Nonparametric Density Estimation under Adversarial Losses with Statistical Convergence Rates for GANs

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, Barnabás Póczos
sssi@cs.cmu.edu

Introduction
- Nonparametric distribution estimation: Given $n$ IID samples $X_1, \ldots, X_n \sim P$ from an unknown distribution $P$ in some large class $\mathcal{P}$ on a sample space $X$, we want to estimate $P$.
- Nonparametric density estimation is usually studied using $L^2$ loss.
- $L^2$ can be very strong
  - Only allows distributions with densities
  - Severe curse of dimensionality
- GANs implicitly use different losses – adversarial losses
- Many other theoretically motivated losses are also adversarial losses (see below)
- We provide unified analysis of optimal rates for distribution estimation with these losses.

Adversarial Losses (Integral Probability Metrics, IPMs)
Fix a sample space $X$. Let $\mathcal{P}$ be a class of probability distributions on $X$, and let $\mathcal{F}$ be a class of (bounded) functions on $X$. Then, the (pseudo)metric $p_{\mathcal{F}} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ on $\mathcal{P}$ is defined by

$$p_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} E_P[f(X)] - E_Q[f(X)].$$

Here, any $f^* \in \arg\max_{f \in \mathcal{F}} E_P[f(X)] - E_Q[f(X)]$

is called a discriminator function.

Upper Bound for Orthogonal Series Estimate
- Consider an orthogonal series estimate $\hat{f}$ (basically, estimate a finite number $\zeta$ of the basis coefficients, with tuning parameter $\zeta \to \infty$ as $n \to \infty$).
- We prove a very general upper bound on $P$ and $\mathcal{P}$ that can be expressed in terms of orthonormal basis approximations (e.g., Fourier, wavelet, etc.)
- Includes all distances in previous table
- Allows distributions without densities!

Corollary (Sobolev IPM). Fix $\zeta \in \mathbb{N}$, define the $\zeta$-Sobolev ball

$$W^{\zeta, 1}(L) = \{ f \in \mathcal{C}(X) : \| f^{(\zeta)} \|_{L^2(X)}^2 \leq L \},$$

where $f^{(\zeta)}$ denotes the $\zeta$-th derivative of $f$. Suppose $P = W^{\zeta, 1}(L)$ and $P = W^{\zeta, 1}(L_d)$. Then, there exists a constant $C > 0$ (depending only on $\zeta$) such that

$$\sup_{P \in \mathcal{P}} p_{\mathcal{F}}(P, P^*) \leq C \frac{L \sqrt{ \log (1 + n^\zeta) + 1}}{n^{\zeta/2}}.$$

(1)

Summary:
1. Upper bounds for wide range of adversarial losses and probability distributions
2. All rates are optimal in $n$ – paper includes minimax lower bounds

Error Bounds for (Perfectly Optimized) GANs
Corollary. Fix a desired precision $\epsilon > 0$. Then, there exists a GAN architecture, in which both the generator $G$ and discriminator $D$ are fully-connected neural networks with ReLU activations, such that:

1. $G$ has at most $O(\log(\zeta) / \epsilon)$ layers and $O(\zeta^2 \log(\zeta))$ total parameters
2. $D$ has at most $O(\log(\zeta) / \epsilon)$ layers and $O(\zeta^2 \log(\zeta))$ total parameters
3. there exists a constant $C$ depending only on $\zeta$ such that:

$$\sup_{P \in \mathcal{P}} \| \mathbb{E}_P f - \mathbb{E}_{P^*} f \|_{L^1} \leq C \| f \|_{L^1} \frac{\sqrt{\log(1 + n^\zeta)} + 1}{n^{\zeta/2}} + \epsilon.$$

when $P \sim \mathcal{P}$, $f \in \mathcal{F}$ where $\mathcal{F}$ is a ball in the RKHS with a Gaussian kernel.

A Statistical Framework for Implicit Generative Modeling
But wait – GANs don’t estimate the distribution – they just generate new samples!
- This task ("sampling") is called implicit generative modeling, as opposed to explicit generative modeling ("density estimation") [1, 4]
- No universally agreed-upon measure of performance for GANs
  - Formally, an implicit generative model is a function $X : \mathbb{X} \to X \to \mathbb{X}$, which maps training data and randomness to a novel sample
- We propose the Implicit Risk:

$$R_I(P, X) = \mathbb{E}_{X \sim P} \left[ \left( P, P_{X(X)}, X \right) \right].$$

(as opposed to the explicit risk

$$R_E(P, F) = \mathbb{E}_{X \sim P} \left[ \left( P, P_{X(X)} \right) \right].$$

Theorem. Let $\mathcal{P}$ be a family of probability distributions on a sample space $X$, and let $\xi : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ be a loss function on $\mathcal{P}$. Suppose

1. $\xi$ satisfies a weak triangle inequality: $\xi(P_1, P_2) \leq C(\xi(P_1, P_3) + \xi(P_2, P_3))$
2. there exists a uniformly consistent estimator $\hat{\xi}(\cdot, \cdot)$ such that $\sup_{P \in \mathcal{P}} \mathbb{E}_{P \sim \mathcal{P}} \xi(\hat{\xi}(\cdot, \cdot)) \to 0$ as $n \to \infty$

We can draw arbitrarily many samples $X_1, \ldots, X_n \sim \mathbb{P}_{X} (\cdot, \cdot)$ of the latent variable.

4. there is a sequence of (nearly) minimax samples $X_1, \ldots, X_n \sim \mathbb{P}_{X} (\cdot, \cdot)$ such that, for each $k \in \mathbb{N}$, almost surely over $X_1, \ldots, X_k \sim \mathbb{P}_{X} (\cdot, \cdot)$:

$$\sup_{P \sim \mathbb{P}_{X}} \xi(\hat{\xi}(\cdot, \cdot)) \leq \epsilon.$$

Proof. Construct a density estimator $\hat{P}$ by feeding $n$ artificial samples from $X$ into a consistent density estimator. Then, $\lim_{n \to \infty, \epsilon \to 0} \xi(\hat{\xi}(\cdot, \cdot)) = \xi(\hat{\xi}(\cdot, \cdot)).$

Some simple tricks:
- Same proof works for other notions (e.g., average-case/Bayesian) of optimality

Summary: Statistically, sampling is no easier than density estimation.
- In many cases, the converse is also true: good density estimators lead to good samplers.
- Justifies applying density estimation result to GANs and applying lower bound to GANs.
- Same discussion applies to other implicit models (variational autoencoders (VAEs), classical MCMC, etc.)

References