Optimal rates of entropy estimation over Lipschitz balls

Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu*

November 12, 2019

Abstract

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 $0 < s \leq 2$ 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 sub-exponential 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.

Contents

*Y. Han and T. Weissman are with the Department of Electrical Engineering, Stanford University, email: {yjhan,tsachy}@stanford.edu. J. Jiao is with the Department of Electrical Engineering and Computer Sciences and Department of Statistics, University of California Berkeley, email: jiantao@eecs.berkeley.edu. Y. Wu is with the Department of Statistics and Data Science, Yale University, email: yihong.wu@yale.edu. Y. Han and T. Weissman are supported in part by the NSF grants CCF-0939370 and CCF-1527105. J. Jiao is supported in part by the NSF grants CCF-0939270, CCF-1527105 and IIS-1901252. Y. Wu is supported in part by the NSF grants CCF-1527105, CCF-1900507, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship.

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 [BKB+93, VdV00], 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, e.g., [HM87, BR88, DN90, Fan91, BM95, KP96, Lau96, Nem00, CL03, CL05, TLRvdV08] and the references therein.

The monograph by [Nem00] 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 [RLTvdV08,RLM+17,MNR17], 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 [RLTvdV08] for a discussion on general classes of "smooth functionals").

In contrast, the results on optimal estimation of non-smooth functionals have been less comprehensive [HI80, Don97, KT12]. Notably, the seminal papers of [LNS99] and [CL11] considered the estimating of L_r-norms in Gaussian mean models. Subsequently, significant progress has been made on testing and estimation of non-smooth functionals, such as the Shannon entropy, support size, total variation and Kullback-Leibler (KL) divergence, for discrete distributions on large domains (see, e.g., [Pan04, VV11, JVHW15, WY16, HJW16, JHW18, BZLV18, WY19]).

An important non-smooth 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:

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

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 (c.f. [KSG04, CH04, HSPVB07] and the survey [BDGVdM97, WKV09]).

A prevalent assumption in nonparametric entropy estimation is that $f(x) \geq c$ everywhere for some constant $c > 0$ [Hal84, Joe89, VE92, SRH12, KKPW15], while others impose various assumptions quantifying on average how close the density is to zero [HM93, Lev78, TVdM96, EHHG09, GOV16, SP16, DF17, GOV18]. Assuming the density is bounded away from zero makes entropy a *smooth* functional, consequently, the general technique for estimating smooth nonparametric functionals [RLTvdV08, RLM+17, MNR17] 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 $d \in \mathbb{N} \triangleq \{1, 2, \cdots\}$. The Lipschitz ball is defined as

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

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

$$\|f\|_{\mathrm{Lip}_{s,p,d}} \triangleq \|f\|_p + \sup_{t>0} t^{-s} \omega_r(f,t)_p, \tag{2}$$

$$\omega_r(f,t)_p \triangleq \sup_{e \in \mathbb{R}^d, |e| \leq 1} \|\Delta_{te}^r f(\cdot)\|_p, \tag{3}$$

$$\Delta_h^r f(x) \triangleq \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} f\left(x + \left(k - \frac{r}{2}\right) h\right), \qquad h \in \mathbb{R}^d. \tag{4}$$

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

$$c((n\ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L) \le \left(\inf_{\hat{H}} \sup_{f \in \mathrm{Lip}_{s,p,d}(L)} \mathbb{E}_f (\hat{H} - H(f))^2 \right)^{\frac{1}{2}} \le C((n\ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L). \tag{5}$$

Moreover, the lower bound part of (5) holds for any $s > 0, 1 \le p < \infty$.

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

$$\Theta\left((n\ln n)^{-\frac{s}{s+d}} (L')^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L\right). \tag{6}$$

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

$$\Theta\left((n\ln n)^{-\frac{s}{s+d}} \left(R^{s+d(1-1/p)} L\right)^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln \left(R^{d(1-1/p)} L\right)\right). \tag{7}$$

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 $s \ge d$. Moreover, when $d \ge 3$, 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 (5) and (6) do not depend on the norm parameter p under the assumption that $2 \le p < \infty$. Another observation from Remark 2 is that the level of smoothness required for the parametric rate is $s \ge d$, which is more than $s \ge d/4$ that suffices for densities bounded away from zero on the support $[0,1]^d$ [Lau96], and also more than $s \ge d/2$ that suffices for densities satisfying a relative version of Hölder smoothness [BSY19].

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 $k \le \ln n$, the k-th power of $f_h(x)$ admits an unbiased estimator using a U-statistic, which enables

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

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 [JVHW15, WY16].

We improve the best known minimax lower bound for estimating non-smooth nonparametric functionals. The well-known lower bound $\Theta(n^{-4s/(4s+d)} + n^{-1/2})$ [BM95], 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, i.e., 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 sub-exponential and sub-Gaussian class, respectively. Consider the following class of densities:

$$\mathrm{Lip}^{\Psi}_{s,p,d}(L) \triangleq \{f : \|f\|_{\mathrm{Lip}_{s,p,d}} \leq L\} \cap \left\{f : \int_{\mathbb{R}^d} \Psi\left(|x|\right) f(x)dx \leq L\right\}, \tag{8}$$

where $\|\cdot\|_{\mathrm{Lip}_{s,p,d}}$ is the Lipschitz norm defined in (2). Note that the second constraint of (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 $d \in \mathbb{N}, 0 < s \leq 2, 2 \leq p < \infty$, there exist constants $c, C, L_0 > 0$ depending on $s, p, d, \kappa(\Psi), \Psi(1)$ such that if $\Psi^{-1}(n) \geq 1$ and $L \geq L_0$, then*

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

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

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

$$\hat{f}_h(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{9}$$

[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 $\frac{1}{n}\sum_{i=1}^{n}\log\hat{f}(X_i)$.

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 (that is, 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 $s \in (0,2], p \geq 2$, let $\hat{f}_h(x)$ be given in (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 $L \leq n^{s/d}$,*

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

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 $L \geq L_0$,

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

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 sub-optimal: 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 $\sup_{x \neq y} |f^{(r)}(x) - f^{(r)}(y)|/|x-y|^\alpha \leq L$. 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 non-integer $s > 0$.[3]

Radius of the Lipschitz ball: The assumption $L \leq (n \ln n)^{s/d}$ ensures that the minimax rate in Theorem 1 is $O(1)$, and $L \geq L_0$ is not superfluous as well. Indeed, if $sp \geq d$, then by standard embedding results of Lipschitz (or Besov) spaces [Jaw77], 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 [RLTvdV08, RLM+17, MNR17], 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 non-negative 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 $0 < s < \infty, 1 \leq p < \infty$.

[3] As opposed to the definition of the Lipschitz ball in (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 $\mathrm{B}^s_{p,\infty,d}$ [DL93]. These two definitions are equivalent for non-integer s, while for integer s the latter is strictly bigger. In this paper we adopt the former definition in (2) to avoid some technical subtleties.

[4] In fact, [Jaw77, Theorem 2.1] states that if $sp \geq d$, $\|f(\cdot) - f(\cdot - t)\|_\infty \leq C(s,p,d)\|f(\cdot) - f(\cdot - t)\|_{\mathrm{Lip}_{s,p,d}}$ for any $t \in [0,1]^d$. Since there must be some $x_0 \in [0,1]^d$ such that $f(x_0) \geq 1$, we conclude that $f(x) \geq 1 - 2C(s,p,d)L$ for almost every $x \in [0,1]^d$, which is bounded from below by a constant if L is sufficiently small.

The norm condition $p \in [2, \infty)$: Our current upper bound requires $p \geq 2$, 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" [KKPW14]. The effect of the tail behavior on the minimax rates is precisely quantified in Theorem 2 for densities with unbounded support.

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 [BDGVdM97], there exist two main approaches, based on either kernel density estimators, e.g. [Joe89, GVdM91, HM93, PY08, KKPW15] or nearest neighbor methods, e.g. [TVdM96, SRH12, SP16, DF17, GOV18, BSY19]. Among these works, some focus on the consistency [GVdM91, PY08], $\sqrt{n}$-consistency [Joe89, TVdM96], or the asymptotic efficiency [HM93, BSY19] of the proposed estimator, while others work on the minimax rate [SRH12, KKPW15, SP16, DF17, GOV18].

Similar estimator constructions have appeared in the literature. Asymptotic efficient estimators are obtained in [INH87, Nem00] for smooth functionals by means of Taylor expansion; [LNS99] and [CL11] estimated the L_1 norm of the mean in Gaussian white noise model using trigonometric polynomial approximation. One related work [HJMW17] deserves special attention. Dealing with the Gaussian white noise model, [HJMW17] 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 [LNS99]. Although both papers use the polynomial approximation technique for the upper and lower bound construction (which trace back to earlier work of [LNS99, CL11, JVHW15, WY16]), there exist significant distinctions between this work and [HJMW17]. 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 [PR16] 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 [Nus96] 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 [RSH18]. In fact, in contrast to the conclusion of Remark 2, it is shown in [HJMW17] 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 non-negative 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, i.e., 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 non-negative 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.

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

$$f_h(x) = \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x-y}{h}\right) f(y) dy, \tag{10}$$

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

$$\hat{f}_h(x) \triangleq \frac{1}{n} \sum_{i=1}^n K_h(x - X_i), \tag{11}$$

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

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 [WY16, JVHW15] 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 [HKPT12, Theorem 8.1]:

Lemma 1. *Fix any $s > 0$ and any kernel $K : \mathbb{R}^d \to \mathbb{R}$ which satisfies $\int_{\mathbb{R}^d} |x|^{\lceil s \rceil} |K(x)| dx < \infty$ and maps any polynomial q in d variables of degree at most $\lceil s \rceil - 1$ to themselves, i.e., $\int_{\mathbb{R}^d} q(y) K(x - y) dy = q(x)$. Then for any $f \in \mathrm{Lip}_{s,p,d}(L)$ and f_h defined in (10), 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. *Non-negativity: $K(t) \geq 0$ 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, e.g., 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, i.e., $\int_{\mathbb{R}^d} (a^\top x + b) K(x - y) dx = a^\top y + b$, and the first requirement (non-negativity) is crucial for proving the concentration result in Section 3. In fact, the non-negativity requirement is the key reason why we need to impose the assumption $s \leq 2$, 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 $s \leq 2$, we have

$$\|f - f_h\|_p \lesssim L h^s. \tag{12}$$

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 (10) and (11). 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 (11). The estimator $\hat{f}_{h,1}(x)$ will be used for classifying smooth versus non-smooth regime, and the other two for estimation.

3. Regime classification and estimator construction:

   - "Non-smooth" 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 $[0, \frac{2c_1 \ln n}{nh^d}]$:

$$Q = \sum_{l=0}^{k} a_l t^l = \arg \min_{P \in \mathsf{Poly}_k} \max_{t \in [0, \frac{2c_1 \ln n}{nh^d}]} | - t \ln t - P(t)|, \tag{13}$$

     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:

$$\hat{H}_1(x) = \sum_{l=0}^{k} a_l \left(\binom{n}{l}^{-1} \sum_{J \in \binom{[n]}{l}} \prod_{j \in J} K_h(x - X_j^{(2)}) \right). \tag{14}$$

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

$$\begin{aligned} \hat{H}_2(x) = \mathbb{1}(\hat{f}_{h,2}(x) \geq \frac{c_1 \ln n}{4nh^d}) \cdot \Bigg\{ & - \hat{f}_{h,2}(x) \ln \hat{f}_{h,2}(x) \\ & - (1 + \ln \hat{f}_{h,2}(x))(\hat{f}_{h,3}(x) - \hat{f}_{h,2}(x)) - \frac{1}{2} \Bigg(\hat{f}_{h,2}(x) \\ & - 2\hat{f}_{h,3}(x) + \frac{1}{\binom{n}{2} \hat{f}_{h,2}(x)} \sum_{i<j} K_h(x - X_i^{(3)}) K_h(x - X_j^{(3)}) \Bigg) \Bigg\}. \end{aligned} \tag{15}$$

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

$$\begin{aligned} \hat{H}(x) \triangleq \min \left\{ \hat{H}_1(x), \frac{1}{n^{1-2\varepsilon} h^d} \right\} \mathbb{1} \left(\hat{f}_{h,1}(x) < \frac{c_1 \ln n}{nh^d} \right) \\ + \hat{H}_2(x) \mathbb{1} \left(\hat{f}_{h,1}(x) \geq \frac{c_1 \ln n}{nh^d} \right). \end{aligned} \tag{16}$$

Finally, choose

$$h = c_0 (Ln \ln n)^{-\frac{1}{s+d}}, \quad k = \lceil c_2 \ln n \rceil, \tag{17}$$

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 Lemma 7–8) and output the estimator

$$\hat{H} = \int_{\mathbb{R}^d} \hat{H}(x) dx. \tag{18}$$

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

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

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

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

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 k-th order U-statistics, note that it is simply the (normalized) k-th 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 $l = 1, \cdots, k$, 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

$$\hat{H}_1(x) = \sum_{l=0}^{k} a_l \cdot \frac{Z_x \cdot (Z_x - 1) \cdot \ldots \cdot (Z_x - l + 1)}{h^{ld} \cdot n \cdot (n-1) \cdot \ldots \cdot (n - l + 1)}, \tag{19}$$

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 non-smooth regime In the non-smooth regime (i.e., $\hat{f}_{h,1}(x) \le \frac{c_1 \ln n}{nh^d}$) 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 non-smooth 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 [JVHW15]). The coefficients can also be pre-computed 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 [HJW16] to correct the bias. Specifically, by Taylor expansion we can write

$$\begin{aligned} \phi(f_h(x)) &\approx \sum_{l=0}^{R} \frac{\phi^{(l)}(\hat{f}_{h,2}(x))}{l!} (f_h(x) - \hat{f}_{h,2}(x))^l \\ &= \sum_{l=0}^{R} \frac{\phi^{(l)}(\hat{f}_{h,2}(x))}{l!} \sum_{j=0}^{l} \binom{l}{j} f_h(x)^j (-\hat{f}_{h,2}(x))^{l-j}. \end{aligned} \tag{20}$$

A natural idea to de-bias is to find an unbiased estimator of the right-hand side in (20). 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 de-biasing purposes).

Choice of bandwidth As will be clarified in Section 3 (cf. (36)), the bandwidth $h \asymp (n \ln n)^{-1/(s+d)}$ in (17) 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 $p \in [1, \infty)$ [Nem00], our choice of the bandwidth results in an "undersmoothed" kernel estimator of the density, which is consistent with the findings in [GM92, PY08] 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 [BR88, KP96, RLTvdV08, MNR17], 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 [LNS99, HJMW17].

Final integration The estimator $\hat{H}(x)$ in (16) 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 (19) 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 (18):*

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

where $C = C(s, p, d) > 0$ is independent of n, L (we omit the dependence of C on the choice of parameters $c_0, c_1, c_2, \varepsilon$ 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 (10). There exists a constant $C > 0$ independent of h, such that*

$$|H(f) - H(f_h)| = \left| \int_{\mathbb{R}^d} f(x) \ln f(x) dx - \int_{\mathbb{R}^d} f_h(x) \ln f_h(x) dx \right| \leq C \cdot Lh^s \tag{21}$$

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. (12)), 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 (21) to an inequality that no longer involves the kernel;

2. By the equivalence between the K-functional and the modulus of continuity [DL93], we approximate f by a non-negative 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:
$$I(A; B) = D(P_{AB} \| P_A \otimes P_B) = \mathbb{E}\left[\ln \frac{dP_{AB}}{dP_A dP_B}\right].$$

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,
$$0 \leq H(f_h) - H(f) = I(U; X + hU). \tag{22}$$

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
$$\begin{aligned}
I(U; X + hU) &= \mathbb{E}_U \left[\int_{[0,1]^d} f(x - hU) \ln \frac{f(x - hU)}{f_h(x)} dx\right] \\
&= \mathbb{E}_U \left[\int_{[0,1]^d} f(x - hU) \ln \frac{f(x - hU)}{w(x)} dx\right] - D(f_h \| w) \\
&\overset{(a)}{\leq} \mathbb{E}_U \left[\int_{[0,1]^d} f(x - hU) \ln \frac{f(x - hU)}{w(x)} dx\right] \\
&\overset{(b)}{\leq} \mathbb{E}_U \left[\int_{[0,1]^d} \frac{(f(x - hU) - w(x))^2}{w(x)} dx\right]
\end{aligned} \tag{23}$$

where (a) follows from the non-negativity 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 $w(x) \ge \max\{f(x)/2, Lh^s/4\}$ for all $x \in [0,1]^d$ and $\|w - f\|_\infty \le Lh^s/4$. Such an existence may be constructed by adding $Lh^s/4$ to $f(x)$ for all x with $f(x) \le Lh^s/4$, and then subtracting $Lh^s/4$ from $f(x)$ for a subset of $\{x \in [0,1]^d : f(x) \ge Lh^s/2\}$ to arrive at a density. As a result,

$$\begin{aligned}
\int_{[0,1]^d} \frac{(f(x-hU) - w(x))^2}{w(x)} dx &= \int_{[0,1]^d} \frac{(f(x-hU) - f(x) + f(x) - w(x))^2}{w(x)} dx \\
&\le 2\int_{[0,1]^d} \frac{(f(x-hU) - f(x))^2}{w(x)} dx + 2\int_{[0,1]^d} \frac{(f(x) - w(x))^2}{w(x)} dx \\
&\le 2\int_{[0,1]^d} \frac{(f(x-hU) - f(x))^2}{w(x)} dx + \frac{2\|f-w\|_\infty^2}{Lh^s/4} \\
&\lesssim \int_{[0,1]^d} \frac{(f(x-hU) - f(x))^2}{w(x)} dx + Lh^s. \qquad (24)
\end{aligned}$$

Combining (22)–(24), 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 $u \in \mathbb{R}$,

$$\int_{[0,1]^d} \frac{(f(x+hu) - f(x))^2}{f(x) + Lh^s} dx \lesssim Lh^s(1 + |u|^2). \qquad (25)$$

Note that (25) no longer involves the kernel $K(\cdot)$.

We provide some insights why (25) 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 (25) clearly holds. However, when $1 < s \le 2$, we will only have $|f(x+hu) - f(x)| \lesssim Lh|u|$ in general, and (25) cannot be derived by this simple approach. The crux of proving (25) for $1 < s \le 2$ is that, when $f(x)$ is close to zero, the difference $|f(x+hu) - f(x)|$ also need to be small to maintain the non-negativity 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 (25).

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 non-negative 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 non-negative function $g \in C^2(\mathbb{R}^d)$ such that*

$$\|f - g\|_p \le CLh^s, \qquad (26)$$

$$\|\|\nabla^2 g(\cdot)\|_{\mathrm{op}}\|_p \le CLh^{s-2}. \qquad (27)$$

Note that Lemma 3 is essentially the equivalence between the K-functional and the modulus of smoothness (Lemma 14 in Appendix A), with an extra constraint that g being non-negative, which turns out to be crucial in proving the inequality (29) below.

