Optimal Rates for Nonparametric Density Estimation Under Communication Constraints

Jayadev Acharya, Member, IEEE, Clément L. Canonne, Aditya Vikram Singh, and Himanshu Tyagi

Abstract—We consider density estimation for Besov spaces when each sample is quantized to only a limited number of bits. We provide a noninteractive adaptive estimator that exploits the sparsity of wavelet bases, along with a simulate-and-infer technique from parametric estimation under communication constraints. We show that our estimator is nearly rate-optimal by deriving minimax lower bounds that hold even when interactive protocols are allowed. Interestingly, while our wavelet-based estimator is almost rate-optimal for Sobolev spaces as well, it is unclear whether the standard Fourier basis, which arise naturally for those spaces, can be used to achieve the same performance.

Index Terms—Density estimation, distributed adaptive estimation, quantization, interactive lower bound.

I. INTRODUCTION

ESTIMATING distributions from samples, i.e., density estimation, is a fundamental statistical task. Modern applications, such as those arising in Federated Learning or the Internet of Things (IoT), often limit access to the true data samples. One common limitation in large scale distributed systems is communication constraints, which require that each data sample must be compressed to a small number of bits.

Most prior work on communication-constrained estimation has focused on parametric problems such as Gaussian mean estimation and discrete distribution estimation. In this work, we study nonparametric density estimation under communication constraints where independent samples from an unknown distribution (whose density \( f \) lies in a suitable function class) are distributed across players (one sample per player), each of whom is restricted to only sending \( \ell \) bits about their sample to a central referee; the referee outputs an estimate of \( f \) based on these \( \ell \)-bit messages. This problem has been considered by [2] for densities in Hölder classes, a relatively simple class for which the normalized histogram with uniform bins is known to be optimal in the centralized setting (that is, when all the samples are available to the estimator). For densities in Hölder classes, a natural method for density estimation in the distributed setting is to quantize each data sample into uniform bins and use the optimal estimator for distributed discrete distribution estimation. Indeed, [2] shows that this is optimal for distributed estimation of densities from the Hölder class under communication constraints. However, this simple estimator does not seem to extend to the richer Sobolev class and the most general Besov classes. In particular, the following question is largely open:

How to quantize samples to estimate densities from Besov classes under communication constraints?

We resolve this question for the cases when the density belongs to a Besov class with known parameters (nonadaptive setting), as well as when the density belongs to a Besov class where only upper and lower bounds on parameters are known (adaptive setting). Specifically, our proposed estimators exploit the sparsity of wavelet basis for the Besov class, and use vector quantization followed by the distributed simulation technique introduced in [3] for distributed parametric estimation. We also establish information-theoretic lower bounds that prove the optimality of our estimators (up to logarithmic terms in the adaptive setting).

A. Problem Setup

\( X_1, \ldots, X_n \) are independent samples from an unknown distribution with density \( f \) supported on \( \mathcal{X} := [0, 1] \) and belonging to the Besov space \( \mathcal{B}(p, q, s) \) (see Section II for details; briefly, if \( f \) is written in terms of wavelets as in (4), then \( f \) belongs to \( \mathcal{B}(p, q, s) \) if the wavelet coefficients decay fast enough to satisfy (5)). There are \( n \) distributed users (players), with player \( i \) having access to sample \( X_i \). Each player can only transmit \( \ell \) bits about their sample to a central server (referee) whose goal, upon observing \( \ell \)-bit messages from \( n \) players, is to estimate \( f \). We consider an interactive setting, where the current player observes the messages from previous players and can use them to design their message.\(^1\) That is, in round \( i \), player \( i \) chooses a communication-constrained channel (randomized mapping) \( W_i : \mathcal{X} \rightarrow \{0, 1\}^\ell \) as a function of prior messages \( Y_1, \ldots, Y_{i-1} \) and randomness \( U \) available to all players; and then passes \( X_i \) through \( W_i \) to generate \( Y_i \in \{0, 1\}^\ell \). The referee observes the messages

\(^1\)Since our lower bounds rely on general results in [4], we borrow the notation from that paper.
Let \( Y_1, \ldots, Y_n \) and outputs an estimate \( \hat{f} \) of \( f \). We term such an \( \hat{f} \) an \((n, \ell)\)-estimate, and let \( \mathcal{E}_{n, \ell} \) denote the set of all \((n, \ell)\)-estimates. Our goal is to design estimators that achieve the minimax expected \( \mathcal{L}_r \)-loss defined, for \( r \geq 1 \), by

\[
\mathcal{L}_r^*(n, \ell, p, q, s) := \inf_{f \in \mathcal{E}_{n, \ell}} \sup_{f \in \mathcal{B}(p, q, s)} \mathbb{E}_f \left[ \| \hat{f} - f \|_r^r \right].
\]

For upper bounds on \( \mathcal{L}_r^*(n, \ell, p, q, s) \), we consider algorithms that use the more restricted noninteractive private-coin protocols, where the channel \( W_i \) of player \( i \) is not allowed to depend on the messages \( Y_1, \ldots, Y_{i-1} \) or on the common randomness \( U \), but may depend on the private randomness \( U_i \) available at player \( i \), where \( U_1, \ldots, U_n \) are independent of each other and jointly of \( U \). Noninteractive protocols are typically easier to implement and result in much simpler engineering for the distributed system, and as such the question of whether they can perform as well as their interactive counterparts in terms of data requirements is of significant interest.

**B. Our Results and Techniques**

Our first result is an information-theoretic lower bound on \( \mathcal{L}_r^* \). We denote \( \max\{a, b\} \) by \( a \lor b \).

**Theorem 1:** For any \( p, q, s, r \), there exist constants \( C = C(p, q, s, r) > 0 \), \( \alpha = \alpha(p, q, s, r) > 0 \) such that

\[
\mathcal{L}_r^*(n, \ell, p, q, s) \geq \begin{cases} 
C \cdot n^{-\frac{r}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}} & \text{if } r \leq (s+1)p; \\
C \cdot n^{-\frac{r}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}} & \text{if } r \in ((s+1)p, (2s+1)p); \\
(n \log n)^{-\frac{r(1+\ell)/p+1)}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}}} & \text{if } r \geq (2s+1)p.
\end{cases}
\]

We emphasize that this lower bound applies to interactive protocols as defined in Section I-A. When the parameters \( p, q, s \) of the Besov space are known, we design a noninteractive estimator that achieves the optimal rate for \( r \leq p \).

