Minimax Rate of Distribution Estimation on Unknown Submanifolds Under Adversarial Loss

Advanced Topics in Statistical Learning, Spring 2023

Sharan Sahu and Raghav Ramanujam (based on paper by Rong Tang and Yun Yang)

1 Introduction

High-dimensional statistical models arise in various areas of computer science, biology, physics, etc and in order to make many of these inference problems tractable, we usually impose some low-dimensional structural assumptions. Sparsity is an example of this, and we have how statistical methods such as LASSO have specific statistical guarantees along with impressive success in various prediction tasks. In other applications, we may see that all variables have some influence on the model, but they all exhibit some low-dimensional structure in the sense that the dimensional of this latent low-dimensional structure, denoted d, is much smaller than the larger example set dimension, denoted D. This low-dimensional structure is seen across machine learning and statistical learning. In fact, as of recently, there have been significant breakthroughs in distribution estimation over complex spaces with generative models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models.

The success of these models relies on the fact that these models extract latent feature representations for reconstruction, and this is essentially using low-dimensional manifold structures for distribution estimation. To fully understand the benefit of low-dimensional manifold structure in generative modeling, it is best to create a general minimax framework for distribution estimation on an unknown submanifold under adversarial losses, with suitable smoothness assumptions on the target distribution and the manifold to make the problem tractable. By going through this minimax analysis using standard techniques in minimax theory, we can illustrate how various problem characteristics, including intrinsic dimensionality of the data and smoothness levels of the target distribution and the manifold, affect the fundamental limit of high-dimensional distribution estimation.

1.1 Generative Models

Mathematically, we can define generative models as follows: A generative model is a pair (ν, G) where ν is a distribution on a low-dimensional latent space $\mathcal{Z} \subset \mathbb{R}^d$ is called the generative distribution, and $G : \mathcal{Z} \to \mathbb{R}^d$ is a map called the generative map. In other words, the target distribution μ can be expressed via a generative model (ν, G) via $\mu = G_{\#}\nu$, the pushforward [1] measure of μ using map G. The set $\mathcal{D}_G = \{G_{\#}\nu : \nu \in \Upsilon, G \in \mathcal{G}\}$ of all generative models (ν, G) with $\nu \in \Upsilon$ and $G \in \mathcal{G}$ for some distribution family Υ on $\mathcal{Z}$ and function class $\mathcal{G}$ is called the generator class. There are a lot of practical benefits from generative models. For one, this decouples distribution estimation into serving two purposes: manifold estimation and density estimation. Additionally, these types of models are very useful because most of the time, we are more concerned with generating samples from an underlying distribution anyways since a known distribution (up to normalizing constant) may still require substantial effort to sample from (for example, sampling from Bayesian posteriors). Furthermore,these generative models are extremely powerful because they can be very good at capturing nonlinear structures which can be hard to capture in density or distribution functions. This may seem very daunting at first, but an example will make this clearer. As a small side note, generative models have the natural adversarial training framework of minimizing certain discrepancy measure between the empirical distributions of the real data and the generated synthetic data.

[1] For any measure ν on $\mathcal{Z}$ and a map $G : \mathcal{Z} \to \mathcal{X}$, the pushforward measure $\mu = G_{\#}\nu$ is defined as the unique measure on $\mathcal{X}$ such that $\mu(A) = \nu(G^{-1}(A))$.

This definition and its abstractness may seem very daunting at first, but a few example will make this clearer.

Example: Generative Adversarial Network. In a typical Generative Adversarial Network, we usually initialize $\nu = \mathcal{N}(0, 1)$, but this can be initialized to any distribution (possibly you know what your target distribution roughly looks like, so you can set ν to be some prior to ensure faster convergence). Our generative map $G : \mathcal{Z} \to \mathcal{X}$ can be a neural network (perhaps you want to create synthetic image data, so you let G be a Convolutional Neural Network). Then, if we take a sample $z \sim \nu$, then $G(z)$ should give us some synthetic data that mimics the training data but is exactly from the training data. In the typical GAN framework, we would then give $G(z)$ to some discriminator model, denoted $D(G(z))$, which would tell us if $G(z)$ is real or fake. We can then construct a loss based on KL-Divergence and coupled with optimization techniques, we can tune ν to be similar to μ.

Example: Variational Autoencoder. In a typical Variational Autoencoder, we have a probability distribution function for random noise defined as $z \sim \mathcal{P}_\phi$. Again, usually, we initialize this to just $\mathcal{N}(0, 1)$, but this can be any distribution family as we discussed above. We have two models: one called the probabilistic encoder, denoted $q_\phi(z|x)$, and another called the probabilistic decoder (also known as the generative model), denoted $p_\theta(x|z)$. In this case, the generative model would be $(\mathcal{P}_\phi, p_\theta)$. These models are just neural network models where the probabilistic encoder tries to learn the best distribution parametrization for the input training data in terms of a learned mean and variance, and the probabilistic decoder takes the latent-representation z and tries to generate the sample x. By using a specific loss function called ELBO, we ensure the input distribution $x \sim \mathcal{P}_\theta$ is similar to the generative distribution $z \sim \mathcal{P}_\phi$.