Let g be given by Lemma 3. Then

$$\begin{aligned}
&\int_{[0,1]^d} \frac{(f(x+hu)-f(x))^2}{f(x)+Lh^s}dx \\
&\le 3\int_{[0,1]^d} \frac{(f(x+hu)-g(x+hu))^2+(f(x)-g(x))^2+(g(x+hu)-g(x))^2}{f(x)+Lh^s}dx \\
&\le 3\int_{[0,1]^d} \frac{(f(x+hu)-g(x+hu))^2+(f(x)-g(x))^2}{Lh^s}dx \\
&\quad + 6\int_{[0,1]^d} \frac{(h\nabla g(x)^\top u)^2+(g(x+uh)-g(x)-h\nabla g(x)^\top u)^2}{f(x)+Lh^s}dx \\
&\le \frac{6\|f-g\|_2^2}{Lh^s} + \frac{6}{Lh^s}\|\rho(\cdot,u)\|_2^2 + 6h^2|u|^2\int_{[0,1]^d}\frac{|\nabla g(x)|^2}{f(x)+Lh^s}dx, \qquad (28)
\end{aligned}$$

where

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

We bound the three terms in (28) separately. By (26), 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

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

where (a) follows from the convexity of norms and (b) follows from (27). Thus the first two terms in (28) are both upper bounded by $O(Lh^s|u|^2)$. Hence, to show (25), it remains to prove that

$$\int_{[0,1]^d}\frac{|\partial_i g(x)|^2}{f(x)+Lh^s}dx \lesssim Lh^{s-2}, \qquad \forall i\in[d] \qquad (29)$$

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

3.1.3 Application of the Hardy-Littlewood maximal inequality

Finally, we use the non-negativity of g and the Hardy-Littlewood maximal inequality [Ste79] to prove (29). Fix any $\tau > 0$ to be optimized later. Since g is non-negative, we have

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

and thus

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

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

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

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

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

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:

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

Plugging this bound into (29) and using the triangle inequality, we have

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

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

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

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

$$\int_0^1 \left(\int_{-1}^1 |\partial_{ii} g(x + t \cdot \tau_x e_i)| dt \right)^2 dx_i \le C \int_0^1 |\partial_{ii} g(x)|^2 dx_i \tag{30}$$

for some constant $C > 0$. In fact, if (30) 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 (29).

The proof of (30) requires the introduction of the maximal inequality. Fix any $x_{\setminus i} \in [0,1]^{d-1}$ and define $h(y) \triangleq |\partial_{ii} g(x_{\setminus i}, y)|$, (30) is equivalent to

$$\int_0^1 \left(\frac{1}{2\tau_x} \int_{x-\tau_x}^{x+\tau_x} h(y) dy \right)^2 dx \le \frac{C}{4} \int_0^1 |h(x)|^2 dx. \tag{31}$$

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

$$M[h](y) \triangleq \sup_{t>0} \frac{1}{2t} \int_{y-t}^{y+t} |h(z)| dz. \tag{32}$$

Next we recall the maximal inequality on the real line [Ste79]:

Lemma 4. *For any non-negative 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 (31) (and thus (30)), 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 (25) 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 (25).

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 non-negative 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 non-smooth regimes in (16), the probability of making an error in classification is negligible.

Now we prove that our estimators $\hat{H}_1$ and $\hat{H}_2$ in (14)–(15) perform well in the corresponding regimes. To bound the variance, we invoke the well-known Efron–Stein–Steele inequality:

Lemma 6 ([Ste86])**.** *Let $X_1, X_2, \cdots, X_n$ be independent random variables, and for $i = 1, \cdots, n$, let X_i' be an independent copy of X_i. Then for any f,*

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

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 (Non-smooth 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*

$$|\mathbb{E}\hat{H}_1(x) + f_h(x) \ln f_h(x)| \lesssim \frac{1}{nh^d \ln n},$$
$$\mathbb{E}[(\hat{H}_1(x) - \hat{H}_1'(x))^2] \lesssim \frac{1}{n^{3-\varepsilon}h^{2d}}.$$

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*

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

For the final pointwise estimator $\hat{H}(x)$ in (16), 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 Lemma 5–8:

Corollary 1. *Under the assumptions of Lemma 5–8, we have*

$$\mathbb{E}\left(\mathbb{E}[\hat{H}(x)|X^{(1)}] + f_h(x) \ln f_h(x)\right)^2 \lesssim \frac{1}{(nh^d \ln n)^2}, \tag{33}$$
$$\mathbb{E}[(\hat{H}(x) - \hat{H}'(x))^2] \lesssim \frac{1}{n^{3-\varepsilon}h^{2d}} + \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n^2 h^d}. \tag{34}$$

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):

$$\begin{aligned}\mathbb{E}(\hat{H} - H(f))^2 &\le 2\left[(H(f_h) - H(f))^2 + \mathbb{E}(\hat{H} - H(f_h))^2\right] \\ &= 2\left[(H(f_h) - H(f))^2 + \mathbb{E}(\mathbb{E}[\hat{H}|X^{(1)}] - H(f_h))^2 + \mathbb{E}[\mathsf{Var}(\hat{H}|X^{(1)})]\right].\end{aligned} \tag{35}$$

We analyze different types of errors in (35) separately:

- First-stage approximation error: by Lemma 2 we know that
$$|H(f) - H(f_h)| \lesssim Lh^s.$$
- Conditional bias (second-stage approximation error): by Corollary 1 and Cauchy–Schwarz,
$$\begin{aligned}\mathbb{E}(\mathbb{E}[\hat{H}|X^{(1)}] - H(f_h))^2 &= \mathbb{E}\left(\int_{[0,1]^d} \left(\mathbb{E}[\hat{H}(x)|X^{(1)}] + f_h(x) \ln f_h(x)\right) dx\right)^2 \\ &\le \int_{[0,1]^d} \mathbb{E}\left(\mathbb{E}[\hat{H}(x)|X^{(1)}] + f_h(x) \ln f_h(x)\right)^2 dx \\ &\lesssim \frac{1}{(nh^d \ln n)^2}.\end{aligned}$$

- 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 $i = 1, 2, \cdots, 2n$, define $\hat{H}_i(x)$ to be the pointwise estimator in (16) with i-th 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

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

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

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

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

$$\begin{aligned}\mathbb{E}[\mathsf{Var}(\hat{H}|X^{(1)})] &\lesssim nh^d\int_{[0,1]^d}\mathbb{E}(\hat{H}(x)-\hat{H}'(x))^2dx \\ &\overset{(34)}{\lesssim} \int_{[0,1]^d}\left(\frac{1}{n^{2-\varepsilon}h^d}+\frac{f_h(x)(1+(\ln f_h(x))^2)}{n}\right)dx \\ &\lesssim \frac{1}{n^{2-\varepsilon}h^d}+\frac{(\ln L)^2}{n},\end{aligned}$$

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. (12)).

Substituting all three types of error bounds into (35), we obtain

$$\left(\sup_{f\in\mathrm{Lip}_{s,p,d}(L)}\mathbb{E}_f\left(\hat{H}-H(f)\right)^2\right)^{\frac{1}{2}} \lesssim Lh^s+\frac{1}{nh^d\ln n}+\frac{1}{n^{1-\varepsilon/2}\sqrt{h^d}}+\frac{\ln L}{\sqrt{n}}. \tag{36}$$

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 (36) to obtain

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

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 (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

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

where the Lipschitz norm $\|\cdot\|_{\mathrm{Lip}^{\star}_{s,p,d}}$ is defined in the same way as (2) to (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 $f(x \mod 1) \triangleq f(x_1 - \lfloor x_1 \rfloor, x_2 - \lfloor x_2 \rfloor, \cdots, x_d - \lfloor x_d \rfloor)$, for $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$. 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 [KKPW14], 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 $n \in \mathbb{N}$,*

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

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 C.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:

[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 $\{x_1, \cdots, x_{3^d}\} = \{-1, 0, 1\}^d$. Consequently, one can choose $C = 3^d$.

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 $s > 2$ [Tsy09]. Concretely, Lemma 5 is closely related to the problem of adapting to the lowest values of density in density estimation [PR16]. It was conjectured in [PR16] 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 non-negative, (71) 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 [MSGHI16, DF17].

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 turn out to be sufficient for non-smooth functional estimation [HJMW17]. 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 [Tsy09]; in smooth functional estimation [BM95, TLRvdV08], the case of $s > 2$ is also not special. It is mainly in estimating non-smooth functionals that designing procedures that can *adapt* to low density regime becomes a crucial challenge [PR16].

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 [Rén59] 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 [JVHW15, WY16]. 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 $S = h^{-d}$ sub-cubes $I_1, \ldots, I_S$ of size h, where h is the "bandwidth" we will choose later. For $i = 1, \cdots, S$, define

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

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 $\hat{H}_{\mathsf{discrete}}$ [WY16, JVHW15] 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

$$\hat{H} = \hat{H}_{\mathsf{discrete}} + \ln h^d. \tag{37}$$

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 $d \in \mathbb{N}, s \in (0,1], p \geq 2$, the performance of the estimator $\hat{H}$ in (37) is given by*

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

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

A Auxiliary lemmas

The first lemma upper bounds the quantity $\int f(x)(\ln f(x))^2 dx$ for any density $f \in \mathrm{Lip}_{s,p,d}(L)$.

Lemma 10. *Let f be a density on $[0,1]^d$ with $\|f\|_p \leq L$ for some $p > 1$ and $L > 0$. Then*

$$\int_{[0,1]^d} f(x)(\ln f(x))^2 dx \leq \frac{4}{e^2} + \frac{1 + e^{1/(p-1)}}{(p-1)^2} + \left(\frac{p}{p-1} \right)^2 (\ln L)^2.$$

The following lemma presents the property of the best approximating polynomial of $-x \ln x$ on the interval $[0, \Delta]$, which follows from [Tim63, Section 7.5.4].

Lemma 11 (Best approximating polynomial [JVHW15, WY16])**.** *Let $\Delta \leq 1$. Denote by $Q(t) = \sum_{l=0}^k a_l t^l$ the best approximating polynomial of $-t \ln t$ on $[0, \Delta]$ in the uniform norm, we have*

$$\sup_{t \in [0,\Delta]} |Q(t) + t \ln t| \leq \frac{C\Delta}{k^2}.$$

where C is a universal constant. Moreover, the coefficients satisfy:

$$|a_l| \leq 2^{3k} \Delta^{1-l}, \qquad l = 0, 2, 3, \ldots, k$$
$$|a_1| \leq 2^{3k} - \ln \Delta.$$

Lemma 12 (Variance of second-order U-statistics)**.** *Let*

$$U_2 \triangleq \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} X_i X_j,$$

where $X_1, \ldots, X_n$ are i.i.d random variables with finite second moment. Then

$$\mathsf{Var}(U_2) \leq \frac{4(n-2)}{n(n-1)} (\mathbb{E}X_1^2)(\mathbb{E}X_1)^2 + \frac{2(\mathbb{E}X_1^2)^2}{n(n-1)}.$$

Lemma 13 (Bennett's inequality [BLM13])**.** *Let $X_1, \ldots, X_n \in [a, b]$ be independent random variables with*

$$\sigma^2 \triangleq \frac{1}{n} \sum_{i=1}^n \mathsf{Var}(X_i).$$

Then we have

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \right| \geq \varepsilon \right) \leq 2 \exp\left(- \frac{n\varepsilon^2}{2(\sigma^2 + (b-a)\varepsilon/3)} \right).$$

Finally, we need the equivalence between Peetre's K-functional and modulus of smoothness on $\mathbb{R}^d$. For a multi-index $\beta = (\beta_1, \ldots, \beta_d)$, we define $|\beta| \triangleq \sum_{i=1}^d |\beta_i|$, and write the differential operator $\prod_{i=1}^d (\partial/\partial x_i)^{\beta_i}$ as $D^\beta(\cdot)$. Now for $r \in \mathbb{N}$, the differential operator D^r is defined as

$$D^r f \triangleq \sup_{|\beta|=r} |D^\beta f|.$$

For $p \in [1, \infty]$ and $r \in \mathbb{N}$, the Peetre's K-functional for f defined on $\mathbb{R}^d$ is defined as

$$K_r(f, t^r)_p \triangleq \inf_g \|f - g\|_p + t^r \|D^r g\|_p$$

where the infimum is taken over all functions g defined on $\mathbb{R}^d$ such that $D^\beta g$ for any $|\beta| = r - 1$ is locally absolutely continuous. Also recall the definition of the modulus of smoothness $\omega_r(f, t)_p$ in (3).

Lemma 14. *For any $p \in [1, \infty]$ and $r \in \mathbb{N}$, there exist universal constants $M = M(r, p)$ and $t_0 = t_0(r, p)$ such that for any $0 < t < t_0$ and f defined on $\mathbb{R}^d$,*

$$M^{-1} K_r(f, t^r)_p \le \omega_r(f, t)_p \le M K_r(f, t^r)_p. \tag{38}$$

Furthermore,

$$\omega_r(f, t)_p \le M t^r \|D^r f\|_p, \quad 0 < t < t_0 \tag{39}$$

and

$$\omega_r(f, t)_p \le 2^r \|f\|_p, \quad t > 0. \tag{40}$$

For a proof this lemma, see [DL93, Chapter 6, Theorem 2.4] for the case of $d = 1$, and [JS77] for the general $\mathbb{R}^d$. Finally, (39) follows from (38) by choosing $g = f$, and (40) follows from the definitions (3)–(4) and the triangle inequality.

B Proof of lower bound

In this section we prove the following lower bound for entropy estimation:

Theorem 7. *Let $s > 0$ and $p \ge 1$. Then there exist constants $L_0 > 1$ and $c_0 > 0$ depending on s, p, d such that the following holds. For $L_0 \le L \le (n \ln n)^{s/d}$, let*

$$R_n \triangleq \left(\inf_{\hat{H}} \sup_{f \in \mathrm{Lip}_{s,p,d}(L)} \mathbb{E}_f \left(\hat{H} - H(f) \right)^2 \right)^{\frac{1}{2}}.$$

Then

- *For any $p \in [1, \infty)$, $R_n \ge c_0((n \ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L)$.*
- *For $p = \infty$, $R_n \ge c_0(n^{-\frac{s}{s+d}} (\ln n)^{-\frac{s+2d}{s+d}} L^{\frac{d}{s+d}} + n^{-\frac{1}{2}} \ln L)$.*

Note that Theorem 7 implies the lower bound part of Theorem 1, which in fact holds for any $s > 0$ and any $1 \le p < \infty$. The second term $\Omega(n^{-1/2} \ln L)$ is essentially the parametric rate, where the only non-trivial part is to establish the proportionality to $\ln L$. This is shown below by means of a two-point argument provided that L is at most polynomial in n:

Lemma 15. *Let $s > 0$, $p \geq 1$. For any $\alpha > 0$, there exist constants $L_0, c_0, n_0 > 0$ depending on s, p, α such that for any $n \geq n_0$ and $L_0 \leq L \leq n^\alpha$,*

$$R_n \geq c_0 n^{-\frac{1}{2}} \ln L.$$

Now our remaining goal is to show that for $p \in [1, \infty)$,

$$R_n \gtrsim (n \ln n)^{-\frac{s}{s+d}} L^{\frac{d}{s+d}}. \tag{41}$$

The outline for the proof of (41) is as follows: In Appendix B.1 we reduce the nonparametric problem to a parametric subproblem in the Poissonized sampling model. The minimax risk in the Poissonized sampling model is essentially no larger than[6] that in the original i.i.d. sampling model (cf. Lemma 17), and the central advantage of the Poissonized sampling model is that the sufficient statistics become independent (cf. (50)), which simplifies the lower bound construction. The minimax lower bound for the parametric submodel is proved by a generalized version of Le Cam's method involving a pair of priors, also known as the method of two fuzzy hypotheses [Tsy09]. In Appendix B.2 we construct the priors using duality to best approximation and Appendix B.3 finishes the proof.

B.1 Reduction to a parametric submodel

Let d_0, d_1, d_2, d_3 be positive constants to be specified later. Set

$$h = (d_0 L n \ln n)^{-\frac{1}{s+d}}, \quad S = (2h)^{-d}. \tag{42}$$

Without loss of generality we assume that S is an integer. Fix

$$\alpha \in \left(0, \frac{1}{S}\right). \tag{43}$$

Let $I_1, \ldots, I_S$ be the partition of the large cube $[\frac{1}{4}, \frac{3}{4}]^d$ into S smaller cubes of edge length h, and t_i denote the leftmost corner of sub-cube I_i, i.e., $I_i = t_i + [0, h]^d$. Let g and w be some fixed smooth probability density functions on $\mathbb{R}^d$ with finite entropy which vanish outside $[0,1]^d$ and $[0,1]^d \backslash [\frac{1}{4}, \frac{3}{4}]^d$, respectively. The smoothness of g and w on $\mathbb{R}^d$ implies that their derivatives of any order vanish at their respective boundary.

To each vector $P = (p_1, \ldots, p_S) \in [0, \frac{d_3 \ln n}{n}]^S$, we associate a function

$$f_P(t) \triangleq \sum_{i=1}^{S} p_i \cdot \frac{1}{h^d} g\left(\frac{t - t_i}{h}\right) + (1 - S\alpha)\, w(t). \tag{44}$$

By the choice of α, f_P is non-negative everywhere. Moreover, since w is smooth on $\mathbb{R}^d$ and vanishes outside $[0,1]^d$, the function f_P connects to zero smoothly on the boundary. The next result is a sufficient condition for f_P to lie in the Lipschitz ball.

Lemma 16. *If $L \geq L_0$ for some constant L_0 depending only on w,*

$$\frac{1}{S} \sum_{i=1}^{S} p_i^p \leq \left(\frac{2C_1}{n \ln n}\right)^p, \tag{45}$$

and $d_0 > 0$ is sufficiently small depending only on (g, L_0, C_1, s, p, d), then $f_P \in \mathrm{Lip}_{s,p,d}(L)$.

[6] In fact, the minimax risks of the i.i.d. sampling and Poissonized sampling models are closely related and it is simpler to show that the estimator constructed in Section 2 satisfies the same risk bound in the Poisson model; see a previous version of the current paper [HJWW17].

