Nonparametric Density Estimation & Convergence of GANs under Besov IPM Losses

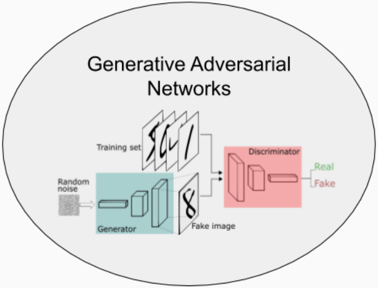
Ananya Uppal, Shashank Singh*, & Barnabás Póczos
Carnegie Mellon University

NeurIPS 2019 Oral Presentation
December 12, Vancouver

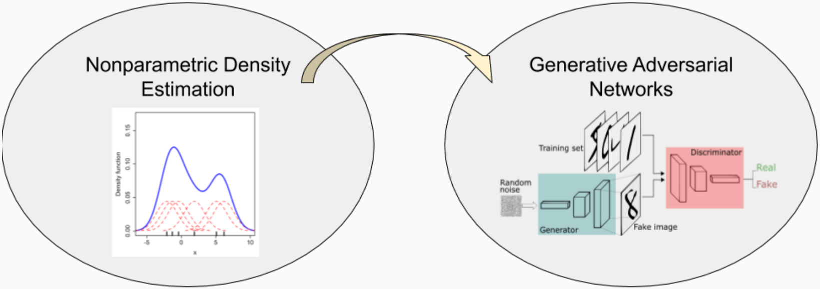
*Now at Google

Big Picture

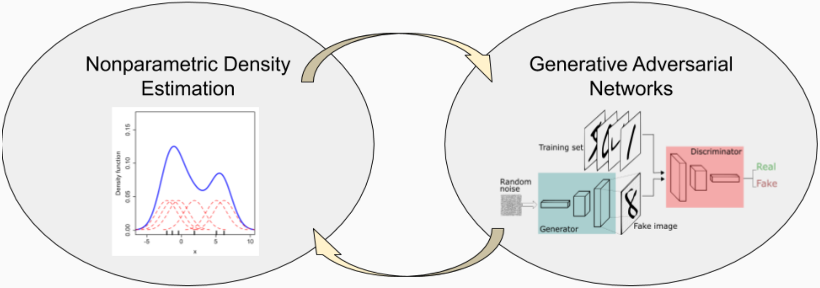
Theoretical Guarantees for GANs



Theoretical Guarantees for GANs
through the lens of nonparametric density estimation



Theoretical Guarantees for GANs
through the lens of nonparametric density estimation



Contributions

1. New generalization of density estimation
    - **Besov IPMs** – new losses motivated partly by GAN discriminators

Contributions

1. New generalization of density estimation
   - Besov IPMs – new losses motivated partly by GAN discriminators

2. New minimax rates under these losses
   - Reduced curse of dimensionality

Contributions

1. New generalization of density estimation
    - Besov IPMs – new losses motivated partly by GAN discriminators
2. New minimax rates under these losses
    - Reduced curse of dimensionality
3. Many classical estimators are provably sub-optimal
    - e.g., kernel density estimator
    - gap increases with dimension

Contributions

1. New generalization of density estimation
   - **Besov IPMs** – new losses motivated partly by GAN discriminators
2. New minimax rates under these losses
   - Reduced curse of dimensionality
3. Many classical estimators are provably sub-optimal
   - e.g., kernel density estimator
   - gap increases with dimension
4. Certain GANs are minimax optimal