1.2 Adversarial Loss

Due to dealing with distributions with different supports, it may not be appropriate to use conventional discrepancy measures between probability measures on $\mathbb{R}^D$. One common set of losses used in this scenario is adversarial loss. For a discriminator class $\mathcal{F}$ of bounded and Borel-measurable functions, we can define the adversarial loss to be

$$d_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left| \int_{\mathbb{R}^D} f(x) d\mu - \int_{\mathbb{R}^D} f(x) d\nu \right|$$

From this general framework, we can get common discrepancy measures based on our discriminator class. For example, taking $\mathcal{F}$ to be the set of all 1-Lipschitz functions, we obtain the Wasserstein-1 metric. Similarly, the total variation is obtained by letting $\mathcal{F}$ be the set of all measurable functions bounded by 1.

Intuitively, since many characteristics of a distribution can be obtained by taking the integral of a function w.r.t the underlying probability measure, choosing our discriminator class wisely can help to compare probability measures in a finite sample sense. Further, adversarial losses have computational benefits as $d_{\mathcal{F}}$ can often be empirically estimated by putting samples from the underlying probability measures into a discriminator network.

In this lecture, we will be focusing on the problem when $\mathcal{F}$ is the D-dimensional γ-smooth Hölder class, and therefore $d_{\mathcal{F}}$ will correspond to the adversarial loss w.r.t the aforementioned function class. As a reminder, the scalar γ-smooth Hölder class with radius $r > 0$ over Ω is given by

$$C_r^\gamma(\Omega) = \{f : \Omega \to \mathbb{R} \mid \|f\|_{C_r^\gamma(\Omega)} = \sum_{|a| \leq \lfloor \gamma \rfloor} \max_{x \in \Omega} |f^{(a)}(x)| + \sum_{|a| = \lfloor \gamma \rfloor} \max_{x, y \in \Omega, x \neq y} \frac{f^{(a)}(x) - f^{(a)}(y)}{\|x - y\|^{\gamma - \lfloor \gamma \rfloor}}\}$$

We define the vector-valued extension to be given by

$$C_r^\gamma(\Omega, \mathbb{R}^d) = \{f = (f_1, \cdots, f_D) : \Omega \to \mathbb{R}^D \mid \forall j \in [D], f_j \in C^\gamma(\Omega)\}$$

In essence, a function is in the vector-valued class if each of its component functions is in the scalar class.

Let us define

$$d_\gamma(\mu,\nu) = \sup_{f\in C_1^\gamma(\mathbb{R}^D)} (\int_{\mathbb{R}^D} f(x)d\mu - \int_{\mathbb{R}^D} f(x)d\nu)$$

Now, we can verify that d_γ is a valid metric on probability measures on $\mathbb{R}^D$ as it satisfies the triangle inequality, and by the Weiestrauss approximation theorem (Stone), $d_\gamma(\mu,\nu) = 0$ iff $\mu = \nu$. Further, when we restrict the metric to distributions over a bounded set, we have with $\gamma = 1$ that it is equivalent to the Wasserstein-1 metric, and as $\gamma \to 0_+$, it approaches the total variation metric d_{TV}.

Further, the smoothness parameter γ in the metric can be viewed as a tuning parameter w.r.t the weightage of supporting manifold recovery and density estimation on the manifold. More specifically, the smaller γ is, the more sensitive d_γ is to support misalignment between μ and ν. To see this, consider two distributions μ and ν with bounded supports. Let us define $\text{dist}(x, A) := \inf_{y\in A} ||x - y||_2$. Then, letting $f(x) = c\ \text{dist}(x, \text{supp}(\nu))^\gamma - c\ \text{dist}(x, \text{supp}(\mu))^\gamma$, where c is some small constant such that $f \in C_1^\gamma(\mathbb{R}^D)$. Then, we have

$$\mathbb{E}_\mu[\text{dist}(X, \text{supp}(\nu))^\gamma] + \mathbb{E}_\nu[\text{dist}(X, \text{supp}(\mu))^\gamma] \le c^{-1} d_\gamma(\mu,\nu)$$

Note that as $\gamma \to 0_+$, the above equation becomes $\mathbb{P}_\mu(X \notin \nu) + \mathbb{P}_\nu(X \notin \mu)$, and therefore gives intuition for the increased sensitivity of d_γ to support misalignment at small γ.