Proof. Let

$$\tilde{f}_P(t) \triangleq \sum_{i=1}^{S} p_i \cdot \frac{1}{h^d} g\left(\frac{t-t_i}{h}\right).$$

Then $f_P = \tilde{f}_P + (1-S\alpha)w$. Since $\|f_P\|_{\mathrm{Lip}} \le \|\tilde{f}_P\|_{\mathrm{Lip}} + (1-S\alpha)\|w\|_{\mathrm{Lip}}$ and $S\alpha \le 1$, after choosing $L_0 = 2\|w\|_{\mathrm{Lip}}$, it suffices to show $\tilde{f}_P \in \mathrm{Lip}_{s,p,d}(L/2)$ for $L \ge L_0$. Observe that

$$\|\tilde{f}_P\|_p = \frac{\|g\|_p}{h^d}\left(\frac{1}{S}\sum_{i=1}^{S} p_i^p\right)^{\frac{1}{p}} = h^s \|g\|_p \cdot d_0 L n \ln n \left(\frac{1}{S}\sum_{i=1}^{S} p_i^p\right)^{\frac{1}{p}}, \tag{46}$$

the condition (45) ensures that $h^{-s}\|\tilde{f}_P\|_p$ is upper bounded by $2d_0C_1\|g\|_p \cdot L$. By (39) in Lemma 14, (46) implies that $\omega_r(\tilde{f}_P, t) \le Lt^s/4$ for any $t \ge h$ and d_0 small. For $t < h$, further observe that for $r = \lceil s \rceil$, we have

$$\|D^r \tilde{f}_P\|_p = \frac{\|D^r g\|_p}{h^{d+r}}\left(\frac{1}{S}\sum_{i=1}^{S} p_i^p\right)^{\frac{1}{p}} = \frac{\|D^r g\|_p}{h^{r-s}} \cdot d_0 L n \ln n \left(\frac{1}{S}\sum_{i=1}^{S} p_i^p\right)^{\frac{1}{p}} \tag{47}$$

Hence, $h^{r-s}\|D^r \tilde{f}_P\|_p$ is upper bounded by $2d_0C_1\|D^r g\|_p \cdot L$. By (40) in Lemma 14, (47) implies that $\omega_r(\tilde{f}_P, t) \le Lt^s/4$ for any $t \le h$ and d_0 sufficiently small, completing the proof. □

Now we would like to reduce the original nonparametric model $f \in \mathrm{Lip}_{s,p,d}(L)$ to some parametric submodel $f \in \{f_P : P \in \mathcal{P}\}$, with a properly chosen $\mathcal{P}$, and impose an iid prior on the coefficient vector P with mean α. However, f_P need not be a valid density, since

$$\int_{[0,1]^d} f_P(t)dt = 1 + \sum_{i=1}^{S}(p_i - \alpha)$$

need not normalize to one. To resolve this issue, we show that the minimax rate remains unchanged as long as $\|f\|_1$ is sufficiently close to one. To this end, let us define a new model: instead of the original sampling model with a fixed sample size n, we draw $N \sim \mathsf{Poi}(n\|f\|_1)$ i.i.d. observations from the density $\frac{f}{\|f\|_1}$. In the new model, the parameter set of f is a parametric one $\{f_P : P \in \mathcal{P}\}$, where

$$\begin{aligned}\mathcal{P} \triangleq &\left\{P \in \left[0, \frac{d_3 \ln n}{n}\right]^S : \left|\sum_{i=1}^{S}(p_i - \alpha)\right| \le \frac{1}{nh^d(\ln n)^3 \ln L}\right\} \\ &\cap \left\{P : \frac{1}{S}\sum_{i=1}^{S}|p_i|^p \le \left(\frac{2C_1}{n \ln n}\right)^p\right\}\end{aligned} \tag{48}$$

with C_1 given in (53). In view of Lemma 16, we have $\{f_P : P \in \mathcal{P}\} \subseteq \mathrm{Lip}_{s,p,d}(L)$. We also extend the definition of entropy to positive functions verbatim: $H(f_P) = \int_{[0,1]^d} -f_P(t)\ln f_P(t)dt$, and denote the minimax risk restricted to the parametric submodel $f \in \{f_P : P \in \mathcal{P}\}$ as

$$R_n^P \triangleq \left(\inf_{\hat{H}} \sup_{P \in \mathcal{P}} \mathbb{E}_{f_P}(\hat{H} - H(f_P))^2\right)^{\frac{1}{2}}.$$

The following lemma relates the new minimax risk R_n^P to the original risk R_n in Theorem 7.

Lemma 17. *There exists a constant $C_2 > 0$ depending on s, p only that*

$$R_n^P \le C_2 \left(R_{\frac{n}{2}} + \frac{1}{nh^d (\ln n)^3} \right).$$

Since our goal (41) is to show $R_n \gtrsim \frac{1}{nh^d \ln n}$, by Lemma 17, it suffices to prove $R_n^P \gtrsim \frac{1}{nh^d \ln n}$. Note that

$$H(f_P) = C_0 + H(P) + (H(g) + d \ln h) \sum_{i=1}^{S} p_i$$

where the constant C_0 does not depend on P, and we denote, with a slight abuse of notation, the discrete entropy $H(P) \triangleq \sum_{i=1}^{S} -p_i \ln p_i$. Hence, given the minimax optimal estimator $\hat{H}$ for $H(f_P)$, the following estimator

$$\tilde{H} \triangleq \hat{H} - C_0 - (H(g) + d \ln h) \cdot S\alpha$$

for $H(P)$ satisfies

$$\begin{aligned}
(\mathbb{E}[(\tilde{H} - H(P))])^{1/2} &\le (\mathbb{E}[(\hat{H} - H(f_P))])^{1/2} + |H(g) + d \ln h| \cdot \left| \sum_{i=1}^{S} (p_i - \alpha) \right| \\
&\lesssim R_n^P + \frac{1}{\ln n} \cdot \frac{1}{nh^d \ln n}.
\end{aligned}$$

Therefore, to show (41), it suffices to prove

$$\inf_{\hat{H}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left(\hat{H} - H(P) \right)^2 \gtrsim \frac{1}{(nh^d \ln n)^2}. \tag{49}$$

Finally, we note that by the factorization theorem, for the Poisson sampling model, to estimate $H(P)$ with $P \in \mathcal{P}$, the histograms

$$h_j = \sum_{i=1}^{N} \mathbb{1}(X_i \in I_j), \qquad j = 1, \ldots, S$$

constitute a sufficient statistic. As a result, we may further assume that our observation model is

$$h_j \overset{\text{ind}}{\sim} \mathsf{Poi}(n \cdot p_j), \qquad j = 1, \ldots, S \tag{50}$$

and the estimator $\hat{H}$ in (49) is a function of $(h_1, \ldots, h_S)$.

B.2 Construction of two priors

The minimax lower bound (49) follows from a generalized version of Le Cam's method involving two priors, also known as the method of two fuzzy hypotheses [Tsy09]. Given a collection of distributions $\{P_\theta : \theta \in \Theta'\}$, suppose the observation $\mathbf{Z}$ is distributed as P_θ with $\theta \in \Theta \subseteq \Theta'$. Let $\hat{T} = \hat{T}(\mathbf{Z})$ be an arbitrary estimator of a function $T(\theta)$ based on $\mathbf{Z}$.

Denote the total variation distance between two probability measures P, Q by

$$V(P, Q) \triangleq \sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \frac{1}{2} \int |p - q| d\nu,$$

where $p = \frac{dP}{d\nu}, q = \frac{dQ}{d\nu}$, and ν is a dominating measure so that $P \ll \nu, Q \ll \nu$. The following general minimax lower bound follows from the same proof as [Tsy09, Theorem 2.15]:

Lemma 18. *Let σ_0 and σ_1 be prior distributions on Θ'. Suppose there exist $\zeta \in \mathbb{R}, s > 0, 0 \le \beta_0, \beta_1 < 1$ such that*

$$\sigma_0(\theta \in \Theta : T(\theta) \le \zeta - s) \ge 1 - \beta_0,$$
$$\sigma_1(\theta \in \Theta : T(\theta) \ge \zeta + s) \ge 1 - \beta_1.$$

Then

$$\inf_{\hat{T}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(|\hat{T} - T(\theta)| \ge s \right) \ge \frac{1 - V(F_1, F_0) - \beta_0 - \beta_1}{2},$$

where $F_i = \int P_\theta \sigma_i(d\theta)$ is the marginal distribution of $\mathbf{Z}$ under the prior σ_i for $i = 0, 1$, respectively.

To apply Lemma 18 to $\Theta' = [0, \frac{d_3 \ln n}{n}]^S$ and $\Theta = \mathcal{P}$, we first describe the construction of the two priors. The following result is simply the duality between the problem of best uniform approximation and moment matching.[7]

Lemma 19. *Given a compact interval $I = [a, b]$ with $a > 0$, integers $q, k > 0$ and a continuous function ϕ on I, let*

$$E_{q-1,k}(\phi; I) \triangleq \inf_{\{a_i\}} \sup_{x \in I} \left| \sum_{i=-q+1}^{k} a_i x^i - \phi(x) \right| \tag{51}$$

denote the best uniform approximation error of ϕ by rational functions spanned by $\{x^{-q+1}, x^{-q+2}, \ldots, x^k\}$. Let $E_k(\phi; I) \triangleq E_{q-1,k}(\phi; I)$ denote the best uniform approximation error of ϕ by degree-k polynomials. Then

$$\begin{aligned} 2E_{q-1,k}(\phi; I) = \max \ & \int \phi(t)\nu_1(dt) - \int \phi(t)\nu_0(dt) \\ \text{s.t.} \ & \int t^l \nu_1(dt) = \int t^l \nu_0(dt), \quad l = -q+1, \ldots, k \end{aligned} \tag{52}$$

where the maximum is taken over pairs of probability measures ν_0 and ν_1 supported on I.

Here we apply this lemma to $\phi_q(t) \triangleq t^{1-q} \ln t$ and

$$\eta = \frac{d_1}{(\ln n)^2}, \quad I = [\eta, 1], \qquad k = \lceil d_2 \ln n \rceil, \qquad q = \lceil p \rceil$$

with constants $d_1, d_2 > 0$ to be specified later. The following lemma provides a lower bound for the approximation error $\phi_q(t) \triangleq t^{1-q} \ln t$:

Lemma 20. *Fix $q \in \mathbb{N}$. There exists constants $c, c' > 0$ depending on q only such that*

$$\liminf_{k \to \infty} k^{-2(q-1)} E_{q-1,k} \left(\phi_q; \left[\frac{c}{k^2}, 1 \right] \right) \ge c'.$$

Choosing $d_1 = c/d_2^2$ with constant $c > 0$ given in Lemma 20 and in view of the definitions of I and k, we conclude that

$$E_{q-1,k}(\phi_q; I) \gtrsim (\ln n)^{2q-2}.$$

[7] The proof of Lemma 19 is identical to that of [WY16, Eqn. (34)], since $\{x^{-q+1}, x^{-q+2}, \ldots, x^k\}$ forms a Haar system [DL93, Section 3.3, Example 2] and hence the Chebyshev alternating theorem holds.

Let ν_0 and ν_1 be the maximizer of (52). We define probability measures $\tilde{\mu}_0, \tilde{\mu}_1$ on $I = [\eta, 1]$ by

$$\tilde{\mu}_i(dt) = \left(1 - \int \left(\frac{\eta}{t}\right)^q \nu_i(dt)\right) \delta_0(dt) + \left(\frac{\eta}{t}\right)^q \nu_i(dt) \qquad i = 0, 1.$$

where δ_0 is the delta measure at zero. It is straightforward to verify that $\tilde{\mu}_0$ and $\tilde{\mu}_1$ are probability measures, satisfying

1. $\int t^l \tilde{\mu}_1(dt) = \int t^l \tilde{\mu}_0(dt)$, for all $l = 0, 1, \ldots, q + k$;
2. $\int t \ln t \tilde{\mu}_1(dt) - \int t \ln t \tilde{\mu}_0(dt) \gtrsim (\ln n)^{-2}$;
3. $\int t^q \tilde{\mu}_i(dt) = \eta^q = d_1^q (\ln n)^{-2q}$, for $i = 0, 1$.

Finally, let μ_i be the dilation of $\tilde{\mu}_i$ by a factor of $\frac{d_3 \ln n}{n}$, that is, if $Y_i \sim \tilde{\mu}_i$, then $\frac{d_3 \ln n}{n} Y_i \sim \mu_i$. Then $\mathsf{supp}(\mu_i) \subseteq [0, \frac{d_3 \ln n}{n}]$ and, furthermore,

1. $\int t^l \mu_1(dt) = \int t^l \mu_0(dt)$, for all $l = 0, 1, \ldots, q + k$;
2. $\int t \ln t \mu_1(dt) - \int t \ln t \mu_0(dt) \gtrsim (n \ln n)^{-1}$;
3. $\int t^q \mu_i(dt) = (\frac{\eta d_3 \ln n}{n})^q = (d_1 d_3)^q (n \ln n)^{-q}$, for $i = 0, 1$.

In the next subsection we will impose the prior where the parameters $(p_1, \ldots, p_S)$ are iid as either μ_0 or μ_1. The utility of the above these properties are as follows: The first condition ensures the priors μ_1 and μ_2 have matching first $q + k$ moments, and the induced Poisson mixtures are exponentially close in total variation; this is exactly the distribution of the sufficient statistics $(h_1, \ldots, h_S)$. The second property ensures the separation of the functional values, while the third property ensures the smoothness of the functions f_P. Indeed, since $q \geq p$, Hölder's inequality yields

$$\int t \mu_i(dt) \leq \left(\int t^p \mu_i(dt)\right)^{\frac{1}{p}} \leq \left(\int t^q \mu_i(dt)\right)^{\frac{1}{q}} \leq \frac{C_1}{n \ln n} \tag{53}$$

where C_1 is a constant depending on d_2, d_3 and p. It implies that controlling the qth moment of μ_i for $q \geq p$ also controls all of its lower-order moments.

B.3 Minimax lower bound in the parametric submodel

In this subsection we invoke Lemma 18 to finish the proof of (49), thereby proving the lower bound in Theorem 1. Consider the probability measures μ_0, μ_1 constructed in Appendix B.2, and define

$$\Delta \triangleq \int t \ln t \mu_1(dt) - \int t \ln t \mu_0(dt) \gtrsim \frac{1}{n \ln n}. \tag{54}$$

Put $\alpha = \int t \mu_0(dt) = \int t \mu_1(dt)$ and recall the parameter space $\mathcal{P}$ defined in (48). By (53) and the assumption $L \leq (n \ln n)^{s/d}$, for large d_2 (and therefore small d_1) we have

$$0 \leq S\alpha \lesssim \frac{d_1 S}{n \ln n} \lesssim \frac{d_1}{n h^d \ln n} \leq 1,$$

which fulfills (43).

Let $\mu_i^{\otimes S}$ denote the S-fold product of μ_i. Consider the following event:

$$E_i \triangleq \{P : P \in \mathcal{P}\} \cap \left\{ P : \left| \frac{1}{S} \sum_{j=1}^{S} p_j \ln p_j - \mathbb{E}_{\mu_i} p \ln p \right| \le \frac{\Delta}{4} \right\}, \qquad i = 0, 1.$$

We first show that $\mu_i^{\otimes S}(E_i) \to 0$ as $n \to \infty$ for any $i = 0, 1$. Recall the definitions of S and h from (42). By Chebyshev's inequality and the fact that μ_i is supported on $[0, \frac{d_3 \log n}{n}]$, we have

$$\begin{aligned}
\mu_i^{\otimes S} \left\{ P : \left| \sum_{i=1}^{S} (p_i - \mathbb{E}_{\mu_i} p_i) \right| > \frac{1}{n h^d (\ln n)^3 \ln L} \right\} &\le (n h^d (\ln n)^3 \ln L)^2 \cdot \mathsf{Var}_{\mu_i^{\otimes S}} \left[\sum_{i=1}^{S} p_i \right] \\
&\le (n h^d (\ln n)^3 \ln L)^2 \cdot S \left(\frac{d_3 \ln n}{n} \right)^2 \\
&\lesssim h^d (\ln n)^8 (\ln L)^2 \to 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mu_i^{\otimes S} \left\{ P : \frac{1}{S} \sum_{i=1}^{S} |p_i|^p > \left(\frac{2C_1}{n \ln n} \right)^p \right\} \\
&\overset{(53)}{\le} \mu_i^{\otimes S} \left\{ P : \frac{1}{S} \sum_{i=1}^{S} (|p_i|^p - \mathbb{E}_{\mu_i} |p_i|^p) > \left(\frac{C_1}{n \ln n} \right)^p \right\} \\
&\le \left(\frac{C_1}{n \ln n} \right)^{-2p} \cdot \mathsf{Var}_{\mu_i^{\otimes S}} \left[\frac{1}{S} \sum_{i=1}^{S} |p_i|^p \right] \\
&\lesssim (n \ln n)^{2p} \cdot \frac{1}{S} \left(\frac{\ln n}{n} \right)^{2p} \\
&\asymp h^d (\ln n)^{4p} \to 0
\end{aligned}$$

and

$$\begin{aligned}
&\mu_i^{\otimes S} \left\{ P : \left| \frac{1}{S} \sum_{j=1}^{S} p_j \ln p_j - \mathbb{E}_{\mu_i} [p \ln p] \right| > \frac{\Delta}{4} \right\} \\
&\le \frac{16}{\Delta^2} \mathsf{Var}_{\mu_i^{\otimes S}} \left[\frac{1}{S} \sum_{j=1}^{S} p_j \ln p_j \right] \\
&\le \frac{16}{S \Delta^2} \cdot \left(\frac{(\ln n)^2}{n} \right)^2 \\
&\overset{(54)}{\lesssim} h^d (\ln n)^6 \to 0.
\end{aligned}$$

Hence, by the union bound, we have $\beta_i \triangleq \mu_i^{\otimes S}(E_i^c) \to 0$ for $i = 0, 1$.

Now we are ready to apply Lemma 18 to

$$\begin{aligned}
T(\theta) = H(P) &= \sum_{i=1}^{S} -p_i \ln p_i \\
\zeta &= \frac{\mathbb{E}_{\mu_1^{\otimes S}} H(P) + \mathbb{E}_{\mu_0^{\otimes S}} H(P)}{2}, \quad s = \frac{S\Delta}{4}
\end{aligned}$$

Under the prior $\sigma_i \triangleq \mu_i^{\otimes S}$, the sufficient statistics $(h_1, \ldots, h_S)$ are i.i.d. with distribution $F_i \triangleq \gamma_i^{\otimes S}$, where $\gamma_i \triangleq \int \mathsf{Poi}(n\lambda)\mu_i(d\lambda)$ is a Poisson mixture, in view of (50). Note that μ_0 and μ_1 have matching first $q+k$ moments. By [WY16, Lemma 3], we have

$$V(\pi_0, \pi_1) \le \left(\frac{2ed_3 \ln n}{q+k}\right)^{q+k} \le n^{-d_2 \log \frac{d_2}{2ed_3}}$$

provided that $d_2 \log \frac{d_2}{2ed_3} \ge 2$. Therefore $V(F_0, F_1) \le S \cdot V(\pi_0, \pi_1) \to 0$ by choosing constant $d_2 > 0$ large enough. Applying Lemma 18 together with Markov's inequality yields

$$\inf_{\hat{H}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left(\hat{H} - H(P)\right)^2 \gtrsim s^2 \cdot \inf_{\hat{H}} \sup_{P \in \mathcal{P}} \mathbb{P}\left(|\hat{H} - H(P)| \ge s\right) \gtrsim \frac{1}{(nh^d \ln n)^2},$$

which is (49), as desired.

