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

sss1@cs.cmu.edu

Carnegie Mellon University

Introduction

- Nonparametric distribution estimation: Given n IID samples $X_{1:n} = X_1, ..., X_n \stackrel{IID}{\sim} P$ from an unknown distribution P in some large class $\mathcal{P}$ on a sample space $\mathcal{X}$, we want to estimate P.
- Nonparametric density estimation is usually studied using $\mathcal{L}^2$ loss.
- $\mathcal{L}^2$ 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 $\mathcal{X}$. Let $\mathcal{P}$ be a class of probability distributions on $\mathcal{X}$, and let $\mathcal{F}$ be a class of (bounded) functions on $\mathcal{X}$. Then, the (pseudo)metric $\rho_{\mathcal{F}} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ on $\mathcal{P}$ is defined by

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

Here, any

$$f^* \in \operatorname{argmax}_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{X \sim P} [f(X)] - \mathop{\mathbb{E}}_{X \sim Q} [f(X)] \right|$$

is called a **discriminator function**.

Examples of Adversarial Losses

Distance	$\mathcal{F}$
$\mathcal{L}^p$	$\mathcal{F} = \{f : \|f\|_q \leq L\}$, where $q = \frac{p}{p-1}$
Total Variation	$\mathcal{F} = \{1_A : A \subseteq \mathcal{X} \text{ measurable}\}$
Wasserstein ("earth-mover")	$\mathcal{F} = \mathcal{W}^{1,\infty}(1)$ (1-Lipschitz class)
Kolmogorov-Smirnov	$\mathcal{F} = \{1_{(-\infty,x]} : x \in \mathbb{R}\})$
Max. mean discrepancy (MMD)	$\mathcal{F}$ is an RKHS ball
GAN Discriminator	$\mathcal{F}$ parameterized by neural network

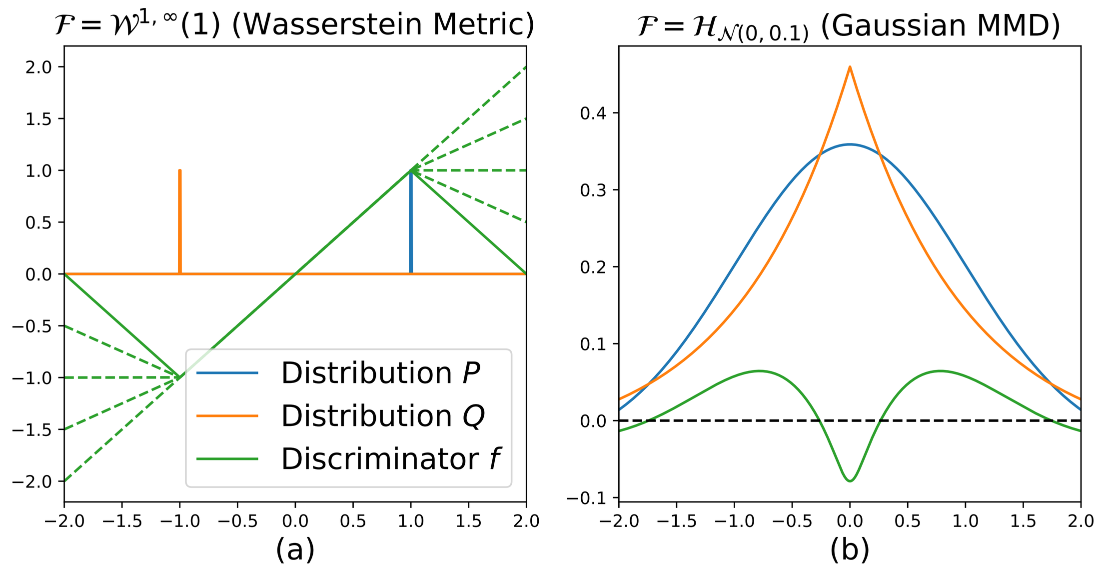

Figure 1: Examples of probability distributions P and Q and corresponding discriminator functions f^*. In (a), P and Q are single Dirac masses at $+1$ and -1, respectively, and $\mathcal{F}$ is the 1-Lipschitz class, so that $d_{\mathcal{F}}$ is the Wasserstein metric. In (b), P and Q are standard Gaussian and standard Laplace distributions, respectively, and $\mathcal{F}$ is a ball in an RKHS with a Gaussian kernel.

Upper Bound for Orthogonal Series Estimate

- Consider an orthogonal series estimate $\hat{P}_\zeta$ (basically, estimate a finite number ζ of the basis coefficients, with tuning parameter $\zeta \to \infty$ as $n \to \infty$).
- We prove a very general upper bound for $\mathcal{F}$ 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!

The general theorem is a bit technical; here are some interesting corollaries:

Corollary (Sobolev IPM). *For $s \in \mathbb{N}$, define the s-Sobolev ball*

$$\mathcal{W}^{s,2}(L) = \left\{ f \in \mathcal{L}^2(\mathcal{X}) : \left\| f^{(s)} \right\|^2_{\mathcal{L}^2(\mathcal{X})} = \int_{\mathcal{X}} \left(f^{(s)}(x) \right)^2 d\mu(x) \leq L^2 \right\},$$

where $f^{(s)}$ denotes the s^{th} derivative of f. Suppose $\mathcal{F} = \mathcal{W}^{s,2}(L_D)$ and $\mathcal{P} = \mathcal{W}^{t,2}(L_G)$. Then, there exists a constant $C > 0$ (depending only on d, s, t) such that

$$\sup_{P \in \mathcal{P}} \mathbb{E} \left[d_{\mathcal{F}} \left(P, \hat{P} \right) \right] \leq C \left(n^{-\frac{s+t}{2t+d}} + n^{-1/2} \right).$$

(note that $n^{-1/2}$ dominates $\Leftrightarrow t \geq 2d$).