1.3 Smooth Submanifolds and The Partition of Unity

From definition, a manifold is a topological space that locally resembles Euclidean space. A submanifold in the ambient space $\mathbb{R}^D$ can be viewed as a nonlinear "subspace". Formally, a β-smooth ($\beta \ge 1$) d-dimensional manifold $\mathcal{M}$ is defined as a topological space satisfying the following:

1. There exists an atlas on $\mathcal{M}$ consisting of d-dimensional charts $\mathscr{A} = \{(U_\lambda, \varphi_\lambda)\}_{\lambda\in\Lambda}$ such that $\mathcal{M} = \bigcup_{\lambda\in\Lambda} U_\lambda$
2. Each chart (U, φ) in atlas $\mathscr{A}$ consists of a homeomorphism [2] $\varphi : U \to \tilde{U}$, called the coordinate map, from an open set $U \subset \mathcal{M}$ to an open set $\tilde{U} \subset \mathbb{R}^d$
3. Any two charts (U, φ) and (V, ψ) in atlas $\mathscr{A}$ are compatible, meaning that the transistion map $\varphi\circ\psi^{-1} : \psi(U \cap V) \to \varphi(U \cap V)$ is a β-smooth diffeomorphism [3]

The manifold structure is an intrinsic property that does not rely on the choice of atlas. For a submanifold embedded in $\mathbb{R}^D$, the second and third condition can be combined into one single condition that the coordinate map φ in each chart is a β-smooth map when identified as a vector-valued function from $U \subset \mathbb{R}^D$ to $\tilde{U} \subset \mathbb{R}^d$. This definition may be a little daunting, so we have included a visualization that will hopefully allow you to better understand the characterization of a manifold. Let us use this definition and our definition of the α-smooth Hölder class as a motivating example.

Example: $C^\alpha(\mathcal{M})$ Over A β-Smooth Manifold $\mathcal{M}$. The α-smooth Hölder (function) class $C^\alpha(\mathcal{M})$ for $\alpha \in (0, \beta]$ over a β-smooth manifold $\mathcal{M}$ consists of all functions $f : \mathcal{M} \to \mathbb{R}$ whose localization $f \circ \varphi^{-1} : \varphi(U) \to \mathbb{R}$ to each local chart (U, φ) is α-Hölder smooth in the usual Euclidean sense. Note that we restrict α to be at most β because it will not be useful to consider higher-order smoothness if it may (and probably will not) be compatible between charts if the atlas is at most β-smooth.

[2] A homeomorphism is a mapping between topological spaces $\varphi : U \to \tilde{U}$ that preserves the structure. More formally, a homeomorphism is a mapping where φ is bijective, continuous, and φ^{-1} is continuous.

[3] A diffeomorphism is like a homeomorphism. The only restriction that we have now is that instead of the mapping and its inverse being continuous, we require them to be differentiable. Sometimes, a homeomorphism is called a C^0-diffeomorphism. In our case, we require the mapping and its inverse be β-times differentiable. This is sometimes called a C^β-diffeomorphism.

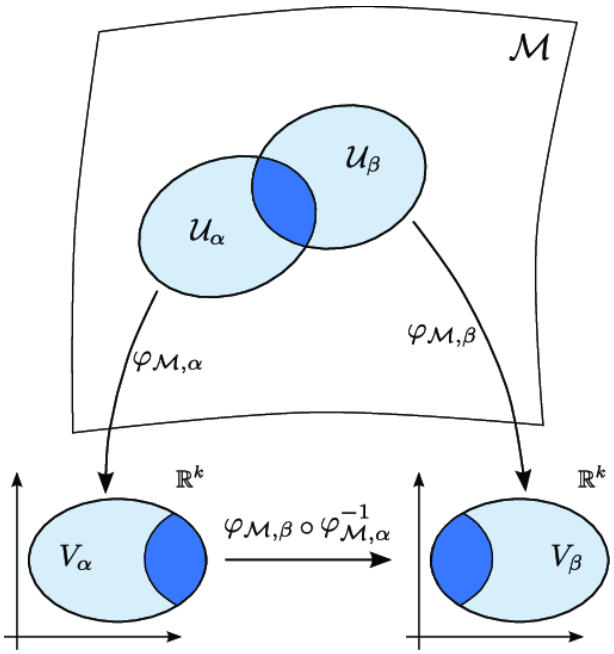

Figure 1: Illustration of Charts on A Manifold $\mathcal{M}$

Most generative model-based distribution estimation procedures use a single generative model (ν, G) to model the underlying target distribution μ. This requirement is too strong since it implicitly requires the underlying submanifold $\mathcal{M}$ that supports the target distribution $\mu = G_{\#}\nu$ to admit a single chart description. This requirement is too strong as many commonly encountered manifolds such as spheres cannot be covered by a single chart in any of its representing atlas. Thus, we allow an underlying data manifold to be supported by multiple charts by using a mathematical technique called the partition of unity:

Definition 1 (Partition of Unity)**.** A partition of unity on a β-smooth manifold $\mathcal{M}$ is a collection of β-smooth functions $\{\rho_\lambda\}_{\lambda\in\Lambda}$ on $\mathcal{M}$ such that:

1. $0 \leq \rho_\lambda \leq 1$ for all $\lambda \in \Lambda$, and $\sum_{\lambda\in\Lambda} \rho_\lambda(x) = 1$ for all $x \in \mathcal{M}$
2. Each point $x \in \mathcal{M}$ has a neighborhood that intersects $\text{supp}(\rho_\lambda)$ for only finitely many $\lambda \in \Lambda$

Using the partition of unity, one can glue constructions in the local charts to form a global construction of the manifold. Such a global construction usually does not rely on the choice of the partition of unity. A partition of unity can be constructed from any open cover $\{U_\lambda\}_{\lambda\in\Lambda}$ of the manifold in a way where the partition of unity $\{\rho_\lambda\}_{\lambda\in\Lambda}$ is indexed over the the same set and $\text{supp}(\rho_\lambda) \subset U_\lambda$ for any $\lambda \in \Lambda$. This leads us to the setting that we will be using for the rest of these notes. Assume that $\mathcal{M}$ is contained in closed ball $\mathbb{B}^D_L$ for some sufficiently large radius L. We shall construct the partition of unity of $\mathcal{M}$ as follows. First, we find some set of points $a_{1:M} = \{a_1, a_2, ..., a_M\}$ in $\mathbb{R}^D$ and set of positive radii $r_{1:M} = \{r_1, ..., r_M\}$ such that $\mathscr{O}_M = \{\mathbb{B}_{r_m}(a_m)^\circ\}_{m\in[M]}$ forms a finite open cover of $\mathbb{B}^D_L$. Let $\chi : \mathbb{R} \Rightarrow [0, \infty)$ be defined as $\chi(t) = e^{-1/t}$ for $t > 0$ and $\chi(t) = 0$ for $t \leq 0$. For each $m \in [M]$, we define a local partition function as

$$\tilde{\rho}_m(x) = \frac{\chi(r_m - ||x - a_m||_2)}{\chi(r_m - ||x - a_m||_2) + \chi(||x - a_m||_2 - r_m/2)}$$

It is a relatively straightforward exercise to check that for $\rho_m = \tilde{\rho}_m / \sum_{m'=1}^M \tilde{\rho}_{m'}$, $\{\rho_m\}$ for a partition of unity for $\mathcal{M}$ for $m \in [M]$.

1.4 Smooth Distributions on Submanifolds and The Generative Model Class

For a smooth submanifold $\mathcal{M}$ with atlas $\mathscr{A} = \{(U_\lambda, \varphi_\lambda)\}_{\lambda\in\Lambda}$, one can define a distribution μ on $\mathcal{M}$ by specifying how it acts on all smooth functions $f \in C^\beta(\mathcal{M})$ through its expectation $\mathbb{E}_\mu[f]$. Usually, we define the expected value of a function f with respect to some measure μ is

$$\mathbb{E}_\mu[f] = \int_{\mathcal{M}} f d\mu$$

However, this also assumes a global parametrization to our submanifold $\mathcal{M}$. However, we can make use of the partition of unity to decompose this integral into the sum of local integrals as follows:

$$\mathbb{E}_\mu[f] = \int_{\mathcal{M}} f d\mu = \sum_{\lambda\in\Lambda} \int_{U_\lambda} f d(\rho_\lambda \mu) = \sum_{\lambda\in\Lambda} \int_{\varphi_\lambda(U_\lambda)} f \circ \varphi_\lambda^{-1} d[(\varphi_\lambda)_\#(\rho_\lambda\mu)]$$

where $\rho_\lambda\mu$ is a non-negative measure whose Radon-Nikodym derivative with respect to μ is ρ_λ. The second equality comes from the the characterization of the charts in the manifold. Again, we restrict α to be at most β because it will not be useful to consider higher-order smoothness if it may (and probably will not) be compatible between charts if the atlas is at most β-smooth. By the change of measure formula

$$[(\varphi_1)_\#(\rho_1\rho_2\mu)](\varphi_1(x)) = [(\varphi_2)_\#(\rho_1\rho_2\mu)](\varphi_2(x))|\det(d[\varphi_2 \circ \varphi_1]_{\varphi_1(x)})|$$

for all $x \in U_1 \cap U_2$. This may lead to incompatible smoothness definitions over the intersection of charts (U_1, φ_1) and (U_2, φ_2) if the atlas is at most β-smooth as the [4] differential $d[\varphi_2 \circ \varphi_1]_y) : \mathbb{R}^d \to \mathbb{R}^d$ at $y \in \varphi(U_1 \cap U_2)$. Now, we can finally define the family of smooth distributions on a smooth compact submanifold without any boundaries on $\mathbb{R}^D$ as the set $\mathcal{P}^* = \mathcal{P}^*(d, D, \alpha, \beta, L^*)$ with $d \le D$, $\beta > 1$, and $\alpha \in (0, \beta - 1]$ composed of all probability measures $\mu \in \mathcal{P}(\mathbb{R}^D)$ satisfying the following:

1. μ is an α-smooth distribution on a β-smooth d-dimensional compact submanifold $\mathcal{M}$ embedded in $\mathbb{R}^D$
2. The density μ relative to the volume measure of $\mathcal{M}$ is uniformly bounded from below by $1/L^*$ on $\mathcal{M}$
3. $\mathcal{M}$ is covered by an atlast $\mathscr{A} = \{(U_\lambda, \varphi_\lambda)\}_{\lambda\in\Lambda}$ on $\mathcal{M}$ such that each chart (U, φ) in atlas $\mathscr{A}$ satisfies $||\varphi^{-1}||_{C^\beta(\varphi(U))} \le L^*$ and $||\mu \circ \varphi^{-1}||_{C^\alpha(\varphi(U))} \le L^*$ as well as for any $z \in \varphi(U)$, the Jacobian of $\varphi^{-1}(z)$ is full rank and all its singular values are lower bounded by $1/L^*$ in absolute values. Moreover, for any $x \in \mathcal{M}$, there exists an $\lambda \in \Lambda$ such that U_λ and $\varphi_\lambda(U_\lambda)$ covers $B_{1/L^*}(x) \cap \mathcal{M}$ and $B_{1/L^*}(\varphi_\lambda(x))$ respectively

1.5 Distribution Estimator Class: Mixture of Generative Models

Using the family of smooth distributions on smooth compact submanifolds without boundaries on $\mathbb{R}^D$ as we defined above, we can begin to describe the statistical model for representing these probability measures μ in these family of distributions. In order to do so, we consider two mixtures of generative models classes $\mathcal{S}^{\text{ap}} = \mathcal{S}^{\text{ap}}(d, D, \alpha, \beta, \mathscr{O}_M, L)$ and $\mathcal{S}^{\text{ap}}_{\nu_0} = \mathcal{S}^{\text{ap}}_{\nu_0}(d, D, \alpha, \beta, \mathscr{O}_M, L)$ where $\mathscr{O}_M = \{\mathbb{B}_{r_m}(a_m)^\circ\}_{m\in[M]}$ is a pre-specified open cover of $\mathbb{B}^D_L$ that contains a submanifold. The first generative model class $\mathcal{S}^{\text{ap}}$ consists of mixtures of generative models with rejection sampling $\mu = \sum_{m=1}^M w_{[m]} \mathcal{A}(G_{[m]}, \nu_{[m]}, \rho_m)$ where $\{\rho_m\}_{m\in[M]}$ is the partition of unity, $\{w_{[m]}\}_{m\in[M]}$ is the set of non-negative weights where $\sum_{m=1}^M w_{[m]} = 1$ and for any $m \in [M]$, each component of $G_{[m]}$ is a β-smooth function over $\mathbb{R}^d$ with β-Hölder norm bounded by L. We have $\nu_{[m]}$ is a α-smooth probability density on $\mathbb{B}^d_1$ with α-Hölder norm bounded by L. $\mathcal{A}(G_{[m]}, \nu_{[m]}, \rho_m)$ denotes the probability measure induced by the data generating process where $X \sim [G_{[m]}]_\#\nu_{[m]}$ is accepted with probability $\rho_m(X) \in [0, 1]$. Now, we want to avoid explicit estimation of the local densities $\nu_{[m]}$, so we consider $\mathcal{S}^{\text{ap}}_{\nu_0}$ which is the same as $\mathcal{S}^{\text{ap}}$ except we replace the local latent variable distribution $\nu_{[m]}$ by a generative model $(\nu_0, V_{[m]})$ for each $m \in [M]$.

[4] There are two tangent spaces $\mathbb{T}_y(\mathbb{R}^d)$ and $\mathbb{T}_{\varphi_2\circ\varphi_1^{-1}(y)}(\mathbb{R}^d)$

2 Minimax Rates of Convergence

Now, we are ready to establish the minimax rate of convergence for the adversarial risk on $\mu^* \in \mathcal{P}^*$ of distribution estimation on an unknown submanifold with i.i.d samples $X_1, ..., X_n \sim \mu^*$. We can summarize the minimax rate of distribution estimation as follows:

Theorem 1 (Minimax Rate of Distribution Estimation)**.** *Fix $L^* > 0$, $\gamma \geq 0$, $0 \leq \alpha \leq \beta - 1$, $\beta > 1$, and $D, d \in \mathbb{N}^+$ with $D > d$. Write $\mathcal{P}^* = \mathcal{P}^*(L^*, \gamma, \alpha, \beta, d, D)$. Then,*

