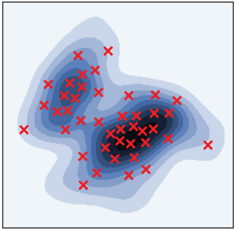
$\implies$ Dependence on (n, d) of $\mathbb{E}W_1\left(P, \mathbb{P}_n\right) \asymp n^{-\frac{1}{d}}$ (for cts. P, $d \geq 3$)

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 $\mathbb{P}_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

$\implies$ Dependence on (n, d) of $\mathbb{E}W_1\left(P, \mathbb{P}_n\right) \asymp n^{-\frac{1}{d}}$ (for cts. P, $d \geq 3$)

Theorem (Goldfeld-Greenewald-Polyanskiy-Weed'20)

For any d, we have $\mathbb{E}W_1\left(P * \mathcal{N}_\sigma, \mathbb{P}_n * \mathcal{N}_\sigma\right) \leq O_{\sigma,d}\left(n^{-\frac{1}{2}}\right)$ provided P is K-subgaussian.

Convergence w.r.t. other distances?

- Question: What about convergence of $\mathbb{P}_n * \mathcal{N}_\sigma \to P * \mathcal{N}_\sigma$ in other distances? Namely:
  - $\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)\right] \asymp ?$
  - $\mathbb{E}\left[D_{KL}(\mathbb{P}_n * \mathcal{N}_\sigma \| P * \mathcal{N}_\sigma)\right] \asymp ?$
  - $\mathbb{E}\left[\chi^2(\mathbb{P}_n * \mathcal{N}_\sigma \| P * \mathcal{N}_\sigma)\right] \asymp ?$

Convergence w.r.t. other distances?

- Question: What about convergence of $\mathbb{P}_n * \mathcal{N}_\sigma \to P * \mathcal{N}_\sigma$ in other distances? Namely:
  - $\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)\right] \asymp ?$
  - $\mathbb{E}\left[D_{KL}(\mathbb{P}_n * \mathcal{N}_\sigma \| P * \mathcal{N}_\sigma)\right] \asymp ?$
  - $\mathbb{E}\left[\chi^2(\mathbb{P}_n * \mathcal{N}_\sigma \| P * \mathcal{N}_\sigma)\right] \asymp ?$
- Surprisingly, the answer is governed by the quantity $I_{\chi^2}(X; X + \sigma Z)$:

$$I_{\chi^2}(X; Y) \triangleq \chi^2(P_{X,Y} \| P_X \otimes P_Y)$$

Convergence of smoothed empirical distributions

Theorem (Goldfeld-Greenewald-Polyanskiy-Weed'20)

For any dimension d: *If $I_{\chi^2}(X;Y) < \infty$*

$$\mathbb{E}[\delta(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)] = e^{O_\sigma(d)} \cdot \frac{1}{n} \qquad \delta \in \{W_2^2, D_{KL}, \chi^2\}$$

Otherwise, *if $I_{\chi^2}(X;Y) = \infty$*

$$\mathbb{E}[\chi^2(\cdots)] = \infty, \quad \mathbb{E}[W_2^2(\cdots)], \mathbb{E}[D_{KL}(\cdots)], = \omega\left(\frac{1}{n}\right).$$

Convergence of smoothed empirical distributions

Theorem (Goldfeld-Greenewald-Polyanskiy-Weed'20)

For any dimension d: *If $I_{\chi^2}(X;Y) < \infty$*

$$\mathbb{E}[\delta(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)] = e^{O_\sigma(d)} \cdot \frac{1}{n} \qquad \delta \in \{W_2^2, D_{KL}, \chi^2\}$$

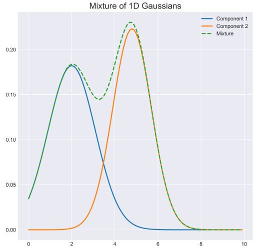
(For W_2^2 also need to assume P is K-subgaussian with $K < \sigma$.)

Otherwise, *if $I_{\chi^2}(X;Y) = \infty$*

$$\mathbb{E}[\chi^2(\cdots)] = \infty, \quad \mathbb{E}[W_2^2(\cdots)], \mathbb{E}[D_{KL}(\cdots)], = \omega\left(\frac{1}{n}\right).$$

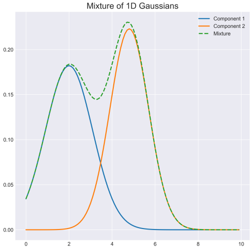
(For W_2^2 also we use $*\mathcal{N}_\tau$ with $\tau < \sigma$).

