Minimax density estimation for growing dimension

Daniel J. McDonald
Department of Statistics
Indiana University
Bloomington, IN, 47405
dajmcdon@indiana.edu

Abstract

This paper presents minimax rates for density estimation when the data dimension d is allowed to grow with the number of observations n rather than remaining fixed as in previous analyses. We prove a non-asymptotic lower bound which gives the worst-case rate over standard classes of smooth densities, and we show that kernel density estimators achieve this rate. We also give oracle choices for the bandwidth and derive the fastest rate d can grow with n to maintain estimation consistency.

1 INTRODUCTION

A convincing argument for the use of sparsity or other structural priors in machine learning and statistics often begins with a discussion of the "curse of dimensionality" (e.g. Donoho, 2000). Unmistakable evidence of this curse is simply demonstrated in the fundamental scenario of non-parametric density estimation: the best estimator has squared L^2 error on the order of $n^{-4/(4+d)}$ given n independent observations in d dimensions, a striking contrast with the parametric rate d/n. If d is even moderately large (but fixed), accurate estimation requires significantly more data than if d were small. In fact, we will show that if d is allowed to increase with n, estimation accuracy degrades even more quickly than the non-parametric rate above indicates.

At first, it may seem that allowing d to grow with n is a rather strange scenario, but the use of "triangular array" asymptotics is exceedingly common in the theory of high-dimensional estimation. Theoretical results for the lasso, beginning at least with (Greenshtein and Ritov, 2004), regularly adopt this framework allowing the number of predictors to grow with n. Bühlmann and van de Geer (2011) introduce the idea at the very beginning of their foundational text, and it has been widely adopted in the literature on regularized linear models (e.g. Belloni et al., 2011; Bickel et al., 2009; Meinshausen, 2007; Nardi and Rinaldo, 2008; Ye and Zhang, 2010). Under this framework, the marginal distribution of the predictors has support whose dimension is increasing with n. In the scenario of high-dimensional regression, the dimension can increase very quickly (often on the order of $d = o(n^\alpha),\ \alpha > 1$) as long as most of these dimensions are irrelevant for predicting the response. The extension of these results for linear models to the non-linear scenario has been studied mainly in the case of generalized (sparse) additive models (Ravikumar et al., 2008, 2009; Yuan and Zhou, 2015) which allow for predictor specific non-linearities as long as the final predictions are merely additive across dimensions. Fully nonparametric regression without the additivity assumption has been completely ignored outside of the fixed-d framework, although it is a natural extension of the work presented here.

Another motivation for appropriating the triangular array framework in non-parametric density estimation is the burgeoning literature on manifold estimation (Genovese et al., 2012a,b; Talwalkar et al., 2008). Given high-dimensional data, a natural assumption is that the data is supported on a low-dimensional manifold embedded in the high-dimensional space. While estimating the manifold is possible, we may also wish to estimate a density or a regression function supported on the manifold. Recent work has focused on density estimation when the dimension of the manifold is fixed and known (Asta, 2013; Bhattacharya and Dunson, 2010; Hendriks, 1990; Pelletier, 2005), but the extension of such results to manifolds of growing di-

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

mension is missing. Such an extension presumes that the minimax framework we present can be extended to manifolds. As pointed out by a reviewer, the short answer is yes. The lower bound we derive applies immediately. The only modification we need relates to our upper bound: the kernel should depend on the metric given by the manifold rather than Euclidean distance as we use here.

A specific application of our setting would be from fMRI data. Given a sequence of 3D resting-state fMRI scans from a patient, researchers seek to estimate the dependence between cubic centimeter voxels (e.g. Bullmore and Sporns, 2009). Each scan can contain on the order of 30,000 voxels, while the number of scans for one individual is smaller. It is too much to estimate the dependence between all voxels, so the data are averaged into a small number (∼20–200) of regions. To estimate the dependence, standard methods assume everything is multivariate Gaussian and estimate the covariance or precision matrix. But the Gaussian assumption cannot be tested without density estimates. Using our results, we could estimate smooth densities. As the number of scans grows, we would want to increase the number of regions. Our work illustrates how quickly the number of regions can grow.

The remainder of this section introduces the statistical minimax framework, discusses the specific data generating model we examine and details notation, presents some background on the estimator we use which achieves the minimax rate, and gives a short overview of related literature. In Section 2, we give our main results and discuss their implications, specifically obtaining the fastest rate at which d can grow with n to yield estimation consistency. Section 3 gives the proof of our lower bound over all possible estimators while the proof of the matching upper bound for the kernel density estimator is given in Section 4. Finally, we discuss these results in Section 5, provide some related results for other loss functions, and suggest avenues for future research.

1.1 The Minimax Framework

In order to evaluate the feasibility of density estimation under the triangular array, we use the statistical minimax framework. In our situation, this framework begins with a specific class of possible densities we are willing to consider and provides a lower bound on the performance of the best possible estimator over this class. With this bound in hand, we have now quantified the difficulty of the problem. If we can then find an estimator which achieves this bound (possibly up to constants), then we can be confident that this estimator performs nearly as well as possible for the given class of densities. Thus, the minimax framework reveals gaps between proposed estimators and the limits of possible inference. Of course if the bounds fail to match, then we won't know whether they are too loose, or the estimator is poor.

1.2 Model and Notation

We specify the following setting for density estimation in a triangular array. Suppose for each $n \geq 1$, $X_i^{(n)} \in \mathbb{R}^{d(n)}$, $i = 1, \ldots, n$ are independent with common density $f^{(n)}$ in some class which we define below. For notational convenience, we will generally suppress the dependence on (n). To be clear, in specifying this model, we do not assume a relationship for some sequence of densities $\{f^{(n)}\}_{n=1}^{\infty}$, but rather we seek to understand the limits of estimation when there is a correspondence between $d(n)$ and n. Thus, we seek non-asymptotic results which characterize this behavior. We will also employ the following notation: given vectors $s, x \in \mathbb{R}^d$, let $|s| = \sum_i s_i$, $s! = \prod_i s_i!$ and $x^s = x_1^{s_1} \cdots x_d^{s_d}$. Then define

$$D^s = \frac{\partial^{|s|}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

Let $\lfloor \beta \rfloor$ denote the largest integer strictly less than β. Throughout, we will use a and A for positive constants whose values may change depending on the context.

