The Annals of Statistics
2020, Vol. 48, No. 6, 3228–3250
https://doi.org/10.1214/19-AOS1927

OPTIMAL RATES OF ENTROPY ESTIMATION OVER LIPSCHITZ BALLS

BY YANJUN HAN[1,*], JIANTAO JIAO[2], TSACHY WEISSMAN[1,†] AND YIHONG WU[3]

[1]*Department of Electrical Engineering, Stanford University,* [*]*yjhan@stanford.edu;* [†]*tsachy@stanford.edu*
[2]*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, jiantao@eecs.berkeley.edu*
[3]*Department of Statistics and Data Science, Yale University, yihong.wu@yale.edu*

We consider the problem of minimax estimation of the entropy of a density over Lipschitz balls. Dropping the usual assumption that the density is bounded away from zero, we obtain the minimax rates $(n \ln n)^{-s/(s+d)} + n^{-1/2}$ for $0 < s \le 2$ for densities supported on $[0, 1]^d$, where s is the smoothness parameter and n is the number of independent samples. We generalize the results to densities with unbounded support: given an Orlicz functions Ψ of rapid growth (such as the subexponential and sub-Gaussian classes), the minimax rates for densities with bounded Ψ-Orlicz norm increase to $(n \ln n)^{-s/(s+d)}(\Psi^{-1}(n))^{d(1-d/p(s+d))} + n^{-1/2}$, where p is the norm parameter in the Lipschitz ball. We also show that the integral-form plug-in estimators with kernel density estimates fail to achieve the minimax rates, and characterize their worst case performances over the Lipschitz ball.

One of the key steps in analyzing the bias relies on a novel application of the Hardy–Littlewood maximal inequality, which also leads to a new inequality on the Fisher information that may be of independent interest.

1. Introduction. Estimation of functionals of data generating distributions is a fundamental problem in statistics. While this problem is relatively well understood in finite dimensional parametric models [3, 63], the corresponding nonparametric counterparts are often much more challenging and have attracted tremendous interest over the last two decades. Initial efforts have focused on inference of linear, quadratic and cubic functionals in Gaussian white noise and density models and have laid the foundation for the ensuing research. We do not attempt to survey the extensive literature in this area, but instead refer to the interested reader to, for example, [4, 5, 7, 8, 14, 16, 22, 36, 40, 45, 58] and the references therein.

The monograph by [45] provides a general treatment of estimating smooth functionals and discusses cases where efficient parametric rate of estimation is possible. Recently, there has been progress toward the understanding of more complex nonparametric functionals over substantially more general observational models. These include causal effect functionals in observational studies and mean functionals in missing data models. For more details, we refer to [44, 52, 53], which considers a general recipe to yield minimax estimation of a large class of nonparametric functionals common in statistical literature. However, among the class of nonparametric functionals considered in literature, most of the research endeavors, at least from the point of view of minimax optimality, have focused on "smooth functionals" (see [52] for a discussion on general classes of "smooth functionals").

In contrast, the results on optimal estimation of nonsmooth functionals have been less comprehensive [13, 28, 37]. Notably, the seminal papers of [41] and [9] considered the estimating of L_r-norms in Gaussian mean models. Subsequently, significant progress has been made on testing and estimation of nonsmooth functionals, such as the Shannon entropy, support size, total variation and Kullback–Leibler (KL) divergence, for discrete distributions on large domains (see, e.g., [6, 25, 32, 33, 47, 61, 65, 66]).

Received October 2018; revised November 2019.
MSC2020 subject classifications. Primary 62G05; secondary 62C20.
Key words and phrases. Nonparametrics, nonsmooth functional estimation, minimax rate, polynomial approximation.

An important nonsmooth functional of probability density function is the *entropy*, which has been the subject of extensive studies. The main goal of this paper is to resolve the minimax rates of entropy estimation in the density model under smoothness constraints, specifically, over Lipschitz classes. To this end, consider the following i.i.d. sampling model:

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f,$$

where f is a probability density function on $\mathbb{R}^d$. The goal is to estimate the entropy (also known as the differential entropy in the information theory literature) of the density f:

$$H(f) \triangleq \int_{\mathbb{R}^d} -f(x) \ln f(x)\, dx.$$

This problem has extensive applications in various fields such as information theory, neuroscience, time series and machine learning (cf. [10, 29, 38] and the survey [1, 64]).

A prevalent assumption in nonparametric entropy estimation is that $f(x) \geq c$ everywhere for some constant $c > 0$ [21, 34, 35, 55, 62], while others impose various assumptions quantifying on average how close the density is to zero [11, 15, 17, 18, 23, 42, 54, 60]. Assuming the density is bounded away from zero makes entropy a *smooth* functional, consequently, the general technique for estimating smooth nonparametric functionals [44, 52, 53] can be directly applied to achieve the minimax rate $\Theta(n^{-4s/(4s+d)} + n^{-1/2})$.

It is well known that smoothness conditions or shape restrictions are often necessary for nonparametric problems. We allow the density to be arbitrarily close to zero and adopt Lipschitz ball $\mathrm{Lip}_{s,p,d}(L)$ smoothness assumptions. Assume smoothness parameter $s > 0$, norm parameter $p \in [2, \infty)$ and dimensionality $d \in \mathbb{N} \triangleq \{1, 2, \ldots\}$. The Lipschitz ball is defined as

$$\mathrm{Lip}_{s,p,d}(L) \triangleq \big\{f : \|f\|_{\mathrm{Lip}_{s,p,d}} \leq L\big\} \cap \big\{f : \mathsf{supp}(f) \subseteq [0,1]^d\big\}, \tag{1.1}$$

where with $r \triangleq \lceil s \rceil$, the Lipschitz norm $\|\cdot\|_{\mathrm{Lip}_{s,p,d}}$ is defined as

$$\|f\|_{\mathrm{Lip}_{s,p,d}} \triangleq \|f\|_p + \sup_{t>0} t^{-s} \omega_r(f,t)_p, \tag{1.2}$$

$$\omega_r(f,t)_p \triangleq \sup_{e \in \mathbb{R}^d, |e| \leq 1} \big\|\Delta_{te}^r f(\cdot)\big\|_p, \tag{1.3}$$

$$\Delta_h^r f(x) \triangleq \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} f\left(x + \left(k - \frac{r}{2}\right)h\right), \quad h \in \mathbb{R}^d. \tag{1.4}$$

Here, $|x|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^d$ and $\|\cdot\|_p$ denotes the L_p norm of measurable functions on $\mathbb{R}^d$. Note that the L_p norm in (1.3) is taken over the whole space $\mathbb{R}^d$ to ensure that the density f vanishes smoothly at the boundary. For example, any density whose derivatives up to order $\lceil s \rceil - 1$ all vanish at the boundary of $[0,1]^d$ suffices.

We characterize the minimax rates of estimating $H(f)$ over the Lipschitz ball $\mathrm{Lip}_{s,p,d}(L)$ in the following theorem.

THEOREM 1 (Compactly supported densities). *For any $d \in \mathbb{N}$, $0 < s \leq 2$ and $2 \leq p < \infty$, there exist constants $L_0 > 1$ and $c, C > 0$ depending on s, p, d, such that for any $L_0 \leq L \leq (n \ln n)^{s/d}$ and any $n \in \mathbb{N}$,*

$$\begin{aligned} c\big((n\ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L\big) &\leq \Big(\inf_{\hat{H}} \sup_{f \in \mathrm{Lip}_{s,p,d}(L)} \mathbb{E}_f\big(\hat{H} - H(f)\big)^2\Big)^{\frac{1}{2}} \\ &\leq C\big((n\ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L\big). \end{aligned} \tag{1.5}$$

Moreover, the lower bound part of (1.5) holds for any $s > 0$, $1 \leq p < \infty$.

REMARK 1. A careful inspection of the proof of Theorem 1 reveals that, for $s \in (0, 2]$, $p \geq 2$ and $L_0 \leq L \leq L' \leq (n \ln n)^{s/d}$, the minimax L_2 risk for entropy estimation over densities supported on $[0, 1]^d$ with $\|f\|_p \leq L$ and $\sup_{t>0} t^{-s}\omega_{\lceil s \rceil}(f, t)_p \leq L'$ is

$$\Theta\big((n \ln n)^{-\frac{s}{s+d}} (L')^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L\big). \tag{1.6}$$

Hence, by scaling,[1] if the density is supported on $[0, R]^d$ with $R \geq 1$ and satisfies $\|f\|_{\mathrm{Lip}_{s,p,d}} \leq L$ with $R^{d(1-1/p)} L \geq L_0$ and $R^{s+d(1-1/p)} L \leq (n \ln n)^{s/d}$, the minimax L_2 risk is

$$\Theta\big((n \ln n)^{-\frac{s}{s+d}} \big(R^{s+d(1-1/p)} L\big)^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln\big(R^{d(1-1/p)} L\big)\big). \tag{1.7}$$

REMARK 2. A direct consequence of Theorem 1 is that, for fixed parameters $s > 0$, $p \in [2, \infty)$ and $L > L_0$, when $d = 1, 2$, the parametric rate $\Theta(n^{-1/2})$ is attainable for entropy estimation over the Lipschitz ball $\mathrm{Lip}_{s,p,d}(L)$ if and only if $s \geq d$. Moreover, when $d \geq 3$, the parametric rate cannot be attained for all $s < d$.

To the best of our knowledge, Theorem 1 is the first characterization of the minimax rate for nonparametric entropy estimation in arbitrary dimensions over Lipschitz balls (or even the simpler Hölder balls) without assuming the density is bounded away from zero. One observes that the exponents of n or L in the minimax rates (1.5) and (1.6) do not depend on the norm parameter p under the assumption that $2 \leq p < \infty$. Another observation from Remark 2 is that the level of smoothness required for the parametric rate is $s \geq d$, which is more than $s \geq d/4$ that suffices for densities bounded away from zero on the support $[0, 1]^d$ [40], and also more than $s \geq d/2$ that suffices for densities satisfying a relative version of Hölder smoothness [2].

We construct the minimax rate-optimal estimator by first approximating the density f by f_h (a locally smoothed version of f), and then designing estimators to estimate $H(f_h)$. The key advantage of estimating $H(f_h)$ over estimating $H(f)$ is that for each $x \in [0, 1]^d$ and positive integer $k \leq \ln n$, the kth power of $f_h(x)$ admits an unbiased estimator using a U-statistic, which enables us to employ the techniques of best polynomial approximation and Taylor expansion to reduce the bias in estimating $H(f_h)$. Moreover, our estimator is directly constructed and proved for the density model rather than the Poissonized model, unlike most prior work based on polynomial approximation [33, 65].

We improve the best known minimax lower bound for estimating nonsmooth nonparametric functionals. The well-known lower bound $\Theta(n^{-4s/(4s+d)} + n^{-1/2})$ [5], which is optimal for smooth functionals such as quadratic functionals, is loose for entropy estimation. Instead, we reduce the nonparametric problem into a parametric submodel, and construct lower bound via the duality between moment matching and best approximation using rational functions.