**Theorem 2:** For any \( r \geq 1 \) and \( p, q, s \) with \( r \leq p \), there exist a constant \( C = C(p, q, s, r) \) and an \((n, \ell)\)-estimate \( \hat{f} \) formed using a noninteractive protocol such that

\[
\sup_{f \in \mathcal{B}(p, q, s)} \mathbb{E}_f \left[ \| \hat{f} - f \|_r^r \right] \leq C \max \{ n^{-\frac{r}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}}, (n^{p})^{-\frac{r}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}} \}.
\]

We finally design an adaptive, noninteractive estimator that only requires bounds on \( s \), and no further knowledge of \( p \) and \( q \). Moreover, this estimator achieves (up to logarithmic factors) the optimal rate for all parameter values.

**Theorem 3:** For any \( N \in \mathbb{N} \), \( r \geq 1 \), and \( p, q, s \) with \( 1/p < s < N \), there exist constants \( C = C(p, q, s, r), \alpha = \alpha(p, q, s, r) \), and an \((n, \ell)\)-estimate \( \hat{f} \) formed using a noninteractive protocol such that

\[
\sup_{f \in \mathcal{B}(p, q, s)} \mathbb{E}_f \left[ \| \hat{f} - f \|_r^r \right] \leq C \max \{ n^{-\frac{r}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}}, (n^{p})^{-\frac{r}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}} \}.
\]

It worth noting that the first term of the maximum in all cases is the standard, unconstrained nonparametric rate (cf. [5], or the discussion below), while the second term reflects the convergence slowdown due to the communication constraints. The effect of communication constraints disappears when \( \ell \) is sufficiently large. In particular, we get back the centralized rates when \( \ell \) satisfies

\[
\ell \geq (\log n) \cdot \begin{cases} 
\frac{1}{2(s+1)p+1} & \text{if } r \leq (s+1)p; \\
\frac{2(s+1)p+1}{2s+1} & \text{if } r \geq (2s+1)p.
\end{cases}
\]

For the standard \( \mathcal{L}_2 \) loss, with, say \( p \geq 2 \), the minimax rate becomes (up to constants) the more interpretable quantity

\[
\mathcal{L}_2^*(n, \ell, p, q, s) \approx \max \{ n^{-\frac{r}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}}, (n^{p})^{-\frac{r}{2(r(1+\ell)/p+1)\log(n^{p})^{-\alpha}}} \},
\]

where we see that the \( \ell \)-bit communication constraint reduces the exponent of the convergence rate from \( \frac{2s}{2s+1} \) to \( \frac{2s}{2s+2} \). As \( \ell \) grows or \( s \) tends to \( \infty \), the two rates coincide. Finally, from (2) we observe qualitative changes at \( r = (s+1)p \) and \( r = (2s+1)p \), where the rate exponent changes slope. This phenomenon, whose analogue is observed in the unconstrained setting [5] as well as under local privacy constraints [6], is sometimes referred to as the elbow effect.

We note that there are works, e.g. [7], that characterize the exact constants in the minimax rates in certain asymptotic regimes and for certain special losses, say \( \mathcal{L}_2 \). In contrast, the focus of our work is on the non-asymptotic regime.
with $\mathcal{L}_r$, loss for arbitrary $r \in [1, \infty]$. This is in line with the classical work [5] that solved this problem in the unconstrained setting. Our main interest is in characterizing the dependence of minmax rate on $\ell$ and our results show an exponential decay of rate with $\ell$ (up to constants) non-asymptotically.

1) **Quantize, Simulate and Infer:** A conceptually simple technique for distributed estimation under communication constraints (“simulate-and-infer”) was proposed in [3], which uses communication to simulate samples from the unknown distribution, and provides an optimal rate estimator for discrete distribution estimation under communication constraints. A natural approach for nonparametric estimation under communication constraints would be to quantize the samples to the available number of bits (i.e., $\ell$) and use this quantized sample to estimate the distribution. However, it is unclear if this leads to optimal rates. Instead, in our approach, we quantize not to approximate the value of the sample but to approximate a sufficient statistic without inducing any “loss of information.” That is, we form an approximately sufficient statistic (based on wavelets) that can be represented using a finite number of bits and does not result in rate loss (up to constants). The number of bits could still be more than $\ell$, and therefore, we apply “simulate-and-infer” to generate samples from the statistic and output an estimate. Thus, the loss of information due to communication constraints only happens in the last step, when we use multiple samples to simulate a sample from the sufficient statistic; the quantization part is just for efficient finite representation.

2) **Sobolev Spaces and Fourier Bases:** A natural first approach for nonparametric estimation over Sobolev spaces is to use the Fourier basis. However, all our attempts with the Fourier basis led to a suboptimal performance either in the dependence on $\ell$ (falling short of the exponential $2^{-\ell}$ dependence) or the exponent of $n$ (i.e., an estimate whose rate featured the right $2^{-\ell}$ dependence, but a suboptimal exponent of $n$). Somewhat surprisingly, the more general wavelet-based approach described above gives us tight bounds for Sobolev space as well, since Sobolev space $S(\beta) = B(2, 2, \beta)$. Thus, it seems necessary to use the wavelet representation even for Sobolev spaces to get an appropriately “small” statistic. One reason for this is that the sparsity induced by the wavelet representation (see Section III) turns out to play a crucial role for inference under communication constraints.

3) **Organization:** After discussing prior works (Section I-C) and preliminaries (Section II), we describe our estimation algorithms in Section III and Section IV. In Section III, we present the nonadaptive single level estimator (comprising Algorithm 2 and Algorithm 3), which achieves guarantees of Theorem 2. In Section IV, we present the adaptive multi-level estimator (comprising Algorithm 4 and Algorithm 5), which achieves guarantees of Theorem 3. Finally, in Section V, we prove information-theoretic lower bounds of Theorem 1.