Finally, the lower bound for $p = \infty$ follows from the same argument, except that we will set $q = 1$, choose the bandwidth differently

$$h = \left(\frac{\ln n}{d_0 L n}\right)^{\frac{1}{s+d}},$$

and replace the RHS of (53) by $\frac{d_3 \ln n}{n}$.

C Proof of Theorems 2, 3 and 6

C.1 Proof of Theorem 2

We prove the upper and lower bounds separately.

C.1.1 Proof of upper bound

First we propose the entropy estimator $\hat{H}$ for the unbounded support case. We begin with a lemma for the tail of density $f \in \mathrm{Lip}^{\Psi}_{s,p,d}(L)$.

Lemma 21. *Let $f \in \mathrm{Lip}^{\Psi}_{s,p,d}(L)$ with $p \in [2, \infty)$ and Orlicz function Ψ of rapid growth. There exist constants $C_0, C_1, C_2 > 0$ depending on $p, d, \kappa(\Psi), \Psi(1), L$ only, such that*

$$\left| \int_{|x| \ge C_0 \Psi^{-1}(n)} f(x) \ln \frac{1}{f(x)} dx \right| \le \frac{C_1}{n^4}, \tag{55}$$

$$\int_{\mathbb{R}^d} f(x) \ln^2 f(x) dx \le C_2. \tag{56}$$

Proof. We first prove the second inequality (56). Since $t \ln^2 t \le t^{1/2} + t^{3/2}$ for all $t \ge 0$,

$$\int_{\mathbb{R}^d} f(x) \ln^2 f(x) dx \le \int_{\mathbb{R}^d} f(x)^{\frac{3}{2}} dx + \int_{\mathbb{R}^d} f(x)^{\frac{1}{2}} dx. \tag{57}$$

By Cauchy–Schwartz and $\|f\|_2 \le \|f\|_p \le L$, we have

$$\int_{\mathbb{R}^d} f(x)^{\frac{3}{2}} dx \le \left(\int_{\mathbb{R}^d} f(x)^2 dx\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} f(x) dx\right)^{\frac{1}{2}} \le L. \tag{58}$$

For the second integral, by the convexity of Ψ and $\Psi(0) = 0$ we know that there is a constant $a > 0$ depending only on $\Psi(1)$ such that $\Psi(a) \geq 2$. Then by Cauchy–Schwartz again,

$$\int_{|x|\leq a} f(x)^{\frac{1}{2}} dx \leq (2a)^{\frac{d}{2}} \left(\int_{|x|\leq a} f(x) dx \right)^{\frac{1}{2}} \leq a^{\frac{d}{2}}. \tag{59}$$

Since for any $x \in \mathbb{R}^d$, we have

$$2f(x)^{\frac{1}{2}} \leq \Psi(|x|) f(x) + \frac{1}{\Psi(|x|)}.$$

Integrating over $|x| > a$ gives

$$\begin{aligned} 2\int_{|x|>a} f(x)^{\frac{1}{2}} dx &\leq \int_{|x|>a} \Psi(|x|) f(x) dx + \int_{|x|>a} \frac{1}{\Psi(|x|)} dx \\ &\leq L + \sum_{k=0}^{\infty} \int_{\kappa^k a < |x| \leq \kappa^{k+1} a} \frac{1}{\Psi(|x|)} dx \\ &\leq L + \sum_{k=0}^{\infty} \frac{(2\kappa^{k+1} a)^d}{\Psi(\kappa^k a)} < \infty, \end{aligned} \tag{60}$$

where the finiteness follows from $\Psi(\kappa^k a) \geq \Psi(a)^{2^k} \geq 2^{2^k}$ by the rapid growth assumption with parameter $\kappa = \kappa(\Psi)$. Combining (57)–(60) gives (56).

To establish the first inequality (55), let $X \sim f$. Then by Cauchy–Schwartz,

$$\begin{aligned} \left| \int_{|x|\geq C_0 \Psi^{-1}(n)} f(x) \ln \frac{1}{f(x)} dx \right| &= \left| \mathbb{E}\left[\ln \frac{1}{f(X)} \mathbb{1}(|X| \geq C_0 \Psi^{-1}(n)) \right] \right| \\ &\leq \sqrt{\mathbb{E}[(\ln f(X))^2] \cdot \mathbb{P}(|X| \geq C_0 \Psi^{-1}(n))}. \end{aligned} \tag{61}$$

The probability $\mathbb{P}(|X| \geq C_0 \Psi^{-1}(n))$ can be upper bounded easily:

$$\mathbb{P}(|X| \geq C_0 \Psi^{-1}(n)) \leq \frac{1}{\Psi(C_0 \Psi^{-1}(n))} \int_{\mathbb{R}^d} \Psi(|x|) f(x) dx \leq \frac{L}{\Psi(C_0 \Psi^{-1}(n))}.$$

By the rapid growth property of Ψ, choosing $C_0 = \kappa^3$ we have $\Psi(C_0 \Psi^{-1}(n)) \geq n^8$, and therefore

$$\mathbb{P}(|X| \geq C_0 \Psi^{-1}(n)) \leq \frac{L}{n^8}. \tag{62}$$

Using the upper bound of $\mathbb{E}[(\ln f(X))^2]$ in (56), a combination of (61) and (62) gives (55) and completes the proof of the lemma. □

Let C_0 be given in Lemma 21, and $R \triangleq C_0 \Psi^{-1}(n) \geq 1$. We define the entropy estimator $\hat{H}$ by

$$\hat{H} = \int_{|x|\leq R} \hat{H}(x) dx,$$

where the pointwise estimator $\hat{H}(x)$ is defined in (16), with the same parameters except that the bandwidth h is chosen to be

$$h = c_0 (n \ln n)^{-\frac{1}{s+d}} \cdot R^{\frac{d}{p(s+d)}}. \tag{63}$$

To show the upper bound, first note that if the density f is indeed supported on $[-R,R]^d$, then the scaled density $g(x) = f(R(2x-1))$ is supported on $[0,1]^d$, with $g \in \mathrm{Lip}_{s,p,d}(L(2R)^{s+d(1-1/p)})$ and $\|g\|_p \le L(2R)^{d(1-1/p)}$. Moreover, the rapid growth property of Ψ ensures that $R = o(n^{\varepsilon})$ for any $\varepsilon > 0$, where the hidden constant depends on $\Psi(1), \kappa(\Psi)$ and ε. Following the same analysis in Section 3, and noting that the scaled bandwidth becomes $h/(2R)$, we have

$$\left(\mathbb{E}_f\left(\hat{H} - H(f)\right)^2\right)^{\frac{1}{2}} \le C\left(R^{s+d(1-1/p)} \cdot (h/R)^s + \frac{1}{n(h/R)^d \ln n} + \frac{1}{\sqrt{n}}\right)$$

for some $C > 0$ depending on $s, p, d, \kappa(\Psi), \Psi(1), L$, where the hidden constant in the $O(n^{-1/2})$ term is thanks to (56) in Lemma 21. Hence, plugging in the bandwidth in (63) gives the desired upper bound. In the general case, we may simply truncate the density f onto the set $\{x : |x| \le R\}$ and apply the same analysis[8], with the additional error upper bounded in Lemma 21 which is negligible.

C.1.2 Proof of lower bound

The proof of the lower bound is similar to Section B, with a slightly different construction. Since the parametric rate $\Omega(n^{-1/2})$ is trivial, we only need to show the first term. Set

$$h = (n \ln n)^{-\frac{1}{s+d}} \cdot [\Psi^{-1}(n)]^{\frac{d}{p(s+d)}}, \qquad R = c_0 \Psi^{-1}(n), \qquad S = \frac{R^d - 1}{h^d} \tag{64}$$

where $c_0 > 0$ is a small numerical constant to be specified later, and without loss of generality we assume that S is an integer. Fix $\alpha \in (0, S^{-1})$, and constants d_1, d_2, d_3 as in Section B.

Let $I_1, \ldots, I_S$ be the partition of $[0,R]^d - [0,1]^d$ into S smaller cubes of edge length h, and t_i denote the leftmost corner of sub-cube I_i, i.e., $I_i = t_i + [0,h]^d$. Let g and w be some fixed smooth probability density functions on $\mathbb{R}^d$ with finite entropy both of which vanish outside $[0,1]^d$. Note that the smoothness of g and w on $\mathbb{R}^d$ implies that their derivatives of any order vanish at their respective boundary.

To each vector $P = (p_1, \ldots, p_S) \in [0, \frac{d_3 \ln n}{n}]^S$, we associate a function (also appeared in (44))

$$f_P(t) \triangleq \sum_{i=1}^{S} p_i \cdot \frac{1}{h^d} g\left(\frac{t - t_i}{h}\right) + (1 - S\alpha)\, w(t).$$

By the choice of α, f_P is non-negative everywhere. We show that with a proper choice of the vector P, the density f_P lies in the Lipschitz ball $\mathrm{Lip}^{\Psi}_{s,p,d}(L_0)$ for a proper constant L_0.

Lemma 22. *Let $\rho > 0$. There exists a numerical constant $c_0 > 0$ depending on $\kappa(\Psi)$ such that if*

$$\frac{1}{S}\sum_{i=1}^{S} p_i^p \le \left(\frac{2C_1}{n \ln n}\right)^p,$$

then $f_P \in \mathrm{Lip}^{\Psi}_{s,p,d}(L_0)$ for some constant $L_0 > 0$ independent of n.

Proof. We first verify the Lipschitz norm condition, whose proof is similar to Lemma 16. Define $\tilde{f}_P$ as in Lemma 16, it suffices to show $\|\tilde{f}_P\|_{\mathrm{Lip}_{s,p,d}} \le L_0$ for some constant L_0 independent of n.

[8]The density after truncation is approximately a density, which integrates to $1 + O(n^{-4})$ following the same line of proofs in Lemma 21.

Observe that

$$\begin{aligned}\|\tilde{f}_P\|_p &= \frac{\|g\|_p}{h^d}\left(h^d\sum_{i=1}^S p_i^p\right)^{\frac{1}{p}}\\ &\le \frac{\|g\|_p R^{\frac{d}{p}}}{2h^d}\left(\frac{1}{S}\sum_{i=1}^S p_i^p\right)^{\frac{1}{p}}\\ &= \frac{c_0^{\frac{d}{p}}h^s\|g\|_p}{2}\cdot n\ln n\left(\frac{1}{S}\sum_{i=1}^S p_i^p\right)^{\frac{1}{p}},\end{aligned} \tag{65}$$

the condition ensures that $h^{-s}\|\tilde{f}_P\|_p$ is upper bounded by a constant depending only on c_0, C_1 and g. By (39) in Lemma 14, (65) implies that there exists a constant L_0 independent of n such that $\omega_r(\tilde{f}_P, t) \le L_0 t^s$ for any $t \ge h$. For $t < h$, further observe that for $r = \lceil s \rceil$, we have

$$\begin{aligned}\|D^r\tilde{f}_P\|_p &= \frac{\|D^r g\|_p}{h^{d+r}}\left(h^d\sum_{i=1}^S p_i^p\right)^{\frac{1}{p}}\\ &\le \frac{\|D^r g\|_p R^{\frac{d}{p}}}{2h^{d+r}}\left(\frac{1}{S}\sum_{i=1}^S p_i^p\right)^{\frac{1}{p}}\\ &= \frac{c_0^{\frac{d}{p}}\|D^r g\|_p}{2h^{r-s}}\cdot n\ln n\left(\frac{1}{S}\sum_{i=1}^S p_i^p\right)^{\frac{1}{p}}\end{aligned} \tag{66}$$

Hence, $h^{r-s}\|D^r\tilde{f}_P\|_p$ is upper bounded by a constant depending on C_1, g. By (40) in Lemma 14, (66) implies that $\omega_r(\tilde{f}_P, t) \le L_0 t^s$ for any $t \le h$, completing the proof of $\|\tilde{f}_P\|_{\mathrm{Lip}_{s,p,d}} \le L_0$.

Next we show that f_P satisfies the Orlicz norm condition. Note that the density f_P is supported on $[0, R]^d$, and

$$f_P(x) \le \frac{d_3 \ln n}{nh^d}\cdot\|g\|_\infty + \|w\|_\infty\cdot\mathbb{1}(x\in[0,1]^d).$$

As a result, there exists a numerical constant C depending on $d, d_3, \Psi(1), \|g\|_\infty, \|w\|_\infty$ such that

$$\int_{\mathbb{R}^d}\Psi(|x|)f_P(x)dx \le C\left(1+\Psi\left(c_0\Psi^{-1}(n)\right)\cdot\frac{R^d\ln n}{nh^d}\right). \tag{67}$$

Since Ψ is of rapid growth, for $c_0 = \kappa(\Psi)^{-m}$ with $m \in \mathbb{N}$ we have $\Psi(c_0\Psi^{-1}(n)) \le n^{2^{-m}}$. Hence, choosing m large enough such that $2^{-m} < \frac{s}{s+d}$, the RHS of (67) is upper bounded by a numerical constant L_0 independent of n, completing the proof. □

By Lemma 22, the construction of two measures in Section B.2 still ensures that f_P lies in the Lipschitz ball $\mathrm{Lip}^{\Psi}_{s,p,d}(L_0)$. Consequently, the remaining arguments (reduction to parametric submodel, fuzzy hypothesis testing, etc.) in Section B can be applied to obtain

$$\inf_{\hat{H}}\sup_{f\in\mathrm{Lip}^{\Psi}_{s,p,d}(L_0)}\mathbb{E}_f\left(\hat{H}-H(f)\right)^{\frac{1}{2}} \gtrsim \frac{S}{n\ln n} \asymp (n\ln n)^{-\frac{s}{s+d}}\cdot[\Psi^{-1}(n)]^{d\left(1-\frac{d}{p(s+d)}\right)},$$

as desired. □

C.2 Proof of Theorem 3

C.2.1 Proof of upper bound

We first prove the achievability part of the plug-in approach. For kernel $K(\cdot)$ satisfying Assumption 1 and $\hat{f}_h(x)$ given by (9), Lemma 2 shows that the first approximation error is at most

$$|H(f) - H(f_h)| \lesssim Lh^s. \tag{68}$$

Next we upper bound the bias $|\mathbb{E}H(\hat{f}_h) - H(f_h)|$ of the plug-in estimator. Let $\phi(x) \triangleq x \ln x$, and $g(x)$ be any twice continuously differentiable function on $[0, \infty)$. By [Tot94, Eqn. (2.5)], for any $u, v \geq 0$,

$$|g(v) - g(u) - g'(u)(v-u)| \leq \frac{4(v-u)^2}{u} \|wg''(w)\|_\infty.$$

Choosing $u = f_h(x), v = \hat{f}_h(x)$ and taking expectation at both sides, and noting that

$$\mathbb{E}(\hat{f}_h(x) - f_h(x))^2 \leq \frac{\|K\|_\infty f_h(x)}{nh^d}$$

we have

$$\begin{aligned}
|\mathbb{E}\phi(\hat{f}_h(x)) - \phi(f_h(x))| &\leq 2\|\phi - g\|_\infty + |\mathbb{E}g(\hat{f}_h(x)) - g(f_h(x))| \\
&\leq 2\|\phi - g\|_\infty + \frac{4\|wg''(w)\|_\infty}{f_h(x)} \cdot \mathbb{E}(\hat{f}_h(x) - f_h(x))^2 \\
&\lesssim \|\phi - g\|_\infty + \frac{1}{nh^d} \cdot \|wg''(w)\|_\infty.
\end{aligned}$$

The previous inequality holds for any g, so we may take infimum over g and reduce the problem to the following lemma:

Lemma 23 ([DT87, Theorem 2.1.1]). *For any real-valued function ϕ defined on $[0, \infty)$ and $t > 0$, there exists a universal constant $C > 0$ independent of ϕ, t such that*

$$\inf_{g \in C^2[0,\infty)} \left(\|\phi - g\|_\infty + t^2 \|wg''(w)\|_\infty \right) \leq C\omega_\varphi^2(f, t)_\infty$$

where $\omega_\varphi^2(f,t)_\infty$ is the second-order modulus of smoothness defined in (99) with $\varphi(x) = \sqrt{x}$.

It follows from [DT87, Example 3.4.2] that for $\phi(x) = x \ln x$, we have

$$\omega_\varphi^2(\phi, t)_\infty \asymp t^2.$$

As a result, by the previous arguments and Lemma 23 we obtain

$$|\mathbb{E}\phi(\hat{f}_h(x)) - \phi(f_h(x))| \lesssim \frac{1}{nh^d},$$

which integrates into

$$|\mathbb{E}H(\hat{f}_h) - H(f_h)| \leq \int_{[0,1]^d} |\mathbb{E}\phi(\hat{f}_h(x)) - \phi(f_h(x))| dx \lesssim \frac{1}{nh^d}. \tag{69}$$

To upper bound the variance $\mathsf{Var}(H(\hat{f}_h))$, we apply the same arguments in Section 3.3 to obtain

$$\mathsf{Var}(H(\hat{f}_h)) \lesssim nh^d \cdot \int_{[0,1]^d} \mathbb{E}[(\hat{f}_h(x) \ln \hat{f}_h(x) - \hat{f}'_h(x) \ln \hat{f}'_h(x))^2] dx,$$

where $\hat{f}'_h(x)$ is the density estimate based on observations $X'_1, X_2, \ldots, X_n$, with X'_1 being an independent copy of X_1. For any $x \in [0,1]^d$, we split into two cases:

1. Case I: $f_h(x) \le \frac{c_1 \ln n}{nh^d}$, where $c_1 > 0$ is the constant appearing in Lemma 5. By Lemma 5 and the union bound, with probability at least $1-2n^{-5d}$ we have $\hat{f}_h(x), \hat{f}_h'(x) \le \frac{2c_1 \ln n}{nh^d}$. Note that for any $\varepsilon \in (0,1)$, there exists some constant $C_\varepsilon > 0$ such that $|x \ln x - y \ln y| \le C_\varepsilon |x-y|^{1-\varepsilon}$ whenever $x, y \in [0, \frac{1}{2}]$. Hence,

$$\begin{aligned}\mathbb{E}(\hat{f}_h(x) \ln \hat{f}_h(x) - \hat{f}_h'(x) \ln \hat{f}_h'(x))^2 \le C_\varepsilon^2 \mathbb{E}|\hat{f}_h(x) - \hat{f}_h'(x)|^{2(1-\varepsilon)} \\ + 2n^{-5d} \cdot \left(2 \cdot \frac{\|K\|_\infty}{h^d} \ln \frac{\|K\|_\infty}{h^d}\right)^2.\end{aligned}$$

   Note that for $\varepsilon \in (0, \frac{1}{2})$, we have

$$\begin{aligned}\mathbb{E}|\hat{f}_h(x) - \hat{f}_h'(x)|^{2(1-\varepsilon)} &= (nh^d)^{-2(1-\varepsilon)} \cdot \mathbb{E}\left|K\left(\frac{x-X_1}{h}\right) - K\left(\frac{x-X_1'}{h}\right)\right|^{2(1-\varepsilon)} \\ &\lesssim (nh^d)^{-2(1-\varepsilon)} \cdot \mathbb{E}K^{2(1-\varepsilon)}\left(\frac{x-X_1}{h}\right) \\ &\lesssim (nh^d)^{-2(1-\varepsilon)} \cdot h^d f_h(x) \lesssim \frac{\ln n}{n^3 h^{2d}} \cdot (nh^d)^{2\varepsilon},\end{aligned}$$

   where the last inequality follows from $f_h(x) \le \frac{c_1 \ln n}{nh^d}$.