In addition to compactly supported densities, Theorem 1 can be extended to densities supported on $\mathbb{R}^d$ with general tail conditions. Let $\Psi : [0, \infty] \to [0, \infty]$ be an Orlicz function, that is, a continuous, increasing and convex function Ψ satisfying $\Psi(0) = 0$, $\Psi(u) > 0$ for any $u > 0$ and $\lim_{u\to\infty} \Psi(u) = \infty$. Moreover, we say Ψ is of *rapid growth* if there is a constant $\kappa = \kappa(\Psi) > 1$ such that $\Psi(\kappa u) \geq \Psi(u)^2$ holds for all $u \geq 0$. Examples of rapidly growing Orlicz functions include $\Psi_q(u) = \exp(u^q) - 1$ for any $q \geq 1$, with $\kappa(\Psi_q) = 2^{1/q}$; in particular, the cases of $q = 1$ and $q = 2$ correspond to the subexponential and sub-Gaussian class, respectively. Consider the following class of densities:

$$\mathrm{Lip}^{\Psi}_{s,p,d}(L) \triangleq \big\{f : \|f\|_{\mathrm{Lip}_{s,p,d}} \leq L\big\} \cap \left\{f : \int_{\mathbb{R}^d} \Psi\big(|x|\big) f(x)\, dx \leq L\right\}, \tag{1.8}$$

[1] Let $\tilde{f}(x) \triangleq R^d f(Rx)$ denote the density of X_i/R. Then $H(\tilde{f}) = H(f) - d \log R$, $\|\tilde{f}\|_p = R^{d(1-1/p)} \|f\|_p$, $\Delta_h \tilde{f}(x) = R^d \Delta^r_{Rh} f(Rx)$, and hence $\sup_{t>0} t^{-s}\omega_r(\tilde{f}, t)_p = R^{d(1-1/p)+s} \sup_{t>0} t^{-s}\omega_r(f, t)_p$.

where $\|\cdot\|_{\mathrm{Lip}_{s,p,d}}$ is the Lipschitz norm defined in (1.2). Note that the second constraint of (1.8) implies that the Ψ-Orlicz norm of the random variable $|X|$ with $X \sim f$ is upper bounded by L.

The following theorem presents the minimax rate for entropy estimation over $\mathrm{Lip}^{\Psi}_{s,p,d}(L)$.

THEOREM 2 (Densities with unbounded support). *Let Ψ be an Orlicz function of rapid growth and Ψ^{-1} its inverse function. For any $d \in \mathbb{N}$, $0 < s \le 2$, $2 \le p < \infty$, there exist constants $c, C, L_0 > 0$ depending on s, p, d, $\kappa(\Psi)$, $\Psi(1)$ such that if $\Psi^{-1}(n) \ge 1$ and $L \ge L_0$, then*

$$c\big((n\ln n)^{-\frac{s}{s+d}}\big[\Psi^{-1}(n)\big]^{d(1-\frac{d}{p(s+d)})} + n^{-\frac{1}{2}}\big) \le \Big(\inf_{\hat{H}} \sup_{f\in \mathrm{Lip}^{\Psi}_{s,p,d}(L)} \mathbb{E}_f\big(\hat{H} - H(f)\big)^2\Big)^{\frac{1}{2}} \le C\big((n\ln n)^{-\frac{s}{s+d}}\big[\Psi^{-1}(n)\big]^{d(1-\frac{d}{p(s+d)})} + n^{-\frac{1}{2}}\big).$$

Moreover, the minimax lower bound works for any $s > 0$, $1 \le p < \infty$.

Comparing Theorem 2 with Remark 1, we see that for general Orlicz function Ψ with rapid growth, any density in $\mathrm{Lip}^{\Psi}_{s,p,d}(L)$ is effectively supported on $[-\Psi^{-1}(n), \Psi^{-1}(n)]^d$. There is also a subtle difference: the hidden constant in the parametric rate $\Theta(n^{-1/2})$ does not involve $\Psi^{-1}(n)$, thanks to the Orlicz norm constraint. Note that for simplicity we assume that L is a constant and omit the dependence on L in Theorem 2.

The estimator that achieves the minimax rates in Theorems 1 and 2 relies on polynomial approximation. It is a natural question to ask whether an integral-form[2] plug-in estimator using kernel density estimate can achieve the minimax rates. Recall that the kernel density estimator takes the form

$$\hat{f}_h(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \tag{1.9}$$

where $K(\cdot)$ is a kernel function, and h is the bandwidth. The next result shows that the answer is negative for any sliding window kernel density estimator with a spatially invariant bandwidth (i.e., the bandwidth h can depend on the sample size n but not on the location x).

THEOREM 3 (Suboptimality of integral-form plug-in estimators). *For $s \in (0, 2]$, $p \ge 2$, let $\hat{f}_h(x)$ be given in (1.9) and define the integral-form plug-in estimator as $H(\hat{f}_h) = \int_{[0,1]^d} -\hat{f}_h(x)\ln\hat{f}_h(x)\,dx$. If the kernel $K(\cdot)$ satisfies Assumption 1 and $h \asymp (Ln)^{-1/(s+d)}$, then for $L \le n^{s/d}$,*

$$\Big[\sup_{f\in \mathrm{Lip}_{s,p,d}(L)} \mathbb{E}_f\big(H(\hat{f}_h) - H(f)\big)^2\Big]^{\frac{1}{2}} \le C\big(n^{-\frac{s}{s+d}}L^{\frac{d}{s+d}} + n^{-\frac{1}{2}}\ln L\big),$$

where $C > 0$ is a constant independent of n, L.

Conversely, for any kernel $K(\cdot)$ satisfying Assumption 1 and any bandwidth $h > 0$, there exist constants $L_0 > 0$, $c > 0$ independent of n, L, h, such that for any $L \ge L_0$,

$$\Big[\sup_{f\in \mathrm{Lip}_{s,p,d}(L)} \mathbb{E}_f\big(H(\hat{f}_h) - H(f)\big)^2\Big]^{\frac{1}{2}} \ge c\big(n^{-\frac{s}{s+d}}L^{\frac{d}{s+d}} + n^{-\frac{1}{2}}\ln L\big).$$

[2]Given a density estimate $\hat{f}$, an integral-form plug-in estimator for the entropy is $\int -\hat{f}(x)\ln\hat{f}(x)\,dx$, as opposed to $\frac{1}{n}\sum_{i=1}^{n}\log\hat{f}(X_i)$.

Theorem 3 presents a tight characterization of the integral-form plug-in approach, and shows that the plug-in idea applied to the integral is strictly suboptimal: the bias of the kernel-based plug-in estimator is $O(n^{-s/(s+d)}L^{d/(s+d)})$, while for the optimal estimator it is $O((n\ln n)^{-s/(s+d)}L^{d/(s+d)})$.

Next, we elaborate on the various assumptions in Theorem 1:

The Lipschitz ball $\mathrm{Lip}_{s,p,d}(L)$. For $s = r + \alpha$ with r integer and $\alpha \in (0, 1]$, the Hölder ball $\mathrm{H}^s_d(L)$ with smoothness s and radius L consists of functions f with $\sup_{x\neq y} |f^{(r)}(x) - f^{(r)}(y)|/|x - y|^\alpha \le L$. The Lipschitz ball is a generalization of the Hölder ball by imposing the smoothness constraint *on average* through the norm parameter p; for example, for $p = \infty$ the Lipschitz ball coincides with the Hölder $\mathrm{Lip}_{s,\infty,d}(L) = \mathrm{H}^s_d(L)$ for any noninteger $s > 0$.[3]

Radius of the Lipschitz ball. The assumption $L \le (n\ln n)^{s/d}$ ensures that the minimax rate in Theorem 1 is $O(1)$, and $L \ge L_0$ is not superfluous as well. Indeed, if $sp \ge d$, then by standard embedding results of Lipschitz (or Besov) spaces [31], there exists $L_1 = L_1(s, p, d) > 0$ such that any $f \in \mathrm{Lip}_{s,p,d}(L_1)$ is bounded from below by a positive constant almost everywhere,[4] which, in view of the previous results [44, 52, 53], implies that the entropy can be estimated at a faster rate $\Theta(n^{-4s/(4s+d)} + n^{-1/2})$ than that in Theorem 1.

The smoothness condition $s \in (0, 2]$. Capturing high-order smoothness $s > 2$ of a function is often challenging in nonparametric statistics, especially for density models. For example, if one would like to apply a kernel density estimator, for $s > 2$ there does not exist a nonnegative kernel to keep all polynomials with degree at most $\lfloor s\rfloor$. We will discuss this phenomenon in details in Section 4.2. We note that the minimax lower bound $\Omega((n\ln n)^{-s/(s+d)}L^{d/(s+d)} + n^{-1/2}\ln L)$ only requires $0 < s < \infty$, $1 \le p < \infty$.

The norm condition $p \in [2, \infty)$. Our current upper bound requires $p \ge 2$, which ensures the difference between the entropy of the true density and its kernel-smoothed version is at the right order. For the lower bound, the case of $p = \infty$ imposes a too strict constraint on the density (i.e., to be smooth everywhere), while $p < \infty$ only imposes an average-case smoothness constraint which can be handled by the current construction. When $p = \infty$, we prove a lower bound of $\Omega(n^{-s/(s+d)}(\ln n)^{-(s+2d)/(s+d)}L^{d/(s+d)} + n^{-1/2}\ln L)$ as shown in Theorem 7.

The support of f. For general nonparametric functional estimation problem, there are essentially three factors contributing to the minimax rates: the tail behavior if f is supported on $\mathbb{R}^d$, the boundary behavior if f is compactly supported, and the behavior of f in the interior of its domain. In Theorem 1, we assume that f is compactly supported and smoothly vanishing at the boundary so that sliding window kernel methods are applicable; this assumption is relaxed in Section 4.1 to the so-called "periodic boundary condition" [39]. The effect of the tail behavior on the minimax rates is precisely quantified in Theorem 2 for densities with unbounded support.

[3] As opposed to the definition of the Lipschitz ball in (1.2), there is another slightly different definition using the modified Lipschitz norm $\mathrm{Lip}^*_{s,p,d}$, which coincides with a special case of the Besov ball $\mathrm{B}^s_{p,\infty,d}$ [12]. These two definitions are equivalent for noninteger s, while for integer s the latter is strictly bigger. In this paper, we adopt the former definition in (1.2) to avoid some technical subtleties.

[4] In fact, [31], Theorem 2.1, states that if $sp \ge d$, $\|f(\cdot) - f(\cdot - t)\|_\infty \le C(s, p, d)\|f(\cdot) - f(\cdot - t)\|_{\mathrm{Lip}_{s,p,d}}$ for any $t \in [0, 1]^d$. Since there must be some $x_0 \in [0, 1]^d$ such that $f(x_0) \ge 1$, we conclude that $f(x) \ge 1 - 2C(s, p, d)L$ for almost every $x \in [0, 1]^d$, which is bounded from below by a constant if L is sufficiently small.

1.1. *Related work.* The problem of estimating the entropy of a density has been investigated extensively in the literature. As discussed in the overview [1], there exist two main approaches, based on either kernel density estimators, for example, [20, 23, 34, 35, 48] or nearest neighbor methods, for example, [2, 11, 18, 54, 55, 60]. Among these works, some focus on the consistency [20, 48], $\sqrt{n}$-consistency [34, 60], or the asymptotic efficiency [2, 23] of the proposed estimator, while others work on the minimax rate [11, 18, 35, 54, 55].