Contributions

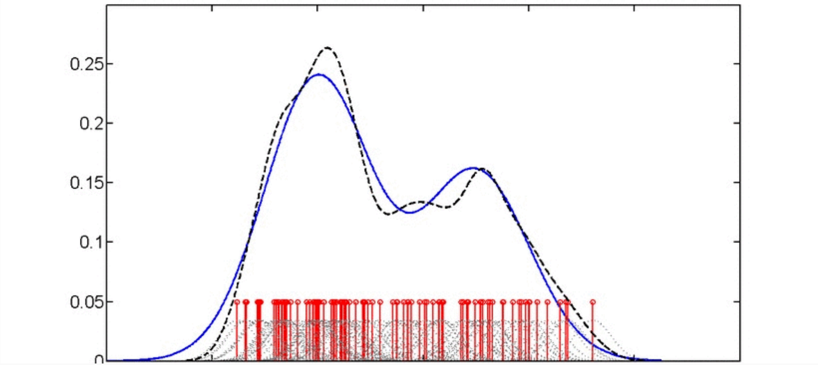
1. New generalization of density estimation
    - Besov IPMs – new losses motivated partly by GAN discriminators
2. New minimax rates under these losses
    - Reduced curse of dimensionality
3. Many classical estimators are provably sub-optimal
    - e.g., kernel density estimator
    - gap increases with dimension
4. Certain GANs are minimax optimal
5. Besov IPMs also have important theoretical roles in math-stats.
    - Unify several previous works in nonparametric density est.

Contributions

1. New generalization of density estimation
    - **Besov IPMs** – new losses motivated partly by GAN discriminators
2. New minimax rates under these losses
    - Reduced curse of dimensionality
3. Many classical estimators are provably sub-optimal
    - e.g., kernel density estimator
    - gap increases with dimension
4. Certain GANs are minimax optimal
5. Besov IPMs also have important theoretical roles in math-stats.
    - Unify several previous works in nonparametric density est.

Density Estimation

- Observe n independent samples $X_1, \ldots, X_n \sim P$.
- Assume $P \in \mathcal{P}$.
- Want to estimate P.



GANs

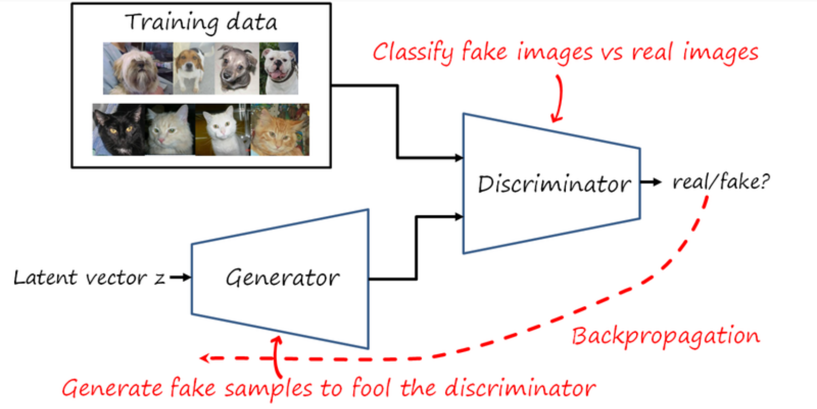
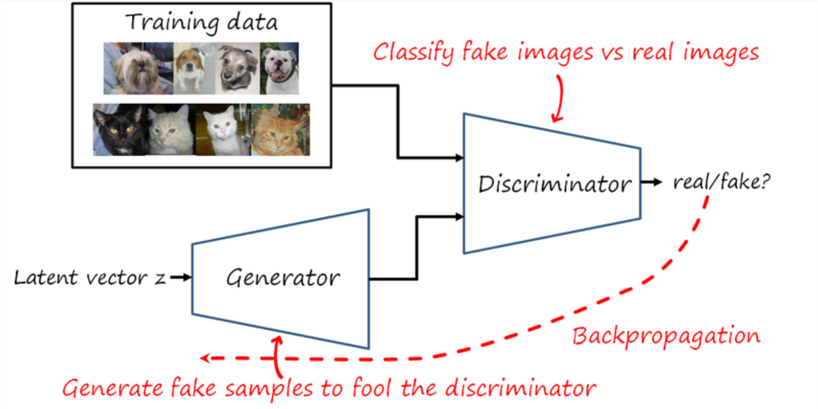


Figure from `http://www.lherranz.org/2018/08/07/imagetranslation/`.