2. Case II: $f_h(x) > \frac{c_1 \ln n}{nh^d}$. By Lemma 5 and the union bound, with probability at least $1-2n^{-5d}$ we have $\hat{f}_h(x), \hat{f}_h'(x) \in [\frac{f_h(x)}{2}, 2f_h(x)]$. Then the mean value theorem gives

$$\begin{aligned}\mathbb{E}(\hat{f}_h(x) \ln \hat{f}_h(x) - \hat{f}_h'(x) \ln \hat{f}_h'(x))^2 &\lesssim (1 + (\ln f_h(x))^2) \cdot \mathbb{E}(\hat{f}_h(x) - \hat{f}_h'(x))^2 \\ &\quad + 2n^{-5d} \cdot \left(2 \cdot \frac{\|K\|_\infty}{h^d} \ln \frac{\|K\|_\infty}{h^d}\right)^2 \\ &\lesssim \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n^2 h^d},\end{aligned}$$

   where in the last step we have used that

$$\begin{aligned}\mathbb{E}(\hat{f}_h(x) - \hat{f}_h'(x))^2 &= \frac{1}{n^2 h^{2d}} \cdot \mathbb{E}\left(K\left(\frac{x-X_1}{h}\right) - K\left(\frac{x-X_1'}{h}\right)\right)^2 \\ &\le \frac{4}{n^2 h^{2d}} \cdot \mathbb{E}K^2\left(\frac{x-X_1}{h}\right) \le \frac{4\|K\|_\infty}{n^2 h^d} \cdot f_h(x).\end{aligned}$$

Combining the previous two cases, for any $x \in [0,1]^d$ and $\varepsilon \in (0, \frac{1}{2})$ we have

$$\mathbb{E}(\hat{f}_h(x) \ln \hat{f}_h(x) - \hat{f}_h'(x) \ln \hat{f}_h'(x))^2 \lesssim \frac{\ln n}{n^3 h^{2d}} \cdot (nh^d)^{2\varepsilon} + \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n^2 h^d}.$$

Now by Lemma 10, an integration yields to

$$\begin{aligned}\mathsf{Var}(H(\hat{f}_h)) &\lesssim \frac{\ln n}{n^2 h^d} \cdot (nh^d)^{2\varepsilon} + \int_{[0,1]^d} \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n} dx \\ &\lesssim \frac{\ln n}{n^2 h^d} \cdot (nh^d)^{2\varepsilon} + \frac{(\ln L)^2}{n}.\end{aligned} \tag{70}$$

In summary, combining (68), (69) and (70), we conclude that

$$\left[\sup_{f\in \mathrm{Lip}_{s,p,d}(L)} \mathbb{E}_f\left(H(\hat{f}_h)-H(f)\right)^2\right]^{\frac{1}{2}} \lesssim Lh^s + \frac{1}{nh^d} + \frac{\ln L}{\sqrt{n}} + \frac{\sqrt{\ln n}}{n\sqrt{h^d}}\cdot (nh^d)^{\varepsilon}.$$

Choosing $h \asymp (Ln)^{-\frac{1}{s+d}}$ and $\varepsilon > 0$ sufficiently small, we arrive at the desired upper bound. □

C.2.2 Proof of lower bound

Next we establish the lower bound. The parametric rate $\Omega(n^{-1/2}\ln L)$ has been established in Theorem 7, so we focus only on the $\Omega(n^{-s/(s+d)}L^{d/(s+d)})$ part. Suppose that some plug-in approach with kernel $K(\cdot)$ and bandwidth h attains a worst-case L_2 risk $o(n^{-s/(s+d)}L^{d/(s+d)})$. We make the following claims:

1. We must have $h \gg (Ln)^{-\frac{1}{s+d}}$. Otherwise, consider $f\equiv 1$, we have $f_h \equiv 1$ and

$$\mathbb{E}(\hat{f}_h(x)-f_h(x))^2 = \frac{1}{n}\left(\frac{1}{h^d}\int_{\mathbb{R}^d} K^2(t)dt - 1\right) \gtrsim \frac{1}{nh^d}.$$

   By Lemma 5, $\hat{f}_h(x) \le 2$ with probability at least $1-n^{-5d}$, Taylor expansion with Lagrange remainder term give

$$\begin{aligned}
&\mathbb{E}[\hat{f}_h(x)\ln\hat{f}_h(x)] - f_h(x)\ln f_h(x)\\
&\ge \frac{1}{2\cdot 2}\mathbb{E}(\hat{f}_h(x)-f_h(x))^2 - n^{-5d}\cdot\left(\frac{\|K\|_\infty f_h(x)}{nh}\ln\frac{\|K\|_\infty f_h(x)}{nh} + |f_h(x)\ln f_h(x)|\right)\\
&\gtrsim \frac{1}{nh^d}.
\end{aligned}$$

   As a result, since $h \lesssim (Ln)^{-1/(s+d)}$, the overall bias is at least

$$\begin{aligned}
H(f) - \mathbb{E}H(\hat{f}_h) &= \int_{[0,1]^d}\left(\mathbb{E}\hat{f}_h(x)\ln\hat{f}_h(x) - f_h(x)\ln f_h(x)\right)dx\\
&\gtrsim \frac{1}{nh^d} \gtrsim n^{-\frac{s}{s+d}}L^{\frac{d}{s+d}},
\end{aligned}$$

   a contradiction to the assumed performance of $H(\hat{f}_h)$.

2. We must have $h \lesssim (Ln)^{-1/(s+d)}$. Otherwise, $h \gg (Ln)^{-1/(s+d)}$, and by (69) and triangle inequality, for any $f \in \mathrm{Lip}_{s,p,d}(L)$ we must have

$$\begin{aligned}
|H(f)-H(f_h)| &\le |H(f)-\mathbb{E}H(\hat{f}_h)| + |\mathbb{E}H(\hat{f}_h) - H(f_h)|\\
&\lesssim |H(f)-\mathbb{E}H(\hat{f}_h)| + \frac{1}{nh^d}\\
&\ll n^{-\frac{s}{s+d}}L^{\frac{d}{s+d}}\\
&\lesssim Lh^s \wedge 1.
\end{aligned}$$

   In the sequel we will construct some $f \in \mathrm{Lip}_{s,p,d}(L)$ such that $|H(f)-H(f_h)| \gtrsim Lh^s \wedge 1$, establishing the desired contradiction. Let f be the density f_P which was constructed in (44) with P consisting of all identical elements (denoted by p), or specifically,

$$f(t) = p\cdot\sum_{i=1}^{S} Lh^s g\left(\frac{t-t_i}{h}\right) + (1-pLh^s)\,w(t) \triangleq f_1(t) + f_2(t).$$

We choose the parameter $p = \Theta((Lh^s)^{-1} \wedge 1)$ so that f is a density, and choose the smooth function g to satisfy $H(g) \neq H(g_1)$ for $g_1(x) \triangleq \int_{[0,1]^d} K(x-y)g(y)dy$. Note that the density $f \in \mathrm{H}_d^s(L)$ lies in the Hölder ball for small p. By the structure of the kernel-smoothed density, we may write

$$f_h(t) = p \cdot \sum_{i=1}^{S} Lh^s g_1\left(\frac{t-t_i}{h}\right) + f_{2h}(t) + \varepsilon(t)$$

where $f_{2h}(t)$ is the kernel-smoothed version of $f_2(t)$, and $\varepsilon(t)$ can be non-zero only in cubes which intersect the boundary of $[\frac{1}{4}, \frac{3}{4}]^d$. Moreover, since f lies in the Hölder ball, we have $\|f - f_h\|_\infty \lesssim Lh^s$ and $\|\varepsilon\|_\infty \lesssim Lh^s$. Noting that the total volume of all cubes which intersect the boundary is $O(h)$, we have

$$H(f_h) - H(f) \gtrsim (H(f_{2h}) - H(f_2)) + p \cdot Lh^s(H(g_1) - H(g)) - Lh^s \ln(1/h) \cdot h.$$

Finally, since convolution increases the entropy, we have $H(f_{2h}) \geq H(f_2), H(g_1) \geq H(g)$. Now our choice of $p = \Theta((Lh^s)^{-1} \wedge 1)$ gives

$$H(f_h) - H(f) \gtrsim p \cdot Lh^s - Lh^{s+1} \ln(1/h) \gtrsim Lh^s \wedge 1,$$

as desired.

The proof of the lower bound is complete noting that these two claims contradict each other. □

C.3 Proof of Theorem 6

The lower bound directly follows from Theorem 1 since we are considering a larger density class. For the upper bound, we define the estimator $\hat{H}$ as in Section 2 except that the kernel convolution (i.e., the definition of f_h in (10)) is understood on the torus $\mathbb{T}^d$. The analyses parallel those in Section 3, where Lemmas 3 and 4 now need some caution when dealing with the torus $\mathbb{T}^d$. For Lemma 3, same result holds simply by changing $\mathbb{R}^d$ into $\mathbb{T}^d$ in the proof of Lemma 3. As for Lemma 4, we remark that the Hardy–Littlewood maximal inequality also holds on the torus[9]. □

D Proof of main lemmas

D.1 Proof of Lemma 3

Let K_d be the following non-negative kernel on $\mathbb{R}^d$:

$$K_d(x) = \prod_{i=1}^{d} (1 - |x_i|)_+, \qquad x = (x_1, \ldots, x_d)$$

where we note that $\int_{\mathbb{R}^d} K_d(x)dx = 1$. Now define

$$K_{d,h}(x) = \frac{1}{h^d} K_d(\frac{x}{h}),$$

[9]For example, see [Tao15]; the proof follows from that of standard maximal inequality on Euclidean space via the Vitali covering lemma [Ste79, Chapter 1, Theorem 1].

and

$$g(x) = \int_{\mathbb{R}^d} f(y) K_{d,h}(x-y) dy.$$

Since both f and $K_{d,h}$ are non-negative, so is g. Also, by the symmetry of $K_d(\cdot)$,

$$\begin{aligned} f(x) - g(x) &= \int_{\mathbb{R}^d} (f(x) - f(y)) K_{d,h}(x-y) dy \\ &= \int_{\mathbb{R}^d} (f(x) - f(x-hu)) K_d(u) du \\ &= \frac{1}{2} \int_{\mathbb{R}^d} (2f(x) - f(x-hu) - f(x+hu)) K_d(u) du \end{aligned}$$

As a result,

$$\begin{aligned} \|f - g\|_p &\le \frac{1}{2} \int_{\mathbb{R}^d} \|2f(x) - f(x-hu) - f(x+hu)\|_p K_d(u) du \\ &\lesssim Lh^s \cdot \frac{1}{2} \int_{\mathbb{R}^d} K_d(u) du \asymp Lh^s. \end{aligned}$$

As for the Hessian of g, we have

$$\begin{aligned} [\nabla^2 g(x)]_{ii} &= \frac{\partial^2}{\partial x_i^2} \int_{\mathbb{R}^d} f(y) K_{d,h}(x-y) dy \\ &= \int_{\mathbb{R}^d} f(y) \frac{\partial^2}{\partial x_i^2} K_{d,h}(x-y) dy \\ &= \frac{1}{h^2} \int_{\mathbb{R}^d} f(x-hu) \frac{\partial^2}{\partial u_i^2} K_d(u) du \\ &= \frac{1}{h^2} \int_{\mathbb{R}^d} f(x-hu) \cdot (\delta(u_i+1) - 2\delta(u_i) + \delta(u_i-1)) K_{d-1}(u_{\setminus i}) du \\ &= \frac{1}{h^2} \int_{\mathbb{R}^{d-1}} \Delta^2_{he_i} f(x_{\setminus i} - hu_{\setminus i}, x_i) K_{d-1}(u_{\setminus i}) du_{\setminus i} \end{aligned}$$

where $\Delta^2_{he_i} f(x) \triangleq f(x+he_i) - 2f(x) + f(x-he_i)$, with e_i being the i-th coordinate vector, and $\delta(\cdot)$ is the delta function. As a result, we have

$$\begin{aligned} \|[\nabla^2 g(x)]_{ii}\|_p &\le \frac{1}{h^2} \int_{\mathbb{R}^{d-1}} \|\Delta^2_{he_i} f(x_{\setminus i} - hu_{\setminus i}, x_i)\|_p K_{d-1}(u_{\setminus i}) du_{\setminus i} \\ &= \frac{1}{h^2} \int_{\mathbb{R}^{d-1}} \|\Delta^2_{he_i} f(x)\|_p K_{d-1}(u_{\setminus i}) du_{\setminus i} \\ &\lesssim \frac{1}{h^2} \int_{\mathbb{R}^{d-1}} Lh^s \cdot K_{d-1}(u_{\setminus i}) du_{\setminus i} \asymp Lh^{s-2}. \end{aligned}$$

Similarly, for $i \neq j$, we have

$$\begin{aligned} [\nabla^2 g(x)]_{ij} &= \frac{1}{h^2} \int_{\mathbb{R}^d} f(x-hu) \frac{\partial^2}{\partial u_i \partial u_j} K_d(u) du \\ &= \frac{1}{h^2} \int_{\mathbb{R}^d} f(x-hu) \cdot (\mathbb{1}_{u_i \in (0,1)} - \mathbb{1}_{u_i \in (-1,0)})(\mathbb{1}_{u_j \in (0,1)} - \mathbb{1}_{u_j \in (-1,0)}) K_{d-2}(u_{\setminus (i,j)}) du \\ &= \frac{1}{h^2} \int_{\mathbb{R}^{d-2}} \int_0^1 \int_0^1 \Delta_{he_i} \Delta_{he_j} f(x-hu) K_{d-2}(u_{\setminus (i,j)}) du_i du_j du_{\setminus (i,j)} \end{aligned}$$

where $\Delta_{he_i} f(x) = f(x + he_i) - f(x)$ denotes the forward difference. Note that

$$\begin{aligned}
\Delta_{he_i}\Delta_{he_j} f(x) &= f(x + he_i + he_j) - f(x + he_i) - f(x + he_j) + f(x) \\
&= \frac{f(x + he_i + he_j) - 2f(x + he_i) + f(x - he_i + he_j)}{2} \\
&\quad + \frac{f(x + he_i + he_j) - 2f(x + he_j) + f(x - he_j + he_i)}{2} \\
&\quad - \frac{f(x - he_i + he_j) - 2f(x) + f(x - he_j + he_i)}{2} \\
&= \frac{\Delta^2_{he_i} f(x + he_i) + \Delta^2_{he_j} f(x + he_j) - \Delta^2_{h(e_i+e_j)} f(x)}{2}
\end{aligned}$$

we conclude that

$$\|\Delta_{he_i}\Delta_{he_j} f\|_p \le \frac{1}{2}\left(\|\Delta^2_{he_i} f\|_p + \|\Delta^2_{he_j} f\|_p + \|\Delta^2_{h(e_i+e_j)} f\|_p\right) \lesssim Lh^s.$$

Then using the same technique, we also have

$$\|[\nabla^2 g(x)]_{ij}\|_p \lesssim Lh^{s-2}.$$

Finally, note that $\|A\|_{\mathrm{op}} \le \sqrt{\mathrm{Tr}(A^T A)} \le \sum_{i,j} |A_{ij}|$, we conclude that

$$\|\|\nabla^2 g\|_{\mathrm{op}}\|_p \le \left\| \sum_{i,j=1}^d |[\nabla^2 g(x)]_{ij}| \right\|_p \le \sum_{i,j=1}^d \|[\nabla^2 g(x)]_{ij}\|_p \lesssim Lh^{s-2},$$

as desired. □

D.2 Proof of Lemma 5

Applying the Bennett inequality (Lemma 13) to independent random variables $Y_i = K_h(x - X_i)$, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{f}_h(x) - f_h(x)\right| \ge \varepsilon\right) \\
&\le 2\exp\left(-\frac{n\varepsilon^2}{2\mathsf{Var}(K_h(x - X_1)) + 2\|K\|_\infty \varepsilon/3h^d}\right).
\end{aligned}$$

For the variance, by the non-negativity of the kernel K and the density f, we have

$$\begin{aligned}
\mathsf{Var}(K_h(x - X_1)) &\le \mathbb{E}[K_h(x - X_1)^2] \\
&= \int K_h(x - y)^2 f(y) dy \\
&\le \frac{\|K\|_\infty}{h^d} \int K_h(x - y) f(y) dy \\
&= \frac{\|K\|_\infty}{h^d} f_h(x).
\end{aligned} \tag{71}$$

Combining these two inequalities yields the desired result. □

D.3 Proof of Lemma 7

Recall that $\mathbb{E}[\hat{H}_1(x)] = Q(f_h(x))$, where $Q(t)$ defined in (13) is the best degree-k polynomial that uniformly approximates $-t\ln t$ on the interval $[0, \frac{2c_1 \ln n}{nh^d}]$. Thus the bias is upper bounded by

$$\begin{aligned}|\mathbb{E}[\hat{H}_1(x)] + f_h(x)\ln f_h(x)| &= \left|\sum_{l=0}^{k} a_l f_h(x)^l + f_h(x)\ln f_h(x)\right| \\ &\lesssim \frac{1}{k^2}\cdot\frac{2c_1\ln n}{nh^d} \lesssim \frac{1}{nh^d \ln n}\end{aligned}$$

provided that $f_h(x) \le \frac{2c_1\ln n}{nh^d}$.