Similar estimator constructions have appeared in the literature. Asymptotic efficient estimators are obtained in [30, 45] for smooth functionals by means of Taylor expansion; [41] and [9] estimated the L_1 norm of the mean in Gaussian white noise model using trigonometric polynomial approximation. One related work [24] deserves special attention. Dealing with the Gaussian white noise model, [24] analyzed the minimax rates of estimating the L_r norms (for all $r \in [1, \infty)$) of the mean function over Besov spaces which was previously studied in [41]. Although both papers use the polynomial approximation technique for the upper and lower bound construction (which trace back to earlier work of [9, 33, 41, 65]), there exist significant distinctions between this work and [24]. First, here we analyze the density model as opposed to the location model, and it is crucial to design estimators to adapt to low-density regions. This specific problem has been investigated in [49] for estimating linear functionals (density at a given point), where it was conjectured that the case of $s > 2$ exhibits significantly different behavior from the case of $0 < s \leq 2$; this is the underlying reason for the assumption $0 < s \leq 2$ for our upper bound, which is discussed in more details in Section 4.2. In contract, in white noise models there is no need to adapt. Moreover, when $d = 1$ these two models are asymptotically equivalent [46] for $s > 1/2$ provided that the density is bounded from below by a positive constant; however, they do *not* imply the minimax rates of a given estimation problem for these two models must coincide, and for small densities the equivalence can break down [50]. In fact, in contrast to the conclusion of Remark 2, it is shown in [24] that the parametric rate is never achievable for "entropy" estimation in the white noise model. Second, the estimator construction in this paper requires more delicate analysis, and bounding the approximation error $H(f_h) - H(f)$ relies on a novel application of the Hardy–Littlewood maximal inequality in conjunction with the nonnegativity of the density function, which also leads to, as a by-product, a new inequality upper bounding the Fisher information in terms of the L_p norm of the second derivative (Theorem 5). Third, in the minimax lower bound, this work carefully chooses nonnegative functions (not required in the Gaussian white noise model), and analyzes the total variation bound instead of the χ^2-divergence bound which is simpler and more suitable for Gaussian models.

1.2. *Notation.* For a finite set A, let $|A|$ denote its cardinality. The norm $|\cdot|$ denotes the Euclidean norm of vectors in $\mathbb{R}^d$, and $\|\cdot\|_p$ denotes the L_p norm (with respect to the Lebesgue measure) of real-valued functions defined on $\mathbb{R}^d$. Let $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm of matrices, that is, the largest singular value. For $x \in \mathbb{R}^d$, let $x_{\setminus i} \triangleq (x_j : j \neq i) \in \mathbb{R}^{d-1}$. For $n \in \mathbb{N}$, let $[n] \triangleq \{1, \ldots, n\}$. Denote by $\binom{[n]}{l} = \{J \subseteq [n] : |J| = l\}$ the collection of all l-subsets of $[n]$. Throughout the paper, for nonnegative sequences $\{a_\gamma\}$ and $\{b_\gamma\}$, we write $a_\gamma \lesssim b_\gamma$ (or $a_n = O(b_n)$) if $a_\gamma \leq C b_\gamma$ for some positive constant C that does *not* depend on the sample size n, the bandwidth h, or the Lipschitz norm L. We use $a_\gamma \gtrsim b_\gamma$ (or $a_\gamma = \Omega(b_\gamma)$) to denote $b_\gamma \lesssim a_\gamma$, and $a_\gamma \asymp b_\gamma$ (or $a_\gamma = \Theta(b_\gamma)$) to denote both $a_\gamma \lesssim b_\gamma$ and $b_\gamma \lesssim a_\gamma$. We use $a_\gamma \ll b_\gamma$ (or $a_\gamma = o(b_\gamma)$) to denote $\lim_\gamma \frac{a_\gamma}{b_\gamma} = 0$, and $a_\gamma \gg b_\gamma$ (or $a_\gamma = \omega(b_\gamma)$) to denote $b_\gamma \ll a_\gamma$. The support set of a probability measure μ is denoted by $\mathsf{supp}(\mu)$. Let P_X denote the distribution of a random variable X. The KL (resp., χ^2) divergence from distribution μ to ν is defined as $D(\mu\|\nu) = \int d\mu \log \frac{d\mu}{d\nu}$ (resp., $\chi^2(\mu\|\nu) = \int d\nu (\frac{d\mu}{d\nu} - 1)^2$) if $\mu \ll \nu$ and $+\infty$ otherwise.

1.3. *Organization.* The rest of this paper is organized as follows. Section 2 presents the construction of the minimax rate-optimal estimator. Section 3 proves the upper bound. In particular, the analysis of the bias incurred by the first-stage approximation relies on a novel application of the Hardy–Littlewood maximal inequality, and the same argument also leads to an inequality on Fisher information, which is presented at the end of Section 3.1 and might be of independent interest. Section 4 discusses generalizations and open problems. In particular, Section 4.1 extends the results to a broader class of densities that satisfy a periodic boundary conditions, and establishes the corresponding minimax rates of entropy estimation. Remaining proofs are relegated to the Appendices in the Supplementary Material [26].

2. Construction of the estimator. Define the smoothed density

$$f_h(x) \triangleq \int_{\mathbb{R}^d} K_h(x-y) f(y)\, dy,$$

where $K_h(\cdot)$ is some kernel function with bandwidth $h > 0$. In the special case of $K_h(x) = \frac{1}{h^d} K(\frac{x}{h})$ for some kernel function $K : \mathbb{R}^d \to \mathbb{R}$, we have

$$f_h(x) = \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x-y}{h}\right) f(y)\, dy, \tag{2.1}$$

which admits the following natural unbiased estimator (kernel density estimate):

$$\hat{f}_h(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{2.2}$$

where $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f$.

The optimal estimator for the entropy $H(f)$ is constructed in two steps: First, by choosing a suitable (in particular, compactly-supported) kernel K, we approximate f by f_h and bound $|H(f_h) - H(f)|$ using functional-analytic properties of the density class. Next, we construct an estimator for $H(f_h)$ based on the kernel density estimator $\hat{f}_h$. The main insight is that $\{\hat{f}_h(x) : x \in [0,1]^d\}$ is essentially a *finite-dimensional* parametric model, in the sense that $\hat{f}_h(x)$ roughly follows the binomial distribution $nh^d \hat{f}_h(x) \sim \mathsf{B}(n, h^d f_h(x))$ (cf. Lemma 5). As a result, we essentially obtain a parametric binomial model with h^{-d} parameters, so that the existing approximation-theoretic techniques for entropy estimation in parametric models [33, 65] can be applied.

We now describe the construction of the optimal entropy estimator for any $s \in (0, 2]$ in any dimension. For the first approximation stage, in order to find a suitable approximation f_h for f, we recall the following property of Lipschitz spaces [27], Theorem 8.1.

LEMMA 1. *Fix any $s > 0$ and any kernel $K : \mathbb{R}^d \to \mathbb{R}$ which satisfies $\int_{\mathbb{R}^d} |x|^{\lceil s \rceil} \times |K(x)|\, dx < \infty$ and maps any polynomial q in d variables of degree at most $\lceil s \rceil - 1$ to themselves, that is, $\int_{\mathbb{R}^d} q(y) K(x-y)\, dy = q(x)$. Then for any $f \in \mathrm{Lip}_{s,p,d}(L)$ and f_h defined in (2.1), we have*

$$\|f_h - f\|_p \triangleq \left(\int_{\mathbb{R}^d} |f_h(x) - f(x)|^p\, dx \right)^{1/p} \lesssim L h^s.$$

To apply Lemma 1, we choose a kernel K with the following properties.

ASSUMPTION 1. Suppose $K : \mathbb{R}^d \to \mathbb{R}$ satisfies the following:

1. Nonnegativity: $K(t) \geq 0$ for any $t \in \mathbb{R}^d$;
2. Unit total mass: $\int_{\mathbb{R}^d} K(t)\, dt = 1$;

3. Zero mean: $\int_{\mathbb{R}^d} t K(t)\,dt = 0$;
4. Finite second moment: $\int_{\mathbb{R}^d} |t|^2 K(t)\,dt < \infty$.
5. Compact support: $\sup\{|t| : K(t) \neq 0\} < \infty$.

There are several kernel functions which fulfill Assumption 1, for example, the box kernel $K(t) = \mathbb{1}(t \in [-1/2, 1/2]^d)$. Note that the second and the third requirements ensure that it keeps all polynomials of degree at most one, that is, $\int_{\mathbb{R}^d} (a^\top x + b) K(x - y)\,dx = a^\top y + b$, and the first requirement (nonnegativity) is crucial for proving the concentration result in Section 3. In fact, the nonnegativity requirement is the key reason why we need to impose the assumption $s \le 2$, and relaxing this requirement appears highly challenging (cf. Section 4.2).

Since the kernel $K(\cdot)$ has a compact support, the approximation f_h is compactly supported as well. By Lemma 1 and our assumption that $s \le 2$, we have

$$\|f - f_h\|_p \lesssim L h^s. \tag{2.3}$$

Later in Section 3.1 we will show that the entropy difference also satisfies $|H(f) - H(f_h)| \lesssim L h^s$.

Fix an appropriate kernel K that fulfills Assumption 1 and define f_h, $\hat{f}_h$ as in (2.1) and (2.2). To construct an estimator $\hat{H}$ for $H(f_h) = \int -f_h(x) \ln f_h(x)\,dx$, we let $\hat{H} = \int \hat{H}(x)\,dx$, where for each $x \in [0,1]^d$, $\hat{H}(x)$ is an estimator for $-f_h(x) \ln f_h(x)$ obtained as follows:

1. For notational convenience, let the sample size be $3n$ as opposed to n. Split the observations into three parts $X^{(1)}$, $X^{(2)}$, $X^{(3)}$, each consisting of n observations.
2. For each part of observations, construct the kernel density estimators $\hat{f}_{h,1}(x)$, $\hat{f}_{h,2}(x)$ and $\hat{f}_{h,3}(x)$ per (2.2). The estimator $\hat{f}_{h,1}(x)$ will be used for classifying smooth versus nonsmooth regime, and the other two for estimation.
3. Regime classification and estimator construction:

- "Nonsmooth" regime: $\hat{f}_{h,1}(x) < \frac{c_1 \ln n}{nh^d}$. Denote by Q the best degree-k polynomial approximation of $-t \ln t$ on $[0, \frac{2c_1 \ln n}{nh^d}]$:

$$Q = \sum_{l=0}^{k} a_l t^l = \arg \min_{P \in \mathsf{Poly}_k} \max_{t \in [0, \frac{2c_1 \ln n}{nh^d}]} |-t \ln t - P(t)|, \tag{2.4}$$

  where Poly_k denotes the collection of all polynomials of degree at most k. Define the following unbiased estimator of $Q(f_h(x))$ in terms of U-statistics:

$$\hat{H}_1(x) = \sum_{l=0}^{k} a_l \Bigg(\binom{n}{l}^{-1} \sum_{J \in \binom{[n]}{l}} \prod_{j \in J} K_h\big(x - X_j^{(2)}\big) \Bigg). \tag{2.5}$$

- "Smooth" regime: $\hat{f}_{h,1}(x) \ge \frac{c_1 \ln n}{nh^d}$. Define the following bias-corrected plug-in estimator:

$$\begin{aligned} \hat{H}_2(x) = \mathbb{1}\Bigg(\hat{f}_{h,2}(x) \ge \frac{c_1 \ln n}{4nh^d} \Bigg) \cdot \Bigg\{ & -\hat{f}_{h,2}(x) \ln \hat{f}_{h,2}(x) \\ & - \big(1 + \ln \hat{f}_{h,2}(x)\big)\big(\hat{f}_{h,3}(x) - \hat{f}_{h,2}(x)\big) - \frac{1}{2}\Bigg(\hat{f}_{h,2}(x) \\ & - 2\hat{f}_{h,3}(x) + \frac{1}{\binom{n}{2} \hat{f}_{h,2}(x)} \sum_{i<j} K_h\big(x - X_i^{(3)}\big) K_h\big(x - X_j^{(3)}\big) \Bigg) \Bigg\}. \end{aligned} \tag{2.6}$$