GANs



$$\widehat{P}_{\text{GAN}} := \underbrace{\operatorname*{argmin}_{Q \in \mathcal{P}}}_{\text{Generator}} \underbrace{\sup_{f \in \mathcal{F}}}_{\text{Discriminator}} \mathop{\mathbb{E}}_{X \sim Q}[f(X)] - \mathop{\mathbb{E}}_{X \sim P_n}[f(X)],$$

Figure from `http://www.lherranz.org/2018/08/07/imagetranslation/`.

GANs as Regularized ERM Density Estimators

$$\widehat{P}_{\text{GAN}} := \operatorname*{argmin}_{Q \in \mathcal{P}} \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{X \sim Q} [f(X)] - \mathop{\mathbb{E}}_{X \sim P_n} [f(X)]$$

GANs as Regularized ERM Density Estimators

$$\widehat{P}_{\mathrm{GAN}} := \operatorname*{argmin}_{Q \in \mathcal{P}} \underbrace{\sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{X \sim Q} [f(X)] - \mathop{\mathbb{E}}_{X \sim P_n} [f(X)]}_{d_{\mathcal{F}}(Q, P_n)}$$

$$= \operatorname*{argmin}_{Q \in \mathcal{P}} d_{\mathcal{F}}(Q, P_n)$$

Empirical Risk Minimization (ERM)

- Hypothesis class $\mathcal{P}$
- Loss $d_{\mathcal{F}}$
- regularize data before feeding to GAN
    - instance noise (Sønderby et al. (2017), ICLR)

Integral Probability Metrics (IPMs)

- $\mathcal{F}$ – class of *discriminator functions*

The metric $d_{\mathcal{F}} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is defined by

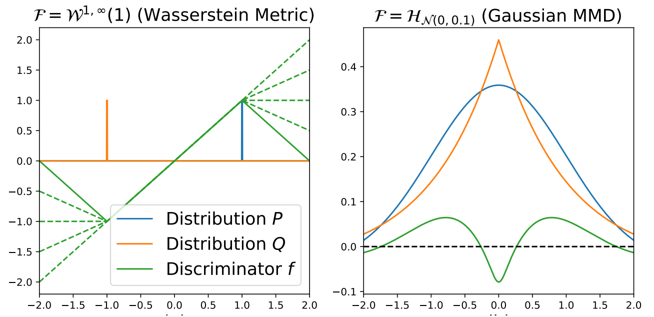
$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{X \sim P} [f(X)] - \mathop{\mathbb{E}}_{X \sim Q} [f(X)] \right|, \quad \text{for all } P, Q \in \mathcal{P}.$$

Integral Probability Metrics (IPMs)

- $\mathcal{F}$ – class of *discriminator functions*

The metric $d_{\mathcal{F}} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is defined by

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{X \sim P} [f(X)] - \mathop{\mathbb{E}}_{X \sim Q} [f(X)] \right|, \quad \text{for all } P, Q \in \mathcal{P}.$$



Examples of IPMs

- 1-Wasserstein[1]
- Max. Mean Discrepancy (MMD)[2]
- L^r distances [3]
- Kolmogorov-Smirnov
- Hilbert-Sobolev distances
- Besov distances
- Neural net distance (GANs)

[1] a.k.a. optimal transport or earthmover's distance
[2] Including energy distances
[3] Including total variation distance

Besov Spaces

- 2-parameter family of function spaces $\mathcal{B}_p^s$
    - $s \in (0, \infty)$, $p \in [1, \infty]$

Besov Spaces

- 2-parameter family of function spaces $\mathcal{B}_p^s$
  - $s \in (0, \infty)$, $p \in [1, \infty]$

For integer s

$$\mathcal{B}_p^s \approx \{f \in \mathcal{L}^p : \|f^{(s)}\|_p \leq C\}$$

where $f^{(s)} = s^{th}$ derivative of f.