To upper bound the second moment of the perturbation, first note that $\hat{H}_1(x)$ does not depend on observations from $X^{(3)}$ and is symmetric in observations from $X^{(2)}$. Consequently, we may assume that the original estimator $\hat{H}_1(x)$ uses the observations $X^{(2)} = (X_1, \ldots, X_n)$ and the perturbed estimator $\hat{H}_1'(x)$ uses observations $(X_1', \ldots, X_n')$, where $X_j' = X_j$ for $j = 1, \ldots, n-1$ and X_n' is an independent copy of X_n. Then

$$\begin{aligned}\hat{H}_1(x) - \hat{H}_1'(x) &= \sum_{l=0}^{k} a_l \frac{1}{\binom{n}{l}} \sum_{J\in\binom{[n]}{l}} \left(\prod_{j\in J} K_h(x-X_j) - \prod_{j\in J} K_h(x-X_j')\right) \\ &= \frac{1}{n}(K_h(x-X_n) - K_h(x-X_n')) \\ &\quad \cdot \underbrace{\sum_{l=1}^{k} l a_l \frac{1}{\binom{n-1}{l-1}} \sum_{J'\in\binom{[n-1]}{l-1}} \prod_{j\in J'} K_h(x-X_j)}_{\triangleq \tilde{H}_1(x)}.\end{aligned}$$

By the independence of (X_n, X_n') and $(X_1, \ldots, X_{n-1})$,

$$\mathbb{E}[(\hat{H}_1(x) - \hat{H}_1'(x))^2] = \frac{1}{n^2}\mathbb{E}[(K_h(x-X_n) - K_h(x-X_n'))^2]\cdot\mathbb{E}[\tilde{H}_1(x)^2]. \tag{72}$$

By (71) and the assumption on $f_h(x)$, we have

$$\begin{aligned}\mathbb{E}[(K_h(x-X_n) - K_h(x-X_n'))^2] &= 2\mathsf{Var}(K_h(x-X_n)) \\ &\le \frac{2\|K\|_\infty f_h(x)}{h^d} \le \frac{4c_1\|K\|_\infty \ln n}{nh^{2d}}.\end{aligned} \tag{73}$$

It remains to upper bound $\mathbb{E}[\tilde{H}_1(x)^2]$. By Cauchy-Schwarz,

$$\mathbb{E}[\tilde{H}_1(x)^2] \le k^3 \cdot \sum_{l=0}^{k-1} \frac{a_{l+1}^2}{\binom{n-1}{l}^2} \sum_{|J|=l}\sum_{|T|=l} \mathbb{E}\left[\prod_{j\in J}\prod_{t\in T} K_h(x-X_j)K_h(x-X_t)\right]. \tag{74}$$

Recall from 71 that

$$\begin{aligned}\mathbb{E}K_h(x-X_i) &= f_h(x), \\ \mathbb{E}K_h(x-X_i)^2 &\le \frac{\|K\|_\infty f_h(x)}{h^d}.\end{aligned}$$

Therefore, if $|J \cap T| = r$, we have

$$\begin{aligned}\mathbb{E}\left[\prod_{j\in J}\prod_{t\in T} K_h(x-X_j)K_h(x-X_t)\right] &\le \left(\frac{\|K\|_\infty f_h(x)}{h^d}\right)^r \cdot f_h(x)^{2(l-r)} \\ &= \|K\|_\infty^r h^{-dr} f_h(x)^{2l-r}.\end{aligned}$$

Moreover, the number of pairs of subsets $J, T \in \binom{[n-1]}{l}$ such that $|J\cap T| = r$ is $\frac{(n-1)!}{r!((l-r)!)^2(n-1-2l+r)!}$. Hence, for $0 \le l \le k-1$,

$$\begin{aligned}&\binom{n-1}{l}^{-2} \sum_{|J|=l,|T|=l} \mathbb{E}\left[\prod_{j\in J}\prod_{t\in T} K_h(x-X_j)K_h(x-X_t)\right] \\ &\le \frac{(l!)^2((n-1-l)!)^2}{(n-1)!}\sum_{r=0}^{l}\frac{\|K\|_\infty^r h^{-dr} f_h(x)^{2l-r}}{r!((l-r)!)^2(n-1-2l+r)!} \\ &= \frac{((n-1-l)!)^2}{(n-1)!(n-1-2l)!}\sum_{r=0}^{l}\binom{l}{r}^2 \cdot \frac{r!\|K\|_\infty^r h^{-dr} f_h(x)^{2l-r}}{(n-1-2l+r)(n-2-2l+r)\cdots(n-2l)} \\ &\le \sum_{r=0}^{l} 2^{2l}\left(\frac{r}{n-2l}\right)^r \|K\|_\infty^r h^{-dr} f_h(x)^{2l-r} \\ &\le 2^{2l}\sum_{r=0}^{l}\left(\frac{2k}{nh^d}\right)^r \|K\|_\infty^r f_h(x)^{2l-r} \\ &\le 2^{2l}\sum_{r=0}^{l}\left(\frac{2\|K\|_\infty c_2 \ln n}{nh^d}\right)^r \left(\frac{2c_1\ln n}{nh^d}\right)^{2l-r} \\ &= \left(\frac{2c_1\ln n}{nh^d}\right)^{2l}\sum_{r=0}^{l}\left(\frac{\|K\|_\infty c_2}{c_1}\right)^r \\ &\le 2\left(\frac{2c_1\ln n}{nh^d}\right)^{2l},\end{aligned} \tag{75}$$

where we have used our assumptions that $n \ge 4k = 4c_2 \ln n$, and $c_1 \ge 2\|K\|_\infty c_2$. By Lemma 11, we have the following coefficient bound:

$$\begin{aligned}|a_{l+1}| &\le \left(\frac{2c_1\ln n}{nh^d}\right)^{-l} \ln\left(\frac{nh^d}{2c_1\ln n}\right)\cdot 2^{3c_2\ln n} \\ &\le \left(\frac{2c_1\ln n}{nh^d}\right)^{-l}\cdot n^{3c_2\ln 2}\ln n.\end{aligned} \tag{76}$$

Combining (74)–(76) yields

$$\begin{aligned}\mathbb{E}[\tilde{H}_1(x)^2] &\le k^3\cdot\sum_{l=0}^{k-1}\left(\frac{2c_1\ln n}{nh^d}\right)^{-2l}(n^{3c_2\ln 2}\ln n)^2\cdot 2\left(\frac{2c_1\ln n}{nh^d}\right)^{2l} \\ &= 2k^4\cdot(n^{3c_2\ln 2}\ln n)^2 \lesssim n^{\varepsilon},\end{aligned} \tag{77}$$

provided that $7c_2\ln 2 < \varepsilon$. Now combining (72), (73) and (77) completes the proof. □

D.4 Proof of Lemma 8

Recall that by Lemma 5, with $c_1 > 0$ large enough we have

$$\mathbb{P}\left(\hat{f}_{h,2}(x) \notin \left(\frac{f_h(x)}{2}, 2f_h(x)\right)\right) \le n^{-5d}. \tag{78}$$

Moreover, since both f_h and $\hat{f}_h$ are convolutions of probability measures with K_h, we have the deterministic upper bound: for all x,

$$\max\{f_h(x), \hat{f}_{h,2}(x), \hat{f}_{h,3}(x)\} \le \|K_h\|_\infty = \frac{\|K\|_\infty}{h^d}. \tag{79}$$

Consequently, the following deterministic upper bound of $|\hat{H}_2(x)|$ (in (15)) holds for all x:

$$|\hat{H}_2(x)| \lesssim 1 + nh^d + \frac{1}{h^d}\ln\frac{n}{h^d}. \tag{80}$$

Recall that throughout this proof we have by assumption

$$f_h(x) \ge \frac{c_1 \ln n}{2nh^d}. \tag{81}$$

For $\phi(z) = -z\ln z$ and $\mathbb{E}_3$ denoting the partial expectation taken only with respect to the third group of observations $X^{(3)}$, we have

$$\begin{aligned}
&\mathbb{E}_3\left[\hat{H}_2(x)\middle|\hat{f}_{h,2}(x) \ge \frac{f_h(x)}{2}\right] \\
&= \sum_{k=0}^{2}\frac{\phi^{(k)}(\hat{f}_{h,2}(x))}{k!}(f_h(x) - \hat{f}_{h,2}(x))^k \\
&= \phi(f_h(x)) - \frac{\phi^{(3)}(\xi_1)}{6}(f_h(x) - \hat{f}_{h,2}(x))^3 \\
&= \phi(f_h(x)) - \frac{\phi^{(3)}(f_h(x))}{6}(f_h(x) - \hat{f}_{h,2}(x))^3 \\
&\quad + \frac{\phi^{(4)}(\xi_2)}{6}(f_h(x) - \hat{f}_{h,2}(x))^3(f_h(x) - \xi_1),
\end{aligned} \tag{82}$$

where the last two steps follow from the Taylor expansion, with ξ_1 lying between $f_h(x)$ and $\hat{f}_{h,2}(x)$, and ξ_2 lying between $f_h(x)$ and ξ_1. The central moments of $\hat{f}_{h,2}(x)$ are computed as follows:

$$\begin{aligned}
\left|\mathbb{E}(\hat{f}_{h,2}(x) - f_h(x))^3\right| &= \left|\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}K_h(x - X_i) - f_h(x)\right)^3\right| \\
&= \frac{1}{n^2}|\mathbb{E}(K_h(x - X_1) - f_h(x))^3| \\
&\le \frac{1}{n^2}\left(\mathbb{E}K_h(x - X_1)^3 + f_h(x)^3\right) \\
&\le \frac{1}{n^2}\left(\frac{\|K\|_\infty^2}{h^{2d}}\mathbb{E}K_h(x - X_1) + f_h(x)^3\right) \\
&= \frac{f_h(x)}{n^2}\left(\frac{\|K\|_\infty^2}{h^{2d}} + f_h(x)^2\right) \overset{(79)}{\le} \frac{2f_h(x)\|K\|_\infty^2}{n^2h^{2d}},
\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(\hat{f}_{h,2}(x)-f_h(x))^4 &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n K_h(x-X_i)-f_h(x)\right)^4\\ &\le \frac{3}{n^2}\mathbb{E}^2[K_h(x-X_1)^2]+\frac{1}{n^3}\mathbb{E}[K_h(x-X_1)^4]\\ &\le \frac{3}{n^2}\left(\frac{\|K\|_\infty f_h(x)}{h^d}\right)^2+\frac{1}{n^3}\frac{\|K\|_\infty^3 f_h(x)}{h^{3d}}.\end{aligned}$$

Consequently, since $\phi^{(3)}(z)=z^{-2}$, we have

$$\begin{aligned}\left|\mathbb{E}\left[\frac{\phi^{(3)}(f_h(x))}{6}(f_h(x)-\hat{f}_{h,2}(x))^3\right]\right| &\le \frac{1}{6f_h(x)^2}\cdot\frac{2f_h(x)\|K\|_\infty^2}{n^2h^{2d}}\\ &\overset{(81)}{\lesssim} \frac{1}{nh^d\ln n}.\end{aligned} \tag{83}$$

Similarly, since $\phi^{(4)}(z)=-2z^{-3}, \xi_2\ge f_h(x)/2$ and $|f_h(x)-\xi_1|\le|f_h(x)-\hat{f}_{h,2}(x)|$, we have

$$\begin{aligned}&\left|\mathbb{E}\left[\frac{\phi^{(4)}(\xi_2)}{6}(f_h(x)-\hat{f}_{h,2}(x))^3(f_h(x)-\xi_1)\right]\right|\\ &\le \frac{1}{3(f_h(x)/2)^3}\cdot\mathbb{E}(\hat{f}_{h,2}(x)-f_h(x))^4\\ &\le \frac{1}{3(f_h(x)/2)^3}\cdot\left[\frac{3}{n^2}\left(\frac{\|K\|_\infty f_h(x)}{h^d}\right)^2+\frac{1}{n^3}\cdot\frac{\|K\|_\infty^3 f_h(x)}{h^{3d}}\right]\\ &\lesssim \frac{1}{n^2h^{2d}f_h(x)}+\frac{1}{n^3h^{3d}f_h(x)^2}\lesssim\frac{1}{nh^d\ln n}.\end{aligned} \tag{84}$$

Then the desired bias bound is a direct consequence of (78)–(84).

Next we upper bound the second moment of the perturbation. First we consider the case where one observation in $X^{(3)}$, say X_n, is replaced by an independent copy X_n'. In this case, by definition (15), we have

$$\begin{aligned}\hat{H}_2(x)-\hat{H}_2'(x) = \mathbb{1}\left(\hat{f}_{h,2}(x)\ge\frac{c_1\ln n}{4nh^d}\right)\Bigg[&\frac{\ln\hat{f}_{h,2}(x)}{n}(K_h(x-X_n')-K_h(x-X_n))\\ &+\frac{K_h(x-X_n')-K_h(x-X_n)}{n(n-1)\hat{f}_{h,2}(x)}\sum_{i=1}^{n-1}K_h(x-X_i)\Bigg].\end{aligned} \tag{85}$$

Note that (X_n,X_n'), $(X_1,\dots,X_{n-1})$ and $\hat{f}_{h,2}(x)$ are mutually independent. Recall from (73) that

$$\mathbb{E}[(K_h(x-X_n')-K_h(x-X_n))^2]\le\frac{2\|K\|_\infty f_h(x)}{h^d}. \tag{86}$$

Moreover,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n-1}\sum_{i=1}^{n-1}K_h(x-X_i)\right)^2\right] &= f_h(x)^2+\frac{1}{n-1}\mathsf{Var}\left(K_h(x-X_i)\right)\\ &\overset{(71)}{\le} f_h(x)^2+\frac{\|K\|_\infty f_h(x)}{(n-1)h^d}\overset{(81)}{\asymp} f_h(x)^2.\end{aligned} \tag{87}$$

To bound the remaining two expectations, let E be the event $\hat{f}_{h,2}(x) \in (f_h(x)/2, 2f_h(x))$, which satisfies $\mathbb{P}(E) \geq 1 - n^{-5d}$ by (78). Then

$$\ln^2 \hat{f}_{h,2}(x) \mathbb{1}(E) \lesssim 1 + (\ln f_h(x))^2, \tag{88}$$

$$\hat{f}_{h,2}(x)^{-2} \mathbb{1}(E) \lesssim f_h(x)^{-2}. \tag{89}$$

Therefore, by (85)–(89), we conclude that

$$\begin{aligned}
\mathbb{E}[(\hat{H}_2(x) - \hat{H}_2'(x))^2 \mathbb{1}(E)] &\lesssim \frac{f_h(x)}{n^2 h^d} \left[1 + (\ln f_h(x))^2 + \frac{1}{f_h(x)^2} \left(f_h(x)^2 + \frac{f_h(x)}{nh^d}\right)\right] \\
&\lesssim \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n^2 h^d} + \frac{1}{n^3 h^{2d}} \\
&\overset{(81)}{\asymp} \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n^2 h^d}.
\end{aligned}$$

Moreover, by (78) and (80),

$$\mathbb{E}[(\hat{H}_2(x) - \hat{H}_2'(x))^2 \mathbb{1}(E^c)] \lesssim \left(1 + nh^d + \frac{1}{h^d} \ln \frac{n}{h^d}\right)^2 \mathbb{P}(E^c) \lesssim n^{2-5d} (\ln n)^2,$$

in view of the assumptions $h \leq 1$ and $nh^d \geq 1$. Combining the previous two inequalities with (81) gives

$$\mathbb{E}[(\hat{H}_2(x) - \hat{H}_2'(x))^2] \lesssim \frac{f_h(x)(1 + (\ln f_h(x))^2)}{n^2 h^d}. \tag{90}$$

Now we consider the case where one observation in $X^{(2)}$ is replaced by an independent copy, and $\hat{f}'_{h,2}(x)$ is the perturbed density estimate. Similar to the event E, let E' be the event $\hat{f}'_{h,2}(x) \in (f_h(x)/2, 2f_h(x))$. Then on the event $E \cap E'$, by the assumption (81), we have $\hat{f}_{h,2}(x) \geq f_h(x)/2 \geq \frac{c_1 \ln n}{4nh^d}$ and similarly for $\hat{f}'_{h,2}(x)$, and hence

$$\hat{H}_2(x) - \hat{H}_2'(x) = \frac{1}{2} D_1 + C_2 D_2 + \frac{1}{2} C_3 D_3,$$

where

$$\begin{aligned}
&D_1 \triangleq \hat{f}_{h,2}(x) - \hat{f}'_{h,2}(x), \\
&D_2 \triangleq \ln \hat{f}'_{h,2}(x - \ln \hat{f}_{h,2}(x)), \quad C_2 \triangleq \hat{f}_{h,3}(x) \\
&D_3 \triangleq \hat{f}'_{h,2}(x)^{-1} - \hat{f}_{h,2}(x)^{-1}, \quad C_3 \triangleq \frac{2}{n(n-1)} \sum_{i<j} K_h(x - X_i^{(3)}) K_h(x - X_j^{(3)}).
\end{aligned}$$

By independence and the triangle inequality, it remains to upper bound the second moments of D_1, D_2, D_3, C_2, C_3 separately. By (86),

$$\mathbb{E}[D_1^2] \leq \frac{2\|K\|_\infty f_h(x)}{n^2 h^d}. \tag{91}$$

By the mean value theorem,

$$\mathbb{E}[D_2^2 \mathbb{1}(E \cap E')] \leq \sup_{\xi \in (f_h(x)/2, 2f_h(x))} \left|\frac{d \ln \xi}{d\xi}\right|^2 \cdot \mathbb{E}[D_1^2] \lesssim \frac{1}{n^2 h^d f_h(x)}, \tag{92}$$

$$\mathbb{E}[D_3^2 \mathbb{1}(E \cap E')] \leq \sup_{\xi \in (f_h(x)/2, 2f_h(x))} \left|\frac{d(\xi^{-1})}{d\xi}\right|^2 \cdot \mathbb{E}[D_1^2] \lesssim \frac{1}{n^2 h^d f_h(x)^3}. \tag{93}$$

Using $\mathbb{E}[X^2] = (\mathbb{E}X)^2 + \mathsf{Var}(X)$, we also have

$$\mathbb{E}[C_2^2] = f_h(x)^2 + \mathsf{Var}(\hat{f}_{h,3}(x)) \overset{(71)}{\lesssim} f_h(x)^2 + \frac{f_h(x)}{nh^d} \overset{(81)}{\asymp} f_h(x)^2, \tag{94}$$

$$\begin{aligned}\mathbb{E}[C_3^2] &\le f_h(x)^4 + \frac{4(n-2)}{n(n-1)} f_h(x)^2 \left(f_h(x)^2 + \frac{f_h(x)}{nh^d}\right) \\ &\quad + \frac{2}{n(n-1)} \left(f_h(x)^2 + \frac{f_h(x)}{nh^d}\right)^2 \\ &\lesssim f_h(x)^4 + \frac{f_h(x)^3}{n^2h^d} + \frac{f_h(x)^2}{n^4h^{2d}} \overset{(81)}{\asymp} f_h(x)^4,\end{aligned} \tag{95}$$

where (95) is due to Lemma 12. Now combining (91)–(95), the triangle inequality gives

$$\begin{aligned}&\mathbb{E}[(\hat{H}_2(x) - \hat{H}_2'(x))^2 \mathbb{1}(E \cap E')] \\ &\lesssim \mathbb{E}[D_1^2] + \mathbb{E}[C_2^2]\mathbb{E}[D_2^2 \mathbb{1}(E \cap E')] + \mathbb{E}[C_3^2]\mathbb{E}[D_3^2 \mathbb{1}(E \cap E')] \\ &\lesssim \frac{f_h(x)}{n^2h^d}.\end{aligned}$$

Moreover, by (78) and (80),

$$\begin{aligned}\mathbb{E}[(\hat{H}_2(x) - \hat{H}_2'(x))^2 \mathbb{1}(E^c \cup E'^c)] &\lesssim \left(1 + nh^d + \frac{1}{h^d} \ln \frac{n}{h^d}\right)^2 \mathbb{P}(E^c \cup E'^c) \\ &\lesssim n^{2-5d} (\ln n)^2,\end{aligned}$$

in view of the assumptions $h \le 1$ and $nh^d \ge 1$. Combining the previous two inequalities with (81) gives

$$\mathbb{E}[(\hat{H}_2(x) - \hat{H}_2'(x))^2] \lesssim \frac{f_h(x)}{n^2h^d}. \tag{96}$$

The proof is then completed by combining (90) and (96). □

D.5 Proof of Lemma 9

To analyze the performance of $\hat{H}$, we consider the following decomposition:

$$\begin{aligned}\hat{H} - H(f) &= H(f_h) - H(f) + \hat{H} - H(f_h) \\ &= H(f_h) - H(f) + \left(\hat{H}_{\mathsf{discrete}} + \sum_{i=1}^S p_i \ln p_i\right) \\ &= h^d \sum_{i=1}^S \left(-\frac{p_i}{h^d} \ln \frac{p_i}{h^d} + \frac{1}{h^d} \int_{I_i} f(t) \ln f(t) dt\right) + \left(\hat{H}_{\mathsf{discrete}} + \sum_{i=1}^S p_i \ln p_i\right).\end{aligned} \tag{97}$$

The first term deals with the approximation error in the first approximation stage. Note that by our Lipschitz ball assumption, we have

$$\|f - f_h\|_2^2 = \sum_{i=1}^S \int_{I_i} \left|f(t) - \frac{p_i}{h^d}\right|^2 dt \lesssim L^2 h^{2s}.$$

Now we deal with the difference $H(f_h) - H(f)$. For $Z_i \sim \mathsf{Unif}(I_i)$, $Y = f(Z_i)$ and $\phi(x) = x \ln x$, we have

$$\frac{1}{h^d} \int_{I_i} f(t) \ln f(t) dt - \frac{p_i}{h^d} \ln \frac{p_i}{h^d} = \mathbb{E}\phi(Y_i) - \phi(\mathbb{E}Y_i)$$

which is the gap in the Jensen's inequality. Using the equivalence between the K-functional and the modulus of smoothness [JHW17] [DL93, Chapter 6, Theorem 2.4], we have the following lemma.

