Nonparametric Density Estimation under IPM Losses

with Statistical Convergence Rates for Generative Adversarial Networks (GANs)

Ananya Uppal, Shashank Singh, Barnabás Póczos

Carnegie Mellon University

↑Link to paper↑

Introduction

- Nonparametric distribution estimation: Given n IID samples $X_{1:n} = X_1, ..., X_n \overset{IID}{\sim} P$ from an unknown distribution P, we want to estimate P.
  - Important sub-routine of many statistical methods
  - Usually analyzed in terms of $\mathcal{L}^2$ loss
    - Severe curse of dimensionality
- **We provide unified minimax-optimal estimation rates under large family of losses called Integral Probability Metrics (IPMs), for many function classes (Sobolev, Besov, RKHS).**
  - Includes most common metrics on probability distributions
  - Implicitly used in Generative Adversarial Networks (GANs)
    - Allows us to derive statistical guarantees for GANs
  - Reduced curse of dimensionality

Integral Probability Metrics (IPMs)

Definition 1 (IPM). *Let $\mathcal{P}$ be a class of probability distributions on a sample space $\mathcal{X}$, and $\mathcal{F}$ a class of (bounded) functions on $\mathcal{X}$. Then, the metric $\rho_{\mathcal{F}} : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ on $\mathcal{P}$ is defined by*

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

Definition 2 (Besov Ball). *Let $\beta_{j,k}$ denote coefficients of a function f in a wavelet basis indexed by $j \in \mathbb{N}, k \in [2^j]$. For parameters $\sigma \geq 0$, $p, q \in [1, \infty]$, $f \in \mathcal{L}^2$ lies in the Besov ball $B^\sigma_{p,q}$ iff*

$$\|f\|_{B^\sigma_{p,q}} := \left\| \left\{ 2^{j(\sigma + D(1/2 - 1/p))} \left\| \{\beta_\lambda\}_{\lambda \in \Lambda_j} \right\|_{l^p} \right\}_{j \in \mathbb{N}} \right\|_{l^q} \leq 1.$$

The parameter q affects convergence rates only by logarithmic factors, so we omit it in sequel.

Examples of IPMs

Distance	$\mathcal{F}$
$\mathcal{L}^p$ (including Total Variation/$\mathcal{L}^1$)	$B^0_{p'}$, with $p' = \frac{p}{p-1}$
Wasserstein ("earth-mover")	B^1_∞ (1-Lipschitz class)
Kolmogorov-Smirnov	B^1_1 (total variation ≤ 1)
Max. mean discrepancy (MMD)	RKHS ball
GAN Discriminator	parameterized by neural network

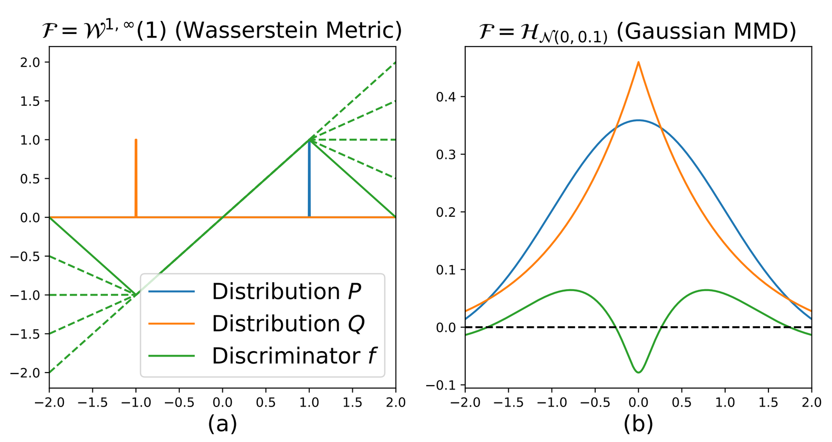


Figure 1: Examples of probability distributions P and Q and corresponding discriminator functions f^*. In (a), P and Q are Dirac masses at $+1$ and -1, resp., and $\mathcal{F}$ is the 1-Lipschitz class, so that $\rho_{\mathcal{F}}$ is the Wasserstein metric. In (b), P and Q are standard Gaussian and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