Besov Spaces

- 2-parameter family of function spaces $\mathcal{B}_p^s$
  - $s \in (0, \infty)$, $p \in [1, \infty]$

For integer s

$$\mathcal{B}_p^s \approx \{f \in \mathcal{L}^p : \|f^{(s)}\|_p \leq C\}$$

where $f^{(s)} = s^{th}$ derivative of f.

Examples:

Ex. 1: Lipschitz/Hölder spaces: $\mathcal{B}_\infty^s \approx \mathcal{C}^s$

Ex. 2: Sobolev spaces: $\mathcal{B}_2^s \approx \mathcal{H}^s$

Neural Network GANs

Ideal ERM:

$$\operatorname*{argmin}_{Q \in \mathcal{P}} d_{\mathcal{F}}(Q, P_n).$$

Neural Network GANs

Ideal ERM:

$$\operatorname*{argmin}_{Q \in \mathcal{P}} d_{\mathcal{F}}(Q, P_n).$$

$\mathcal{P}$ and $\mathcal{F}$ are ∞-dimensional... How to approximate?

Neural Network GANs

Ideal ERM:

$$\operatorname*{argmin}_{Q \in \mathcal{P}} d_{\mathcal{F}}(Q, P_n).$$

$\mathcal{P}$ and $\mathcal{F}$ are ∞-dimensional... How to approximate?

ReLU Neural Networks (Suzuki (2019), ICLR):

$$\mathcal{B}_p^s \approx \Phi(L, W, S, B)$$

$\Phi(L, W, S, B) =$ class of fully-connected ReLU networks of size:

- $L = \#$ of layers (depth)
- $W = \#$ neurons/layer (width)
- $S = \#$ nonzero weights/layer (sparsity)
- $B =$ largest weight value

Neural Network GANs

Ideal ERM:

$$\operatorname*{argmin}_{Q \in \mathcal{P}} d_{\mathcal{F}}(Q, P_n).$$

$\mathcal{P}$ and $\mathcal{F}$ are ∞-dimensional... How to approximate?

ReLU Neural Networks (Suzuki (2019), ICLR):

$$\mathcal{B}_p^s \approx \Phi(L, W, S, B)$$

$\Phi(L, W, S, B) =$ class of fully-connected ReLU networks of size

$L \in O(\log n)$, $W, S, B \in O(\text{poly}(n))$.

$$\widehat{P}_{\text{GAN}} = \operatorname*{argmin}_{Q \in \Phi(L_g, W_g, S_g, B_g)} d_{\Phi(L_d, W_d, S_d, B_d)}(Q, P_n).$$

GANs are optimal***

$$\widehat{P}_{\text{GAN}} = \operatorname*{argmin}_{Q \in \Phi(L_g, W_g, S_g, B_g)} d_{\Phi(L_d, W_d, S_d, B_d)}(Q, P_n).$$

is optimal for estimating Besov distributions under Besov IPMs.

GANs are optimal***

$$\widehat{P}_{\text{GAN}} = \underset{Q \in \Phi(L_g, W_g, S_g, B_g)}{\text{argmin}} d_{\Phi(L_d, W_d, S_d, B_d)}(Q, P_n).$$

is optimal for estimating Besov distributions under Besov IPMs.

***Caveats:

1. Well-optimized (maybe computationally challenging)

GANs are optimal***

$$\widehat{P}_{\text{GAN}} = \underset{Q \in \Phi(L_g, W_g, S_g, B_g)}{\text{argmin}} \; d_{\Phi(L_d, W_d, S_d, B_d)}(Q, P_n).$$

is optimal for estimating Besov distributions under Besov IPMs.