The result is interesting already in $d = 1$



- Consider $P = \mathrm{Ber}(\frac{1}{2})$. Then $\mathbb{P}_n = \mathrm{Ber}(\frac{1}{2} + \frac{Z}{\sqrt{n}})$

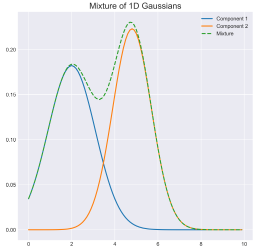
The result is interesting already in $d = 1$



- Consider $P = \mathrm{Ber}(\frac{1}{2})$. Then $\mathbb{P}_n = \mathrm{Ber}(\frac{1}{2} + \frac{Z}{\sqrt{n}})$
- Since $\frac{Z}{\sqrt{n}}$ mass must travel distance-1, we have

$$\mathbb{E}[W_2^2(\mathbb{P}_n, P)] \gtrsim \frac{1}{\sqrt{n}}$$

The result is interesting already in $d = 1$



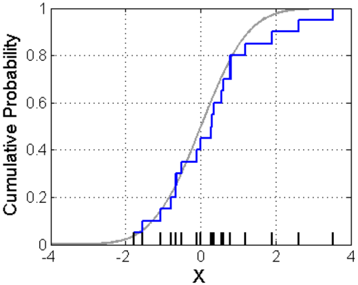
- Consider $P = \mathrm{Ber}(\frac{1}{2})$. Then $\mathbb{P}_n = \mathrm{Ber}(\frac{1}{2} + \frac{Z}{\sqrt{n}})$
- Since $\frac{Z}{\sqrt{n}}$ mass must travel distance-1, we have

$$\mathbb{E}[W_2^2(\mathbb{P}_n, P)] \gtrsim \frac{1}{\sqrt{n}}$$

At the same time for arbitrarily small $\sigma > 0$:

$$\mathbb{E}[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)] = O_\sigma(\frac{1}{n})$$

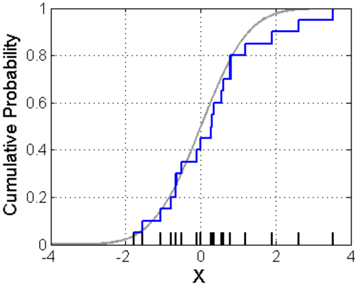
The result is interesting already in $d = 1$



- Even for $P = \mathcal{N}(0,1)$ we have [Bobkov-Ledoux'16]:

$$\mathbb{E}[W_2^2(\mathbb{P}_n, P)] \asymp \frac{\log \log n}{n}$$

The result is interesting already in $d = 1$



- Even for $P = \mathcal{N}(0,1)$ we have [Bobkov-Ledoux'16]:

$$\mathbb{E}[W_2^2(\mathbb{P}_n, P)] \asymp \frac{\log\log n}{n}$$

- while for any $\sigma > 0$:

$$\mathbb{E}[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)] = O_\sigma(\frac{1}{n})$$

(indeed, $I_\chi^2(X; Y) < \infty$ for $X \sim \mathcal{N}$)

2020 and 2021: When is $I_{\chi^2}(X;Y) < \infty$?

Theorem (Goldfeld-Greenewald-Polyanskiy-Weed'20)

1. *If P_X has bounded support, then $I_{\chi^2}(X;Y) < \infty$;*
2. *If P_X is 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.*

Recall: X is K-subgaussian iff

$$\mathbb{E}[e^{\lambda^T(X-\mathbb{E}[X])}] \le e^{\frac{K^2}{2}\|\lambda\|_2^2} \qquad \forall \lambda \in \mathbb{R}^d$$

2020 and 2021: When is $I_{\chi^2}(X;Y) < \infty$?

Theorem (Goldfeld-Greenewald-Polyanskiy-Weed'20)

1. *If P_X has bounded support, then $I_{\chi^2}(X;Y) < \infty$;*
2. *If P_X is 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.*

Recall: X is K-subgaussian iff

$$\mathbb{E}[e^{\lambda^T(X-\mathbb{E}[X])}] \le e^{\frac{K^2}{2}\|\lambda\|_2^2} \qquad \forall \lambda \in \mathbb{R}^d$$

Theorem (Jia-Block-Polyanskiy-Rakhlin'21)

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

2020 and 2021: When is $I_{\chi^2}(X;Y) < \infty$?

Theorem (Goldfeld-Greenewald-Polyanskiy-Weed'20)

1. *If P_X has bounded support, then $I_{\chi^2}(X;Y) < \infty$;*
2. *If P_X is 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.*

Recall: X is K-subgaussian iff

$$\mathbb{E}[e^{\lambda^T(X-\mathbb{E}[X])}] \le e^{\frac{K^2}{2}\|\lambda\|_2^2} \qquad \forall \lambda \in \mathbb{R}^d$$

Theorem (Jia-Block-Polyanskiy-Rakhlin'21)

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

Closes entire range (except $K = \sigma$).

$I_{\chi^2} < \infty$: proof idea ($K < \sigma$)

- When $K < \sigma$, we write

$$I_{\chi^2}(S;Y) = \mathbb{E}_{S\sim\mathbb{P}} \int_{\mathbb{R}^d} \frac{\varphi^2_{\sigma^2 I_d}(z-S)}{\mathbb{E}_{\tilde{S}\sim P}\varphi_{\sigma^2 I_d}(z-\tilde{S})} dz - 1,$$

where $\varphi_{\sigma^2 I_d}(\cdot)$ is the PDF of $\mathcal{N}(0, \sigma^2 I_d)$.