1. *There exists a constant L_0 such that when $L^* \geq L_0$, then*

$$\inf_{\hat{\mu} \in \mathcal{P}(\mathbb{R}^D)} \sup_{\mu \in \mathcal{P}^*} \mathbb{E}\left[d_\gamma(\hat{\mu}, \mu)\right] \geq C n^{-\frac{1}{2}} \vee n^{\frac{-\alpha+\gamma}{2\alpha+d}} \vee n^{-\frac{\gamma\beta}{d}}$$

2. *There exists positive constants L_1, L_2 such that for any $L \geq L_1$ and open cover $\mathscr{O}_M = \{\mathbb{B}_{r_m}(a_m)^\circ\}_{m \in [M]}$ of $\mathbb{B}_L^D$ with $\max\{r_1, r_2, ..., r_M\} \leq L_2$, it holds that*

$$\inf_{\hat{\mu} \in S_{\nu_0}^{\text{ap}}} \sup_{\mu \in \mathcal{P}^*} \mathbb{E}\left[d_\gamma(\hat{\mu}, \mu)\right] \leq C \left(\frac{n}{\log n}\right)^{-\frac{1}{2}} \vee \left(\frac{n}{\log n}\right)^{\frac{-\alpha+\gamma}{2\alpha+d}} \vee \left(\frac{n}{\log n}\right)^{-\frac{\gamma\beta}{d}}$$

Proof: We shall only prove the lower bound and will leave the proof of the upper bound to Theorem 2. In order to prove the lower bound, we will make use of the standard minimax techniques such as Le Cam's method and Fano's method. We will proceed by finding a subset of distributions within the considered distribution family $\mathcal{P}^*(d, D, \alpha, \beta, L^*)$ that are statistically hard to distinguish. We will not completely specify the proof. Instead, we will outline what techniques and methods are used in proving each part of the lower bound.

Lower Bound of $n^{-\frac{\gamma\beta}{d}}$. Under the assumption that $\beta > 1$, we only need to consider $\gamma \in [0, 1)$. To see this, we will show that otherwise, $n^{-\frac{\gamma\beta}{d}}$ is always dominated by the other terms $n^{-\frac{1}{2}} \vee n^{\frac{-\alpha+\gamma}{2\alpha+d}}$. If $\gamma \geq d/2$, then by $\beta > 1$, we have that $n^{-\frac{\gamma\beta}{d}} \leq n^{-1/2}$. If $1 \leq \gamma < d/2$, then using $\alpha \leq \beta - 1$, we get that

$$n^{\frac{-\alpha+\gamma}{2\alpha+d}} \geq n^{-\frac{\beta-1+\gamma}{2(\beta-1)+d}} \geq n^{-\frac{\beta-1+\gamma}{d}} \geq n^{-\frac{\gamma\beta}{d}}$$

where the last inequality is due to the fact that $\gamma\beta - (\beta - 1 + \gamma) = (\beta - 1)(\gamma - 1) \geq 0$. Thus, we focus on $\gamma \in [0, 1)$. Now, we construct a subset of distributions that is statistically hard to distinguish, meaning the the KL Divergence between the two distributions is bounded by a constant while maintaining that their mutual distances d_γ distances are at least $O(n^{\frac{-\gamma\beta}{d}})$. This way, we can apply the standard reduction from an estimation problem to a multiple testing problem which will allow us to use Fano's lemma. We will not detail how we construct this subset of distributions. If you are interested, please take a look at the proof of Theorem 1 in the paper. Once we have constructed this subset of distributions, we are able to obtain

$$\inf_{\hat{h}} \sup_{h \in H} \mathbb{E}\left[d_\gamma(\mu_{\hat{h}}, \mu_h)\right] \geq \frac{1}{2} \inf_{h,l \in [H], h \neq l} d_\gamma(\mu_h, \mu_l) \geq \frac{c}{2} n^{-\frac{\gamma\beta}{d}}$$

Lower Bound of $n^{\frac{-\alpha+\gamma}{2\alpha+d}}$. Again, we must construct a subset of distributions that is statistically hard to distinguish. We will not detail how we construct this subset of distributions. Once we have constructed this subset of distributions, we are able to obtain, we apply Fano's lemma to obtain

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathcal{P}^*} \mathbb{E}\left[d_\gamma(\hat{\mu}, \mu)\right] \geq \frac{1}{2} \inf_{h,l \in [H], h \neq l} d_\gamma(\mu_h, \mu_l) \cdot \left(1 - \frac{\log 2 + \frac{n}{H^2} \sum_{h,l=1}^{H} D_{\text{KL}}(\mu_h || \mu_l)}{\log H}\right) \geq c' n^{\frac{-\alpha+\gamma}{2\alpha+d}}$$

Lower Bound of $n^{-\frac{1}{2}}$. The $n^{-\frac{1}{2}}$ can be obtained by Le Cam's method of reducing the estimation problem into a two-point hypothesis testing problem. The proof relies on the existence of two distributions μ_0 (uniform distribution) and μ_1 supported on $\mathcal{M}_0$ with the following properties: For two distributions ν and μ such that $\nu << \mu$, the chi-squared distribution ν to μ is defined as $d_{\chi^2}(\nu, \mu) = \int \left(\frac{d\nu}{d\mu} - 1\right) d\mu$.