***Caveats:

1. Well-optimized (maybe computationally challenging)
2. Well-tuned (neural network sizes)

GANs are optimal***

$$\widehat{P}_{\text{GAN}} = \operatorname*{argmin}_{Q \in \Phi(L_g, W_g, S_g, B_g)} d_{\Phi(L_d, W_d, S_d, B_d)}(Q, P_n).$$

is optimal for estimating Besov distributions under Besov IPMs.

***Caveats:

1. Well-optimized (maybe computationally challenging)
2. Well-tuned (neural network sizes)
3. Assumes fully-connected ReLU networks

Summary

1. New generalization of density estimation
    - **Besov IPMs** – new losses motivated partly by GAN discriminators
2. New minimax rates under these losses
    - Reduced curse of dimensionality
3. Many classical estimators are provably sub-optimal
    - e.g., kernel density estimator
4. Certain GANs are minimax optimal
5. Besov IPMs also have important theoretical roles in math. stats.
    - Unify several previous works in nonparametric density est.

Poster #243 tonight

Further Reading

- Liang, Tengyuan. (2019) "On how well generative adversarial networks learn densities: Nonparametric and parametric results." arXiv
- Bauer & Kohler. "On deep learning as a remedy for the curse of dimensionality in nonparametric regression." *Annals of Statistics*.
- Johannes Schmidt-Hieber. "Nonparametric regression using deep neural networks with ReLU activation function." *Annals of Statistics*.

A Minimax Framework for Implicit Generative Modeling

Implicit generative model (sampler):

$$\widehat{X} : \underbrace{\mathcal{X}^n}_{\text{Training Data}} \times \underbrace{\mathcal{Z}}_{\text{Randomness}} \to \underbrace{\mathcal{X}}_{\text{Novel Sample}}$$

Output distribution: conditional distribution $P_{\widehat{X}(X_1,...,X_n,Z)|X_1,...,X_n}$ of novel sample X_{n+1} given training data $X_1, ..., X_n$.

Define the *implicit risk of* $\widehat{X}$ *at* P by

$$R_I(P, \widehat{X}) := \mathop{\mathbb{E}}_{X_1,...,X_n \overset{IID}{\sim} P} \left[\ell(P, P_{\widehat{X}(X_1,...,X_n,Z)|X_1,...,X_n}) \right].$$

Implicit versus Explicit Generative Modeling

Theorem (When do good samplers imply good density estimators?)
Let $\mathcal{F}_G$ be a family of probability distributions on a sample space $\mathcal{X}$. Suppose

1. *Loss $\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ satisfies a weak triangle inequality*
2. *$M_D(\mathcal{F}_G, \ell, m) \to 0$ as $m \to \infty$. (i.e., there exists a uniformly consistent density estimator)*
3. *we can draw arbitrarily many IID samples $Z_1, Z_2, \ldots$ of the latent variable Z*
4. *Output distributions of (nearly) minimax samplers lie in $\mathcal{F}_G$*

Then, $M_D(\mathcal{F}_G, \ell, n) \lesssim M_I(\mathcal{F}_G, \ell, n)$.

Implicit versus Explicit Generative Modeling

Theorem (When do good samplers imply good density estimators?)
Let $\mathcal{F}_G$ be a family of probability distributions on a sample space $\mathcal{X}$. Suppose

1. *Loss $\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ satisfies a weak triangle inequality*
2. *$M_D(\mathcal{F}_G, \ell, m) \to 0$ as $m \to \infty$. (i.e., there exists a uniformly consistent density estimator)*
3. *we can draw arbitrarily many IID samples $Z_1, Z_2, \ldots$ of the latent variable Z*
4. *Output distributions of (nearly) minimax samplers lie in $\mathcal{F}_G$*

Then, $M_D(\mathcal{F}_G, \ell, n) \lesssim M_I(\mathcal{F}_G, \ell, n)$.

Proof: Train a new density estimator $\widehat{P}$ with m IID samples drawn from the sampler $\widehat{X}$. Then, $R(\widehat{P}) \leq R(\widehat{X}) + \varepsilon_m$.