$I_{\chi^2} < \infty$: proof idea ($K < \sigma$)

- When $K < \sigma$, we write

$$I_{\chi^2}(S;Y) = \mathbb{E}_{S\sim\mathbb{P}} \int_{\mathbb{R}^d} \frac{\varphi^2_{\sigma^2 I_d}(z-S)}{\mathbb{E}_{\tilde{S}\sim P}\varphi_{\sigma^2 I_d}(z-\tilde{S})} dz - 1,$$

  where $\varphi_{\sigma^2 I_d}(\cdot)$ is the PDF of $\mathcal{N}(0, \sigma^2 I_d)$.
- Divide the domain of $\mathbb{E}_{S\sim P}\int_{\mathbb{R}^d}$ into the following three parts:
  1. $A = \{\|S\|_2 \le 1\}$;
  2. $B = \{\|S\|_2 > 1 \text{ and } \|z-S\|_2 \ge \delta\|S\|_2\}$;
  3. $C = \{\|z-S\|_2 < \delta\|S\|_2\}$;

  and proved $\mathbb{E}_{S\sim P}\int_{\mathbb{R}^d}$ in each parts is less than infinity.

$I_{\chi^2} = \infty$ counter-example ($K > \sigma$)

- Choose the hard case

$$\mathbb{P} = p_0\delta_0 + \sum_{k=1}^{\infty} p_k \delta_{r_k},$$

with $r_k = c^{k-1}, p_k = c_0 \exp\left(-\frac{r_k^2}{2K^2}\right)$ for some constant c_0, c and $p_0 = 1 - \sum_{k=1}^{\infty} p_k$.

$I_{\chi^2} = \infty$ counter-example ($K > \sigma$)

- Choose the hard case

$$\mathbb{P} = p_0\delta_0 + \sum_{k=1}^{\infty} p_k \delta_{r_k},$$

  with $r_k = c^{k-1}, p_k = c_0 \exp\left(-\frac{r_k^2}{2K^2}\right)$ for some constant c_0, c and $p_0 = 1 - \sum_{k=1}^{\infty} p_k$.
- $\mathbb{P}$ is K-subgaussian.

$I_{\chi^2} = \infty$ counter-example $(K > \sigma)$

- Choose the hard case

$$\mathbb{P} = p_0 \delta_0 + \sum_{k=1}^{\infty} p_k \delta_{r_k},$$

with $r_k = c^{k-1}, p_k = c_0 \exp\left(-\frac{r_k^2}{2K^2}\right)$ for some constant c_0, c and $p_0 = 1 - \sum_{k=1}^{\infty} p_k$.

- $\mathbb{P}$ is K-subgaussian.
- When $\sigma < K$, $\delta_{r_j} * \mathcal{N}_\sigma$ for $j \neq k$ hardly affect the density of $\mathbb{P} * \mathcal{N}_\sigma$ in comparison to $\delta_{r_k} * \mathcal{N}_\sigma$ if c is chosen large enough.

$I_{\chi^2} = \infty$ counter-example

- WLOG, we assume $\sigma = 1$;

$I_{\chi^2} = \infty$ counter-example

- WLOG, we assume $\sigma = 1$;
- $I_{\chi^2}(S; Y)$ can be decomposed into

$$I_{\chi^2}(S; Y) = \sum_{k=0}^{\infty} \int_{\mathbb{R}} \frac{\varphi_{\frac{1}{\sqrt{2}}}(z - r_k)}{\varphi_1(z - r_k)} \cdot \frac{1}{1 + \sum_{j \neq k} \frac{p_j}{p_k} \frac{\varphi_1(z - r_j)}{\varphi_1(z - r_k)}} dz - 1.$$

$I_{\chi^2} = \infty$ counter-example

- WLOG, we assume $\sigma = 1$;
- $I_{\chi^2}(S;Y)$ can be decomposed into

$$I_{\chi^2}(S;Y) = \sum_{k=0}^{\infty} \int_{\mathbb{R}} \frac{\varphi_{\frac{1}{\sqrt{2}}}(z-r_k)}{\varphi_1(z-r_k)} \cdot \frac{1}{1+\sum_{j\neq k} \frac{p_j}{p_k} \frac{\varphi_1(z-r_j)}{\varphi_1(z-r_k)}} dz - 1.$$

- When z is in a small neighborhood of r_k, $\frac{\varphi_{1/\sqrt{2}}(z-r_k)}{\varphi_1(z-r_k)}$ is uniformly lower bounded for all k.

$I_{\chi^2} = \infty$ counter-example

- WLOG, we assume $\sigma = 1$;
- $I_{\chi^2}(S; Y)$ can be decomposed into

$$I_{\chi^2}(S;Y) = \sum_{k=0}^{\infty} \int_{\mathbb{R}} \frac{\varphi_{\frac{1}{\sqrt{2}}}(z - r_k)}{\varphi_1(z - r_k)} \cdot \frac{1}{1 + \sum_{j \neq k} \frac{p_j}{p_k} \frac{\varphi_1(z - r_j)}{\varphi_1(z - r_k)}} dz - 1.$$

- When z is in a small neighborhood of r_k, $\frac{\varphi_{1/\sqrt{2}}(z - r_k)}{\varphi_1(z - r_k)}$ is uniformly lower bounded for all k.
- When z is in a small neighborhood of r_k $j \neq k$ we have

$$\frac{\varphi_1(z - r_j)}{\varphi_1(z - r_k)} \leq \exp(-j/2).$$

Summary for K-Subgaussian P

$$\sup_{P \in \mathrm{SubG}(K)} \mathbb{E}\left[\delta\left(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \textbf{???}$$

In All Dimensions:

Summary for K-Subgaussian P

$$\sup_{P \in \mathrm{SubG}(K)} \mathbb{E}\left[\delta\left(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \textbf{???}$$

In All Dimensions:

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

Summary for K-Subgaussian P

$$\sup_{P \in \mathrm{SubG}(K)} \mathbb{E}\left[\delta\left(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \textbf{???}$$

In All Dimensions:

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

Summary for K-Subgaussian P

$$\sup_{P \in \mathrm{SubG}(K)} \mathbb{E}\left[\delta\left(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \mathbf{???}$$

In All Dimensions:

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

Summary for K-Subgaussian P

$$\sup_{P \in \mathrm{SubG}(K)} \mathbb{E}\left[\delta\left(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \textbf{???}$$

In All Dimensions:

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

$$\sup_{P \in \mathrm{SubG}(K)} \mathbb{E}\left[\delta\left(\mathbb{P}_n * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right)\right] \asymp \textbf{???}$$

In All Dimensions:

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

Threshold: In all cases the alternative is governed by $K < \sigma$ vs $K > \sigma$

Convergence of Smoothed W2 Convergence in 1D

Question: When rate is $\omega(\frac{1}{n})$ does it switch to $\frac{1}{\sqrt{n}}$ right away?

Convergence of Smoothed W2 Convergence in 1D

Question: When rate is $\omega(\frac{1}{n})$ does it switch to $\frac{1}{\sqrt{n}}$ right away? No!

Convergence of Smoothed W2 Convergence in 1D

Question: When rate is $\omega(\frac{1}{n})$ does it switch to $\frac{1}{\sqrt{n}}$ right away? No!

Theorem (Jia-Block-Polyanskiy-Rakhlin'21)

In dimension $d = 1$ we have:

- *For any K-subgaussian distribution $\mathbb{P}$, we have*

$$\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma)\right] = \tilde{O}\left(n^{-\frac{K^2}{2K^2-\sigma^2}}\right).$$

- *There exists a K-subgaussian distribution $\mathbb{P}$ such that*

$$\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma)\right] = \tilde{\Omega}\left(n^{-\frac{(\sigma^2+K^2)^2}{2(\sigma^4+K^4)}}\right).$$

Convergence of Smoothed W2 Convergence in 1D

Question: When rate is $\omega(\frac{1}{n})$ does it switch to $\frac{1}{\sqrt{n}}$ right away? No!

Theorem (Jia-Block-Polyanskiy-Rakhlin'21)

In dimension $d = 1$ we have:

- *For any K-subgaussian distribution $\mathbb{P}$, we have*

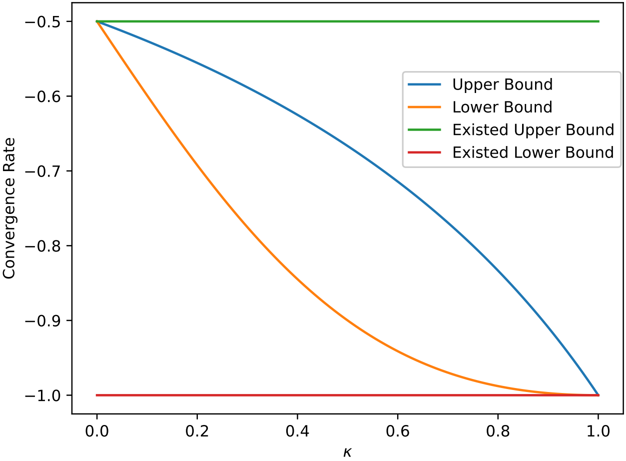
$$\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma)\right] = \tilde{O}\left(n^{-\frac{K^2}{2K^2-\sigma^2}}\right).$$

- *There exists a K-subgaussian distribution $\mathbb{P}$ such that*

$$\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma)\right] = \tilde{\Omega}\left(n^{-\frac{(\sigma^2+K^2)^2}{2(\sigma^4+K^4)}}\right).$$

Proof ideas: 1. use optimal (quantile-quantile) coupling
2. use dyadic haircomb c/ex.
3. $\tilde{O}(n^{-E})$ is in fact $O(n^{-E+\epsilon})$

W2 Convergence in 1D: illustration



$$\kappa \triangleq \frac{\sigma^2}{K^2}$$

Convergence of Smoothed KL Divergence

- [GGNWP20]: If $\sigma > K$ then

$$\mathbb{E}\left[D_{KL}(\mathbb{P}_n * \mathcal{N}_\sigma || \mathbb{P} * \mathcal{N}_\sigma)\right] = \mathcal{O}\left(n^{-1}\right)$$

Convergence of Smoothed KL Divergence

- [GGNWP20]: If $\sigma > K$ then

$$\mathbb{E}\left[D_{KL}(\mathbb{P}_n * \mathcal{N}_\sigma \| \mathbb{P} * \mathcal{N}_\sigma)\right] = \mathcal{O}\left(n^{-1}\right)$$

- When $\sigma < K$, there exists a distribution $\mathbb{P}$ such that

$$\mathbb{E}\left[D_{KL}(\mathbb{P}_n * \mathcal{N}_\sigma \| \mathbb{P} * \mathcal{N}_\sigma)\right] = \omega(n^{-1})\,.$$

(but $O(n^{-1/2})$, as we know)