### C. Prior Work

The seminal work of Donoho, Johnstone, Kerkhaycharian, and Picard [5] proposed wavelet estimators for Besov class in the centralized setting, and showed that these estimators achieve near-optimal rates of convergence (up to logarithmic factors),

$$
\mathcal{L}_r^*(n, \infty, p, q, s) = \begin{cases} 
n^{-\frac{r}{q+s}}, & r < (2s+1)p, \\
n^{-\frac{r}{s+1} + \frac{1}{2(s-1)p+1/r}}, & r \geq (2s+1)p.
\end{cases}
$$

(3)

Their results highlight the fact that linear estimators are inherently suboptimal for estimation with respect to $\mathcal{L}_r$ losses, when $r$ is large; that is, some nonlinearity in the estimator is required to achieve optimal rates for $r > p$. In particular, they show that nonlinearity in the form of thresholding achieves optimal rates for $r > p$ (see Section II-C below for details). Further, they use thresholding to design adaptive estimators that achieve near-optimal rates. These minimax rates exhibit the aforementioned elbow effect, where the error exponent is only piecewise linear, and changes slope at $r = (2s+1)p$. We refer the reader to [5] for a further discussion of these phenomena.

Butucea et al. [6] recently extended these ideas to obtain near-rate optimal estimators for Besov spaces under local differential privacy constraints. Their adaptive estimator, as well as the information-theoretic lower bounds they establish, show that similar phenomena occur in the context of locally private nonparametric estimation. Our work, specifically the analysis of our adaptive estimator, draws upon some of the ideas of [6], with some crucial differences. In particular, the key ideas underlying our estimators – the wavelet-induced sparsity (Claim 4), the use of distributed simulation, and vector quantization – are neither present in nor applicable to the setting of [6] (where the introduction of random noise to ensure differential privacy effectively removes wavelet sparsity). Furthermore, our lower bounds even apply to interactive protocols, unlike the lower bounds from [6] which are restricted to the noninteractive setting.

In summary, our paper is the first to derive the counterpart of the nonparametric density estimation results of [5] and [6] under communication constraints, and shows that the analogue of the phenomena observed in [5] holds in the communication-constrained setting.

1) **Other Works on Distributed Estimation:** We briefly discuss the related literature on distributed (communication-constrained) estimation problems. References [8], [9], [10], and [11] have studied the problem of distributed nonparametric function estimation (regression) under a Gaussian white noise model in a noninteractive setting with $n$ players, where each player observes an independent copy of the stochastic process $f(t) = f(t)dt + (1/\sqrt{N})dW(t)$, $0 \leq t \leq 1$. Here $W(t)$ is the standard Wiener process, and $f$ is the function to be estimated. Reference [8] derive minimax rates for $f$ in Sobolev space under $\mathcal{L}_\alpha^*$ loss, where each player can send at most $\ell$ bits. Reference [9] derive minimax rates for $f$ in the Besov space $B(2, \infty, s)$ (“Besov type”) under $\mathcal{L}_2^*$ loss, and $f$ in $B(\infty, \infty, s)$ (“Hölder type”) under $\mathcal{L}_\infty$ loss, where each player can send at most $\ell$ (assumed to be at least $\log N$) bits on average. Further, the paper proposes near-optimal adaptive estimators (based on Lepski’s method) that adapt to the smoothness parameter $s$, provided that $s \in [s_{\min}, s_{\max}]$, where $s_{\min}$ depends on $n, \ell$ and $s_{\max}$ can be arbitrary. Reference [10] further study the problem of adaptivity for
The papers discussed above consider the nonparametric regression (i.e., function estimation) problem, while we consider the nonparametric density estimation problem. In the classical setting, these problems have similar rates. However, we observe an interesting divergence between them in the communication-constrained setting. Namely, when the communication budget $\ell$ is insufficient to achieve centralized rates, the minimax rates in the distributed nonparametric regression problems decay polynomially in $\ell$, while we show that the minimax rate for distributed nonparametric density estimation problem decays exponentially in $\ell$.

In another related work, [12] consider estimation of Hölder smooth density function $p_{XY}$ at a given point in a two-party setting, where $X$ and $Y$ are $d$-dimensional. One party observes only the $X$ coordinate of the samples and the other party only the $Y$ coordinate of the samples, and the parties must communicate to estimate the unknown density. They show that interactive communication strictly helps over one-way communication.

We now discuss related works on distributed parametric estimation problems. Reference [2] establish lower bounds on parametric density estimation, and on some restricted nonparametric families (Hölder classes) by bounding the Fisher information. Reference [3] obtain upper and lower bounds for discrete distribution estimation; our algorithms leverage the concept of distributed simulation ("simulate-and-infer") introduced in that context. Reference [13] derive lower bounds for various parametric estimation tasks, including discrete distributions and continuous (parametric) families such as high-dimensional Gaussians (including the sparse case). Reference [4], building on [14] (which focused on learning and testing discrete distributions), developed a general technique to prove estimation lower bounds for parametric families; our lower bounds rely on their framework, by suitably extending it to handle the nonparametric case.

We note that there are other approaches for establishing lower bounds under communication constraints such as the early works [15], [16], [17], and [18], where bounds for specific inference problems under communication constraints were obtained; and [19], where Cramér–Rao bounds for this setting were developed. We found the general approach of [4] best fits our specific application, where we needed to handle interactive communication as well as a nonuniform prior on the parameter in the lower bound construction.

II. Preliminaries

In this section, we first set out the notation used in the paper, before formally defining Besov spaces. Then, we list the assumptions that we make on the density function $f$, and briefly recall some aspects the existing estimators for density estimation in the centralized setting.

Given two integers $m \leq n$, we write $[m, n]$ for the set $\{m, m + 1, \ldots, n\}$ and $[n]$ for $\{1, n\}$. For two sequences or functions $(a_n)_n$, $(b_n)_n$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ (independent of $n$) such that $a_n \leq C b_n$ for all $n$, and $a_n \asymp b_n$ (or $a_n \approx b_n$) if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For a function $g$, $\text{supp}(g)$ denotes the support of $g$.

A. Besov Spaces

Our exposition here is based on [5] and [20]. We start with a discussion on wavelets.

1) Wavelets: A wavelet basis for $L^2(\mathbb{R})$ is generated using two functions: $\phi$ (father wavelet) and $\psi$ (mother wavelet). The main feature that distinguishes a wavelet basis from the Fourier basis is that the functions $\phi$ and $\psi$ can have compact support.