Lemma 24. *[ST77] There exists a universal constant $C > 0$ such that for any real-valued function ϕ defined on an interval $I \subset \mathbb{R}$ and any random variable X supported on I with a finite variance,*

$$|\mathbb{E}\phi(X) - \phi(\mathbb{E}X)| \le C\omega_I^2(\phi, \sqrt{\mathsf{Var}(X)})$$

where $\omega_I^2(\phi, t)$ is the second-order modulus of smoothness defined in [DL93].

To apply Lemma 24, we note that [DL93, Chapter 2.9, Example 1]

$$\omega_{\mathbb{R}_+}^2(-x \ln x, t) \asymp t$$

and

$$\mathsf{Var}(Y_i) = \frac{1}{h^d} \int_{I_i} \left| f(t) - \frac{p_i}{h^d} \right|^2 dt.$$

As a result,

$$\begin{aligned} |H(f_h) - H(f)| &\le \frac{1}{S} \sum_{i=1}^S |\mathbb{E}\phi(Y_i) - \phi(\mathbb{E}Y_i)| \lesssim \frac{1}{S} \sum_{i=1}^S \omega_{\mathbb{R}_+}^2(\phi, \sqrt{\mathsf{Var}(Y_i)}) \\ &\lesssim \frac{1}{S} \sum_{i=1}^S \sqrt{\mathsf{Var}(Y_i)} \le \sqrt{\frac{1}{S} \sum_{i=1}^S \mathsf{Var}(Y_i)} \\ &= \sqrt{\sum_{i=1}^S \int_{I_i} \left| f(t) - \frac{p_i}{h^d} \right|^2 dt} \lesssim Lh^s. \end{aligned}$$

As for the second term in (97), we have $p_i \le \frac{L}{S}$ for any $i = 1, \ldots, S$ since $\|f\|_\infty \le L$. Note that the discrete entropy can be related to the KL divergence as

$$\sum_{i=1}^S -p_i \ln p_i = -\sum_{i=1}^S p_i \ln \frac{p_i}{S^{-1}} + \ln S,$$

where the likelihood ratio between $(p_1, \ldots, p_S)$ and the uniform distribution is upper bounded by L, the result in [HJW16] yields the following risk bound:

$$\left(\sup_f \mathbb{E}_f \left(\hat{H}_{\mathsf{discrete}} + \sum_{i=1}^S p_i \ln p_i \right)^2 \right)^{\frac{1}{2}} \lesssim \frac{1}{nh^d \ln n} + \frac{\ln L}{\sqrt{n}}.$$

Combining the previous inequalities together, we arrive at

$$\left(\sup_{f \in \mathrm{Lip}_{s,p,d}(L), \|f\|_\infty \le L} \mathbb{E}_f (\hat{H} - H(f))^2 \right)^{\frac{1}{2}} \lesssim Lh^s + \frac{1}{nh^d \ln n} + \frac{\ln L}{\sqrt{n}}.$$

Now choosing $h \asymp (Ln \ln n)^{-\frac{1}{s+d}}$ completes the proof. □

D.6 Proof of Lemma 10

Let $X \sim f$, then

$$\begin{aligned}\int_{[0,1]^d} f(x)(\ln f(x))^2 dx &= \mathbb{E}[(\ln f(X))^2] \\ &= \underbrace{\mathbb{E}[(\ln f(X))^2 \mathbb{1}(f(X) \le e^{1/(p-1)})]}_{\triangleq A_1} \\ &\quad + \underbrace{\mathbb{E}[(\ln f(X))^2 \mathbb{1}(f(X) > e^{1/(p-1)})]}_{\triangleq A_2}.\end{aligned}$$

Since $\max_{t\in[0,e^{1/(p-1)}]} t \ln^2 t = \max\{4e^{-2}, e^{1/(p-1)}/(p-1)^2\}$, we have

$$A_1 = \int_{x\in[0,1]^d: f(x)\le e^{p-1}} f(x)(\ln f(x))^2 dx \le \frac{4}{e^2} + \frac{e^{1/(p-1)}}{(p-1)^2}.$$

As for A_2, note that whenever $t > e$,

$$\frac{d^2}{dt^2}\left[(\ln t)^2\right] = \frac{2 - 2\ln t}{t^2} < 0.$$

Hence, by conditional Jensen's inequality, we have

$$\begin{aligned}A_2 &= \frac{1}{(p-1)^2}\mathbb{E}[(\ln f(X)^{p-1})^2 \mathbb{1}(f(X)^{p-1} > e)] \\ &\le \frac{\mathbb{P}(f(X)^{p-1} > e)}{(p-1)^2} \cdot \left(\ln \mathbb{E}[f(X)^{p-1} | f(X)^{p-1} > e]\right)^2 \\ &\le \frac{\mathbb{P}(f(X)^{p-1} > e)}{(p-1)^2} \cdot \left(\ln \frac{\int_{[0,1]^d} f(x)^p dx}{\mathbb{P}(f(X)^{p-1} > e)}\right)^2 \\ &\le \frac{\mathbb{P}(f(X)^{p-1} > e)}{(p-1)^2} \cdot \left(\ln \frac{L^p}{\mathbb{P}(f(X)^{p-1} > e)}\right)^2 \\ &\le \frac{1 + p^2(\ln L)^2}{(p-1)^2},\end{aligned}$$

where the last inequality follows from

$$\max_{t\in[0,1]} t\left(\ln \frac{a}{t}\right)^2 = \begin{cases} a/e & \text{if } 1 \le a \le e \\ (\ln a)^2 & \text{if } a > e \end{cases} \le 1 + (\ln a)^2$$

whenever $a \ge 1$. Combining the upper bounds of A_1 and A_2 completes the proof. □

D.7 Proof of Lemma 12

It is straightforward to see $\mathbb{E}U = (\mathbb{E}X_1)^2$. As a result,

$$\mathsf{Var}(U_2) = \frac{4}{n^2(n-1)^2} \sum_{1\le i<j\le n, 1\le i'<j'\le n} \mathbb{E}[X_i X_j X_{i'} X_{j'}] - (\mathbb{E}X_1)^4.$$

Denote by I_0, I_1, I_2 the set of indices (i, j, i', j') where (i, j) and (i', j') have $0, 1$ and 2 elements in common, respectively. Then

$$\begin{aligned}
|I_0| &= \frac{n!}{2!2!(n-4)!} = \frac{n(n-1)(n-2)(n-3)}{4}, \\
|I_1| &= \frac{n!}{1!1!1!(n-3)!} = n(n-1)(n-2), \\
|I_2| &= \binom{n}{2} = \frac{n(n-1)}{2}.
\end{aligned}$$

As a result,

$$\begin{aligned}
\mathsf{Var}(U_2) &= \frac{4}{n^2(n-1)^2}\left(|I_0| \cdot (\mathbb{E}X_1)^4 + |I_1| \cdot (\mathbb{E}X_1^2)(\mathbb{E}X_1)^2 + |I_2| \cdot (\mathbb{E}X_1^2)^2\right) - (\mathbb{E}X_1)^4 \\
&= -\frac{4n-6}{n(n-1)}(\mathbb{E}X_1)^4 + \frac{4(n-2)}{n(n-1)}(\mathbb{E}X_1^2)(\mathbb{E}X_1)^2 + \frac{2(\mathbb{E}X_1^2)^2}{n(n-1)},
\end{aligned}$$

and the result follows. □

D.8 Proof of Corollary 1

Recall that

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

To establish (33), we first show that for $f_h(x) \leq \frac{2c_1 \ln n}{nh^d}$,

$$\left|\mathbb{E}\min\left\{\hat{H}_1(x), \frac{1}{n^{1-2\varepsilon}h^d}\right\} + f_h(x)\ln f_h(x)\right| \lesssim \frac{1}{nh^d \ln n}. \tag{98}$$

In fact, by Lemma 7 and the Efron–Stein inequality (cf. Lemma 6),

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

Then by the bias bound in Lemma 7,

$$\mathbb{E}[\hat{H}_1(x)^2] = (\mathbb{E}[\hat{H}_1(x)])^2 + \mathsf{Var}(\hat{H}_1(x)) \lesssim \frac{1}{n^{2-\varepsilon}h^{2d}}.$$

Using $\mathbb{E}[X\mathbb{1}(X \geq c)] \leq c^{-1}\mathbb{E}[X^2]$ for any random variable X and $c > 0$, we have

$$0 \leq \mathbb{E}\hat{H}_1(x) - \mathbb{E}\min\left\{\hat{H}_1(x), \frac{1}{n^{1-2\varepsilon}h^d}\right\} \leq n^{1-2\varepsilon}h^d \cdot \mathbb{E}[\hat{H}_1(x)^2] \lesssim \frac{1}{n^{1+\varepsilon}h^d},$$

which together with the bias bound in Lemma 7 establishes (98).

By (98) and Lemma 8, we have

$$\mathbb{E}[\hat{H}(x)|X^{(1)}] + f_h(x)\ln f_h(x) = I_1(x)\mathbb{1}\left(\hat{f}_{h,1}(x) < \frac{c_1 \ln n}{nh^d}\right) + I_2(x)\mathbb{1}\left(\hat{f}_{h,1}(x) < \frac{c_1 \ln n}{nh^d}\right),$$

where

$$|I_1(x)| \triangleq \left|\mathbb{E}\min\left\{\hat{H}_1(x), \frac{1}{n^{1-2\varepsilon}h^d}\right\} + f_h(x)\ln f_h(x)\right| \overset{(79)}{\lesssim} \frac{1}{n^{1-2\varepsilon}h^d} + \frac{1}{h^d}\ln\frac{1}{h^d},$$
$$|I_2(x)| \triangleq |\mathbb{E}[\hat{H}_2(x)] + f_h(x)\ln f_h(x)| \overset{(79),(80)}{\lesssim} 1 + nh^d + \frac{1}{h^d}\ln\frac{n}{h^d}$$

for all $x \in \mathbb{R}^d$. Moreover, by (98) and Lemma 8, we also have $|I_1(x)| \lesssim (nh^d \ln n)^{-1}$ whenever $f_h(x) \le \frac{2c_1 \ln n}{nh^d}$ and $|I_2(x)| \lesssim (nh^d \ln n)^{-1}$ whenever $f_h(x) \ge \frac{c_1 \ln n}{2nh^d}$. Note that

$$\mathbb{E}\left|\mathbb{E}[\hat{H}(x)|X^{(1)}] + f_h(x)\ln f_h(x)\right|^2 = I_1(x)^2 \cdot \mathbb{P}\left(\hat{f}_{h,1}(x) < \frac{c_1 \ln n}{nh^d}\right) + I_2(x)^2 \cdot \mathbb{P}\left(\hat{f}_{h,1}(x) \ge \frac{c_1 \ln n}{nh^d}\right).$$

Hence, by considering the three cases of $f_h(x) \in (0, \frac{c_1 \ln n}{2nh^d}), [\frac{c_1 \ln n}{2nh^d}, \frac{2c_1 \ln n}{nh^d}], (\frac{2c_1 \ln n}{nh^d}, \infty)$ separately and applying Lemma 5 in the first and third cases, we arrive at the desired bound (33).

The proof of the upper bound on the second moment of the perturbation is similar. If one observation in $X^{(2)} \cup X^{(3)}$ is perturbed, we have

$$\mathbb{E}|\hat{H}(x) - \hat{H}'(x)|^2 = J_1(x)^2 \cdot \mathbb{P}\left(\hat{f}_{h,1}(x) < \frac{c_1 \ln n}{nh^d}\right) + J_2(x)^2 \cdot \mathbb{P}\left(\hat{f}_{h,1}(x) \ge \frac{c_1 \ln n}{nh^d}\right),$$

where

$$|J_1(x)| \triangleq \mathbb{E}\left|\min\left\{\hat{H}_1(x), \frac{1}{n^{1-2\varepsilon}h^d}\right\} - \min\left\{\hat{H}_1'(x), \frac{1}{n^{1-2\varepsilon}h^d}\right\}\right|^2 \lesssim \frac{1}{(n^{1-2\varepsilon}h^d)^2},$$
$$|J_2(x)| \triangleq \mathbb{E}|\hat{H}_2(x) - \hat{H}_2'(x)|^2 \overset{(80)}{\lesssim} \left(1 + nh^d + \frac{1}{h^d}\ln\frac{n}{h^d}\right)^2$$

for all x. Moreover, Lemma 7 and 8 give the improvement $|J_1(x)| \le \mathbb{E}|\hat{H}_1(x) - \hat{H}_1'(x)|^2 \lesssim \frac{1}{n^{3-\varepsilon}h^{2d}}$ whenever $f_h(x) \le \frac{2c_1 \ln n}{nh^d}$ and $|J_2(x)| \lesssim \frac{f_h(x)(1+(\ln f_h(x))^2)}{n^2h^d}$ whenever $f_h(x) \ge \frac{c_1 \ln n}{2nh^d}$. Again, by considering the three regimes $f_h(x) \in (0, \frac{c_1 \ln n}{2nh^d}), [\frac{c_1 \ln n}{2nh^d}, \frac{2c_1 \ln n}{nh^d}], (\frac{2c_1 \ln n}{nh^d}, \infty)$ and applying Lemma 5 in the first and third cases, we arrive at the desired bound (34).

□

D.9 Proof of Lemma 15

We use the following two-point argument. Fix $A > 0, \varepsilon > 0$ to be specified later. Let $f \in \mathrm{Lip}_{s,p,d}(L)$ be some fixed density supported on $[1/4, 3/4]^d$ which is bounded from below by some constant, say 0.1. Let g be a dilation of f defined by $g(x) = A^d f(A(x - \mathbf{1}/2))$ such that g is a density supported on $[1/2 - 1/(4A), 1/2 + 1/(4A)]^d$. We take $A \le L^{1/(s+d)}$ so that g is an element of $\mathrm{Lip}_{s,p,d}(L)$.

Consider testing the hypothesis $H_0 : X_i \overset{\text{i.i.d.}}{\sim} f_0 = \frac{f+g}{2}$ against $H_1 : X \overset{\text{i.i.d.}}{\sim} f_1 = \frac{1-\varepsilon}{2}f + \frac{1+\varepsilon}{2}g$. Then

$$\chi^2(f_1\|f_0) = \chi^2\left(\frac{f+g}{2}\middle\|\frac{1-\varepsilon}{2}f + \frac{1+\varepsilon}{2}g\right) \le \chi^2(\mathrm{Bern}((1+\varepsilon)/2)\|\mathrm{Bern}(1/2)) = \varepsilon^2,$$

where the inequality follows from the data processing inequality for χ^2-divergence. Thus, choosing $\varepsilon = 1/\sqrt{n}$ ensures the indistinguishability of the two hypotheses with n observations.