Convergence of Smoothed KL Divergence

- [GGNWP20]: If $\sigma > K$ then

$$\mathbb{E}\left[D_{KL}(\mathbb{P}_n * \mathcal{N}_\sigma \| \mathbb{P} * \mathcal{N}_\sigma)\right] = \mathcal{O}\left(n^{-1}\right)$$

- When $\sigma < K$, there exists a distribution $\mathbb{P}$ such that

$$\mathbb{E}\left[D_{KL}(\mathbb{P}_n * \mathcal{N}_\sigma \| \mathbb{P} * \mathcal{N}_\sigma)\right] = \omega(n^{-1})\,.$$

  (but $O(n^{-1/2})$, as we know)
- **Question:** What happens to KL rate when $\sigma < K$? From W_2^2 we might guess the exponent in n drops.

Convergence of Smoothed KL Divergence when $\sigma < K$

Theorem (Jia-Block-Polyanskiy-Rakhlin'21)

Suppose $\mathbb{P}$ is a d-dimensional K-subgaussian distribution, then:

$$\mathbb{E}\left[D_{KL}\left(\mathbb{P}_n * \mathcal{N}(0,\sigma^2 I_d) \middle\| \mathbb{P} * \mathcal{N}(0,\sigma^2 I_d)\right)\right] = \mathcal{O}\left(\frac{(\log n)^{d+1}}{n}\right).$$

Convergence of Smoothed KL Divergence when $\sigma < K$

Theorem (Jia-Block-Polyanskiy-Rakhlin'21)

Suppose $\mathbb{P}$ is a d-dimensional K-subgaussian distribution, then:

$$\mathbb{E}\left[D_{KL}\left(\mathbb{P}_n * \mathcal{N}(0,\sigma^2 I_d) \middle\| \mathbb{P} * \mathcal{N}(0,\sigma^2 I_d)\right)\right] = \mathcal{O}\left(\frac{(\log n)^{d+1}}{n}\right).$$

- Recall that for $K < \sigma$ we know $D_{KL} \leq O(\frac{1}{n})$.
- Thus, only a polylog(n) slowdown!

Implication: LSI non-existence

- **T2 Transportation Inequality:** If $\mathbb{P} * \mathcal{N}_\sigma$ satisfies log-Sobolev inequality with constant $C_{P,\sigma}$, then for any distribution Q

$$W_2^2(Q, \mathbb{P} * \mathcal{N}_\sigma) \leq C_{P,\sigma} D_{KL}(Q \| \mathbb{P} * \mathcal{N}_\sigma).$$

- [WW16] When $K < \sigma$, $\mathbb{P} * \mathcal{N}_\sigma$ satisfies log-Sobolev inequality. Extends the case of compact-support in [Zim13].
- [WW16] also proposed open problem: when $K \geq \sigma$, will $\mathbb{P} * \mathcal{N}_\sigma$ also satisfies log-Sobolev inequality?

Implication: LSI non-existence

- **T2 Transportation Inequality:** If $\mathbb{P} * \mathcal{N}_\sigma$ satisfies log-Sobolev inequality with constant $C_{P,\sigma}$, then for any distribution Q

$$W_2^2(Q, \mathbb{P} * \mathcal{N}_\sigma) \le C_{P,\sigma} D_{KL}(Q \| \mathbb{P} * \mathcal{N}_\sigma).$$

- [WW16] When $K < \sigma$, $\mathbb{P} * \mathcal{N}_\sigma$ satisfies log-Sobolev inequality. Extends the case of compact-support in [Zim13].
- [WW16] also proposed open problem: when $K \ge \sigma$, will $\mathbb{P} * \mathcal{N}_\sigma$ also satisfies log-Sobolev inequality?
- Comparing results for KL divergence and (lower bd) for W_2^2:
  $\exists$ K-subgaussian P such that T2 transportation inequality does not hold for $P * \mathcal{N}_\sigma$, $\sigma < K$.
- ... $\Rightarrow$ when $K > \sigma$ no LSI is possible.

Summary of new results (2021)

- $I_\chi^2(S; Y) < \infty$ vs $= \infty$ dichotomy: $K < \sigma$ vs $K > \sigma$.
- For 1D cases: prove sharper lower and upper bounds on the convergence rate under W_2^2 distance.
- Convergence in KL: $O(\frac{1}{n})$ vs $O(\frac{\mathrm{polylog}(n)}{n})$ for $K < \sigma$ vs $K > \sigma$.
- Corollary: no LSI for $\mathbb{P} * \mathcal{N}_\sigma$ when $K > \sigma$ (and $\mathbb{P}$ is a K-subgaussian).

Summary of new results (2021)

- $I^2_\chi(S;Y) < \infty$ vs $= \infty$ dichotomy: $K < \sigma$ vs $K > \sigma$.
- For 1D cases: prove sharper lower and upper bounds on the convergence rate under W_2^2 distance.
- Convergence in KL: $O(\frac{1}{n})$ vs $O(\frac{\mathrm{polylog}(n)}{n})$ for $K < \sigma$ vs $K > \sigma$.
- Corollary: no LSI for $\mathbb{P} * \mathcal{N}_\sigma$ when $K > \sigma$ (and $\mathbb{P}$ is a K-subgaussian).