More precisely, there exists a function $\phi: \mathbb{R} \rightarrow \mathbb{R}$ such that

1) $\{\phi(-k) : k \in \mathbb{Z}\}$ forms an orthonormal family of $L^2(\mathbb{R})$. Let $V_0 = \text{span} \{\phi(-k) : k \in \mathbb{Z}\}$.

2) For $j \in \mathbb{Z}$, let $V_j = \text{span} \{\phi_{j,k} : k \in \mathbb{Z}\}$, where $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$. Then $V_j \subseteq V_{j+1}$.

3) $\phi \in L^2(\mathbb{R})$, $\int \phi(x)dx = 1$.

We note that (1), (2), and (3) together ensure that $\cap_{j \in \mathbb{Z}} V_j = \{0\}$ and $\cup_{j \in \mathbb{Z}} V_j = L^2(\mathbb{R})$.

4) $\phi$ satisfies the following regularity conditions for a given $N \in \mathbb{Z}^+$:

a) There exists a bounded non-increasing function $F$ such that $\int F(|x|)|x|^N dx < \infty$, and $|\phi(x)| \leq F(|x|)$ almost everywhere.

b) $\phi$ is $N+1$ times (weakly) differentiable and $\phi^{(N+1)}$ satisfies $\sup_{x} \sum_{k \in \mathbb{Z}} |\phi^{(N+1)}(x-k)| < \infty$.

Any $\phi$ satisfying (a) and (b) is said to be $N$-regular.

Let $W_j \subseteq L^2(\mathbb{R})$ be a subspace such that $V_{j+1} = V_j \bigoplus W_j$ (i.e. $V_{j+1} = V_j + W_j$ and $V_j \cap W_j = \{0\}$). Then, there exists a function $\psi: \mathbb{R} \rightarrow \mathbb{R}$ such that

1) $\{\psi(-k) : k \in \mathbb{Z}\}$ forms an orthonormal basis of $W_0$.

2) $\sup_{j \in \mathbb{Z}} \|\psi_{j,k}\|_{L^2(\mathbb{R})} = 1$, where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$.

3) $\psi$ satisfies the same regularity conditions as $\phi$.

For any $L \in \mathbb{Z}$, we can decompose $L^2(\mathbb{R})$ as

$L^2(\mathbb{R}) = V_L \bigoplus W_L \bigoplus W_{L+1} \bigoplus \cdots$

That is, for any $f \in L^2(\mathbb{R})$

\[ f = \sum_{k \in \mathbb{Z}} \alpha_{L,k} \phi_{L,k} + \sum_{j \geq L, k \in \mathbb{Z}} \beta_{j,k} \psi_{j,k}, \]

where

\[ \alpha_{L,k} = \int f(x) \phi_{L,k}(x)dx, \quad \beta_{j,k} = \int f(x) \psi_{j,k}(x)dx \]
are called the wavelet coefficients of $f$. The convergence in (4) is to be understood in the $L^2$ sense in general; however, when $\phi, \psi$ satisfies the regularity conditions, then the convergence holds in $L^p$ sense for $p \in [1, \infty]$ (see Fact 15). Moreover, for a father wavelet $\phi$, there is a canonical mother wavelet $\psi$ (Section V-B in [20]) corresponding to $\phi$.

2) Besov Spaces: We now define the Besov space $B(p,q,s)$ with parameters $p,q,s$, where $1 \leq p, q \leq \infty$, $s > 0$. Let $\phi$ be a father wavelet satisfying properties 1–4 above, with $\phi, \psi$ are called the wavelet coefficients of $\phi, \psi$. The definition (5) of $B(p,q,s)$ is invariant to the choice of $\phi$ as long as $N > s$. For the purposes of defining Besov norm, we fix a particular $\phi, \psi$, where $\phi$ is $N$-regular with $N > s$. Then, the Besov norm of a function $f$ is defined as

$$
\|f\|_{p,q,s} := \|\alpha_0\|_p + \left( \sum_{j=0}^\infty 2^{\frac{1}{2}+\frac{s}{2}} \|\beta_j\|_p \right)^{1/q}.
$$

(5)

where $\|\alpha_0\|_p$ and $\|\beta_j\|_p$ are the $\ell_p$ norms of the sequences $\{\alpha_0,k\}_{k \in \mathbb{Z}}$ and $\{\beta_j,k\}_{k \in \mathbb{Z}}$ respectively. The sequences $\{\alpha_0,k\}_{k \in \mathbb{Z}}, \{\beta_j,k\}_{k \in \mathbb{Z}}$ come from the wavelet expansion of $f$ using the father wavelet $\phi$ and the corresponding mother wavelet $\psi$. The definition (5) of $B(p,q,s)$ is invariant to the choice of $\phi$ as long as $N > s$. For the purposes of defining Besov norm, we fix a particular $\phi, \psi$, where $\phi$ is $N$-regular with $N > s$. Then, the Besov norm of a function $f$ is defined as

$$
\|f\|_{p,q,s} := \|\alpha_0\|_p + \left( \sum_{j=0}^\infty 2^{\frac{1}{2}+\frac{s}{2}} \|\beta_j\|_p \right)^{1/q}.
$$

(6)

B. Assumptions

We make the following assumptions on the density $f$:

1) $f$ is compactly supported: without loss of generality, $\text{supp}(f) \subseteq [0,1]$.

2) Besov norm of $f$ is bounded: without loss of generality, $\|f\|_{p,q,s} \leq 1$.

Our algorithm works with any father and mother wavelets $\phi$ and $\psi$ satisfying the following conditions:

1) $\phi$ and $\psi$ are $N$-regular, where $N > s$, and

2) $\text{supp}(\phi), \text{supp}(\psi) \subseteq [-A,A]$ for some integer $A > 0$ (which may depend on $N$).

As an example, Daubechies’ family of wavelets [21] satisfies these assumptions.

C. Density Estimation in Centralized Setting

In the centralized setting, $X_1, \ldots, X_n$ from an unknown density $f \in B(p,q,s)$ are accessible to the estimator. Let the wavelet expansion of $f$ be

$$
f = \sum_{k \in \mathbb{Z}} \alpha_{0,k} \phi_{0,k} + \sum_{j=0}^\infty \sum_{k \in \mathbb{Z}} \beta_{j,k} \psi_{j,k},
$$

(7)

where $\phi_{j,k}(x) = 2^{j/2} \phi(2^j x - k)$, $\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k)$. The unknown $f$ is estimated by estimating its wavelet coefficients up to a certain scale.