- The final point estimate of $H(x) = -f_h(x)\ln f_h(x)$ is

$$\hat{H}(x) \triangleq \min\left\{\hat{H}_1(x), \frac{1}{n^{1-2\varepsilon}h^d}\right\}\mathbb{1}\left(\hat{f}_{h,1}(x) < \frac{c_1 \ln n}{nh^d}\right) + \hat{H}_2(x)\mathbb{1}\left(\hat{f}_{h,1}(x) \geq \frac{c_1 \ln n}{nh^d}\right). \tag{2.7}$$

Finally, choose

$$h = c_0(Ln\ln n)^{-\frac{1}{s+d}}, \quad k = \lceil c_2 \ln n \rceil, \tag{2.8}$$

where $c_0 > 0$ is any constant, $0 < 7c_2 \ln 2 < \varepsilon < \frac{s}{s+d}$ and $c_1 > 0$ is sufficiently large (per Lemmas 7–8) and output the estimator

$$\hat{H} = \int_{\mathbb{R}^d} \hat{H}(x)\,dx. \tag{2.9}$$

Note that the integration only need to be taken over the support of f_h, which is slightly larger than the unit cube $[0,1]^d$. This completes the construction of our estimator. A few remarks are in order:

Choice of the U-statistics. The following U-statistic

$$U_m = \frac{1}{\binom{n}{m}} \sum_{1\leq i_1<i_2<\cdots<i_m\leq n} \prod_{j=1}^{m} K_h(x - X_{i_j})$$

has appeared several times in the estimator construction, which is the natural unbiased estimator for powers of $f_h(x)$:

$$\mathbb{E}[U_m] = \frac{1}{\binom{n}{m}} \sum_{1\leq i_1<i_2<\cdots<i_m\leq n} \prod_{j=1}^{m} \mathbb{E}\big[K_h(x - X_{i_j})\big] = f_h(x)^m.$$

The reason why we average over all possible subsets of size m is to reduce the variance to the correct order (cf. Lemma 12). In practice, to compute the kth order U-statistics, note that it is simply the (normalized) kth elementary symmetric polynomial of $K_h(x - X_i)$. Hence, it suffices to compute the power sum $\sum_{i=1}^n (K_h(x - X_i))^l$ for all $l = 1, \ldots, k$, and then invoke Newton's identity to compute elementary symmetric polynomials; this has overall time complexity $O(nk + k^2) = O(n \log n)$. For the special case of the box kernel $K(t) = \mathbb{1}(t \in [-1/2, 1/2]^d)$, which can be used to achieve the upper bound in Theorem 4, $\hat{H}_1(x)$ reduces to

$$\hat{H}_1(x) = \sum_{l=0}^{k} a_l \cdot \frac{Z_x \cdot (Z_x - 1) \cdot \ldots \cdot (Z_x - l + 1)}{h^{ld} \cdot n \cdot (n-1) \cdot \ldots \cdot (n - l + 1)}, \tag{2.10}$$

where $Z_x = \sum_{i=1}^n h^d K_h(x - X_i^{(2)})$. Hence, the computational cost can be further reduced to $O(n + k^2) = O(n)$ in this simple example.

Polynomial approximation in the nonsmooth regime. In the nonsmooth regime (i.e., $\hat{f}_{h,1}(x) \leq \frac{c_1 \ln n}{nh^d}$) a suitable linear combination of the U-statistics is applied, where the coefficients come from the best approximating polynomial of our target functional $-x\ln x$. By the previous property of the U-statistic, in the nonsmooth regime we estimate $Q(f_h(x))$ without any bias, and thus the bias in this regime becomes the polynomial approximation error. The coefficients of the polynomial $Q(\cdot)$ can be efficiently computed via the Remez algorithm, which converges double exponentially fast (see discussions in [33]). The coefficients can also be precomputed and stored so that there is no need to recompute the coefficients when applying the estimator.

Bias correction based on Taylor expansion in the smooth regime. In the smooth regime (i.e., $\hat{f}_{h,1}(x) > \frac{c_1 \ln n}{nh^d}$), we use the idea in [25] to correct the bias. Specifically, by Taylor expansion we can write

$$\begin{aligned}\phi\big(f_h(x)\big) &\approx \sum_{l=0}^{R} \frac{\phi^{(l)}(\hat{f}_{h,2}(x))}{l!}\big(f_h(x) - \hat{f}_{h,2}(x)\big)^l \\ &= \sum_{l=0}^{R} \frac{\phi^{(l)}(\hat{f}_{h,2}(x))}{l!} \sum_{j=0}^{l} \binom{l}{j} f_h(x)^j \big(-\hat{f}_{h,2}(x)\big)^{l-j}.\end{aligned} \tag{2.11}$$

A natural idea to debias is to find an unbiased estimator of the right-hand side in (2.11). Indeed, this can be done by sample splitting: we can split observations to obtain $\hat{f}_{h,3}(x)$, an independent copy of $\hat{f}_{h,2}(x)$, and then apply the previous U-statistics to $\hat{f}_{h,3}(x)$ to obtain an unbiased estimator $f_h(x)^j$. Our estimator construction uses this idea with $\phi(z) = -z \ln z$ and $R = 2$ (which suffices for our debiasing purposes).

Choice of bandwidth. As will be clarified in Section 3 (cf. (3.16)), the bandwidth $h \asymp (n \ln n)^{-1/(s+d)}$ in (2.8) is chosen in order to balance between two types of biases of our estimator. Compared with the optimal bandwidth $h \asymp n^{-1/(2s+d)}$ in estimating the density under L_p risk for $p \in [1, \infty)$ [45], our choice of the bandwidth results in an "undersmoothed" kernel estimator of the density, which is consistent with the findings in [19, 48] that an undersmoothed kernel estimator should be used in estimating nonparametric functionals. However, our specific choice of $h \asymp (n \ln n)^{-1/(s+d)}$ is different from the optimal bandwidths for other problems, such as $h \asymp n^{-2/(4s+d)}$ for estimating quadratic, cubic and general smooth functionals [4, 36, 44, 52], and the optimal bandwidth $h \asymp (n \ln n)^{-1/(2s+d)}$ for estimating the L_r norm of the function in Gaussian white noise in one dimension with $r \in [1, \infty)$ not an even integer [24, 41].

Final integration. The estimator $\hat{H}(x)$ in (2.7) provides the pointwise estimation of $-f_h(x) \ln f_h(x)$ for all $x \in \mathsf{supp}(f_h)$, and an integration is required to produce the final entropy estimator. If the box kernel is used, notice that n small cubes of equal size can partition the unit cube $[0, 1]^d$ into $O(n^d)$ pieces, the mapping $x \mapsto Z_x$ in (2.10) is piecewise constant on $O(n^d)$ pieces. Hence, for exact integration it suffices to evaluate $\hat{H}(x)$ at $O(n^d)$ points, which yields an overall $O(n^{d+1} \log n)$ time complexity of our estimator. For practical implementation with general kernels, numerical integration methods and quadrature formulas can be used to evaluate the integral, and then we only need to evaluate $\hat{H}(x)$ at finitely many points.

In the next section, we prove the following result, which completes the proof of the upper bound in Theorem 1.

THEOREM 4. *For $s \in (0, 2]$, $p \geq 2$ and $L \leq (n \ln n)^{s/d}$, the following holds for the estimator $\hat{H}$ defined in (2.9):*

$$\Big(\sup_{f \in \mathrm{Lip}_{s,p,d}(L)} \mathbb{E}_f\big(\hat{H} - H(f)\big)^2\Big)^{\frac{1}{2}} \leq C\big((n \ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L\big),$$

where $C = C(s, p, d) > 0$ is independent of n, L (we omit the dependence of C on the choice of parameters c_0, c_1, c_2, ε and the kernel $K(\cdot)$).

3. Proof of upper bound. The error of our estimator $\hat{H}$ can be decomposed into three terms: the approximation error of $|H(f_h) - H(f)|$, the bias and the variance of the estimation error of $\hat{H}$ in estimating $H(f_h)$. Next, we deal with these terms separately.

3.1. *First-stage approximation error.* The approximation error between $H(f_h)$ and $H(f)$ is summarized in the following lemma, which is one of the key results in this paper.

LEMMA 2. *Let $s \in (0, 2]$, $p \geq 2$. For any $f \in \mathrm{Lip}_{s,p,d}(L)$ and bandwidth h with $0 < Lh^s \leq 1$, let f_h be defined in (2.1). There exists a constant $C > 0$ independent of h, such that*

$$|H(f) - H(f_h)| = \left| \int_{\mathbb{R}^d} f(x) \ln f(x)\, dx - \int_{\mathbb{R}^d} f_h(x) \ln f_h(x)\, dx \right| \leq C \cdot Lh^s \tag{3.1}$$

whenever $0 < h < h_0$, where h_0 is a constant depending only on s.

In view of the fact that $\|f - f_h\|_p \lesssim Lh^s$ (cf. (2.3)), Lemma 2 essentially says that the entropy functional $H(\cdot)$ is "Lipschitz" with respect to convolution. The proof of Lemma 2 consists of three steps:

1. By the convolution property, we first express the entropy difference $H(f_h) - H(f)$ as a mutual information term. Then using the variational representation of the mutual information and χ^2-divergence, we reduce (3.1) to an inequality that no longer involves the kernel;
2. By the equivalence between the K-functional and the modulus of continuity [12], we approximate f by a nonnegative C^2-function g and further reduce the goal to an estimate of the form

$$\int_{[0,1]^d} \frac{|\nabla g(x)|^2}{f(x) + h^s}\, dx;$$

3. We invoke the Hardy–Littlewood maximal inequality to control the above integral using the L_2-norm of the second-order derivative of g. This is the crux of the proof. The same proof technology also leads to a new upper bound on Fisher information, which we summarize at the end of this subsection.

3.1.1. *Mutual information and χ^2-divergence.* Recall the mutual information between random variables A and B is defined as the KL divergence between the joint distribution and product of the marginal distributions:

$$I(A; B) = D(P_{AB} \parallel P_A \otimes P_B) = \mathbb{E}\left[\ln \frac{dP_{AB}}{dP_A\, dP_B} \right].$$

Recall that, by Assumption 1, the kernel satisfies $K \geq 0$ and $\int_{\mathbb{R}^d} K(x)\, dx = 1$. Let X and U be independent random variables with density function f and K, respectively. Then by the convolution property, the density of $X + hU$ is f_h, and as a result,

$$0 \leq H(f_h) - H(f) = I(U; X + hU). \tag{3.2}$$

Note that by the compact support of the kernel K, the density f_h is supported on a cube slightly larger than $[0, 1]^d$ (i.e., with edge size $1 + O(h)$), and by a proper scaling we assume without loss of generality that both f and f_h are supported on $[0, 1]^d$.

Next, we reduce the desired inequality into a simpler one independent of the kernel $K(\cdot)$. Let w be an arbitrary density supported on $[0, 1]^d$. Then

$$\begin{aligned}
I(U; X + hU) &= \mathbb{E}_U \left[\int_{[0,1]^d} f(x - hU) \ln \frac{f(x - hU)}{f_h(x)}\, dx \right] \\
&= \mathbb{E}_U \left[\int_{[0,1]^d} f(x - hU) \ln \frac{f(x - hU)}{w(x)}\, dx \right] - D(f_h \parallel w) \\
&\overset{(a)}{\leq} \mathbb{E}_U \left[\int_{[0,1]^d} f(x - hU) \ln \frac{f(x - hU)}{w(x)}\, dx \right] \\
&\overset{(b)}{\leq} \mathbb{E}_U \left[\int_{[0,1]^d} \frac{(f(x - hU) - w(x))^2}{w(x)}\, dx \right],
\end{aligned} \tag{3.3}$$

where (a) follows from the nonnegativity of the KL divergence, and (b) is due to the fact that the KL divergence is upper bounded by the χ^2-divergence.