Thanks!

References

Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy.
Convergence of smoothed empirical measures with applications to entropy estimation.
IEEE Transactions on Information Theory, 66(7):4368–4391, 2020.

Feng-Yu Wang and Jian Wang.
Functional inequalities for convolution probability measures.
In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 898–914. Institut Henri Poincaré, 2016.

David Zimmermann.
Logarithmic sobolev inequalities for mollified compactly supported measures.
Journal of Functional Analysis, 265(6):1064–1083, 2013.

$\boxed{\textbf{Proofs}}$

W2 in 1D: Lower Bound Part

Theorem

For any $K > \sigma > 0$ and $\epsilon > 0$, there exists some K-subgaussian distribution $\mathbb{P}$ such that

$$\liminf_{n\to\infty} \frac{\mathbb{E}\left[W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma)\right]}{n^{(\sigma^2+K^2)^2/(2(\sigma^4+K^4))+\epsilon}} > 0.$$

W2 in 1D: Lower Bound Part

- When $\mathbb{P}, \mathbb{P}_n$ are both 1D distributions, we can write

$$W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma) = \int_{-\infty}^{\infty} \rho_\sigma(x) \left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right|^2 dx,$$

where ρ_σ is PDF of $\mathbb{P} * \mathcal{N}_\sigma$, and $F_\sigma, \tilde{F}_{n,\sigma}$ are CDFs of $\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P}_n * \mathcal{N}_\sigma$.

W2 in 1D: Lower Bound Part

- When $\mathbb{P}, \mathbb{P}_n$ are both 1D distributions, we can write

$$W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma) = \int_{-\infty}^{\infty} \rho_\sigma(x) \left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right|^2 dx,$$

  where ρ_σ is PDF of $\mathbb{P} * \mathcal{N}_\sigma$, and $F_\sigma, \tilde{F}_{n,\sigma}$ are CDFs of $\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P}_n * \mathcal{N}_\sigma$.
- If $\tilde{F}_{n,\sigma}(z) \geq F_\sigma(z+2)$, then $\forall x \in [z+1, z+2]$ we have $F_\sigma(x) \leq F_\sigma(z+2) \leq \tilde{F}_{n,\sigma}(z) \leq \tilde{F}_{n,\sigma}(x-1)$. Hence

$$\left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right| \geq 1.$$

W2 in 1D: Lower Bound Part

- When $\mathbb{P}, \mathbb{P}_n$ are both 1D distributions, we can write

$$W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma) = \int_{-\infty}^{\infty} \rho_\sigma(x) \left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right|^2 dx,$$

  where ρ_σ is PDF of $\mathbb{P} * \mathcal{N}_\sigma$, and $F_\sigma, \tilde{F}_{n,\sigma}$ are CDFs of $\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P}_n * \mathcal{N}_\sigma$.
- If $\tilde{F}_{n,\sigma}(z) \geq F_\sigma(z+2)$, then $\forall x \in [z+1, z+2]$ we have $F_\sigma(x) \leq F_\sigma(z+2) \leq \tilde{F}_{n,\sigma}(z) \leq \tilde{F}_{n,\sigma}(x-1)$. Hence

$$\left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right| \geq 1.$$

- $W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma) \geq \mathbf{P}\left(X \in [z+1, z+2]\right), \quad X \sim \mathbb{P} * \mathcal{N}_\sigma.$

W2 in 1D: Lower Bound Part

- Choose
$$\mathbb{P} = \sum_{k=1}^{\infty} c_0 \exp\left(-\frac{r_k^2}{2K^2}\right) \delta_{r_k},$$
with $r_k = c^{k-1}$ for $k \geq 1$.

W2 in 1D: Lower Bound Part

- Choose
$$\mathbb{P} = \sum_{k=1}^{\infty} c_0 \exp\left(-\frac{r_k^2}{2K^2}\right) \delta_{r_k},$$
with $r_k = c^{k-1}$ for $k \geq 1$.
- For $\kappa = \frac{\sigma^2}{K^2}$ and $t = 1/2(c+1)(\kappa+1)$ and $X \sim \mathbb{P} * \mathcal{N}_\sigma$,
$$\mathbf{P}(X \in [tr_k, tr_k + 2]) \asymp \exp\left(-\left(t^2 - \kappa c - c\right) \cdot \frac{r_k^2}{2\sigma^2}\right),$$
i.e. δ_{r_k} in $\mathbb{P}$ determines the probability of $\mathbb{P} * \mathcal{N}_\sigma$ within the interval $[tr_k, tr_k + 2]$.