For any $L, H \in \mathbb{Z}$ with $H \geq L$, we have (by construction of wavelet basis) $\sum_{k \in \mathbb{Z}} \alpha_{L,k} \phi_{L,k} + \sum_{j=L}^{H-1} \sum_{k \in \mathbb{Z}} \beta_{j,k} \psi_{j,k} = \sum_{k \in \mathbb{Z}} \alpha_{H,k} \phi_{H,k}$. Also, note that for a given $j, k,$

$$
\hat{\alpha}_{j,k} := \frac{1}{n} \sum_{i=1}^n \phi_{j,k}(X_i)
$$

is an unbiased estimate of $\alpha_{j,k}$. Thus, for some $H \in \mathbb{Z}_+$, an estimate of $f$ is

$$
\hat{f}_{\text{lin}} = \sum_{k \in \mathbb{Z}} \hat{\alpha}_{H,k} \phi_{H,k}, \quad \hat{\alpha}_{H,k} := \frac{1}{n} \sum_{i=1}^n \phi_{H,k}(X_i),
$$

(8)

where $H$ is chosen depending on $n$ and parameters $p,q,s$ to minimize the worst-case $L^p$ loss. This simple estimator (with appropriate choice of $H$) is rate-optimal when $1 \leq r \leq p$, but is sub-optimal when $r > p$ [5]. Moreover, setting $H$ requires knowing the Besov parameters $p,q,s$, which renders this estimator nonadaptive. The main contribution of [5] was to demonstrate that thresholding leads to estimators that are (i) near-optimal for every $r \geq 1$; (ii) adaptive, in the sense that the estimator does not use the values of parameters $p,q,s$ as long as $s$ lies in a certain range. For a given $L,H \in \mathbb{Z}_+$, $L \leq H$, a thresholded estimator outputs the estimate

$$
\hat{f}_{\text{thresh}} = \sum_{k \in \mathbb{Z}} \hat{\alpha}_{L,k} \phi_{L,k} + \sum_{j=L}^{H-1} \sum_{k \in \mathbb{Z}} \tilde{\beta}_{j,k} \psi_{j,k}, \quad \tilde{\beta}_{j,k} = \begin{cases} \beta_{j,k} & \text{if } t_j \end{cases}
$$

where $\hat{\alpha}_{L,k} := \frac{1}{n} \sum_{i=1}^n \phi_{L,k}(X_i)$, $\tilde{\beta}_{j,k} = \frac{1}{n} \sum_{i=1}^n \psi_{j,k}(X_i)$, and $t_j$ is a fixed threshold proportional to $\sqrt{j}/n$. Here, $L, H$ depend on $n$, but not on parameters $p,q,s$. Our proposed estimators draw upon these classical estimators.

III. Single-Level Estimator

In this section and the next, we propose algorithms for density estimation under communication constraints that achieve optimal/near-optimal performance in terms of $n$ (number of players) and $\ell$ (number of bits each player can send). Recall that designing a density estimator in the communication-constrained setting consists of: (i) specifying the sample-dependent $\ell$-bit message that a player sends to the referee; (ii) specifying the density estimate that the referee outputs based on the $\ell$-bit messages from the $n$ players. As in the centralized setting, our algorithm estimates $f$ by estimating its wavelet coefficients. We start by discussing the main ideas used in designing our estimators.

A. Algorithmic Ideas

Our estimators use three main ingredients: wavelet-induced sparsity, vector quantization, and distributed simulation. We describe these next.

1) Wavelet-Induced Sparsity: Let the wavelet expansion of the density function $f$ be given by (7). For a given $J \in \mathbb{Z}_+$, partition the interval $[0,1]$ into $2^J$ uniform bins as

$$
[0,1] = \bigcup_{t=0}^{2^J-1} E_t^{(J)}
$$

where $E_t^{(J)} := \begin{cases} [2^{-j}, (t+1)2^{-j}) & \text{if } t \in [0,2^j-2], \\ [1-2^{-j}, 1] & \text{if } t = 2^J-1. \end{cases}$

(10)
For a bin $E_t^{(j)}$, $t \in [0, 2^j - 1]$, let
\begin{align}
\mathcal{A}_t^{(j)} := \{ k \in \mathbb{Z} : E_t^{(j)} \cap \text{supp} \phi_{J,k} \text{ is non-empty} \} \quad (11) \\
\mathcal{B}_t^{(j)} := \{ k \in \mathbb{Z} : E_t^{(j)} \cap \text{supp} \psi_{J,k} \text{ is non-empty} \} \quad (12)
\end{align}
That is, for $x \in E_t^{(j)}$, we have $\phi_{J,k}(x) = 0$ for $k \notin \mathcal{A}_t^{(j)}$, and $\psi_{J,k}(x) = 0$ for $k \notin \mathcal{B}_t^{(j)}$. By “wavelet-induced sparsity,” we mean the following:

Claim 4: Let $[0, 1] = \bigcup_{t=0}^{2^j-1} E_t^{(j)}$ as in (10). Then, for each $t \in [0, 2^j - 1]$, 
\[
|\mathcal{A}_t^{(j)}| \leq 2(A + 2), \quad |\mathcal{B}_t^{(j)}| \leq 2(A + 2),
\]
where $A$ is the assumed bound for points in the support of $\phi$ and $\psi$.

In other words, if $x \in E_t^{(j)}$, then there are at most a constant number of translation indices $k$ such that $\phi_{J,k}(x) \neq 0$ or $\psi_{J,k}(x) \neq 0$. The claim follows from the observation that $\phi_{J,k}$ (resp., $\psi_{J,k}$) is obtained by translating $\phi_{J,0}$ (resp., $\psi_{J,0}$) in steps of size $2^{-j}$, and that supp$(\phi_{J,k}) \subseteq [-A2^{-j}, A2^{-j}]$ (resp., supp$(\psi_{J,k}) \subseteq [-A2^{-j}, A2^{-j}]$).