Even were d fixed at 1, it is clear that density estimation is impossible were we to allow f to be arbitrary.[1] For this reason, we will restrict the class of densities we are willing to allow.

Definition 1 (Nikol'skii class)**.** *Let $p \in [2, \infty)$. The* isotropic Nikol'skii class $\mathcal{N}_p(\beta, C)$ *is the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that:*
(i) $f \geq 0$ a.e.
(ii) $\int f = 1$.
(iii) partial derivatives $D^s f$ exist whenever $|s| \leq \lfloor \beta \rfloor$
(iv) $\left[\int (D^s f(x+t) - D^s f(x))^p \, dx\right]^{1/p} \leq C \|t\|_1^{\beta - |s|}$,
for all $t \in \mathbb{R}^d$.

This definition essentially characterizes the smoothness of the densities in a natural way. It can be shown easily that the Nikol'skii class generalizes Sobolev and Hölder classes under similar conditions (see e.g. Tsybakov, 2009, p. 13).

1.3 Parzen-Rosenblatt Kernel Estimator

Given a sample $X_1, \ldots, X_n$, the Parzen-Rosenblatt kernel density estimator on $\mathbb{R}^d$ at a point x is given

[1] In the sense that, an adversary can choose a density and give us a finite amount of data on which our estimators will perform arbitrarily poorly.

by

$$\widehat{f}_h(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

We will consider only certain functions K.

Definition 2. *We say that* $K : \mathbb{R}^d \to \mathbb{R}$ *is an* isotropic kernel of order β *if* $K(u) = G(u_1)G(u_2)\cdots G(u_d)$ *for* $G : \mathbb{R} \to \mathbb{R}$ *satisfying* $\int G = 1$, $\int |u|^\beta |G(u)| du < \infty$, *and* $\int u^j G(u) du = 0$, *for* $0 < j \le \lfloor \beta \rfloor$.

For the standard case $\beta = 2$, the Epanechnikov kernel $G(u) = 0.75(1-u^2)I(|u| \le 1)$ satisfies these conditions and is often the default in software. The Gaussian kernel, $G(u) = (2\pi)^{-1/2}e^{-u^2/2}$, is also a member of this class. For $\beta > 2$, the kernel must take negative values, possibly resulting in negative density estimates, although, using the positive-part estimator eliminates this pathology without affecting the results. Kernels for such β can be constructed using an orthonormal basis (see Tsybakov, 2009, p. 11).

The intuition for this estimator is that it can be seen as a smooth generalization of the histogram density estimator which uses local information rather than fixed bins. Thus, if we believe the density is smooth, using such a smoothed out version is natural. Another way to see this is to observe that the kernel estimator is the convolution of K with the empirical density function f_n, defined implicitly via $\int_{-\infty}^{x} f_n(y)dy = F_n(x) = \frac{1}{n}\sum I(x_i \le x)$. Using the empirical density itself is an unbiased estimator of the true density (and it satisfies the central limit theorem for fixed d), but by adding bias through the kernel, we may be able to reduce variance, and achieve lower estimation risk for densities which “match” the kernel in a certain way.

In this work, we have chosen, for simplicity, to use isotropic kernels and the isotropic Nikol’skii class of densities. Basically, densities $f \in \mathcal{N}_p(\beta, C)$ have the same degree of smoothness in all directions. The same is true of the kernels which satisfy Definition 2. Allowing anisotropic smoothness is a natural extension, although the notation becomes complicated very quickly. For the anisotropic case under fixed-d asymptotics, see for example Goldenshluger and Lepski (2011).

1.4 Related Work

Density estimation in the minimax framework is a well-studied problem with many meaningful contributions over the last six decades, and we do not pretend to give a complete overview here. Recent advances tend to build on one of four frameworks: (1) the support of f, (2) the smoothness of f, (3) whether the loss is adapted to the nature of the smoothness, and (4) whether the estimator can adapt to different degrees of smoothness. For a comprehensive overview of these and other concerns, an excellent resource is Goldenshluger and Lepski (2014) which presents results for adaptive estimators over classes of varying smoothness when the loss is not necessarily adapted to the smoothness. It also contextualizes and compares existing work. For previous results most similar to those we present in terms of function classes and losses, see Hasminskii and Ibragimov (1990). Other important work is given in Devroye and Györfi (1985); Goldenshluger and Lepski (2011); Juditsky and Lambert-Lacroix (2004); Kerkyacharian et al. (1996).

Unlike in the density estimation setting, there are some related results in the information theory literature which endeavor to address the limits of estimation under the triangular array. Essentially, this work examines the estimation of the joint distribution of a d-block of random variables observed in sequence from an ergodic process supported on a finite set of points. Marton and Shields (1994) show that if d grows like $\log n$, then these joint distributions can be estimated consistently. An extension of these results to the case of a Markov random field embedded in a higher dimension is given by Steif (1997). Our results are slightly slower than these (see Corollary 5), but estimating continuous densities rather than finitely supported distributions is more difficult.

2 MAIN RESULTS

Our main results give non-asymptotic rates for density estimation under growing dimension. It generalizes existing results in that, had d been fixed, we recover the usual rate. Deriving the minimax rate for density estimation requires two components: (1) finding the risk of the best possible estimator for the hardest density in our class and (2) exhibiting an estimator which achieves this risk. Our results are only *rate minimax* in that the upper and lower bounds match in d and n, but constants may be different.

We first present the lower bound. Our proof is given in Section 3.