Since $Lh^s \le 1$, there exists another density w on $[0,1]^d$ such that $w(x) \ge f(x)/2 + Lh^s/2$ for all $x \in [0,1]^d$ and $|w(x) - f(x)| \le Lh^s(1+f(x))/2$. Such an existence may be constructed by choosing $w(x) = Lh^s/2 + (1 - Lh^s/2) f(x)$. As a result,

$$
\begin{aligned}
&\int_{[0,1]^d} \frac{(f(x-hU) - w(x))^2}{w(x)}\,dx \\
&\quad= \int_{[0,1]^d} \frac{(f(x-hU) - f(x) + f(x) - w(x))^2}{w(x)}\,dx \\
&\quad\le 2\int_{[0,1]^d} \frac{(f(x-hU) - f(x))^2}{w(x)}\,dx + 2\int_{[0,1]^d} \frac{(f(x) - w(x))^2}{w(x)}\,dx \\
&\quad\le 2\int_{[0,1]^d} \frac{(f(x-hU) - f(x))^2}{w(x)}\,dx + \int_{[0,1]^d} \frac{(Lh^s)^2(1+f(x)^2)}{f(x)/2 + Lh^s/2}\,dx \\
&\quad\lesssim \int_{[0,1]^d} \frac{(f(x-hU) - f(x))^2}{w(x)}\,dx + Lh^s.
\end{aligned} \tag{3.4}
$$

Combining (3.2)–(3.4), we have

$$
0 \le H(f_h) - H(f) \lesssim \mathbb{E}_U\left[\int_{[0,1]^d} \frac{(f(x-hU) - f(x))^2}{f(x) + Lh^s}\,dx\right] + Lh^s.
$$

Recall that the finite second moment of the kernel K ensures that $\mathbb{E}|U|^2 < \infty$. Therefore, for Lemma 2 to hold, it suffices to prove that for any $u \in \mathbb{R}$,

$$
\int_{[0,1]^d} \frac{(f(x+hu) - f(x))^2}{f(x) + Lh^s}\,dx \lesssim Lh^s\bigl(1 + |u|^2\bigr). \tag{3.5}
$$

Note that (3.5) no longer involves the kernel $K(\cdot)$.

We provide some insights why (3.5) is expected to hold. When $s \le 1$ and $p = \infty$, the Lipschitz ball condition ensures that $|f(x+hu) - f(x)| \lesssim Lh^s|u|^s \le Lh^s(1+|u|)$, and (3.5) clearly holds. However, when $1 < s \le 2$, we will only have $|f(x+hu) - f(x)| \lesssim Lh|u|$ in general, and (3.5) cannot be derived by this simple approach. The crux of proving (3.5) for $1 < s \le 2$ is that, when $f(x)$ is close to zero, the difference $|f(x+hu) - f(x)|$ also needs to be small to maintain the nonnegativity of $f(x-hu)$. In Section 3.1.3, we will essentially show that $|f(x+hu) - f(x)| \lesssim L\sqrt{f(x)h^s}(1+|u|)$, which leads to (3.5).

3.1.2. *Approximation by C^2 functions.* We need the following lemma to replace f with a smoother function g.

LEMMA 3. *Let $f \in \mathrm{Lip}_{s,p,d}(L)$ be a nonnegative function, with $s \in (0,2]$ and $p \ge 1$. Then there exists $C = C(s,p,d)$, such that for any $h > 0$, there exists a nonnegative function $g \in C^2(\mathbb{R}^d)$ such that*

$$
\|f - g\|_p \le CLh^s, \tag{3.6}
$$

$$
\bigl\|\|\nabla^2 g(\cdot)\|_{\mathrm{op}}\bigr\|_p \le CLh^{s-2}. \tag{3.7}
$$

Note that Lemma 3 is essentially the equivalence between the K-functional and the modulus of smoothness (Lemma 14 in Appendix 5), with an extra constraint that g being nonnegative, which turns out to be crucial in proving the inequality (3.9) below.

Let g be given by Lemma 3. Then

$$\begin{aligned}
&\int_{[0,1]^d} \frac{(f(x+hu)-f(x))^2}{f(x)+Lh^s}\,dx \\
&\quad\le 3\int_{[0,1]^d} \frac{(f(x+hu)-g(x+hu))^2+(f(x)-g(x))^2+(g(x+hu)-g(x))^2}{f(x)+Lh^s}\,dx \\
&\quad\le 3\int_{[0,1]^d} \frac{(f(x+hu)-g(x+hu))^2+(f(x)-g(x))^2}{Lh^s}\,dx \\
&\qquad + 6\int_{[0,1]^d} \frac{(h\nabla g(x)^\top u)^2+(g(x+uh)-g(x)-h\nabla g(x)^\top u)^2}{f(x)+Lh^s}\,dx \\
&\quad\le \frac{6\|f-g\|_2^2}{Lh^s}+\frac{6}{Lh^s}\|\rho(\cdot,u)\|_2^2+6h^2|u|^2\int_{[0,1]^d}\frac{|\nabla g(x)|^2}{f(x)+Lh^s}\,dx,
\end{aligned} \tag{3.8}$$

where

$$\rho_h(x,u)\triangleq g(x+uh)-g(x)-h\nabla g(x)^\top u.$$

We bound the three terms in (3.8) separately. By (3.6), the first term is upper bounded by

$$\frac{6\|f-g\|_2^2}{Lh^s}\le\frac{6\|f-g\|_p^2}{Lh^s}\lesssim Lh^s.$$

For the second term, by the integral representation of the Taylor remainder term, we have

$$\begin{aligned}
|\rho_h(x,u)| &= \left|\int_0^1 (1-t)u^\top\nabla^2 g(x+t\cdot hu)u\cdot dt\right| \\
&\le h^2|u|^2\cdot\int_0^1(1-t)\|\nabla^2 g(x+t\cdot hu)\|_{\mathrm{op}}\,dt
\end{aligned}$$

and hence

$$\begin{aligned}
\|\rho_h(\cdot,u)\|_2 &\le h^2|u|^2\cdot\left\|\int_0^1(1-t)\|\nabla^2 g(x+t\cdot hu)\|_{\mathrm{op}}\,dt\right\|_2 \\
&\overset{\text{(a)}}{\le} h^2|u|^2\cdot\int_0^1(1-t)\|\|\nabla^2 g(x+t\cdot hu)\|_{\mathrm{op}}\|_2\,dt \\
&\overset{\text{(b)}}{\lesssim} h^2|u|^2\cdot Lh^{s-2}=Lh^s|u|^2,
\end{aligned}$$

where (a) follows from the convexity of norms and (b) follows from (3.7). Thus the first two terms in (3.8) are both upper bounded by $O(Lh^s|u|^2)$. Hence, to show (3.5), it remains to prove that

$$\int_{[0,1]^d}\frac{|\partial_i g(x)|^2}{f(x)+Lh^s}\,dx\lesssim Lh^{s-2}\quad\forall i\in[d], \tag{3.9}$$

where $\partial_i g=\frac{\partial g}{\partial x_i}$.

3.1.3. *Application of the Hardy–Littlewood maximal inequality.* Finally, we use the nonnegativity of g and the Hardy–Littlewood maximal inequality [57] to prove (3.9). Fix any $\tau>0$ to be optimized later. Since g is nonnegative, we have

$$\begin{aligned}
0&\le g(x+\tau e_i)\\
&= g(x)+\tau\cdot\partial_i g(x)+\big(g(x+\tau e_i)-g(x)-\tau\cdot\partial_i g(x)\big),
\end{aligned}$$

and thus

$$-\tau \cdot \partial_i g(x) \le g(x) + \big(g(x+\tau e_i) - g(x) - \tau \cdot \partial_i g(x)\big).$$

Replacing $x + \tau e_i$ by $x - \tau e_i$, we also have

$$\tau \cdot \partial_i g(x) \le g(x) + \big(g(x-\tau e_i) - g(x) + \tau \cdot \partial_i g(x)\big).$$

Combining these two inequalities, we arrive at the following pointwise bound:

$$\begin{aligned}\tau \cdot \big|\partial_i g(x)\big| &\le 2g(x) + \big|g(x+\tau e_i) - g(x) - \tau \cdot \partial_i g(x)\big| \\ &\quad + \big|g(x-\tau e_i) - g(x) + \tau \cdot \partial_i g(x)\big| \\ &\le 2g(x) + \tau^2 \int_{-1}^{1} \big|\partial_{ii} g(x + t \cdot \tau e_i)\big|\, dt\end{aligned}$$

where for the second inequality we have used the integral representation of the Taylor remainder term again.

Since the previous inequality holds for any $\tau > 0$, we choose $\tau = \tau_x = \sqrt{h^{2-s} f(x)/L + h^2}$ to obtain an upper bound on the derivative:

$$\big|\partial_i g(x)\big| \le \frac{2g(x)}{\sqrt{h^{2-s} f(x)/L + h^2}} + \tau_x \int_{-1}^{1} \big|\partial_{ii} g(x + t \cdot \tau_x e_i)\big|\, dt.$$

Plugging this bound into (3.9) and using the triangle inequality, we have

$$\begin{aligned}&\int_{[0,1]^d} \frac{|\partial_i g(x)|^2}{f(x) + Lh^s}\, dx \\ &\le 2\Bigg(\underbrace{Lh^{s-2} \int_{[0,1]^d} \frac{4g(x)^2}{(f(x) + Lh^s)^2}\, dx}_{\triangleq A_1} \\ &\qquad + \underbrace{(Lh^{s-2})^{-1} \int_{[0,1]^d} \bigg(\int_{-1}^{1} \big|\partial_{ii} g(x + t \cdot \tau_x e_i)\big|\, dt\bigg)^2 dx}_{\triangleq A_2}\Bigg).\end{aligned}$$

Next, we upper bound A_1 and A_2 separately. For A_1, we use the triangle inequality again to obtain

$$\begin{aligned}A_1 &= Lh^{s-2} \cdot \int_{[0,1]^d} \frac{4g(x)^2}{(f(x) + Lh^s)^2}\, dx \\ &\le 8Lh^{s-2} \cdot \int_{[0,1]^d} \frac{(g(x) - f(x))^2 + f(x)^2}{(f(x) + Lh^s)^2}\, dx \\ &\le 8Lh^{s-2} \cdot \bigg(\int_{[0,1]^d} \frac{(g(x) - f(x))^2}{L^2 h^{2s}}\, dx + \int_{[0,1]^d} \frac{(f(x))^2}{(f(x))^2}\, dx\bigg) \\ &= 8Lh^{s-2} \cdot \bigg(\frac{\|g - f\|_2^2}{L^2 h^{2s}} + 1\bigg) \lesssim Lh^{s-2}.\end{aligned}$$

Hence, it remains to upper bound A_2, and it further suffices to prove that for any $x_{\setminus i} \in [0,1]^{d-1}$,

$$\int_0^1 \bigg(\int_{-1}^{1} \big|\partial_{ii} g(x + t \cdot \tau_x e_i)\big|\, dt\bigg)^2 dx_i \le C \int_0^1 \big|\partial_{ii} g(x)\big|^2 dx_i \tag{3.10}$$

for some constant $C > 0$. In fact, if (3.10) holds, then integrating both sides over $x_{\setminus i} \in [0,1]^{d-1}$ together with the fact $\|\partial_{ii} g\|_2 \le \|\partial_{ii} g\|_p \lesssim L h^{s-2}$ completes the proof of (3.9).