2) Vector Quantization: Consider the problem of designing a randomized algorithm that takes as input an arbitrary $x \in \mathbb{R}^d$ satisfying $\|x\|_\infty \leq B$ and outputs a random vector $Q(x) \in \mathbb{R}^d$ chosen from an alphabet of finite cardinality, such that $\mathbb{E}[Q(x)] = x$. Our vector quantization algorithm (Algorithm 1) achieves this, and is based on the following idea: Let $\mathcal{P}$ be a convex polytope with vertices $\{v_1, v_2, \ldots\}$ such that $\{x \in \mathbb{R}^d : \|x\|_\infty \leq B\} \subseteq \mathcal{P}$. Given $x$ (with $\|x\|_\infty \leq B$), express $x$ as a convex combination of vertices of $\mathcal{P}$ (say, $x = \sum_i \theta_i v_i$) and output a random vertex $V$, where $V = v_i$ with probability $\theta_i$. Clearly, $\mathbb{E}[V] = x$.

Specifically, Algorithm 1 uses the polytope $\mathcal{P} = \mathcal{P}_d$ formed by the vertex set $\mathcal{V} = \{\pm(Bd)e_1, \ldots, \pm(Bd)e_d\}$, where $e_i$ is the $i$-th standard basis vector (i.e., $\mathcal{P}_d$ is the unit ball of radius $Bd$). Note that $|\mathcal{V}| = 2d$ and that $\{x \in \mathbb{R}^d : \|x\|_\infty \leq B\} \subseteq \mathcal{P}_d$. This leads to the following claim.

Claim 5: Given $x \in \mathbb{R}^d$ with $\|x\|_\infty \leq B$ as input, Algorithm 1 outputs a random variable $Q(x) \in \mathcal{V}$ that is an unbiased estimate of $x$, with $|\mathcal{V}| = 2d$.

Remark: A more direct approach to quantization would be to do it coordinate-wise, i.e., quantize (independently) each coordinate to $\pm B$ with appropriate probability to make it unbiased. This can equivalently be seen as quantizing the vector using the $\ell_\infty$ ball (of radius $B$) as the polytope. Here, the alphabet size becomes $2^d$ instead of $2d$ in Algorithm 1; on the plus side, the coordinate-wise variance of the quantized vector becomes $\approx B^2$, instead of $\approx (Bd)^2$ in Algorithm 1. In our estimators, we will be quantizing vectors of constant length (i.e. $d$ will be a constant because of “wavelet-induced sparsity” discussed above), so these dependencies on $d$ do not affect the rate (up to constants).

Algorithm 1 Vector Quantization
Let $\mathcal{V} = \{\pm(Bd)e_1, \ldots, \pm(Bd)e_d\} \subset \mathbb{R}^d$. Note that $|\mathcal{V}| = 2d$.

Input: $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ with $\|x\|_\infty \leq B$.

1: Write $x$ as convex combination of vectors in $\mathcal{V}$: 
\[
x = \sum_{i=1}^{d} \frac{|x_i|}{Bd} \text{sign}(x_i)(Bd)e_i.
\]

2: Choose $I \in \{1, \ldots, d\}$ randomly where $I = i$ with probability $\frac{|x_i|}{Bd}$ and return $Q(x) = (\text{sign}(x_I)(Bd)e_I) \in \mathcal{V}$.

3) Distributed Simulation: The problem of distributed simulation is the following: There are $n$ players, each having an i.i.d. sample from an unknown probability distribution $p$ supported on a (known) discrete set $Z$ with $|Z| = d$. Each player can only send $\ell$ bits to a central referee, where $\ell < \log d$ (so, a player cannot directly send its sample to the referee). Can the referee simulate a sample from $p$ using $\ell$-bit messages from the players?

Reference [3] considers this problem and proposed a noninteractive communication protocol [3, Algorithm 3] using which the referee can simulate one sample from $p$ using $\ell$-bit messages from $O(d/\ell^2)$ players. More precisely, in [3, Theorem IV.9], after looking at $\ell$-bit messages from $4d/\ell^2$ players, the referee outputs either a symbol $z \in Z$ or an abort symbol \( \bot \) such that $\Pr\{\text{referee output is} \bot \} \leq 1/4$ and $\Pr\{\text{referee output is} \bot \} = \Pr\{z \in \text{referee's output} \}$.

As a corollary, upon dividing $n$ players into blocks of $4d/\ell^2$ players and using Hoeffding’s inequality, we get the following.

Fact 6: For any $1 \leq \ell < \log d$, there exists a (randomized) simulation protocol (denoted $\text{DISTSIM}_d$) that lets the referee simulate (with probability at least $1 - e^{-\frac{\ell^2}{16}}$) a multiniset of $n2^\ell/(8d)$ i.i.d. samples from an unknown $d$-ary probability distribution $p$ using $\ell$-bit messages from $n$ players, where each player holds an independent sample from $p$. Moreover, the protocol is deterministic at the players’ sides, and only requires private randomness at the referee’s side.

Notation: In the $\text{DISTSIM}_d$ protocol, we denote by $\text{DISTSIM}_d(i, Z_i)$ the $\ell$-bit message sent by player $i$, where $Z_i$ is the sample available to player $i$.

4) Combining the Ingredients: We now discuss how the three ideas come together. To mimic the classical estimator (8), a player with sample $X$ would ideally like to communicate $\{\phi_{H,k}(X)\}_{k \in \mathbb{Z}}$, but cannot do so due to communication constraints. Wavelet-induced sparsity (Claim 4) ensures that communicating the bin (out of $2H$ possible bins) in which $X$ lies is tantamount to identifying the set of at most $d := 2(A + 2)$ indices $k$ for which $\phi_{H,k}(X)$ is possibly non-zero. Moreover, the player can quantize (unbiasedly) the vector containing values of $\phi_{H,k}(X)$ at these indices using Algorithm 1, whose output is one of $2d$ possibilities (Claim 5).