W2 in 1D: Lower Bound Part

- Berry-Esseen Theorem indicates that with certain probability uniformly for all k, we have

$$\tilde{F}_{n,\sigma}(tr_k) - F_\sigma(tr_k) \succeq \sqrt{\frac{p_{k+1}}{n}}.$$

W2 in 1D: Lower Bound Part

- Berry-Esseen Theorem indicates that with certain probability uniformly for all k, we have

$$\tilde{F}_{n,\sigma}(tr_k) - F_\sigma(tr_k) \succeq \sqrt{\frac{p_{k+1}}{n}}.$$

- Chosen n and k, we have $\tilde{F}_{n,\sigma}(tr_k) - F_\sigma(tr_k) \geq \mathbf{P}(X \in [tr_k, tr_k + 2])$ and hence

$$\tilde{F}_{n,\sigma}(tr_k) \geq F_\sigma(tr_k + 2).$$

W2 in 1D: Upper Bound Part

Theorem

Suppose $\mathbb{P}$ *is a 1D* K*-subgaussian random variable, i.e. for some* $C > 0$,

$$\mathbf{P}(|X| \geq x) \leq C \exp\left(-\frac{x^2}{2K^2}\right), \quad x \sim \mathbb{P},$$

then for any $\sigma < K, \epsilon > 0$ *we have*

$$\mathbb{E}\left[W_2^2(\mathbb{P} * \mathcal{N}_\sigma, \mathbb{P}_n * \mathcal{N}_\sigma)\right] = \tilde{\mathcal{O}}\left(n^{-\frac{K^2}{2K^2-\sigma^2}+\epsilon}\right).$$

W2 in 1D: Upper Bound Part

- Recall the formula

$$W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma) = \int_{-\infty}^{\infty} \rho_\sigma(x) \left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right|^2 dx.$$

W2 in 1D: Upper Bound Part

- Recall the formula
$$W_2^2(\mathbb{P}_n * \mathcal{N}_\sigma, \mathbb{P} * \mathcal{N}_\sigma) = \int_{-\infty}^{\infty} \rho_\sigma(x) \left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right|^2 dx.$$

- For those x with large $\rho_\sigma(x)$, one can show that $\left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right|^2$ is small and will decay with $1/\rho_\sigma(x)$.
- For those x with small $\rho_\sigma(x)$, one can show that $\left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right|^2$ is bounded with high probability.

W2 in 1D: Upper Bound Part

We divide $x \in \mathbb{R}$ into the following two cases:

1. $\rho_\sigma(x) = \mathcal{O}\left(n^{-\frac{K^2}{2K^2-\sigma^2}-\epsilon}\right)$, indicating the density is small;
2. $\rho_\sigma(x) = \Omega\left(n^{-\frac{K^2}{2K^2-\sigma^2}-\epsilon}\right)$, indicating the density is large.

W2 in 1D: (When $\rho_\sigma(x)$ is large)

W2 in 1D: (When $\rho_\sigma(x)$ is large)

- $\rho_\sigma(t)$ does not deviate too much from $\rho_\sigma(x)$ for those t in a small neighborhood of x.

W2 in 1D: (When $\rho_\sigma(x)$ is large)

- $\rho_\sigma(t)$ does not deviate too much from $\rho_\sigma(x)$ for those t in a small neighborhood of x.

Lemma

Suppose ρ_σ to be the density function of $P * \mathcal{N}(0, \sigma^2)$. If for some x and $a \geq 0$ we have $\rho_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{a^2}{2\sigma^2}\right)$, then for any δ we have

$$\rho_\sigma(x+\delta) \geq \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a+|\delta|+4\sigma)^2}{2\sigma^2}\right)$$

$$\rho_\sigma(x+\delta) \leq \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\max\{0, a-|\delta|-4\sigma\}^2}{2\sigma^2}\right).$$

W2 in 1D: (When $\rho_\sigma(x)$ is large)

- $\rho_\sigma(t)$ does not deviate too much from $\rho_\sigma(x)$ for those t in a small neighborhood of x.

Lemma

Suppose ρ_σ to be the density function of $P * \mathcal{N}(0, \sigma^2)$. If for some x and $a \geq 0$ we have $\rho_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{a^2}{2\sigma^2}\right)$, then for any δ we have

$$\rho_\sigma(x+\delta) \geq \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a+|\delta|+4\sigma)^2}{2\sigma^2}\right)$$
$$\rho_\sigma(x+\delta) \leq \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\max\{0, a-|\delta|-4\sigma\}^2}{2\sigma^2}\right).$$

- Therefore, is $\rho_\sigma(x)$ is large, then $\mathbf{P}(X \in [x-\delta, x+\delta])$ can be showed to be large as well.

W2 in 1D: (When $\rho_\sigma(x)$ is large)

- The CDF between $P * \mathcal{N}(0, \sigma^2)$ and $P_n * \mathcal{N}(0, \sigma^2)$ can be upper bounded uniformly.

W2 in 1D: (When $\rho_\sigma(x)$ is large)

- The CDF between $P * \mathcal{N}(0, \sigma^2)$ and $P_n * \mathcal{N}(0, \sigma^2)$ can be upper bounded uniformly.

Lemma

Suppose $F_\sigma, \tilde{F}_{\sigma,n}$ are CDF of $P * \mathcal{N}(0, \sigma^2)$ and $P_n * \mathcal{N}(0, \sigma^2)$. Define

$$G(t) = \frac{1}{n} \vee \left(\frac{1}{2} - \left| t - \frac{1}{2} \right| \right), \quad t \in [0, 1].$$

Then with probability at least $1 - \delta$,

$$\sup_{x \in \mathbb{R}} \frac{|F_\sigma(x) - \tilde{F}_{\sigma,n}(x)|}{\sqrt{G(F(x))}} \leq \frac{16}{\sqrt{n}} \log\left(\frac{2n}{\delta} \right).$$

W2 in 1D: (When $\rho_\sigma(x)$ is large)

- One can show that when $\rho_\sigma(x)$ is large, $\left|\tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x\right|$ is small.