Minimax Rates for General Estimators

Theorem 1. *Suppose $\sigma_g \geq D/p_g$, $p'_d > p_g$. Then, up to polylog factors in n,*

$$M\left(B^{\sigma_g}_{p_g}, B^{\sigma_d}_{p_d}\right) := \inf_{\widehat{p}} \sup_{p \in B^{\sigma_g}_{p_g}} \mathop{\mathbb{E}}_{X_{1:n}} \left[\rho_{B^{\sigma_d}_{p_d}}(p, \widehat{p}(X_{1:n}))\right] \asymp n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}} + n^{-\frac{\sigma_g + \sigma_d + D - D/p_g - D/p'_d}{2\sigma_g + D - 2D/p_g}} + n^{-\frac{1}{2}}.$$

Moreover, this rate is achieved by the wavelet thresholding estimator of Donoho et al. [1].

Minimax Rates for Linear Estimators

Definition 3 (Linear Estimator). *A distribution estimate $\widehat{P}$ is said to be* linear *if there exist measures $T_i(X_i, \cdot)$ such that for all measurable A,*

$$\widehat{P}(A) = \sum_{i=1}^n T_i(X_i, A).$$

Examples: empirical distribution, kernel density estimate, or orthogonal series estimate.

Theorem 2. *Suppose $r > \sigma_g \geq D/p_g$. Then, up to polylog factors in n,*

$$M_{lin}\left(B^{\sigma_g}_{p_g}, B^{\sigma_d}_{p_d}\right) := \inf_{\substack{\widehat{p} \\ lin}} \sup_{p \in B^{\sigma_g}_{p_g}} \mathop{\mathbb{E}}_{X_{1:n}} \left[\rho_{B^{\sigma_d}_{p_d}}(p, \widehat{p}(X_{1:n}))\right] \asymp n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}} + n^{-\frac{\sigma_g + \sigma_d - D/p_g + D/p'_d}{2\sigma_g + D - 2D/p_g + 2D/p'_d}} + n^{-\frac{1}{2}},$$

where the $\inf$ *is over all linear estimates of $p \in \mathcal{F}_g$, and μ_p is the distribution with density p.*

Error Bounds for GANs

A natural statistical model for a perfectly optimized GAN as a distribution estimator is

$$\widehat{P} := \mathop{\mathrm{argmin}}_{Q \in \mathcal{F}_g} \sup_{f \in \mathcal{F}_d} \mathop{\mathbb{E}}_{X \sim Q} [f(X)] - \mathop{\mathbb{E}}_{X \sim \widetilde{P}_n} [f(X)], \tag{1}$$

where $\mathcal{F}_d$ and $\mathcal{F}_g$ are function classes parametrized by the discriminator and generator, resp [2].

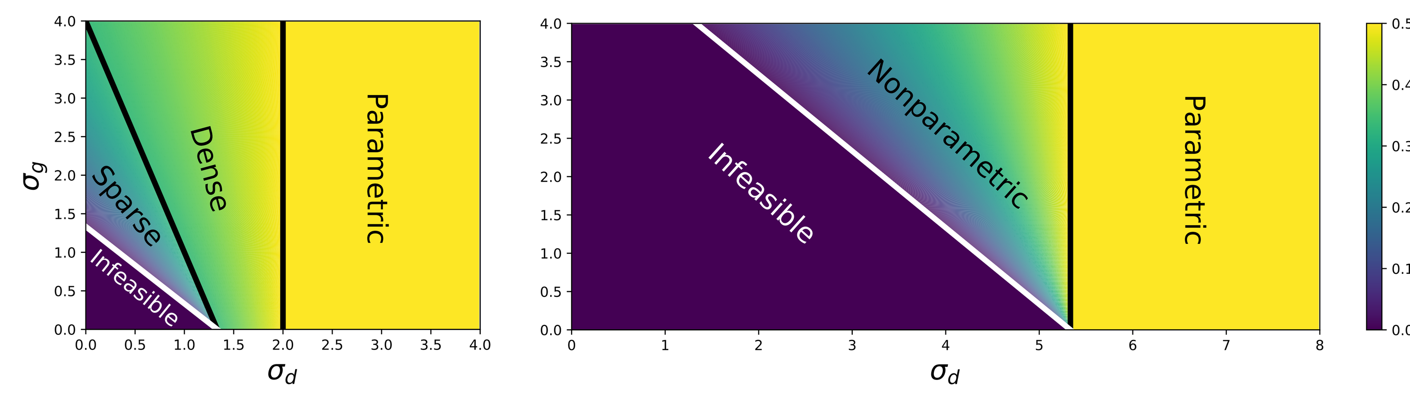
Theorem 3 (Convergence Rate of a Regularized GAN). *Fix a Besov density class $B^{\sigma_g}_{p_g}$ with $\sigma_g > D/p_g$ and discriminator class $B^{\sigma_d}_{p_d}$. Then, for some constant $C > 0$ depending only on $B^{\sigma_d}_{p_d}$ and $B^{\sigma_g}_{p_g}$, for any desired approximation error $\epsilon > 0$, one can construct a GAN $\widehat{P}$ of the form* (1) *(with $\widetilde{P}_n$ denoting the wavelet-thresholded distribution) whose discriminator network N_d and generator network N_g are fully-connected ReLU networks, such that*