Theorem 3 (Lower bound for density estimation)**.** *For any* $d \in \mathbb{Z}^+$, $\beta > 1$, $p \in [2, \infty)$, *choose* $n > n^*$ *with*

$$n^* = 64 \left\|\Gamma_0\right\|_2^{-2d} \times \left[\left\|\Gamma_0\right\|_p^{-(d+1)(2\beta+d)} C^{4\beta+d} \left(\frac{\sigma}{\varphi(1/\sigma)}\right)^{d(d+\beta)}\right]^{1/\beta}.$$

Then,

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{N}_p(\beta, C)} \mathbb{E}_f \left[\left(\frac{n^\beta}{d^d} \right)^{\frac{1}{2\beta+d}} \left\| f - \widehat{f} \right\|_p \right] \geq c \left(\frac{1}{8} \right) \frac{C}{2} \frac{\kappa^{-\beta}}{8^{1/p}},$$

for $c(v)$ a function only of v and $\kappa := \frac{\varphi(1/\sigma)}{\sigma \|\Gamma_0\|_2^2}$. The infimum is over all estimators $\widehat{f}$.

This result says that there exists a triangular array $\{f^{(n)}\}$ of densities in $\mathcal{N}_p(\beta, C)$ so that the best risk we can hope to achieve over all possible estimators $\widehat{f}$ is

$$\mathbb{E}_f \left[\left\| f - \widehat{f} \right\|_p \right] = O\left(\left(\frac{d^d}{n^\beta} \right)^{\frac{1}{2\beta+d}} \right).$$

The specific constant κ as well as the minimum n^* are properties of the proof technique, so their forms are not really relevant (except that κ is independent of n and d). Specifically, $\varphi(u) = (2\pi)^{-1/2} e^{-u^2/2}$ is the standard normal density, $\sigma > 0$ is the standard deviation to be chosen, and Γ_0 is a small perturbation we make explicit below. One could make other choices for the "worst case" density which result in different values. We also note that here C is the same constant in each equation (and in the remainder of the paper): it quantifies the smoothness of the class $\mathcal{N}_p(\beta, C)$.

Our second result shows that, for an oracle choice of the bandwidth h, kernel density estimators can achieve this rate. That is, for *any* density in $\mathcal{N}_p(\beta, C)$, the risk of the kernel density estimator is optimal. The proof is given in Section 4.

Theorem 4 (Upper bound for kernels)**.** *Let $f \in \mathcal{N}_p(\beta, C)$. Let $K(u)$ be an isotropic kernel of order $\ell = \lfloor \beta \rfloor$ which satisfies $\int K^2(u) du < \infty$. Take $d \in \mathbb{Z}^+$, $p \in [2, \infty)$. Finally, take $h = A(d^2 n)^{-1/(2\beta+d)}$ for some constant A. Then, for n large enough,*

$$\sup_{f \in \mathcal{N}_p(\beta, C)} \mathbb{E}_f \left[\left\| \widehat{f}_h(x) - f(x) \right\|_p \right] = O\left(\left(\frac{d^d}{n^\beta} \right)^{\frac{1}{2\beta+d}} \right).$$

Our results so far have been finite sample bounds (which nonetheless depend on d and n). However, we also wish to know how quickly d can increase so that the estimation risk can still go to zero asymptotically (estimation consistency). Clearly, to have any hope that kernel density estimators are consistent, d must increase quite slowly with n.

Corollary 5. *If $d = o\left(\frac{\beta \log n}{W(\beta \log n)} \right)$, then*

$$\sup_{f \in \mathcal{N}_p(\beta, C)} \mathbb{E}_f \left[\left\| \widehat{f}_h(x) - f(x) \right\|_p \right] = o(1).$$

Here W is the Lambert W function, implicitly defined as the inverse of $u \mapsto u \exp(u)$. For n large, one can show using a series expansion that $W(\log n) = \log \log n - \log \log \log n + o(1)$. So essentially, we require d to grow just slightly slower than $\log n$, the information theoretic rate for estimating finite distributions with a sample from an ergodic process (see Section 1.4).

While we have stated both main theorems in terms of expectations, analogous high-probability bounds can be derived similarly without extra effort.

3 LOWER BOUND FOR DENSITY ESTIMATION

The technique we use for finding the lower bound is rather standard. The idea is to convert the problem of density estimation into one of hypothesis testing. This proceeds by first noting that the probability that the error exceeds a constant is a lower bound for the risk. We then further reduce this lower bound by searching over only a finite class rather than all possible densities. Finally, we ensure that there are sufficiently many members in this class which are well-separated from each other but difficult to distinguish from the true density. Relative to previous techniques for minimax lower bounds for density estimation, the main difference in our proof is that we must choose different members of our finite class such that they have the right dependence on d. Our construction will make use of the Kullback-Leibler divergence.

Definition 6 (KL divergence)**.** *The Kullback-Leibler divergence between two probability measures P and P' is*

$$KL(P, P') = \begin{cases} \int dP \log \frac{dP}{dP'} & P \ll P' \\ \infty & \text{else.} \end{cases}$$

If both P and P' have Radon-Nikodym derivatives with respect to the same dominating measure μ, then we can replace distributions with densities and integrate with respect to μ. As long as the KL divergence between the true density and the alternatives is small on average, it will be difficult to discriminate between them. Therefore, the probability of falsely rejecting the true density will be large. The following lemma makes the process explicit.

Lemma 7 (Tsybakov 2009)**.** *Let $\mathcal{L} : \mathbb{R}^+ \to \mathbb{R}^+$ which is monotone increasing with $\mathcal{L}(0) = 0$ and $\mathcal{L} \not\equiv 0$, and let $A > 0$ such that $\mathcal{L}(A) > 0$.*

1. *Choose elements $\theta_0, \theta_1, \ldots, \theta_M$, $M \geq 1$ in some class Θ;*
2. *Show that $\rho(\theta_j,\ \theta_k) \geq 2\tau > 0$, $\forall 0 \leq j < k \leq M$ for some semi-distance ρ;*

3. Show that $P_{\theta_j} \ll P_{\theta_0}$, $\forall j = 1, \ldots, M$ and

$$\frac{1}{M}\sum_{j=1}^{M} KL(P_{\theta_j},\ P_{\theta_0}) \leq \alpha \log M,$$

with $0 < \alpha < 1/8$.