Thus, overall, using an alphabet of size at most $2H \cdot (2d) = O(2^{H})$, a player can communicate an unbiased estimate of $\{\phi_{H,k}(X)\}_{k \in \mathbb{Z}}$. It can be shown that a density estimate based on these unbiased estimates from $n$ players still achieves centralized minimax rates (up to constants). However, if $2^\ell <
(2d)2^H$, the players cannot send these estimates directly to the referee. In this case, the players and the referee use the distributed simulation protocol $\text{DistrSim}_t$ (Theorem 6), which, effectively, enables the referee to simulate $\Theta(n2^d / 2^H)$ i.i.d. realizations of unbiased estimates of $\{\phi_{H,k}(X)\}_{k \in \mathbb{Z}}$. The referee can now output a density estimate based on these simulated estimates. The degradation in minimax rates under communication constraints is due to the fact that the referee has only $\Theta(n2^d / 2^H)$ realizations of unbiased estimates of $\{\phi_{H,k}(X)\}_{k \in \mathbb{Z}}$, instead of $n$.

We now give details of the idea outlined above. The resulting estimator (“single level estimator”) is a communication-constrained version of the classical estimator given in (8). In Section IV, we propose an adaptive estimator (“multi level estimator”), which is a communication-constrained version of the classical adaptive estimator (9).

### B. Single Level Estimator

1) **Shared Context:** The $n$ players and the referee agree beforehand on the following: wavelet functions $\phi, \psi; H \in \mathbb{Z}_+$; partition $[0, 1] = \bigcup_{i=0}^{2^H-1} E_i^{(H)}$ as in (10); collections of indices $A_t^{(H)}$, $t \in [0, 2^H-1]$ as in (11). For every $t$, the indices in $A_t^{(H)}$ are arranged in ascending order.

2) **Player’s Side (Algorithm 2):** Each player carries out two broad steps: (i) quantization; and (ii) simulation (if $\ell$ bits are insufficient to send quantized message). The quantization step involves computing $Z_t = (B_t, Q(V_t))$ (step 2 in Algorithm 2), which involves two quantizations: $B_t$ can be seen as a quantized version of $X_t \in [0, 1]$; $Q(V_t)$ is a quantized version of $\{\phi_{H,k}(X_t)\}_k$. The scaling by $2^{-H/2}$ in the definition of $V_t$ ensures that $\|V_t\|_\infty \leq \|\phi\|_\infty$. This constant. This enables the use of Algorithm 1 to compute quantization of $V_t$.

### Algorithm 2 Single Level Estimator (Players)

**Input:** Player $i$ has input $X_i, i \in [n]$.

1. for $i = 1, \ldots, n$ do

2. Player $i$ computes $Z_t := (B_t, Q(V_t)) \in \mathcal{Z}^{(H)}$, where: (i) $B_t$ is the bin in which $X_t$ lies; (ii) $Q(V_t)$ is an unbiased quantization of the vector $V_t := \{2^{-H/2} \phi_{H,k}(X_t)\}_k \in A_t^{(H)}$.

3. Player $i$ computes $\ell$-bit message $Y_i$ where

$$Y_i = \begin{cases} Z_i, & \text{if } 2^\ell \geq |\mathcal{Z}^{(H)}|, \\ \text{DistrSim}_t(i, Z_t), & \text{if } 2^\ell < |\mathcal{Z}^{(H)}| \text{ (see Fact 6).} \end{cases}$$

4. Player $i$ sends $Y_i$ to the referee.

5. end for

For each $i \in [n]$, $Z_i \in \mathcal{Z}^{(H)}$, where

$$\mathcal{Z}^{(H)} := [0, 2^H - 1] \times \{\pm(\|\phi\|_\infty d)e_1, \ldots, \pm(\|\phi\|_\infty d)e_d\}.$$  

Since $d \leq 2(A + 2)$ by Claim 4, we have $|\mathcal{Z}^{(H)}| \leq 4(A + 2)2^H = O(2^H)$. 

Thus, $Z_1, \ldots, Z_n$ are i.i.d. samples (since $X_1, \ldots, X_n$ are i.i.d.) from a $|\mathcal{Z}^{(H)}|$-ary distribution (call it $p_{\mathcal{Z}^{(H)}}$), where player $i$ has sample $Z_i$, and $|\mathcal{Z}^{(H)}| = O(2^H)$. Since a player can send only $\ell$ bits, player $i$ cannot send $Z_i$ directly if $2^\ell < |\mathcal{Z}^{(H)}|$. In this case, player $i$ computes the $\ell$-bit message $Y_i$ according to the distributed simulation protocol $\text{DistrSim}_t$, and sends $Y_i$ to the referee.

3) **Referee’s Side (Algorithm 3):** The referee obtains $m$ i.i.d. samples $Z_1', \ldots, Z_m' \sim p_{\mathcal{Z}^{(H)}}$, where $m$ is determined by the distributed simulation protocol $\text{DistrSim}_t$ if $2^\ell < |\mathcal{Z}^{(H)}|$. If it turns out that $m < n2^d / 8|\mathcal{Z}^{(H)}|$ (which happens with exponentially small probability; see Fact 6), then the referee disregards all the samples and outputs the constant function as the density estimate (step 3 in Algorithm 3). Otherwise, in the high probability event that $m \geq n2^d / 8|\mathcal{Z}^{(H)}|$, the referee computes the density estimate similar to the centralized estimate (8). This is possible because the $m$ simulated samples are i.i.d. realizations of unbiased quantization of $\{\phi_{H,k}(X)\}_{k \in \mathbb{Z}}$.

### Algorithm 3 Single Level Estimator (Referee)

**Input:** $Y_1, \ldots, Y_n$ ($\ell$-bit messages from $n$ players).

1. From $Y_1, \ldots, Y_n$, referee obtains $m$ i.i.d. samples $Z_1', \ldots, Z_m' \sim p_{\mathcal{Z}^{(H)}}$, where $Z_i' = (B_i', Q_i') \in \mathcal{Z}^{(H)}$ (recall that $B_i'$ represents the bin, and $Q_i'$ represents the vector quantization), and

$$m = n, \quad \text{if } 2^\ell \geq |\mathcal{Z}^{(H)}|.$$  

2. $m$ is determined by $\text{DistrSim}_t$, if $2^\ell < |\mathcal{Z}^{(H)}|$. 

3. Referee outputs density estimate $\hat{f}$, where

$$\hat{f}(x) = 1 \text{ for every } x \in [0, 1].$$

4. else

5. for $i = 1, \ldots, m$ do

6. Referee computes

$$\hat{f}^{(i)} := \begin{cases} 2^{H/2} Q_i'(k) & \text{if } k \in A_{B_i'}^{(H)} \\ 0 & \text{otherwise}, \end{cases}$$

where $Q_i'(k)$ is the entry in $Q_i'$ corresponding to index $k \in A_{B_i'}^{(H)}$.  