$$\sup_{P \in B^{\sigma_g}_{p_g}} \mathbb{E}\left[d_{B^{\sigma_d}_{p_d}}\left(\widehat{P}, P\right)\right] \lesssim \epsilon + n^{-\eta(D, \sigma_d, p_d, \sigma_g, p_g)},$$

where $\eta(D, \sigma_d, p_d, \sigma_g, p_g)$ is the optimal exponent in Theorem 1.

- N_d and N_g have (rate-optimal) depth $\mathrm{polylog}(1/\epsilon)$ and width, max weight, and sparsity $\mathrm{poly}(1/\epsilon)$.
- Proof uses recent fully-connected ReLU network for approximating Besov functions [3].

Example Phase Diagrams



General Estimators　　　　Linear Estimators

Figure 2: Minimax convergence rates as functions of discriminator smoothness σ_d and distribution function smoothness σ_g, in the case $D = 4$, $p_d = 1.2$, $p_g = 2$. Color shows exponent of minimax convergence rate (i.e., $\alpha(\sigma_d, \sigma_g)$ such that $M\left(B^{\sigma_d}_{1.2}(\mathbb{R}^D), B^{\sigma_g}_2(\mathbb{R}^D)\right) \asymp n^{-\alpha(\sigma_d, \sigma_g)}$), ignoring polylogarithmic factors.

Applications/Examples

Example 1. ***(Total variation/Wasserstein-type losses)*** *If, for some $\sigma_d > 0$, $\mathcal{F}$ is a ball in B^s_∞, we obtain generalizations of total variation ($\sigma_d = 0$) and Wasserstein ($\sigma_d = 1$) losses. For these losses, we always have "Dense" rate*

$$M\left(B^{\sigma_g}_{p_g}, B^{\sigma_d}_{p_d}\right) \asymp n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}} + n^{-1/2}.$$

Example 2. ***(Kolmogorov-Smirnov-type losses)*** *If, for $\sigma_d > 0$, $\mathcal{F}$ is a ball in $B^{\sigma_d}_1$, we obtain generalizations of Kolmogorov-Smirnov loss ($\sigma_d = 0$). For these losses, we have "Sparse" rate*

$$M\left(B^{\sigma_g}_{p_g}, B^{\sigma_d}_1\right) \asymp n^{-\frac{\sigma_g + \sigma_d - D/p_g}{2\sigma_g + D - 2D/p_g}} + n^{-1/2}.$$

Example 3. ***(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,*

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

Example 4. ***(Sobolev IPMs)*** *For $\sigma \in \mathbb{N}$, B^σ_2 is the σ-order Hilbert-Sobolev ball $B^\sigma_2 = \left\{ f \in \mathcal{L}^2(\mathcal{X}) : \int_{\mathcal{X}} \left(f^{(\sigma)}(x)\right)^2 dx \leq \infty \right\}$, where $f^{(\sigma)}$ is the σ^{th} derivative of f. For these losses, we always have the rate*

$$M\left(B^{\sigma_g}_2, B^{\sigma_d}_2\right) \asymp n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}} + n^{-1/2}.$$

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

References

[1] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.

[2] Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabas Poczos. Nonparametric density estimation under adversarial losses. In *NeurIPS*, 2018.

[3] Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.