Then for $\psi = \tau/A$ we have

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \left[\mathcal{L}(\psi^{-1}\rho(\widehat{\theta},\ \theta))\right] \geq c(\alpha)\mathcal{L}(A),$$

where $\inf_{\widehat{\theta}}$ denotes the infimum over all estimators and $c(\alpha) > 0$ is a constant depending only on α.

To use this result, we first choose a base density f_0 and M alternative densities in $\mathcal{N}_p(\beta, C)$. We then show that these densities are sufficiently well-separated from each other in the L^p-norm, $p \in [2, \infty)$, that is we take $\rho(u, u') = \rho(u-u') = \|u - u'\|_p$. Finally, we show that the KL-divergence between the alternatives and f_0 is uniformly small, and therefore small on average. Our proof will use $\mathcal{L}(u) = u$, though, as discussed following the proof, other choices of monotone increasing functions (e.g. $\mathcal{L}(u) = u^2$) simply modify the conclusion but not the proof.

In order to get the "right" rate, we need to choose a base density and a series of small perturbations to create a large collection of alternatives. Getting the perturbations to be the right size and allow sufficiently many of them is the main trick to derive tight bounds. In our case, it is the choice $\Gamma(\mathbf{u})$ (described below) that has this effect. The multiplicative dependence on d turns out to be the necessary deviation from existing lower bounds. Determining that this is the appropriate modification is an exercise in trial-and-error, and even this seemingly minor one is enough to compel a complete overhaul of the proof.

The densities. Define $f_0(\mathbf{x}) = \frac{1}{\sigma^d}\prod_{i=1}^d \varphi(x_i/\sigma)$ where $\varphi(u)$ is the standard Gaussian density.

Let $\Gamma_0 : \mathbb{R} \to \mathbb{R}^+$ satisfy

(*i*) $|\Gamma_0^{(\ell)}(u) - \Gamma_0^{(\ell)}(u')| \leq |u - u'|^{\beta-\ell}/2$, $\forall u, u',\ \ell \leq \lfloor\beta\rfloor$,

(*ii*) $\Gamma_0 \in C^\infty(\mathbb{R})$,

(*iii*) $\Gamma_0(u) > 0 \Leftrightarrow u \in (-1/2, 1/2)$.

There exist many functions satisfying these conditions: e.g. $\Gamma_0(u) = a^{-1}\frac{d}{du}\exp(-1/(1-4u^2))I(|u| < 1/2)$ for some $a > 0$, since it is infinitely continuously differentiable and $\left\|\Gamma_0^{(s)}\right\|_\infty$ is decreasing in s.

Define $\Gamma(\mathbf{u}) = dC\prod_{i=1}^d \Gamma_0(u_i)$, and for any integer $m > 0$, let

$$\gamma_{m,\mathbf{j}}(\mathbf{x}) = m^{-\beta}\Gamma(m\mathbf{x} - \mathbf{j}), \quad j \in \{1, \ldots, m\}^d.$$

Note that $\gamma_{m,\mathbf{j}}(\mathbf{x}) > 0 \Leftrightarrow \|\mathbf{x}\|_\infty \leq 1$. Finally, take $f_\omega(\mathbf{x}) = f_0(\mathbf{x}) + \sum_{\mathbf{j}} \omega(\mathbf{j})\gamma_{m,\mathbf{j}}(\mathbf{x})$ where for any $\mathbf{j}$, $\omega(\mathbf{j}) \in \{0, 1\}$ so that $\omega = \{\omega(\mathbf{j})\}_{\mathbf{j}}$ is a binary vector in $\mathbb{R}^{(m-1)^d}$.

Now, we show that f_0, f_ω are densities in $\mathcal{N}_p(\beta, C)$. For f_0, this is a density which is infinitely differentiable, so we choose $\sigma > 0$ such that $\left\|f_0^{(s)}(\mathbf{x})\right\|_p \leq C/2$. We also have that for any $\mathbf{j}$, the functions $\gamma_{m,\mathbf{j}}$ are non-zero only on non-intersecting intervals of the form $(0, \ldots, \frac{j_i}{m} \pm \frac{1}{2m}, \ldots, 0)$, so for any $|s| < \beta$,

$$\begin{aligned}
&\left\|\sum_{\mathbf{j}} \omega(\mathbf{j}) \left[\gamma_{m,\mathbf{j}}^{(s)}(\mathbf{x}+\mathbf{t}) - \gamma_{m,\mathbf{j}}^{(s)}(\mathbf{x})\right]\right\|_p \\
&\leq dCm^{-\beta+|s|} \sup_{|z|<t} \left\|\Gamma_0^{(s)}(x+z) - \Gamma_0^{(s)}(x)\right\|_p^d \\
&\leq 2^{-d}dCm^{-\beta+|s|} \sup_{z\in[0,1]} |z|^{d(\beta-|s|)} < C/2,
\end{aligned}$$

$\forall m > 0$, so, f_ω is sufficiently smooth by the triangle inequality. As long as f_ω is a density, we will have $f_\omega \in \mathcal{N}_p(\beta, C)$. First, $\int \Gamma_0 = 0$, so $\int f_\omega = 1$. It remains to show that $f_\omega \geq 0$. We have

$$\left\|\sum_{\mathbf{j}} \omega(\mathbf{j})\gamma_{m,\mathbf{j}}\right\|_\infty \leq m^{-\beta}\|\Gamma\|_\infty \leq dCm^{-\beta}\|\Gamma_0\|_\infty^d. \tag{1}$$

The smallest value taken by f_0 on the interval $[-1, 1]$ where we are adding perturbations is $\inf_{u\in[-1,1]} f_0(u) = (\varphi(1/\sigma)/\sigma)^d$. So, it is sufficient to require (1) to be smaller. Therefore, we require

$$m > \left[dC\left(\frac{\sigma\|\Gamma_0\|_\infty}{\varphi(1/\sigma)}\right)^d\right]^{1/\beta}.$$

Sufficient separation of alternatives. We have for any f_ω, $f_{\omega'}$,

$$\begin{aligned}
\|f_\omega - f_{\omega'}\|_p &= \left\|\sum_{\mathbf{j}} (\omega(\mathbf{j}) - \omega'(\mathbf{j}))\gamma_{m,\mathbf{j}}\right\|_p \\
&= m^{-\beta-d/p}H^{1/p}(\omega, \omega')\|\Gamma\|_p,
\end{aligned}$$

where H is the Hamming distance between binary vectors. We will use some of the f_ω as our collection of M alternatives. But we need to know how many there are in the collection that are far enough apart. The following theorem tells us about the size of such a collection.