Finally, for the separation of entropy values, by the Taylor expansion of the function $t \mapsto t \ln t$ and that $f(x) \geq 0.1$ everywhere on its support, we have

$$|H(f_0) - H(f_1)| = \frac{\varepsilon}{2} \left| \int_{[\frac{1}{4}, \frac{3}{4}]^d} (f(x) - g(x)) \ln \frac{f(x) + g(x)}{2} dx \right| + O\left(\varepsilon^2 \int_{[\frac{1}{4}, \frac{3}{4}]^d} (f(x) - g(x))^2 dx \right).$$

Using $f(x) \geq 0.1$ and $\|f\|_\infty + \|g\|_\infty = O(A^d)$, the above inequality further gives $|H(f_0) - H(f_1)| = O(\varepsilon \ln A + \varepsilon^2 A^{2d})$. Finally, in view of $\varepsilon = 1/\sqrt{n}$, taking $A = \min\{L^{1/(s+d)}, n^{1/(4d)}\}$ yields the desired lower bound $n^{-1/2} \ln A \asymp n^{-1/2} \ln L$. □

D.10 Proof of Lemma 17

For any $\delta > 0$, suppose $\hat{H}_n$ nearly attains the minimax risk in the usual sampling model with

$$\sup_{f \in \mathrm{Lip}_{s,p,d}(L_0)} \mathbb{E}_f (\hat{H}_n - H(f))^2 \leq R_n^2 + \delta.$$

Now we use $\hat{H}_n$ to construct an estimator in the Poisson sampling model. Let $N \sim \mathsf{Poi}(n\|f\|_1)$ be the number of observations drawn in the Poisson sampling model, our estimator is just constructed as $\hat{H} = \hat{H}_N$. Note that conditioned on the event that $N = m$, the Poisson sampling model is just the usual sampling model with sample size m and density $\frac{f}{\|f\|_1}$. As a result, for any $P \in \mathcal{P}$, we have

$$\begin{aligned}
&\mathbb{E}_{f_P} (\hat{H} - H(f_P))^2 \\
&\leq 2\mathbb{E}_{f_P} \left(\hat{H} - H\left(\frac{f_P}{\|f_P\|_1} \right) \right)^2 + 2 \left(H(f_P) - H\left(\frac{f_P}{\|f_P\|_1} \right) \right)^2 \\
&= 2 \sum_{m=0}^{\infty} \mathbb{E}_{f_P} \left[\left(\hat{H} - H\left(\frac{f_P}{\|f_P\|_1} \right) \right)^2 \middle| N = m \right] \cdot \mathbb{P}(N = m) + 2 \left(H(f_P) - H\left(\frac{f_P}{\|f_P\|_1} \right) \right)^2 \\
&\leq 2 \sum_{n=0}^{\infty} (R_n^2 + \delta) \mathbb{P}(N = n) + 2 \left(H(f_P) - H\left(\frac{f_P}{\|f_P\|_1} \right) \right)^2 \\
&\leq 2c_0 \cdot \mathbb{P}(N \leq n/2) + 2R_{\frac{n}{2}}^2 \cdot \mathbb{P}(N > n/2) + 2\delta + 2 \left(H(f_P) - H\left(\frac{f_P}{\|f_P\|_1} \right) \right)^2 \\
&\leq 2c_0 \exp(-\frac{n}{5}) + 2R_{\frac{n}{2}}^2 + 2\delta + 4(\|f_P\|_1 - 1)^2 c_0 + 4(\|f_P\|_1 \ln \|f_P\|_1)^2
\end{aligned}$$

where $c_0 \triangleq \sup_{f \in \mathrm{Lip}_{s,p,d}(L)} H(f)^2 \lesssim (\ln L)^2$ in view of Lemma 10, and in the last step we have used the Poisson tail bound. Note that by definition of $\mathcal{P}$,

$$|\|f_P\|_1 - 1| \lesssim \frac{1}{nh^d (\ln n)^3 \ln L},$$

and thus we conclude that

$$R_n^P \lesssim R_{\frac{n}{2}} + \frac{1}{nh^d (\ln n)^3} + \sqrt{\delta}.$$

The desired result follows by the arbitrariness of $\delta > 0$. □

D.11 Proof of Lemma 20

We introduce some necessary definitions and results from approximation theory first. For functions defined on $[0,1]$, define the r-th order Ditzian–Totik modulus of smoothness by [DT87]

$$\omega_\varphi^r(f,t)_\infty \triangleq \sup_{0<h\le t} \|\Delta_{h\varphi(x)}^r f(x)\|_\infty \tag{99}$$

where $\varphi(x) \triangleq \sqrt{x(1-x)}$. This quantity is related to the polynomial approximation error via the following lemma.

Lemma 25 ([DT87]). *For any integer $u>0$ and $n>u$, there exists some constant M_u depending only on u, such that for all $t\in(0,1)$ and f,*

$$E_n(f;[0,1]) \le M_u \omega_\varphi^u(f,1/n)_\infty$$
$$\frac{M_u}{n^u}\sum_{l=0}^{n}(l+1)^{u-1}E_l(f;[0,1]) \ge \omega_\varphi^u(f,1/n)_\infty.$$

Defining $r=q-1\ge 0$, we will make use of the second inequality in Lemma 25 to prove that $E_{r,n}(x^{-r}\ln x;[cn^{-2},1]) \gtrsim n^{2r}$ for some proper constant c. The case $r=0$ has been handled in [WY16, Lemma 5], and we assume that $r>0$. Note that by definition (51), for any function f and interval I, the rational approximation error can be related to the polynomial approximation error as

$$E_{r,n}(f;I) = \inf_{a_1,\dots,a_r} E_n\left(f(x)+\sum_{l=1}^{r} a_l x^{-l}; I\right).$$

Thus it suffices to prove that for any coefficients $a_1,\dots,a_r\in\mathbb{R}$, $g(x)\triangleq x^{-r}\ln x+\sum_{l=1}^r a_l x^{-l}$ and $\tilde g(x)\triangleq g(cn^{-2}+(1-cn^{-2})x)$, the inequality

$$E_n(\tilde g;[0,1]) \ge c'$$

holds with proper universal constants c,c' independent of the choices of $a_1,\dots,a_r$. For the sake of simplicity, in the sequel we use the following equivalent definition of $g(x)$ (by a proper translation of a_r):

$$g(x) \triangleq x^{-r}\ln(c^{-1}n^2x) + \sum_{l=1}^{r} a_l x^{-l}.$$

Let $u=1$ and $m=c^{-\frac{1}{2}}n$ with $c<1$ in Lemma 25, we have

$$\begin{aligned}
E_n(\tilde g;[0,1]) &\ge \frac{1}{m}\sum_{l=n}^{m} E_l(\tilde g;[0,1]) \\
&\ge \frac{1}{M_1}\omega_\varphi^1(\tilde g,\frac{1}{m})_\infty - \frac{1}{m}\sum_{l=0}^{n-1} E_l(\tilde g;[0,1]) \\
&\ge \frac{1}{M_1}\omega_\varphi^1(\tilde g,\frac{1}{m})_\infty - \frac{n}{m}E_0(\tilde g;[0,1]).
\end{aligned}$$

We distinguish into two cases. If $E_0(\tilde{g};[0,1]) \le 2m^{2r}$, by definition of the Ditzian–Totik modulus of smoothness, there exists universal constants $0 < A < B$ (independent of c) such that

$$\begin{aligned}
\omega_\varphi^1(\tilde{g}, \frac{1}{m}) &\ge \max_{z\in[A,B]} \left| \left(\frac{z+1}{m^2}\right)^{-r} \ln(z+1) - \left(\frac{z}{m^2}\right)^{-r} \ln z + \sum_{l=1}^{r} a_l \left(\left(\frac{z+1}{m^2}\right)^{-l} - \left(\frac{z}{m^2}\right)^{-l} \right) \right| \\
&= m^{2r} \cdot \max_{z\in[A,B]} \left| h(z) - \sum_{l=1}^{r} a_l m^{2(l-r)} h_l(z) \right| \\
&\ge m^{2r} \cdot \inf_{b_1,\ldots,b_r} \max_{z\in[A,B]} \left| h(z) - \sum_{l=1}^{r} b_l h_l(z) \right|
\end{aligned}$$

where

$$\begin{aligned}
h(z) &\triangleq (z+1)^{-r} \ln(z+1) - z^{-r} \ln z, \\
h_l(z) &\triangleq (z+1)^{-l} - z^{-l}, \qquad l = 1, \ldots, r.
\end{aligned}$$

It is clear that the functions $h_l, l = 1, \ldots, r$ and h are linearly independent over $[A, B]$, and hence

$$\inf_{b_1,\ldots,b_r} \max_{z\in[A,B]} \left| h(z) - \sum_{l=1}^{r} b_l h_l(z) \right| \ge C_1$$

for some universal constant $C_1 > 0$. Hence, in this case we have

$$E_n(\tilde{g};[0,1]) \ge \left(\frac{C_1}{M_1} - 2\sqrt{c} \right) m^{2r}. \tag{100}$$

Now we consider the second case where $E_0(\tilde{g};[0,1]) > 2m^{2r}$, which implies that

$$\begin{aligned}
E_0(\tilde{g};[0,1]) &\le \max_{x\in[m^{-2},1]} \left| x^{-r} \ln(m^2 x) + \sum_{l=1}^{r} a_l x^{-l} \right| \\
&\le m^{2r} \cdot \max_{z\in[1,\infty)} z^{-r} \ln z + \sum_{l=1}^{r} |a_l| m^{2l} \\
&\le m^{2r} + \sum_{l=1}^{r} |a_l| m^{2l} \\
&\le \frac{1}{2} E_0(\tilde{g};[0,1]) + r \cdot \max_{1\le l\le r} |a_l| m^{2l}.
\end{aligned}$$

As a result,

$$\max_{1\le l\le r} |a_l| m^{2l} \ge \frac{E_0(\tilde{g};[0,1])}{2r}.$$

Suppose the maximum on the LHS is achieved by l^*. Then we have

$$\begin{aligned}
\omega_\varphi^1(\tilde{g}, \frac{1}{m}) &\geq \max_{z\in[A,B]} \left| \left(\frac{z+1}{m^2}\right)^{-r} \ln(z+1) - \left(\frac{z}{m^2}\right)^{-r} \ln z + \sum_{l=1}^{r} a_l \left(\left(\frac{z+1}{m^2}\right)^{-l} - \left(\frac{z}{m^2}\right)^{-l} \right) \right| \\
&= |a_{l^*}| m^{2l^*} \cdot \max_{z\in[A,B]} \left| h_{l^*}(z) - \frac{m^{2(r-l^*)}}{a_{l^*}} h(z) - \sum_{l\neq l^*} \frac{m^{2(l-l^*)} a_l}{a_{l^*}} h_l(z) \right| \\
&\geq |a_{l^*}| m^{2l^*} \cdot \inf_{b, b_l : l\neq l^*} \max_{z\in[A,B]} \left| h_{l^*}(z) - bh(z) - \sum_{l\neq l^*} b_l h_l(z) \right| \\
&\geq C_2 |a_{l^*}| m^{2l^*} \\
&\geq \frac{C_2 E_0(\tilde{g}; [0,1])}{2r}
\end{aligned}$$

where, again by the linear independence,

$$C_2 \triangleq \min_{1\leq l^*\leq r} \inf_{b, b_l : l\neq l^*} \max_{z\in[A,B]} \left| h_{l^*}(z) - bh(z) - \sum_{l\neq l^*} b_l h_l(z) \right| > 0$$

is a universal constant. Hence

$$\begin{aligned}
E_n(\tilde{g}; [0,1]) &\geq \left(\frac{C_2}{2rM_1} - \sqrt{c} \right) E_0(\tilde{g}; [0,1]) \\
&\geq 2 \left(\frac{C_2}{2rM_1} - \sqrt{c} \right) m^r. \qquad (101)
\end{aligned}$$

Combining (100) and (101), we conclude that

$$\begin{aligned}
E_{r,n}(x^{-r} \ln x; [cn^{-2}, 1]) &\geq \min \left\{ \frac{C_1}{M_1} - 2\sqrt{c}, 2 \left(\frac{C_2}{2rM_1} - \sqrt{c} \right) \right\} \cdot m^r \\
&= \min \left\{ \frac{C_1}{M_1} - 2\sqrt{c}, 2 \left(\frac{C_2}{2rM_1} - \sqrt{c} \right) \right\} \cdot (\frac{n}{\sqrt{c}})^r.
\end{aligned}$$

Now choosing $c > 0$ small enough completes the proof of the lemma. □

References

[BDGVdM97] Jan Beirlant, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

[BKB+93] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Y Ritov, J Klaassen, Jon A Wellner, and Ya'acov Ritov. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[BM95] Lucien Birgé and Pascal Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, pages 11–29, 1995.

[BR88] Peter J Bickel and Ya'acov Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

[BSY19] Thomas B Berrett, Richard J Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, 2019.

[BZLV18] Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Estimation of kl divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4):2648–2674, 2018.

[CH04] Jose A Costa and Alfred O Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004.

[CL03] T Tony Cai and Mark G Low. A note on nonparametric estimation of linear functionals. *Annals of statistics*, 31(4):1140–1153, 2003.

[CL05] T Tony Cai and Mark G Low. Nonquadratic estimators of a quadratic functional. *The Annals of Statistics*, 33(6):2930–2956, 2005.

[CL11] T Tony Cai and Mark G Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[DF17] Sylvain Delattre and Nicolas Fournier. On the Kozachenko–Leonenko entropy estimator. *Journal of Statistical Planning and Inference*, 185:69–93, 2017.

[DL93] Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer, 1993.

[DN90] David L Donoho and Michael Nussbaum. Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, 6(3):290–323, 1990.

[Don97] David L Donoho. Renormalizing experiments for nonlinear functionals. *Festschrift for Lucien Le Cam*, pages 167–181, 1997.

[DT87] Zeev Ditzian and Vilmos Totik. *Moduli of smoothness*. Springer, 1987.

[EHHG09] Fidah El Haje Hussein and Yu Golubev. On entropy estimation by m-spacing method. *Journal of Mathematical Sciences*, 163(3):290–309, 2009.

[Fan91] Jianqing Fan. On the estimation of quadratic functionals. *The Annals of Statistics*, 19(3):1273–1294, 1991.

[GM92] Larry Goldstein and Karen Messer. Optimal plug-in estimators for nonparametric functional estimation. *The Annals of Statistics*, 20(3):1306–1328, 1992.

[GOV16] Weihao Gao, Sewoong Oh, and Pramod Viswanath. Breaking the bandwidth barrier: Geometrical adaptive entropy estimation. In *Advances in Neural Information Processing Systems*, pages 2460–2468, 2016.

[GOV18] Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*, 64(8):5629–5661, 2018.

[GVdM91] László Györfi and Edward C Van der Meulen. On the nonparametric estimation of the entropy functional. In *Nonparametric functional estimation and related topics*, pages 81–95. Springer, 1991.

[Hal84] Peter Hall. Limit theorems for sums of general functions of m-spacings. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 96, pages 517–532. Cambridge University Press, 1984.

[HI80] Rafail Z Hasminskii and Ildar A Ibragimov. Some estimation problems for stochastic differential equations. In *Stochastic Differential Systems Filtering and Control*, pages 1–12. Springer, 1980.

[HJMW17] Yanjun Han, Jiantao Jiao, Rajarshi Mukherjee, and Tsachy Weissman. On estimation of L_r-norms in Gaussian white noise models. *arXiv preprint arXiv:1710.03863*, 2017.

[HJW16] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax rate-optimal estimation of divergences between discrete distributions. *arXiv preprint arXiv:1605.09124*, 2016.

[HJWW17] Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over Lipschitz balls. *arXiv preprint*, Nov 2017. https://arxiv.org/abs/1711.02141v2.

[HKPT12] Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 2012.

[HM87] Peter Hall and James Stephen Marron. Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 6(2):109–115, 1987.

[HM93] Peter Hall and Sally C Morton. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45(1):69–88, 1993.

[HSPVB07] Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.

[INH87] Ildar A Ibragimov, Arkadi S Nemirovskii, and Rafail Z Hasminskii. Some problems on nonparametric estimation in Gaussian white noise. *Theory of Probability & Its Applications*, 31(3):391–406, 1987.

[Jaw77] Björn Jawerth. Some observations on Besov and Lizorkin–Triebel spaces. *Mathematica Scandinavica*, 40(1):94–104, 1977.

[JHW17] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Bias correction with jackknife, bootstrap, and taylor series. *arXiv preprint arXiv:1709.06183*, 2017.

[JHW18] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

[Joe89] Harry Joe. Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 41(4):683–697, 1989.

[JS77] Hans Johnen and Karl Scherer. On the equivalence of the K-functional and moduli of continuity and some applications. In *Constructive theory of functions of several variables*, pages 119–140. Springer, 1977.

[JVHW15] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[KKPW14] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman. Nonparametric estimation of Rényi divergence and friends. In *International Conference on Machine Learning*, pages 919–927, 2014.

[KKPW15] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, and Larry Wasserman. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pages 397–405, 2015.

[KP96] Gérard Kerkyacharian and Dominique Picard. Estimating nonquadratic functionals of a density using haar wavelets. *The Annals of Statistics*, 24(2):485–507, 1996.

[KSG04] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[KT12] Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media, 2012.

[Lau96] Béatrice Laurent. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

[Lev78] Boris Ya Levit. Asymptotically efficient estimation of nonlinear functionals. *Problemy Peredachi Informatsii*, 14(3):65–72, 1978.

[LNS99] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the L_r norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.

[MNR17] Rajarshi Mukherjee, Whitney K Newey, and James M Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.

[MSGHI16] Kevin R Moon, Kumar Sricharan, Kristjan Greenewald, and Alfred O Hero III. Nonparametric ensemble estimation of distributional functionals. *arXiv preprint arXiv:1601.06884*, 2016.

[Nem00] Arkadi Nemirovski. Topics in non-parametric statistics. *Ecole dEté de Probabilités de Saint-Flour*, 28, 2000.

[Nus96] Michael Nussbaum. Asymptotic equivalence of density estimation and gaussian white noise. *The Annals of Statistics*, pages 2399–2430, 1996.

[Pan04] Liam Paninski. Estimating entropy on m bins given fewer than m samples. *Information Theory, IEEE Transactions on*, 50(9):2200–2203, 2004.

[PR16] Tim Patschkowski and Angelika Rohde. Adaptation to lowest density regions with application to support recovery. *The Annals of Statistics*, 44(1):255–287, 2016.

[PY08] Liam Paninski and Masanao Yajima. Undersmoothed kernel entropy estimators. *IEEE Transactions on Information Theory*, 54(9):4384–4388, 2008.

[Rén59] Alfréd Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Hungarica*, 10(1–2), Mar 1959.

[RLM+17] James Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Higher order estimating equations for high-dimensional models. *Annals of statistics*, 45(5):1951–1987, 2017.

[RLTvdV08] James Robins, Lingling Li, Eric Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

[RSH18] Kolyan Ray and Johannes Schmidt-Hieber. Asymptotic nonequivalence of density estimation and gaussian white noise for small densities. *arXiv preprint arXiv:1802.03425*, 2018.

[SP16] Shashank Singh and Barnabás Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Advances in Neural Information Processing Systems*, pages 1217–1225, 2016.

[SRH12] Kumar Sricharan, Raviv Raich, and Alfred O Hero. Estimation of nonlinear functionals of densities with confidence. *IEEE Transactions on Information Theory*, 58(7):4135–4159, 2012.

[ST77] LI Strukov and AF Timan. Mathematical expectation of continuous functions of random variables. smoothness and variance. *Siberian Mathematical Journal*, 18(3):469–474, 1977.

[Ste79] Elias M Stein. *Singular integrals and differentiability properties of functions (PMS-30)*. Princeton university press, 1979.

[Ste86] J Michael Steele. An Efron-Stein inequality for nonsymmetric statistics. *The Annals of Statistics*, pages 753–758, 1986.

[Tao15] Terrence Tao. The Hardy-Littlewood maximal inequality and applications. Lecture notes for MATH 247A: Fourier analysis. 2015. http://www.math.ucla.edu/~tao/247a.1.06f/notes3.pdf.

[Tim63] Aleksandr Filippovich Timan. *Theory of approximation of functions of a real variable*. Pergamon Press, 1963.

[TLRvdV08] Eric Tchetgen, Lingling Li, James Robins, and Aad van der Vaart. Minimax estimation of the integral of a power of a density. *Statistics & Probability Letters*, 78(18):3307–3311, 2008.

[Tot94] Vilmos Totik. Approximation by Bernstein polynomials. *American Journal of Mathematics*, pages 995–1018, 1994.

[Tsy09] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, 2009.

[TVdM96] Alexandre B Tsybakov and EC Van der Meulen. Root-n consistent estimators of entropy for densities with unbounded support. *Scandinavian Journal of Statistics*, pages 75–83, 1996.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VE92] Bert Van Es. Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics*, pages 61–72, 1992.

[VV11] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 403–412. IEEE, 2011.

[WKV09] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Universal estimation of information measures for analog sources. *Foundations and Trends in Communications and Information Theory*, 5(3):265–353, 2009.

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[WY19] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019.