Lemma 1 (Perturbation of Uniform Distribution)**.** *There exists two distributions μ_0 and μ_1, belonging both to $\mathcal{P}^*(d, D, \alpha, \beta, L^*)$ such that $\mu_1 << \mu_0$, $d_{\chi^2}(\mu_1, \mu_0) \leq \frac{1}{n}$, and $d_\gamma(\mu_1, \mu_0) \geq \frac{c}{\sqrt{n}}$*

Using this lemma paired with Le Cam's method, we get that

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathcal{P}^*} \mathbb{E}\left[d_\gamma(\hat{\mu}, \mu)\right] \geq d_\gamma(\mu_1, \mu_0) \left(1 - \sqrt{\frac{e-1}{2}}\right) \geq \left(1 - \sqrt{\frac{e-1}{2}}\right) \frac{c}{\sqrt{n}}$$

□

Intuitively, the term $n^{-\frac{\gamma\beta}{d}}$ reflects the statistical hardness of estimating an unknown β-smooth submanifold whereas the first two terms $n^{-\frac{1}{2}} \vee n^{\frac{-\alpha+\gamma}{2\alpha+d}}$ reflects the statistical hardness of estimating an unknown α-smooth density as if the submanifold is known. A figure which shows the terms in the minimax rates and where they start to dominate can be seen below.

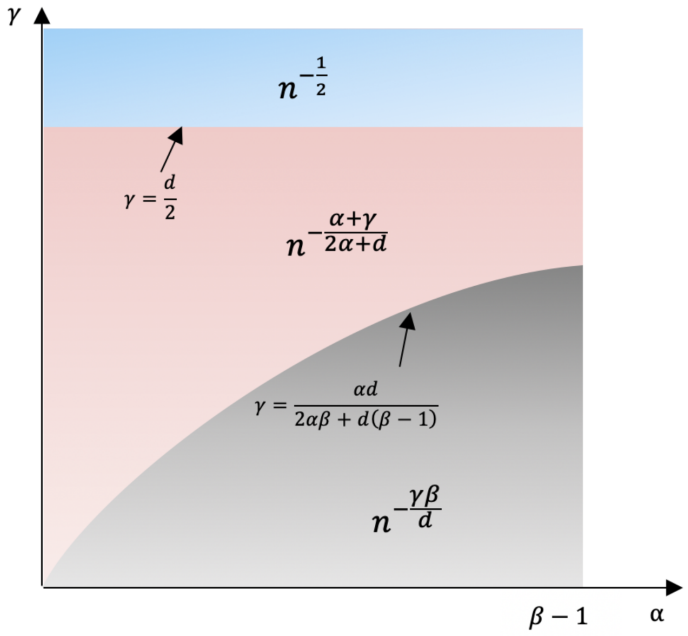

Figure 2: Illustration of Minimax Rate $n^{-\frac{1}{2}} \vee n^{\frac{-\alpha+\gamma}{2\alpha+d}} \vee n^{-\frac{\gamma\beta}{d}}$ for fixed $d \in \mathbb{N}^+, \beta > 1$

From the diagram, there exist transitions at $\gamma = d/2$ and $\gamma = \alpha d/[2\alpha\beta + d(\beta - 1)]$. When the discriminator smoothness level γ satisfies $\gamma \geq d/2$ so that discriminator class is relatively small, the rate is the $n^{-1/2}$. When the discriminator smoothness level is modernate or $\alpha d/[2\alpha\beta + d(\beta - 1)] \leq \gamma < d/2$, the term $n^{\frac{-\alpha+\gamma}{2\alpha+d}}$ dominates the minimax rate. When $0 \leq \gamma < \alpha d/[2\alpha\beta + d(\beta - 1)]$, the minimax rate becomes $n^{-\frac{\gamma\beta}{d}}$ since when γ is small, the adversarial loss $d_\gamma(\mu, \nu)$ between the two distributions μ and ν tends to be more sensitive to misalignment between their supports than the probabilty miss allocations on their supports. The last part of Theorem 1 can be generalized into another Theorem which can be seen below:

Theorem 2 (Minimax Upper Bound of Generative Model Class)**.** *Let the approximation family $\mathcal{S}$ to be $\mathcal{S}^{ap}(d, D, \alpha, \beta, \mathscr{O}_M, L)$. Suppose $\mu^* \in \mathcal{S}^*(d, D, \alpha, \beta, \mathscr{O}_M, L)$ and $X_{1:n}$ are i.i.d. samples from μ^*. If $D > d$,*