Lemma 8 (Varshamov-Gilbert; Tsybakov 2009)**.** *Let $m \geq 8$. Then there is a subset $\mathcal{D}$ of densities f_ω such that for all $\omega, \omega' \in \mathcal{D}$, $H(\omega, \omega') \geq m^d/8$ and $|\mathcal{D}| \geq \exp\{m^d/8\}$.*

We now restrict our collection of densities to be only those corresponding to the set $\mathcal{D}$. Then,

$$\begin{aligned}
&m^{-\beta-d/p} H^{1/p}(\omega,\omega') \|\Gamma\|_p \\
&\geq m^{-\beta-d/p}\left(\frac{m^d}{8}\right)^{1/p} dC \|\Gamma_0\|_p^d \\
&= 8^{-1/p} d m^{-\beta} C \|\Gamma_0\|_p^d .
\end{aligned}$$

Constant likelihood ratio. We have that for distributions P_0 with density f_0 and P_ω with density $f_\omega \in \mathcal{D}$,

$$\begin{aligned}
&KL(P_\omega, P_0) \\
&= n \int_{\mathbb{R}^d} d\mathbf{x} f_\omega(\mathbf{x}) \log \frac{f_\omega(\mathbf{x})}{f_0(\mathbf{x})} \\
&= n \int_{\mathbb{R}^d} d\mathbf{x} \left(\frac{1}{\sigma^d}\prod_{i=1}^d \varphi(x_i/\sigma) + \sum_{\mathbf{j}} \omega(\mathbf{j})\gamma_{m,\mathbf{j}}(\mathbf{x})\right) \\
&\quad \times \left[\log\left(\frac{1}{\sigma^d}\prod_{i=1}^d \varphi(x_i/\sigma) + \sum_{\mathbf{j}} \omega(\mathbf{j})\gamma_{m,\mathbf{j}}(\mathbf{x})\right)\right. \\
&\quad \left. - \log\left(\frac{1}{\sigma^d}\prod_{i=1}^d \varphi(x_i/\sigma)\right)\right] \\
&\leq n \int_{\mathbb{R}^d} d\mathbf{x} \left(\frac{1}{\sigma^d}\prod_{i=1}^d \varphi(x_i/\sigma) + \sum_{\mathbf{j}} \omega(\mathbf{j})\gamma_{m,\mathbf{j}}(\mathbf{x})\right) \\
&\quad \times \left[\frac{\sum_{\mathbf{j}} \omega(\mathbf{j})\gamma_{m,\mathbf{j}}(\mathbf{x})}{\frac{1}{\sigma^d}\prod_{i=1}^d \varphi(x_i/\sigma)}\right] \\
&= \int_{[0,1]^d} d\mathbf{x} \frac{\left(\sum_{\mathbf{j}} \omega(\mathbf{j})\gamma_{m,\mathbf{j}}(\mathbf{x})\right)^2}{\frac{1}{\sigma^d}\prod_{i=1}^d \varphi(x_i/\sigma)} \\
&\leq n \left(\frac{\sigma \|\Gamma_0\|_2^2}{\varphi(1/\sigma)}\right)^d d^2 C^2 m^{-2\beta}
\end{aligned}$$

Therefore, we must choose m so that for n, d, large enough,

$$n \left(\frac{\sigma \|\Gamma_0\|_2^2}{\varphi(1/\sigma)}\right)^d C^2 m^{-2\beta} \leq \alpha \log |\mathcal{D}|$$

with $0 < \alpha < 1/8$. This is equivalent to requiring

$$8 \left(\frac{\sigma \|\Gamma_0\|_2^2}{\varphi(1/\sigma)}\right)^d d^2 C^2 m^{-2\beta-d} \leq \frac{1}{8n}$$

which is equivalent to

$$m \leq \left[\frac{1}{(8C)^2}(d^2 n)\left(\frac{\sigma \|\Gamma_0\|_2^2}{\varphi(1/\sigma)}\right)^d\right]^{\frac{1}{2\beta+d}} .$$

Completing the result. Combining the results of the previous two sections gives us the following lower bound on density estimators in increasing dimensions.

Proof of Theorem 3. Choose an integer $m = \|\Gamma_0\|_p^{(d+1)/\beta} \kappa_d^{-1} (d^2 n)^{1/(2\beta+d)}$ where for convenience we define $\kappa_d := (64C^2)^{1/2\beta+d}\left(\frac{\varphi(1/\sigma)}{\sigma\|\Gamma_0\|_2^2}\right)^{d/(2\beta+d)} \xrightarrow{d\to\infty} \kappa = \frac{\varphi(1/\sigma)}{\sigma\|\Gamma_0\|_2^2}$. Note that $\kappa_d < \kappa$ for all d so $\kappa_d^{-1} > \kappa^{-1}$. Then, we have the following:

1. The functions f_0, f_ω are densities in $\mathcal{N}_p(\beta, C)$ as, for $n > n^*$, $m > \left[dC\left(\frac{\sigma\|\Gamma_0\|_\infty}{\varphi(1/\sigma)}\right)^d\right]^{1/\beta}$.