- The case $s = 1$ corresponds to the Wasserstein metric
- Improves on previous rate of order $\asymp n^{-\frac{s+t}{2(s+t)+d}}$ [2].
- Bonus Result: Optimal ζ is the same as under $\mathcal{L}^2$ loss, so we can use classic results for cross-validation under $\mathcal{L}^2$ loss [3] to obtain adaptive minimax estimators under adversarial losses.

Corollary (Maximum Mean Discrepancy). *If $\mathcal{F}$ is a ball of radius L in a reproducing kernel Hilbert space with translation invariant kernel $K(x, y) = \kappa(x - y)$ for some $\kappa \in \mathcal{L}^2(\mathcal{X})$, then,*

$$\sup_{P \text{ Borel}} \mathbb{E} \left[d_{\mathcal{F}} \left(P, \hat{P} \right) \right] \leq \frac{L \|\kappa\|_{\mathcal{L}^2(\mathcal{X})}}{\sqrt{n}}.$$

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 F_G and discriminator $\mathcal{F}_D$ are fully-connected neural networks with ReLU activations, such that:*

1. $\mathcal{F}_D$ has at most $O(\log(1/\epsilon))$ layers and $O(\epsilon^{d/s} \log \epsilon)$ total parameters

2. $\mathcal{F}_G$ has at most $O(\log(1/\epsilon))$ layers and $O(\epsilon^{d/t} \log \epsilon)$ total parameters

3. there exists a constant C depending only on d, s, t such that, if

$$\hat{P}_* := \operatorname{argmin}_{\hat{P} \in \mathcal{F}_G} d_{\mathcal{F}_D} \left(P_n, \hat{P} \right)$$

(where $P_n = \frac{1}{n} \sum_{i=1}^n 1_{X_i}$ denotes the empirical distribution of the data), then, for a constant C depending on only d, s, t,

$$\sup_{P \in \mathcal{W}^{t,2}} \mathbb{E} \left[d_{\mathcal{W}^{s,2}} \left(P, \hat{P}_* \right) \right] \leq C \left(\epsilon + n^{-\frac{s+t}{2t+d}} \right).$$

- Proof builds on construction by [5] of fully-connected ReLU network for approximating Sobolev functions.

Summary: Under an appropriate loss, Sobolev GANs are statistically optimal for Sobolev densities, provided:

(a) the networks are allowed to converge to a global optimum, and

(b) the size of the networks is allowed to grow with the sample size.

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 $\hat{X} : \mathcal{X}^n \times \mathcal{Z} \to \mathcal{X}$, which maps training data and randomness to a novel sample
- We propose the **Implicit Risk:**

$$R_I(P, \hat{X}) := \mathop{\mathbb{E}}_{X_{1:n} \stackrel{IID}{\sim} P} \left[\ell \left(P, P_{\hat{X}(X_{1:n}, Z) | X_{1:n}} \right) \right]$$

(as opposed to the **explicit risk**

$$R_E(P, \hat{P}) := \mathop{\mathbb{E}}_{X_{1:n} \stackrel{IID}{\sim} P} \left[\ell \left(P, \hat{P}(X_{1:n}) \right) \right]).$$

Theorem. *Let $\mathcal{P}$ be a family of probability distributions on a sample space $\mathcal{X}$, and let $\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ be a loss function on $\mathcal{P}$. Suppose*

(A1) *ℓ satisfies a weak triangle-inequality: $\ell(P_1, P_3) \leq C (\ell(P_1, P_2) + \ell(P_2, P_3))$*

(A2) *there exists a uniformly consistent estimator $\hat{P}$ (i.e., $\sup_{P \in \mathcal{P}} R_E(\hat{P}) \to 0$ as $n \to \infty$)*

(A3) *we can draw arbitrarily many samples $Z_1, ..., Z_m \stackrel{IID}{\sim} Q_Z$ of the latent variable.*

(A4) *there exists a sequence of (nearly) minimax samplers $\hat{X}_k : \mathcal{X}^n \times \mathcal{Z} \to \mathcal{X}$ such that, for each $k \in \mathbb{N}$, almost surely over $X_{1:n}$, $P_{\hat{X}_k(X_{1:n}, Z) | X_{1:n}} \in \mathcal{P}$.*

Then,

$$\inf_{\hat{P}} \sup_{P \in \mathcal{P}} R_E(P, \hat{P}) \leq \inf_{\hat{P}} \sup_{P \in \mathcal{P}} R_I(P, \hat{X}).$$

Proof. Construct a density estimator $\hat{P}$ by feeding m artificial samples from $\hat{X}$ into a consistent density estimator. Then, $\lim_{m \to \infty} R_E(P, \hat{P}) \leq R_I(P, \hat{X})$. □

- 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

[1] Peter J Diggle and Richard J Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

[2] Tengyuan Liang. How well can generative adversarial networks (GAN) learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

[3] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[4] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[5] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.