γ, $\alpha \geq 0$, and $\beta > 1$, then for any positive constant c, there exist positive constants c_1 and n_0 such that when $n \geq n_0$, it holds with probability greater than $1 - n^{-c}$ that

$$\sup_{f \in C_1^\gamma(\mathbb{R}^D)} (\mathbb{E}_{\mu^*}[f(X)] - \sum_{m=1}^{M} \hat{\mathcal{J}}_m(f)) \leq c_1 (\frac{n}{\log n})^{-\frac{1}{2}} \vee (\frac{n}{\log n})^{-\frac{\gamma+\alpha}{2\alpha+d}} \vee (\frac{n}{\log n})^{-\frac{\gamma\beta}{d}}$$

As a result,

$$\mathbb{E}[d_\gamma(\hat{\mu}, \mu)] \leq C (\frac{n}{\log n})^{-\frac{1}{2}} \vee (\frac{n}{\log n})^{-\frac{\gamma+\alpha}{2\alpha+d}} \vee (\frac{n}{\log n})^{-\frac{\gamma\beta}{d}}$$

Proof: For obtaining the upper bound, we make use of wavelets. We create a surrogate loss $\hat{\mathcal{J}} = \hat{\mathcal{J}}_l + \hat{\mathcal{J}}_s + \hat{\mathcal{J}}_h$ to approximate $\mathbb{E}\left[d_\gamma(\hat{\mu}, \mu)\right]$ where $\hat{\mathcal{J}}_l = \sum_{m=1}^{M} \hat{\mathcal{J}}_{m,l}$, $\hat{\mathcal{J}}_h = \sum_{m=1}^{M} \hat{\mathcal{J}}_{m,h}$, and $\hat{\mathcal{J}}_s = \sum_{m=1}^{M} \hat{\mathcal{J}}_{m,s}$ where each $m \in [M]$ corresponds to an index in the partition of unity. In particular, $\hat{\mathcal{J}}_l f$ estimates the expectation of $\Pi_J f$ that collects the low frequency components in the wavelet expansion of f where $\Pi_J f$ is the projection of f onto the first scale J wavelet coefficients which is given by

$$\Pi_J f = \sum_{k \in \mathbb{Z}^d} b_k \phi_k(x) + \sum_{l=1}^{2^d-1} \sum_{j=1}^{J} \sum_{k \in \mathbb{Z}^d} f_{ljk} \psi_{ljk}(x)$$

$\hat{\mathcal{J}}_h f$ estimates the expectation of $\Pi_J^\perp f$ that collects the high frequency components. Lastly, $\hat{\mathcal{J}}_s f$ corresponds to a high-order smoothness correction due to the submanifold error from the local coordinate map estimator. By applying Bernstein's inequality for a binomial random variable and a union bound argument, we can get terms $\sqrt{\frac{\log n}{n}}$. Now, similar to Homework 1 where we upper bounded the L^2 risk of a wavelet smoothing operator by its truncation, discretization, and estimation error, we can upper bound the low-frequency components, high-frequency components, and higher-order smoothness correctness components in the wavelet expansion of f. In doing so, we get the following upper bound:

$$\sup_{f \in C_1^\gamma(\mathbb{R}^D)} (\mathbb{E}_{\mu^*}[f(X)] - \sum_{m=1}^{M} \hat{\mathcal{J}}_m(f)) \leq c_1 (\frac{n}{\log n})^{-\frac{1}{2}} \vee (\frac{n}{\log n})^{-\frac{\gamma+\alpha}{2\alpha+d}} \vee (\frac{n}{\log n})^{-\frac{\gamma\beta}{d}}$$

The full details of the proof are omitted here. The full proof is detailed in the paper. □

There are some things to notice in these bounds. First, the logarithmic terms appearing in the upper bound of Theorem 1 enable us to obtain a high probability bound for further bounding the expected loss. Second, the ambient space dimension D does not appear in the exponents of the minimax rate, so the problem of estimating a distribution on a low-dimensional submanifold does not suffer from the "curse of dimensionality" due to a large D. Third, recall that the adversarial loss d_γ employed in distribution estimation captures two aspects of the data generating process: supporting manifold recovery and density estimation on the manifold. These were successfully captured in our minimax rate.

3 Conclusion

In this paper, we studied the minimax rate of distribution estimation on unknown submanifold under adversarial losses, covering cases where the manifold, the density,and the discriminator class have various Hölder regularities. In conclusion, the minimax rate shows that the curse of dimensionality can be overcome for data with low intrinsic dimension, smooth density and regular support, which partly explains the empirical successes of generative model based approaches for generating realistic objects in real applications.

References

[1] Tang, Rong and Yang, Yun. (2022). Minimax Rate of Distribution Estimation on Unknown Submanifold under Adversarial Losses.

[2] Mimetic framework on curvilinear quadrilaterals of arbitrary order - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Coordinate-charts-on-a-manifold fig 51956714 [accessed 2 May, 2023]