2. For all $f_\omega,\, f_{\omega'} \in \mathcal{D}$,

$$\begin{aligned}
&\|f_\omega - f_{\omega'}\|_p \geq 8^{-1/p} d m^{-\beta} C \|\Gamma_0\|_p^d \\
&= \frac{dC}{8^{1/p}}\left(\|\Gamma_0\|_p^{(d+1)/\beta} \kappa_d^{-1}(d^2 n)^{1/(2\beta+d)}\right)^{-\beta} \|\Gamma_0\|_p^d \\
&= \frac{2C}{8^{1/p}} \|\Gamma_0\|_p \kappa_d^{-\beta} d^{d/(2\beta+d)} n^{-\beta/(2\beta+d)} \\
&\geq \frac{2C}{2} 8^{-1/p} \kappa^{-\beta} d^{d/(2\beta+d)} n^{-\beta/(2\beta+d)} =: 2A\psi_{nd},
\end{aligned}$$

   where $A = \frac{C}{2} 8^{-1/p}\kappa^{-\beta}$ and $\psi_{nd} = (d^d n^{-\beta})^{1/(2\beta+d)}$.

3. $\frac{1}{M}\sum_{\omega\in\mathcal{D}} KL(P_\omega, P_0) \leq \alpha \log|\mathcal{D}|$ since $\|\Gamma_0\|_p^{(d+1)/\beta} < 1$ for all d, β by construction of Γ_0. Therefore,

$$m \leq \left[\frac{1}{8C^2}(d^2 n)\left(\frac{\sigma\|\Gamma_0\|_2^2}{\varphi(1/\sigma)}\right)^d\right]^{\frac{1}{2\beta+d}} .$$

Therefore, all the conditions of Lemma 7 are satisfied. □

We note that Lemma 7 actually allows more general lower bounds which are immediate consequences of those presented here. In particular, we are free to choose ρ to be other distances than L^p-norms, and we may take powers of those norms or apply other monotone-increasing functions $\mathcal{L}$. For example, this gives the standard lower bound under the mean-squared error. We will not pursue these generalities further here, however, as finding matching upper bounds is often more difficult, requiring specific constructions for each combination $\mathcal{L}$ and ρ. Deriving lower bounds for $1 \leq p < 2$ is also of interest, although this requires more complicated proof techniques. The case of $p = \infty$ is actually a fairly straightforward extension, and we discuss it briefly in Section 5.

4 UPPER BOUND FOR KERNELS

To prove Theorem 4, we first use the triangle inequality to decompose the loss into a bias component and a variance component:

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{f}_h - f\right\|_p\right] \\
&\leq \mathbb{E}\left[\left\|\widehat{f}_h - \mathbb{E}\widehat{f}_h\right\|_p\right] + \left\|\mathbb{E}\widehat{f}_h - f]\right\|_p \\
&=: \mathbb{E}\left[\left(\int |\sigma(x)|^p\right)^{1/p}\right] + \left(\int |b(x)|^p\right)^{1/p}.
\end{aligned}
$$

We now give two lemmas which bound these components separately. For the bias, we will need a well known preliminary result.

Lemma 9 (Minkowski's integral inequality)**.** *Let $(\Omega_1, \Sigma_1, \mu_1)$, $(\Omega_2, \Sigma_2, \mu_2)$ be measure spaces, and let $g : \Omega_1 \times \Omega_2 \to \mathbb{R}$. Then for $p \in [1, \infty]$*

$$
\begin{aligned}
&\left[\int_{\Omega_2} \left|\int_{\Omega_1} g(x_1, x_2) d\mu_1(x_1)\right|^p d\mu_2(x_2)\right]^{1/p} \\
&\leq \int_{\Omega_1} \left[\int_{\Omega_2} |g(x_1, x_2)|^p \, d\mu_2(x_2)\right]^{1/p} d\mu_1(x_1),
\end{aligned}
$$

with appropriate modifications for $p = \infty$.

Lemma 10. *Let $f \in \mathcal{N}_p(\beta, C)$ for $p \in [1, \infty)$ and let K be an isotropic Kernel of order $\ell = \lfloor \beta \rfloor$. Then for all $h > 0$, $d \geq 1$, and $n \geq 1$,*

$$
\int |b(x)|^p dx := \int \left|\mathbb{E}\widehat{f}_h(x) - f(x)\right|^p dx = O\left(d^p h^{p\beta}\right).
$$

For the bias, the proof technique depends on the smoothness of the density f as well as the smoothness of the kernel. It also holds for any $p \in [1, \infty)$.

Proof. By Taylor's theorem

$$
\begin{aligned}
&f(x + uh) = f(x) + \sum_{|s|=1} u^s h D^s f(x) + \cdots + \\
&\frac{h^\ell}{(\ell-1)!} \sum_{|s|=\ell} u^s \int_0^1 (1-\tau)^{\ell-1} D^s f(x + \tau uh) d\tau.
\end{aligned}
$$

Since the kernel is of order ℓ, lower order polynomials in u are equal to 0, so

$$
\begin{aligned}
&|b(x)| \\
&= \Bigg| \int du \Omega_\ell(u) \times \\
&\left[\sum_{|s|=\ell} u^s \int_0^1 d\tau (1-\tau)^{\ell-1} D^s f(x + \tau uh)\right] \Bigg| \\
&= \left| \int du \Omega_\ell(u) \left[\sum_{|s|=\ell} u^s \int_0^1 d\tau (1-\tau)^{\ell-1} \Delta(x, \tau)\right] \right|,
\end{aligned}
$$

where $\Delta(x, \tau) = D^s f(x + \tau uh) - D^s f(x)$ and $\Omega_\ell(u) = K(u)\frac{h^\ell}{(\ell-1)!}$. Now applying Lemma 9 twice,

$$
\begin{aligned}
&\int |b(x)|^p dx \\
&\leq \int dx \Bigg(\int du |\Omega_\ell(u)| \, \|u\|_1^\ell \times \\
&\int_0^1 d\tau (1-\tau)^{\ell-1} |\Delta(x, \tau)| \Bigg)^p \\
&\leq \Bigg(\int du |\Omega_\ell(u)| \, \|u\|_1^\ell \times \\
&\left[\int dx \left(\int_0^1 d\tau (1-\tau)^{\ell-1} |\Delta(x, \tau)|\right)^p\right]^{1/p} \Bigg)^p \\
&\leq \Bigg(\int du |\Omega_\ell(u)| \, \|u\|_1^\ell \times \\
&\int_0^1 d\tau (1-\tau)^{\ell-1} \left(\int dx \Delta(x, \tau)^p\right)^{1/p} \Bigg)^p.
\end{aligned}
$$