The proof of (3.10) requires the introduction of the maximal inequality. Fix any $x_{\setminus i} \in [0,1]^{d-1}$ and define $h(y) \triangleq |\partial_{ii} g(x_{\setminus i}, y)|$, (3.10) is equivalent to

$$\int_0^1 \left(\frac{1}{2\tau_x} \int_{x-\tau_x}^{x+\tau_x} h(y)\, dy \right)^2 dx \le \frac{C}{4} \int_0^1 |h(x)|^2\, dx. \tag{3.11}$$

For any function h on the real line, recall the Hardy–Littlewood maximal function $M[h]$ is defined as

$$M[h](y) \triangleq \sup_{t>0} \frac{1}{2t} \int_{y-t}^{y+t} |h(z)|\, dz. \tag{3.12}$$

Next, we recall the maximal inequality on the real line [57].

LEMMA 4. *For any nonnegative real-valued measurable function h on the real line $\mathbb{R}$, the following tail bound holds: for any $t > 0$, there exists a universal constant $C_1 > 0$ such that for $p > 1$ we have*

$$\|M[h]\|_p \le C_1 \left(\frac{p}{p-1} \right)^{\frac{1}{p}} \|h\|_p.$$

Applying this lemma with $p = 2$ yields (3.11) (and thus (3.10)), as desired, completing the proof of Lemma 2.

We finish this subsection by noting that the proof technology developed based on the maximal inequality in fact leads to the following upper bound on Fisher information, which may be of independent interest.

THEOREM 5. *Let $f \in C^1(\mathbb{R}^d)$ be a density function supported on $[0,1]^d$ with an absolute continuous gradient. Denote its Fisher information by*

$$J(f) \triangleq \int_{\mathbb{R}^d} \frac{|\nabla f|^2}{f}.$$

Then for any $p > 1$, there exists a constant $C_p > 0$, such that

$$J(f) \le C_p \sum_{i=1}^d \|\partial_{ii} f\|_p.$$

The connection between this result and the previous proof of Lemma 2 is the well-known fact that the local expansion of χ^2-divergence is given by the Fisher information. Indeed, by Taylor expansion (assuming for simplicity that $d = 1$ and $f = g$), the LHS of the main estimate (3.5) behaves as $h^2 J(f)$. Thanks to Theorem 5, we can control the Fisher information by $J(f) = O(\|f''\|_2)$, which, by the smoothness assumption, is $O(L h^{s-2})$, and leads to the desired (3.5).

3.2. *Second-stage approximation error and variance.* In this subsection, we analyze the performance of our pointwise estimator $\hat{H}(x)$ in estimating $-f_h(x) \ln f_h(x)$ for any $x \in \mathbb{R}^d$. Recall that

$$f_h(x) = \int_{\mathbb{R}^d} K_h(x-y) f(y)\, dy$$

is the smoothed density, and

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

is our estimator of $f_h(x)$, and $K_h(t) \triangleq h^{-d}K(h^{-1}t)$. In addition to the unbiasedness of $\hat{f}_h(x)$ in estimating $f_h(x)$, it satisfies some more properties: the random variable $h^d\hat{f}_h(x)$ roughly follows a binomial distribution $\mathsf{B}(n, h^d f_h(x))$. This property confines to the one-dimensional simple example that $h^d f_h(x)$ can be viewed as the discrete probability in the bin containing x, and $h^d\hat{f}_h(x)$ is the empirical frequency. Specifically, we prove the following lemma.

LEMMA 5. *If the kernel K is nonnegative everywhere, then there exists a constant $c_1 > 0$ depending on d and $\|K\|_\infty$ only such that*:

1. *If $f_h(x) \le \frac{c_1 \ln n}{2nh^d}$, we have*

$$\mathbb{P}\left(\hat{f}_h(x) < \frac{c_1 \ln n}{nh^d}\right) \ge 1 - n^{-5d};$$

2. *If $f_h(x) \ge \frac{c_1 \ln n}{2nh^d}$, we have*

$$\mathbb{P}\left(\frac{f_h(x)}{2} \le \hat{f}_h(x) \le 2f_h(x)\right) \ge 1 - n^{-5d}.$$

Note that the concentration property of $\hat{f}_h(x)$ in Lemma 5 behaves as if $n\hat{f}_h(x)$ is distributed binomially as $\mathsf{B}(n, h^d f_h(x))$. It shows that, based on our threshold to split the smooth and nonsmooth regimes in (2.7), the probability of making an error in classification is negligible.

Now we prove that our estimators $\hat{H}_1$ and $\hat{H}_2$ in (2.5)–(2.6) perform well in the corresponding regimes. To bound the variance, we invoke the well-known Efron–Stein–Steele inequality.

LEMMA 6 ([56]). *Let $X_1, X_2, \ldots, X_n$ be independent random variables, and for $i = 1, \ldots, n$, let X_i' be an independent copy of X_i. Then for any f,*

$$\mathsf{Var}\big(f(X_1, \ldots, X_n)\big) \le \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\big(f(X_1, \ldots, X_n) - f\big(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\big)\big)^2.$$

To apply Lemma 6, we need to bound the difference between the original estimator and the perturbed one where one observation is substituted by a fresh copy. Recall that $\hat{H}_1$ depends only on the second part of observations $X^{(2)}$, and $\hat{H}_2$ depends on the last two parts $X^{(2)} \cup X^{(3)}$. Hence, we may define $\hat{H}_1'$, $\hat{H}_2'$ to be the perturbed estimators where exactly one observation chosen uniformly at random from $X^{(2)} \cup X^{(3)}$ is replaced by its independent copy. The following lemmas summarize the upper bounds of the bias and the second moment of the perturbations.

LEMMA 7 (Nonsmooth regime). *If $f_h(x) \le \frac{2c_1 \ln n}{nh^d}$, $c_1 \ge 2\|K\|_\infty c_2$, $0 < 7c_2 \ln 2 < \varepsilon$ and $n \ge 4c_2 \ln n$, we have*

$$\big|\mathbb{E}\hat{H}_1(x) + f_h(x)\ln f_h(x)\big| \lesssim \frac{1}{nh^d \ln n},$$

$$\mathbb{E}\big[\big(\hat{H}_1(x) - \hat{H}_1'(x)\big)^2\big] \lesssim \frac{1}{n^{3-\varepsilon}h^{2d}}.$$

LEMMA 8 (Smooth regime). *If $f_h(x) \ge \frac{c_1 \ln n}{2nh^d}$ with sufficiently large constant $c_1 > 0$ (as in Lemma 5), $h \le 1$ and $nh^d \ge 1$, we have*

$$\big|\mathbb{E}\hat{H}_2(x) + f_h(x)\ln f_h(x)\big| \lesssim \frac{1}{nh^d \ln n},$$
$$\mathbb{E}\big[\big(\hat{H}_2(x) - \hat{H}_2'(x)\big)^2\big] \lesssim \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n^2 h^d}.$$

For the final pointwise estimator $\hat{H}(x)$ in (2.7), let $\hat{H}'(x)$ be its perturbed version where exactly one observation chosen uniformly at random from $X^{(2)} \cup X^{(3)}$ is replaced by its independent copy (note that $X^{(1)}$ is excluded). The following guarantee for $\hat{H}(x)$ follows from Lemmas 5–8.

COROLLARY 1. *Under the assumptions of Lemmas 5–8, we have*

$$\mathbb{E}\big(\mathbb{E}\big[\hat{H}(x)|X^{(1)}\big] + f_h(x)\ln f_h(x)\big)^2 \lesssim \frac{1}{(nh^d \ln n)^2}, \tag{3.13}$$

$$\mathbb{E}\big[\big(\hat{H}(x) - \hat{H}'(x)\big)^2\big] \lesssim \frac{1}{n^{3-\varepsilon}h^{2d}} + \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n^2 h^d}. \tag{3.14}$$

3.3. *Overall performance.* Now we are ready to analyze the overall performance of the integrated estimator $\hat{H} = \int_{\mathbb{R}^d} \hat{H}(x)\,dx$. As argued in Section 3.1.1, we assume without loss of generality that $f_h(\cdot)$ is supported on $[0,1]^d$, so that $\hat{H} = \int_{[0,1]^d} \hat{H}(x)\,dx$. By the triangle inequality, we have the following decomposition of the mean squared error (recall that $X^{(1)}$ is the first part of observations for regime classification):

$$\begin{aligned}\mathbb{E}\big(\hat{H} - H(f)\big)^2 &\le 2\big[\big(H(f_h) - H(f)\big)^2 + \mathbb{E}\big(\hat{H} - H(f_h)\big)^2\big] \\ &= 2\big[\big(H(f_h) - H(f)\big)^2 + \mathbb{E}\big(\mathbb{E}\big[\hat{H}|X^{(1)}\big] - H(f_h)\big)^2 \\ &\quad + \mathbb{E}\big[\mathsf{Var}\big(\hat{H}|X^{(1)}\big)\big]\big].\end{aligned} \tag{3.15}$$

We analyze different types of errors in (3.15) separately:

- First-stage approximation error: by Lemma 2, we know that
$$\big|H(f) - H(f_h)\big| \lesssim Lh^s.$$
- Conditional bias (second-stage approximation error): by Corollary 1 and Cauchy–Schwarz,
$$\begin{aligned}\mathbb{E}\big(\mathbb{E}\big[\hat{H}|X^{(1)}\big] - H(f_h)\big)^2 &= \mathbb{E}\bigg(\int_{[0,1]^d}\big(\mathbb{E}\big[\hat{H}(x)|X^{(1)}\big] + f_h(x)\ln f_h(x)\big)\,dx\bigg)^2 \\ &\le \int_{[0,1]^d}\mathbb{E}\big(\mathbb{E}\big[\hat{H}(x)|X^{(1)}\big] + f_h(x)\ln f_h(x)\big)^2\,dx \\ &\lesssim \frac{1}{(nh^d \ln n)^2}.\end{aligned}$$
- Conditional variance: conditioned on $X^{(1)}$, the estimator $\hat{H}$ is a deterministic function of $(X^{(2)}, X^{(3)})$ consisting of mutually independent observations. We now apply Lemma 6 to bound the variance. For $i = 1, 2, \ldots, 2n$, define $\hat{H}_i(x)$ to be the pointwise estimator in (2.7) with ith observation in $(X^{(2)}, X^{(3)})$ replaced by an independent copy, and let $\hat{H}_i =$

$\int_{[0,1]^d} \hat{H}_i(x)\,dx$. Then by Lemma 6, we have

$$\begin{aligned}\mathsf{Var}\big(\hat{H}|X^{(1)}\big) &\le \frac{1}{2}\sum_{i=1}^{2n}\mathbb{E}\big[(\hat{H}-\hat{H}_i)^2|X^{(1)}\big]\\ &= \frac{1}{2}\sum_{i=1}^{2n}\mathbb{E}\bigg[\bigg(\int_{[0,1]^d}\big(\hat{H}(x)-\hat{H}_i(x)\big)\,dx\bigg)^2\Big|X^{(1)}\bigg].\end{aligned}$$

Since K has compact support (cf. Assumption 1), by our estimator construction we have $\mathsf{Leb}(\{x\in[0,1]^d:\hat{H}(x)\neq\hat{H}_i(x)\})\lesssim h^d$. Hence, by Cauchy–Schwarz, we have

$$\begin{aligned}&\bigg(\int_{[0,1]^d}\big(\hat{H}(x)-\hat{H}_i(x)\big)\,dx\bigg)^2\\ &\quad\le \int_{[0,1]^d}\big(\hat{H}(x)-\hat{H}_i(x)\big)^2\,dx\cdot\int_{[0,1]^d}\mathbb{1}\big(\hat{H}(x)\neq\hat{H}_i(x)\big)\,dx\\ &\quad\lesssim h^d\int_{[0,1]^d}\big(\hat{H}(x)-\hat{H}_i(x)\big)^2\,dx.\end{aligned}$$

Combining the previous two displays, the conditional variance can be upper bounded as (recall the definition of $\hat{H}'(x)$ before Corollary 1)

$$\begin{aligned}\mathbb{E}\big[\mathsf{Var}\big(\hat{H}|X^{(1)}\big)\big] &\lesssim nh^d\int_{[0,1]^d}\mathbb{E}\big(\hat{H}(x)-\hat{H}'(x)\big)^2\,dx\\ &\overset{(3.14)}{\lesssim}\int_{[0,1]^d}\bigg(\frac{1}{n^{2-\varepsilon}h^d}+\frac{f_h(x)(1+(\ln f_h(x))^2)}{n}\bigg)\,dx\\ &\lesssim \frac{1}{n^{2-\varepsilon}h^d}+\frac{(\ln L)^2}{n},\end{aligned}$$

where the last inequality follows from Lemma 10 and $\|f_h\|_p\le\|f\|_p+\|f-f_h\|_p\lesssim L(1+h^s)$ (cf. (2.3)).