Lemma

Consider two 1D-distributions $\mathbb{P}, \mathbb{Q}$. We denote the PDF of $\mathbb{P}$ as $\rho_p(\cdot)$, and the CDFs of $\mathbb{P}, \mathbb{Q}$ as F_p, F_q respectively. If for some $\sigma > 0$ we have

$$\alpha(t, \sigma) \triangleq \frac{\sup_{t \in [x-\sigma, x+\sigma]} |F_p(t) - F_q(t)|}{\inf_{t \in [x-\sigma, x+\sigma]} \rho_p(t)} \leq \sigma,$$

then

$$\left|F_q^{-1}(F_p(t)) - t\right| \leq \alpha(t, \sigma) .$$

W2 in 1D: (When $\rho_\sigma(x)$ is small)

W2 in 1D: (When $\rho_\sigma(x)$ is small)

- Given $R > 0$, then for $\forall |x| \leq R$, with high probability we have $\left|\tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x\right| = \tilde{\mathcal{O}}(R)$.

W2 in 1D: (When $\rho_\sigma(x)$ is small)

- Given $R > 0$, then for $\forall |x| \leq R$, with high probability we have $\left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right| = \tilde{\mathcal{O}}(R)$.
- $\mathbf{P}(|X| \geq R) \leq C \exp\left(-\frac{R^2}{2K^2}\right)$;

W2 in 1D: (When $\rho_\sigma(x)$ is small)

- Given $R > 0$, then for $\forall |x| \leq R$, with high probability we have $\left|\tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x\right| = \tilde{\mathcal{O}}(R)$.
- $\mathbf{P}(|X| \geq R) \leq C \exp\left(-\frac{R^2}{2K^2}\right)$;
- For those $|x| \leq R$ and $\rho_\sigma(x) \leq \epsilon$, the measure of the set of such x is at most $2R\epsilon$.

W2 in 1D: (When $\rho_\sigma(x)$ is small)

- Given $R > 0$, then for $\forall |x| \leq R$, with high probability we have $\left| \tilde{F}_{n,\sigma}^{-1}(F_\sigma(x)) - x \right| = \tilde{\mathcal{O}}(R)$.
- $\mathbf{P}(|X| \geq R) \leq C \exp\left(-\frac{R^2}{2K^2}\right)$;
- For those $|x| \leq R$ and $\rho_\sigma(x) \leq \epsilon$, the measure of the set of such x is at most $2R\epsilon$.
- If choosing R, ϵ properly, one can also upper bounded the integral over those x with small $\rho_\sigma(x)$ with $\mathcal{O}\left(n^{-\frac{K^2}{2K^2-\sigma^2}-\epsilon}\right)$.

KL-convergence: Proof Idea

- The expected KL-divergence can be upper bounded using **Rényi-mutual information**:

KL-convergence: Proof Idea

- The expected KL-divergence can be upper bounded using **Rényi-mutual information**:

Lemma

We suppose $(X, Y) \sim P_{X,Y}$*, and its marginal distribution to be* P_X, P_Y*, respectively. We let* $\hat{P}_n$ *to be an empirical version of* P_X *generated with* n *samples. Then for every* $1 < \lambda \leq 2$*, we have*

$$\mathbb{E}[D_{KL}(P_{Y|X} \circ \hat{P}_n \| P_Y)] \leq \frac{1}{\lambda - 1} \log(1 + \exp\{(\lambda - 1)(I_\lambda(X; Y) - \log n)\}).$$

KL-convergence: Proof Idea

- The expected KL-divergence can be upper bounded using **Rényi-mutual information**:

Lemma

We suppose $(X, Y) \sim P_{X,Y}$, and its marginal distribution to be P_X, P_Y, respectively. We let $\hat{P}_n$ to be an empirical version of P_X generated with n samples. Then for every $1 < \lambda \leq 2$, we have

$$\mathbb{E}[D_{KL}(P_{Y|X} \circ \hat{P}_n \| P_Y)] \leq \frac{1}{\lambda - 1} \log(1 + \exp\{(\lambda - 1)(I_\lambda(X; Y) - \log n)\}).$$

This lemma indicates a convergence rate of $\mathcal{O}\left(n^{-(\lambda-1)}\right)$ provided $I_\lambda(X; Y) < \infty$, where $X \sim \mathbb{P}, Z \sim \mathcal{N}_\sigma$ are independent and $Y = X + Z$.

KL-convergence: Proof Idea

- $I_\lambda(X; Y)$ can be proved to be finite for any $\lambda < 2$.

KL-convergence: Proof Idea

- $I_\lambda(X;Y)$ can be proved to be finite for any $\lambda < 2$.

Lemma

Suppose $\mathbb{P}$ is a d-dimensional K-subgaussian distribution and random variables $X \sim \mathbb{P}, Z \sim \mathcal{N}(0, \sigma^2 I_d)$ are independent to each other. We let $Y = X + Z$. Then for any $\sigma > 0$ and $1 < \lambda < 2$, there exists a positive constant C only depending on $\mathbb{P}$ and K, σ such that

$$I_\lambda(X;Y) \leq \frac{1}{\lambda - 1} \log \left(\frac{C}{(2-\lambda)^{d+1}} \right).$$