Because $f \in \mathcal{N}_p(\beta, C)$, we have

$$
\left(\int dx \Delta(x, \tau)^p\right)^{1/p} \leq C(\tau h \|u\|_1)^{\beta-\ell}.
$$

So,

$$
\begin{aligned}
&\int |b(x)|^p \\
&\leq \Bigg(\int du |\Omega_\ell(u)| \, \|u\|_1^\ell \times \\
&\left[\int_0^1 d\tau (1-\tau)^{\ell-1} C(\tau h \|u\|_1)^{\beta-\ell}\right] \Bigg)^p \\
&= \left(\int du |K(u)| \frac{C \|u\|_1^\beta h^\beta}{(\ell-1)!} \left[\int_0^1 d\tau (1-\tau)^{\ell-1} \tau^{\beta-\ell}\right]\right)^p \\
&= ACh^{p\beta} \left(\sum_{i=1}^d \int |u_i|^\beta |G(u_i)| du_i\right)^p = O(d^p h^{p\beta}).
\end{aligned}
$$

□

Next we find an upper bound on the variance component. This result does not depend on the smoothness of the density, only on properties of the kernel. It does however depend strongly on p. Finally, note that the result is non-random, so we can ignore the outer expectation.

Lemma 11. *Let $K : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying $\int K^2(u) du < \infty$. Then for any $h > 0$, $n \geq 1$ and any probability density f, and $p \geq 1$,*

$$
\begin{aligned}
\int |\sigma(x)|^p dx &= \int \left(\widehat{f}_h(x) - \mathbb{E}\widehat{f}_h(x)\right)^p dx \\
&= O\left(\left(\frac{1}{nh^d}\right)^{p/2}\right).
\end{aligned}
$$

Proof. The proof is an easy generalization of Proposition 1 in (Masaon, 2009) and is omitted. For intuition, we simply present the case of $p = 2$.

$$\begin{aligned}\int |\sigma(x)|^2 dx &\leq \frac{1}{nh^{2d}} \mathbb{E}\left[K^2\left(\frac{X_i - x}{h}\right)\right] \\ &= \frac{1}{nh^{2d}} \int \left[\int f(z) K^2\left(\frac{z-x}{h}\right) dz\right] dx \\ &= \frac{1}{nh^{2d}} \int f(z) \left[\int K^2\left(\frac{z-x}{h}\right) dx\right] dz \\ &= \frac{1}{nh^d} \int K^2(u) du = O\left(1/nh^d\right).\end{aligned}$$

□

With these results in hand, we can now prove Theorem 4.

Proof of Theorem 4. Applying Lemma 10 and Lemma 11 gives

$$\sup_{f \in \mathcal{N}_p(\beta, C)} \mathbb{E}\left[\left\|\widehat{f}_h - f\right\|_p\right] = O\left(dh^\beta\right) + O\left(\left(\frac{1}{nh^d}\right)^{1/2}\right).$$

Taking $h = A(d^2 n)^{-\frac{1}{2\beta+d}}$ balances the terms and gives the result. □

5 DISCUSSION

In this paper we have developed the first results for density estimation under triangular array asymptotics, where both the number of observations n and the ambient dimension d are allowed to increase. Our results generalize existing, fixed-d minimax results, in that, were d fixed rather than increasing, we would recover previously known minimax rates (both lower and upper bounds). Our results also show that kernel density estimators are minimax optimal, which should come as no surprise, since they are minimax optimal for fixed d.

The results presented in this paper say essentially that, for n large enough there exist constants $0 < a < A < \infty$ independent of d, n such that for n large enough,

$$\begin{aligned}a\left(\frac{d^d}{n^\beta}\right)^{\frac{1}{2\beta+d}} &\leq \inf_{\widehat{f}} \sup_{f \in \mathcal{N}_p(\beta, C)} \mathbb{E}\left[\left\|\widehat{f} - f\right\|_p\right] \\ &\leq \sup_{f \in \mathcal{N}_p(\beta, C)} \mathbb{E}\left[\left\|\widehat{f}_h - f\right\|_p\right] \leq A\left(\frac{d^d}{n^\beta}\right)^{\frac{1}{2\beta+d}},\end{aligned}$$

for $p \in [2, \infty)$ when $\widehat{f}_h$ is the kernel density estimator with oracle h. This result generalizes immediately to a result for $\mathbb{E}\left[\left\|\widehat{f} - f\right\|_p^p\right]$. With longer proofs, we can generalize this result to $\mathbb{E}\left[\left\|\widehat{f} - f\right\|_p^s\right]$ for some $s \neq p$ and to the case $p \in [1, 2)$. Another extension is to the case $p = \infty$ which picks up a factor of $\log n$ in the numerator of the rate.

With the same techniques used here, we could also give results for nonparametric regression under triangular array asymptotics. Given pairs (y_i, x_i), kernel regression $g(x)$ can be written in terms of densities as $g(x) = \mathbb{E}[Y \mid X = x] = \int y f(x, y) dy / f(x)$ for joint and marginal densities $f(x, y)$ and $f(x)$ respectively. So results for the Nadaraya-Watson kernel estimator

$$\widehat{g}_h(x) = \frac{\sum_{i=1}^n y_i K((x - x_i)/h)}{\sum_{i=1}^n K((x - x_i)/h)}$$

can be obtained with similar proof techniques to those presented here.