Substituting all three types of error bounds into (3.15), we obtain

$$\bigg(\sup_{f\in\mathrm{Lip}_{s,p,d}(L)}\mathbb{E}_f\big(\hat{H}-H(f)\big)^2\bigg)^{\frac{1}{2}}\lesssim Lh^s+\frac{1}{nh^d\ln n}+\frac{1}{n^{1-\varepsilon/2}\sqrt{h^d}}+\frac{\ln L}{\sqrt{n}}. \tag{3.16}$$

Finally, we choose $h=(Ln\ln n)^{-\frac{1}{s+d}}$ (note that the condition $Lh^s\le 1$ in Lemma 2 holds due to the assumption $L\le(n\ln n)^{s/d}$) and $\varepsilon<\frac{s}{s+d}$ in (3.16) to obtain

$$\bigg(\sup_{f\in\mathrm{Lip}_{s,p,d}(L)}\mathbb{E}_f\big(\hat{H}-H(f)\big)^2\bigg)^{\frac{1}{2}}\lesssim(n\ln n)^{-\frac{s}{s+d}}L^{\frac{d}{s+d}}+n^{-\frac{1}{2}}\ln L,$$

completing the proof of Theorem 4.

4. Further discussions.

4.1. *Extensions to densities satisfying periodic boundary conditions.* In this section, we relax the assumptions that the underlying density f is supported on $[0,1]^d$ and smoothly vanishing at the boundary, and establish the corresponding minimax rates in entropy estimation.

Note that since the L_p norm in (1.2) is taken in $\mathbb{R}^d$, the definition of the Lipschitz norm requires that the density f connects to zero smoothly at the boundary of $[0,1]^d$, which may exclude some well-known densities such as the uniform distribution. This assumption can be

relaxed by considering the "periodic boundary condition," which requires that the periodic extension of the density f lies in the Lipschitz ball. Specifically, we define a new Lipschitz ball

$$\mathrm{Lip}^{\star}_{s,p,d}(L) = \{f : \|f\|_{\mathrm{Lip}^{\star}_{s,p,d}} \le L\} \cap \{f : \mathrm{supp}(f) \subseteq [0,1]^d\},$$

where the Lipschitz norm $\|\cdot\|_{\mathrm{Lip}^{\star}_{s,p,d}}$ is defined in the same way as (1.2) to (1.4), with the only exceptions that the L_p norm is taken over the unit cube $[0,1]^d$ instead of $\mathbb{R}^d$, and f is periodically extended to the entire space $\mathbb{R}^d$ via $f(x \mod 1) \triangleq f(x_1 - \lfloor x_1 \rfloor, x_2 - \lfloor x_2 \rfloor, \ldots, x_d - \lfloor x_d \rfloor)$, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Note that in this case we may also identify the unit cube $[0,1]^d$ as the d-dimensional torus $\mathbb{T}^d$.

The periodic boundary condition is weaker than the previous Lipschitz ball condition, in the sense of the norm comparison $\|f\|_{\mathrm{Lip}^{\star}_{s,p,d}} \le C\|f\|_{\mathrm{Lip}_{s,p,d}}$ for some constant $C > 0$.[5] This assumption has already appeared in the literature [39], and the special case $s = 2$, $d = 1$ corresponds to $f(0) = f(1)$, $f'(0) = f'(1)$. The next theorem shows that, the minimax rate remains unchanged in this weaker setting.

THEOREM 6. *For any $d \in \mathbb{N}$, $0 < s \le 2$ and $2 \le p < \infty$, there exists $L_0 > 1$ depending on s, p, d, such that for any $L_0 \le L \le (n \ln n)^{s/d}$ and any $n \in \mathbb{N}$,*

$$c\big((n\ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L\big) \le \Big(\inf_{\hat{H}} \sup_{f \in \mathrm{Lip}^{\star}_{s,p,d}(L)} \mathbb{E}_f\big(\hat{H} - H(f)\big)^2\Big)^{\frac{1}{2}}$$
$$\le C\big((n\ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L\big),$$

where $c, C > 0$ are constants depending on s, p, d.

Theorem 6 is a straightforward extension of Theorem 1. Since $\|\cdot\|_{\mathrm{Lip}^{\star}_{s,p,d}}$ is a weaker norm than $\|\cdot\|_{\mathrm{Lip}_{s,p,d}}$, the minimax lower bound in Theorem 1 continues to hold for the new Lipschitz ball. As for the upper bound, we use the same estimator construction as in Section 2, with the understanding that the kernel convolution is taken with respect to the periodic extension of f (or equivalently, is taken on the torus $\mathbb{T}^d$). For the analysis of this estimator, we apply a version of the maximal inequality on the torus $\mathbb{T}^d$ in Section 3.1, and the remaining arguments in Section 3 are essentially the same. We postpone the detailed proof to Section 7.3 in the Appendix.

4.2. *The case of $s > 2$.* Note that the minimax lower bound in Theorem 7 holds for all smoothness parameters $s > 0$, but our current proof techniques of upper bound only work for the smoothness regime of $0 < s \le 2$. There are two main reasons:

1. Classifying smooth/nonsmooth regime (Lemma 5) fails when $s > 2$;
2. Bias correction based on Taylor expansion (Lemma 8) in the smooth regime does not extend to $s > 2$.

The failures of Lemma 5 and 8 are intrinsically related to the fact that one has to use kernels with negative parts to take advantage of smoothness $s > 2$ [59]. Concretely, Lemma 5 is closely related to the problem of adapting to the lowest values of density in density estimation [49]. It was conjectured in [49] that the case of $s > 2$ exhibit significantly different behavior from the case of $0 < s \le 2$. Regarding bias correction, when the kernel is no longer

[5] This can be shown by applying the triangle inequality to $\|\sum_{i=1}^{3^d} f_i\|_{\mathrm{Lip}_{s,p,d}}$, where $f_i(x) = f(x - x_i)$ is the translation of f with $\{x_1, \ldots, x_{3^d}\} = \{-1, 0, 1\}^d$. Consequently, one can choose $C = 3^d$.

nonnegative, (8.1) can fail even when $f_h(x) > 0$, which makes the proof of Lemma 8 break down. It is possible that bias correction based on Jackknife may achieve better performances when $s > 2$. For the application of this approach in entropy estimation, we refer to [11, 43].

Finally, we remark that the high smoothness regime of $s > 2$ may not pose significant challenge for other problems of nonparametric statistics. For example, in the Gaussian white noise model, since it is a location model, the concentration of kernel estimators can be directly guaranteed using concentration inequality for sums of independent bounded random variables, which turns out to be sufficient for nonsmooth functional estimation [24]. Even in the density model, the case of $s > 2$, which indeed calls for kernels with negative parts, can be easily handled. For example, to estimate density itself under L_2 risk, we can simply truncate the negative density estimates to obtain a better performance [59]; in smooth functional estimation [5, 58], the case of $s > 2$ is also not special. It is mainly in estimating nonsmooth functionals that designing procedures that can *adapt* to low density regime becomes a crucial challenge [49].

4.3. *Connections to discrete entropy estimation.* Another intuitive idea for estimating the entropy of densities is to reduce it to a discrete entropy estimation problem. The motivation is that for a continuous random variable X with density f, it is well known [51] that the Shannon entropy of its quantized version $[X]_k \triangleq \lfloor kX \rfloor / k$ satisfies $H([X]_k) = d \log k + H(f) + o(1)$ as the quantization level $k \to \infty$. Thus, to estimate $H(f)$, we can choose an appropriate k, quantize all the observations, and apply the optimal Shannon entropy estimator developed in [33, 65]. Below we show that this approach achieves the minimax rate if $s \le 1$.

For the ease of exposition, we consider $s \in (0, 1]$ and $p \ge 2$, with general dimension $d \in \mathbb{N}$. We split the unit cube $[0, 1]^d$ into $S = h^{-d}$ subcubes $I_1, \dots, I_S$ of size h, where h is the "bandwidth" we will choose later. For $i = 1, \dots, S$, define

$$p_i = \int_{I_i} f(t)\, dt, \qquad \hat{p}_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(X_j \in I_i),$$

as the "probability" and the "empirical frequency" of the cube I_i, respectively. Since the entropy of the piecewise constant density $f_h(x) = \sum_{i=1}^{S} p_i \mathbb{1}(x \in I_i)$ is $H(f_h) = \sum_{i=1}^{S} p_i \ln \frac{1}{p_i} + \ln h^d$, the problem of estimating $H(f_h)$ is reduced to a discrete Shannon entropy estimation problem. We can then use the minimax rate-optimal estimators $\hat{H}_{\mathsf{discrete}}$ [33, 65] for the discrete entropy $\sum_{i=1}^{S} -p_i \ln p_i$ to define the estimator $\hat{H}$ for the entropy of the density $H(f)$ as

$$\hat{H} = \hat{H}_{\mathsf{discrete}} + \ln h^d. \tag{4.1}$$

One can show that $|H(f) - H(f_h)| = O(Lh^s)$. The optimal bandwidth is $h \asymp (Ln \ln n)^{-1/(s+d)}$, leading to the following risk bound, with an additional mild assumption that the density is bounded.

LEMMA 9. *For $d \in \mathbb{N}$, $s \in (0, 1]$, $p \ge 2$, the performance of the estimator $\hat{H}$ in (4.1) is given by*

$$\Big(\sup_{f \in \mathrm{Lip}_{s,p,d}(L), \|f\|_\infty \le L} \mathbb{E}_f \big(\hat{H} - H(f)\big)^2 \Big)^{\frac{1}{2}} \le C\big((n \ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L \big),$$

where $C > 0$ is a constant independent of n, L.

Acknowledgments. Yanjun Han and Tsachy Weissman were supported in part by the NSF Grants CCF-0939370 and CCF-1527105.

Jiantao Jiao was supported in part by the NSF Grants CCF-0939270, CCF-1527105 and IIS-1901252.

Yihong Wu was supported in part by the NSF Grants CCF-1900507, CCF-1527105, an NSF CAREER award CCF-1651588 and an Alfred Sloan fellowship.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal rates of entropy estimation over Lipschitz balls" (DOI: 10.1214/19-AOS1927SUPP; .pdf). We provide auxiliary lemmas used in this paper and proofs of minimax lower bound (Theorem 7), Theorems 2, 3, 6, Lemmas 3, 5, 7, 8, 9 and Corollary 1.