A related extension would consider the problem of conditional density estimation directly. Using a similar form,

$$\widehat{q}_h(x, y) = \frac{\sum_{i=1}^n K_1((y_i - y)/h) K_2((x - x_i)/h)}{\sum_{i=1}^n K_2((x - x_i)/h)}$$

estimates the conditional density $q(Y|X)$. If $X \in \mathbb{R}^d$, this estimator has been shown to converge at a rate of $O(n^{-\beta/(2\beta+1+d)})$ under appropriate smoothness assumptions (see, e.g. Hall et al., 2004).

Our results also suggest some open questions. Wavelet density estimators and projection estimators are known to be rate-minimax for d fixed in that upper bounds match those of kernels in n, though constants may be larger or smaller. Whether these methods also match for increasing d remains to be seen (the class of densities examined is usually slightly different). Histograms are also useful density estimators, and for fixed d, they are minimax over Lipschitz densities with a slower rate than that for kernels, again because the class of allowable densities is different. Upper bounds under the triangular array with a similar form to those presented here were shown in (McDonald et al., 2011, 2015), but deriving minimax lower bounds for this class remains an open problem. Extending our results to the manifold setting (as mentioned in §1) is the most obvious path toward fast rates for large d and is left as future work.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DMS–1407439 and the Institute for New Economic Thinking under Grant No. INO14–00020. The author thanks the anonymous referees and the program committee for their insightful comments and Cosma Shalizi for comments on an early draft.

References

Asta, D. (2013), "Nonparametric density estimation on hyperbolic space," in *Neural Information Processing Systems (NIPS) workshop: Modern Nonparametric Methods in Machine Learning*, eds. A. Gretton, M. Kolar, S. Kpotufe, J. Lafferty, and H. Liu.

Belloni, A., Chernozhukov, V., and Wang, L. (2011), "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, **98**(4), 791–806.

Bhattacharya, A., and Dunson, D. B. (2010), "Nonparametric bayesian density estimation on manifolds with applications to planar shapes," *Biometrika*, **97**(4), 851–855.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, **37**(4), 1705–1732.

Bühlmann, P., and van de Geer, S. (2011), *Statistics for high-dimensional data: Methods, theory and applications*, Springer, New York.

Bullmore, E., and Sporns, O. (2009), "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, **10**(3), 186–198.

Devroye, L., and Györfi, L. (1985), *Nonparametric Density Estimation: The L_1 View*, John Wiley & Sons, Inc., New York.

Donoho, D. L. (2000), "High-dimensional data analysis: The curses and blessings of dimensionality," in *AMS Conference on Math Challenges of the 21st Century*.

Genovese, C., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012a), "Minimax manifold estimation," *Journal of Machine Learning Research*, **13**, 1263–1291.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012b), "Manifold estimation and singular deconvolution under Hausdorff loss," *The Annals of Statistics*, **40**(2), 941–963.

Goldenshluger, A., and Lepski, O. (2011), "Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality," *The Annals of Statistics*, **39**(3), 1608–1632.

Goldenshluger, A., and Lepski, O. (2014), "On adaptive minimax density estimation on R^d," *Probability Theory and Related Fields*, **159**(3), 479–543.

Greenshtein, E., and Ritov, Y. (2004), "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization," *Bernoulli*, **10**(6), 971–988.

Hall, P., Racine, J., and Li, Q. (2004), "Cross-validation and the estimation of conditional probability densities," *Journal of the American Statistical Association*, **99**(468), 1015–1026.

Hasminskii, R., and Ibragimov, I. (1990), "On density estimation in the view of Kolmogorov's ideas in approximation theory," *The Annals of Statistics*, **18**(3), 999–1010.

Hendriks, H. (1990), "Nonparametric estimation of a probability density on a riemannian manifold using fourier expansions," *The Annals of Statistics*, **18**, 832–849.

Juditsky, A., and Lambert-Lacroix, S. (2004), "On minimax density estimation on R," *Bernoulli*, **10**(2), 187–220.

Kerkyacharian, G., Picard, D., and Tribouley, K. (1996), "L^p adaptive density estimation," *Bernoulli*, **2**(3), 229–247.

Marton, K., and Shields, P. C. (1994), "Entropy and the consistent estimation of joint distributions," *Annals of Probability*, **22**, 960–977.

Masaon, D. M. (2009), "Risk bounds for kernel density estimators," *Journal of Mathematical Sciences*, **163**(3), 238–261.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2011), "Estimating β-mixing coefficients," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, eds. G. Gordon, D. Dunson, and M. Dudík, vol. 15, JMLR W&CP.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2015), "Estimating beta-mixing coefficients via histograms," *Electronic Journal of Statistics*, **9**, 2855–2883.

Meinshausen, N. (2007), "Relaxed lasso," *Computational Statistics & Data Analysis*, **52**(1), 374–393.

Nardi, Y., and Rinaldo, A. (2008), "On the asymptotic properties of the group lasso estimator for linear models," *Electronic Journal of Statistics*, **2**, 605–633.

Pelletier, B. (2005), "Kernel density estimation on riemannian manifolds," *Statistics & Probability Letters*, **73**(3), 297–304.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2008), "Spam: Sparse additive models," in *Advances in Neural Information Processing Systems 20*, eds. J. Platt, D. Koller, Y. Singer, and S. Roweis, pp. 1201–1208, MIT Press, Cambridge, MA.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse additive models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(5), 1009–1030.

Steif, J. E. (1997), "Consistent estimation of joint distributions for sufficiently mixing random fields," *Annals of Statistics*, **25**, 293–304.

Talwalkar, A., Kumar, S., and Rowley, H. (2008), "Large-scale manifold learning," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008*, IEEE.

Tsybakov, A. (2009), *Introduction to Nonparametric Estimation*, Springer Verlag.

Ye, F., and Zhang, C. (2010), "Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls," *The Journal of Machine Learning Research*, **11**, 3519–3540.

Yuan, M., and Zhou, D.-X. (2015), "Minimax optimal rates of estimation in high dimensional additive models: Universal phase transition," .