REFERENCES

[1] Beirlant, J., Dudewicz, E. J., Györfi, L. and van der Meulen, E. C. (1997). Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **6** 17–39. MR1471870

[2] Berrett, T. B., Samworth, R. J. and Yuan, M. (2019). Efficient multivariate entropy estimation via k-nearest neighbour distances. *Ann. Statist.* **47** 288–318. MR3909934 https://doi.org/10.1214/18-AOS1688

[3] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[4] Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhya, Ser. A* **50** 381–393. MR1065550

[5] Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29. MR1331653 https://doi.org/10.1214/aos/1176324452

[6] Bu, Y., Zou, S., Liang, Y. and Veeravalli, V. V. (2018). Estimation of KL divergence: Optimal minimax rate. *IEEE Trans. Inform. Theory* **64** 2648–2674. MR3782280 https://doi.org/10.1109/TIT.2018.2805844

[7] Cai, T. T. and Low, M. G. (2003). A note on nonparametric estimation of linear functionals. *Ann. Statist.* **31** 1140–1153. MR2001645 https://doi.org/10.1214/aos/1059655908

[8] Cai, T. T. and Low, M. G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33** 2930–2956. MR2253108 https://doi.org/10.1214/009053605000000147

[9] Cai, T. T. and Low, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346 https://doi.org/10.1214/10-AOS849

[10] Costa, J. A. and Hero, A. O. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. Signal Process.* **52** 2210–2221. MR2085582 https://doi.org/10.1109/TSP.2004.831130

[11] Delattre, S. and Fournier, N. (2017). On the Kozachenko–Leonenko entropy estimator. *J. Statist. Plann. Inference* **185** 69–93. MR3612672 https://doi.org/10.1016/j.jspi.2017.01.004

[12] DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **303**. Springer, Berlin. MR1261635 https://doi.org/10.1007/978-3-662-02888-9

[13] Donoho, D. L. (1997). Renormalizing experiments for nonlinear functionals. In *Festschrift for Lucien Le Cam* 167–181. Springer, New York. MR1462945 https://doi.org/10.1007/978-1-4612-1880-7_11

[14] Donoho, D. L. and Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323. MR1081043 https://doi.org/10.1016/0885-064X(90)90025-9

[15] El Haje Hussein, F. and Golubev, Y. (2009). On entropy estimation by m-spacing method. *J. Math. Sci.* **163** 290–309.

[16] Fan, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294. MR1126325 https://doi.org/10.1214/aos/1176348249

[17] Gao, W., Oh, S. and Viswanath, P. (2018). Breaking the bandwidth barrier: Geometrical adaptive entropy estimation. *IEEE Trans. Inform. Theory* **64** 3313–3330. MR3798379 https://doi.org/10.1109/TIT.2018.2810313

[18] GAO, W., OH, S. and VISWANATH, P. (2018). Demystifying fixed k-nearest neighbor information estimators. *IEEE Trans. Inform. Theory* **64** 5629–5661. MR3832327 https://doi.org/10.1109/TIT.2018.2807481

[19] GOLDSTEIN, L. and MESSER, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.* **20** 1306–1328. MR1186251 https://doi.org/10.1214/aos/1176348770

[20] GYÖRFI, L. and VAN DER MEULEN, E. C. (1991). On the nonparametric estimation of the entropy functional. In *Nonparametric Functional Estimation and Related Topics* (*Spetses*, 1990). *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **335** 81–95. Kluwer Academic, Dordrecht. MR1154321 https://doi.org/10.1007/978-94-011-3222-0_6

[21] HALL, P. (1984). Limit theorems for sums of general functions of m-spacings. *Math. Proc. Cambridge Philos. Soc.* **96** 517–532. MR0757846 https://doi.org/10.1017/S0305004100062459

[22] HALL, P. and MARRON, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115. MR0907270 https://doi.org/10.1016/0167-7152(87)90083-6

[23] HALL, P. and MORTON, S. C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.* **45** 69–88. MR1220291 https://doi.org/10.1007/BF00773669

[24] HAN, Y., JIAO, J., MUKHERJEE, R. and WEISSMAN, T. (2017). On estimation of L_r-norms in Gaussian white noise models. Preprint. Available at arXiv:1710.03863.

[25] HAN, Y., JIAO, J. and WEISSMAN, T. (2016). Minimax rate-optimal estimation of divergences between discrete distributions. Preprint. Available at arXiv:1605.09124.

[26] HAN, Y., JIAO, J., WEISSMAN, T. and WU, Y. (2020). Supplement to "Optimal rates of entropy estimation over Lipschitz balls." https://doi.org/10.1214/19-AOS1927SUPP.

[27] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (2012). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics* **129**. Springer, New York. MR1618204 https://doi.org/10.1007/978-1-4612-2222-4

[28] HAS'MINSKIĬ, R. Z. and IBRAGIMOV, I. A. (1980). Some estimation problems for stochastic differential equations. In *Stochastic Differential Systems* (*Proc. IFIP-WG 7/1 Working Conf., Vilnius*, 1978). *Lecture Notes in Control and Information Sci.* **25** 1–12. Springer, Berlin. MR0609167

[29] HLAVÁČKOVÁ-SCHINDLER, K., PALUŠ, M., VEJMELKA, M. and BHATTACHARYA, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* **441** 1–46.

[30] IBRAGIMOV, I. A., NEMIROVSKIĬ, A. S. and KHAS'MINSKIĬ, R. Z. (1987). Some problems of nonparametric estimation in Gaussian white noise. *Theory Probab. Appl.* **31** 391–406.

[31] JAWERTH, B. (1977). Some observations on Besov and Lizorkin–Triebel spaces. *Math. Scand.* **40** 94–104. MR0454618 https://doi.org/10.7146/math.scand.a-11678

[32] JIAO, J., HAN, Y. and WEISSMAN, T. (2018). Minimax estimation of the L_1 distance. *IEEE Trans. Inform. Theory* **64** 6672–6706. MR3860754 https://doi.org/10.1109/TIT.2018.2846245

[33] JIAO, J., VENKAT, K., HAN, Y. and WEISSMAN, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inform. Theory* **61** 2835–2885. MR3342309 https://doi.org/10.1109/TIT.2015.2412945

[34] JOE, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.* **41** 683–697. MR1039399 https://doi.org/10.1007/BF00057735

[35] KANDASAMY, K., KRISHNAMURTHY, A., POCZOS, B. and WASSERMAN, L. (2015). Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems* 397–405.

[36] KERKYACHARIAN, G. and PICARD, D. (1996). Estimating nonquadratic functionals of a density using Haar wavelets. *Ann. Statist.* **24** 485–507. MR1394973 https://doi.org/10.1214/aos/1032894450

[37] KOROSTELËV, A. P. and TSYBAKOV, A. B. (2012). *Minimax Theory of Image Reconstruction. Lecture Notes in Statistics* **82**. Springer, New York. MR1226450 https://doi.org/10.1007/978-1-4612-2712-0

[38] KRASKOV, A., STÖGBAUER, H. and GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E* (3) **69** 066138, 16. MR2096503 https://doi.org/10.1103/PhysRevE.69.066138

[39] KRISHNAMURTHY, A., KANDASAMY, K., POCZOS, B. and WASSERMAN, L. (2014). Nonparametric estimation of Rényi divergence and friends. In *International Conference on Machine Learning* 919–927.

[40] LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659–681. MR1394981 https://doi.org/10.1214/aos/1032894458

[41] LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields* **113** 221–253. MR1670867 https://doi.org/10.1007/s004409970006

[42] LEVIT, B. Y. (1978). Asymptotically efficient estimation of nonlinear functionals. *Probl. Inf. Transm.* **14** 65–72. MR0533450

[43] MOON, K. R., SRICHARAN, K., GREENEWALD, K. and HERO, A. O. III (2016). Nonparametric ensemble estimation of distributional functionals. Preprint. Available at arXiv:1601.06884.

[44] MUKHERJEE, R., NEWEY, W. K. and ROBINS, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. Preprint. Available at arXiv:1705.07577.
[45] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math*. **1738** 85–277. Springer, Berlin. MR1775640
[46] NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist*. **24** 2399–2430. MR1425959 https://doi.org/10.1214/aos/1032181160
[47] PANINSKI, L. (2004). Estimating entropy on m bins given fewer than m samples. *IEEE Trans. Inform. Theory* **50** 2200–2203. MR2097210 https://doi.org/10.1109/TIT.2004.833360
[48] PANINSKI, L. and YAJIMA, M. (2008). Undersmoothed kernel entropy estimators. *IEEE Trans. Inform. Theory* **54** 4384–4388. MR2451978 https://doi.org/10.1109/TIT.2008.928251
[49] PATSCHKOWSKI, T. and ROHDE, A. (2016). Adaptation to lowest density regions with application to support recovery. *Ann. Statist*. **44** 255–287. MR3449768 https://doi.org/10.1214/15-AOS1366
[50] RAY, K. and SCHMIDT-HIEBER, J. (2019). Asymptotic nonequivalence of density estimation and Gaussian white noise for small densities. *Ann. Inst. Henri Poincaré Probab. Stat*. **55** 2195–2208. MR4029152 https://doi.org/10.1214/18-AIHP946
[51] RÉNYI, A. (1959). On the dimension and entropy of probability distributions. *Acta Math. Acad. Sci. Hung*. **10** 193–215. MR0107575 https://doi.org/10.1007/BF02063299
[52] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David A. Freedman. Inst. Math. Stat.* (*IMS*) *Collect*. **2** 335–421. IMS, Beachwood, OH. MR2459958 https://doi.org/10.1214/193940307000000527
[53] ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist*. **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515
[54] SINGH, S. and PÓCZOS, B. (2016). Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Advances in Neural Information Processing Systems* 1217–1225.
[55] SRICHARAN, K., RAICH, R. and HERO, A. O. III (2012). Estimation of nonlinear functionals of densities with confidence. *IEEE Trans. Inform. Theory* **58** 4135–4159. MR2943080 https://doi.org/10.1109/TIT.2012.2195549
[56] STEELE, J. M. (1986). An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist*. **14** 753–758. MR0840528 https://doi.org/10.1214/aos/1176349952
[57] STEIN, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions*. *Princeton Mathematical Series* **30**. Princeton Univ. Press, Princeton, NJ. MR0290095
[58] TCHETGEN, E., LI, L., ROBINS, J. and VAN DER VAART, A. (2008). Minimax estimation of the integral of a power of a density. *Statist. Probab. Lett*. **78** 3307–3311. MR2479495 https://doi.org/10.1016/j.spl.2008.07.001
[59] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794
[60] TSYBAKOV, A. B. and VAN DER MEULEN, E. C. (1996). Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Stat*. **23** 75–83. MR1380483
[61] VALIANT, G. and VALIANT, P. (2011). The power of linear estimators. In 2011 *IEEE* 52*nd Annual Symposium on Foundations of Computer Science—FOCS* 2011 403–412. IEEE Computer Soc., Los Alamitos, CA. MR2932716 https://doi.org/10.1109/FOCS.2011.81
[62] VAN ES, B. (1992). Estimating functionals related to a density by a class of statistics based on spacings. *Scand. J. Stat*. **19** 61–72. MR1172967
[63] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256
[64] WANG, Q., KULKARNI, S. R. and VERDÚ, S. (2009). Universal estimation of information measures for analog sources. *Found. Trends Commun. Inf. Theory* **5** 265–353.
[65] WU, Y. and YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory* **62** 3702–3720. MR3506758 https://doi.org/10.1109/TIT.2016.2548468
[66] WU, Y. and YANG, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist*. **47** 857–883. MR3909953 https://doi.org/10.1214/17-AOS1665