7. Scaling by $2^{-H/2}$ is to negate the scaling by $2^{-H/2}$ used in definition of $V_t$ on the players’ side.

8. Referee outputs the density estimate

$$\hat{f} = \sum_{k \in \mathbb{Z}} \hat{a}_{H,k} \phi_{H,k}, \quad \text{where } \hat{a}_{H,k} = \frac{1}{m} \sum_{i=1}^{m} \hat{f}^{(i)}, \quad k \in \mathbb{Z}.$$  

end if

For $H$ such that $2^H \asymp \min\{(n2^d)^{1+a}, n^{1+a}\}$, the single level estimator recovers the guarantees in Theorem 2. The estimator is nonadaptive because setting $H$ requires knowing Besov parameter $s$. Further, note that the estimator is indeed
noninteractive, as player \(i\)'s message \(Y_i\) does not depend on messages \(Y_1, \ldots, Y_{i-1}\).

C. Analysis of the Single Level Estimator

Our goal is to upperbound the worst-case \(L^r\) loss
\[
\mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \right],
\]
where \(\hat{f}\) is the estimate output by the referee in Algorithm 3. For now, we assume that the referee can always simulate \(m = \Theta(n2^H/|Z(H)|)\) samples using \textsc{DistrSim}_r, if \(2^r < |Z(H)|\). At the end of this section, we will show that this assumption does not affect the rate since the event that the referee simulates fewer than \(n2^H/(8|Z(H)|)\) samples happens with exponentially small probability (Fact 6).

We present an outline of the analysis here, leaving the supporting facts and claims to Appendices A and B. We have
\[
\mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \right] = \mathbb{E}\left[ \left\| \hat{f} - f^H + f^H - f \right\|_r^r \right] \leq 2^{r-1} \left( \mathbb{E}\left[ \left\| \hat{f} - f^H \right\|_r^r \right] + \left\| f^H - f \right\|_r^r \right),
\]
(15)

where \(f^H = \sum_{k \in \mathbb{Z}} \alpha_{H,k} \phi_{H,k}\) (truncation of wavelet expansion of \(f\) at scale \(H\)).

From Fact 15, we have
\[
\left\| f - f^H \right\|_r^r \lesssim 2^{-Hr\sigma}, \text{ where } \sigma = \begin{cases} 8 & \text{if } r \leq p, \\ (s - 1/p + 1/r) & \text{if } r > p. \end{cases}
\]

Moreover,
\[
\mathbb{E}\left[ \left\| \hat{f} - f^H \right\|_r^r \right] = \mathbb{E}\left[ \sum_{k \in \mathbb{Z}} (\hat{\alpha}_{H,k} - \alpha_{H,k}) \phi_{H,k} \right] \lesssim 2^{H(r/2 - 1)} \sum_{k \in \mathbb{Z}} |\hat{\alpha}_{H,k} - \alpha_{H,k}|^r
\approx 2^{H(r/2 - 1)} \left( \frac{2H}{m} \right)^{r/2}
= \left( \frac{2H}{m} \right)^{r/2}.
\]
(18)

Fact 23)

In our case, \(m\) and \(H\) will be such that \(m \geq 2^H\) holds (see (16)), which is why we can use Corollary 23. Substituting these in (15), we get
\[
\mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \right] \lesssim 2^{-Hr\sigma} + \left( \frac{2H}{m} \right)^{r/2}.
\]

Recall that \(m\) is the number of quantized samples available with the referee, where, for a constant \(C\),
\[
m = \begin{cases} n & \text{if } 2^H \leq n^a, \\ Cn^{2^H/2H} & \text{if } 2^H > n^{a}. \end{cases}
\]

In other words, \(m \approx n^{2^{H}/2H\sigma}\), where \(a \vee b = \max\{a, b\}\). Thus,
\[
\mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \right] \lesssim 2^{-Hr\sigma} + \left( \frac{2H}{n^{2^{H}/2H}} \right)^{r/2} + \left( \frac{2H}{n} \right)^{r/2}.
\]

Setting \(H\) such that
\[
2^H = (n2^f)^{\frac{1}{2^{H}/2H}} \wedge n^{\frac{2H}{2^{H}/2H}}
\]
gives us
\[
\mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \right] \lesssim (n2^f)^{-\frac{2^{H}/2H}{2^{H}/2H}} \vee n^{-\frac{2H}{2^{H}/2H}}.
\]

For \(r \leq p\), we have \(\sigma = s\). Since \(f \in B(p, q, s)\) was arbitrary, we get that, for \(1 \leq r \leq p\),
\[
\sup_{f \in B(p, q, s)} \mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \right] \leq C \max\{n^{-\frac{2^H}{2H}}, (n2^f)^{-\frac{2^{H}/2H}{2^{H}/2H}}\}.
\]

1) Error From Distributed Simulation: We now show that the small probability of failure of \textsc{DistrSim}_r to simulate more than \(n2^f/(8|Z(H)|)\) samples does not affect the rate. Let \(G\) be the “good” event that the number \(m\) of simulated samples satisfies \(m \geq n2^f/(8|Z(H)|)\). Then we have the following:
\[
\mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \right] = \left[ \mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \middle| G \right] \mathbb{P}\{G\} + \mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \middle| G^c \right] \mathbb{P}\{G^c\} \right]
\leq \left[ \mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \middle| G \right] \mathbb{P}\{G\} + \left(1 + \sup_{f \in B(p, q, s)} \left\| f \right\|_r^r \right) \mathbb{P}\{G^c\} \right]
= \left[ \mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \middle| G \right] \mathbb{P}\{G\} + C \cdot \mathbb{P}\{G^c\} \right]
\]

(16)

(1) for a constant \(C\) depending on \(p, q, s\); see Fact 16
\[
\leq \left[ \mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \middle| G \right] + C \cdot e^{-n2^f/32|Z(H)|} \right]
\leq \left[ \mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \middle| G \right] + C \cdot e^{-C(n2^H)^{2/2H}} \right]
= O \left( \mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \middle| G \right] \right),
\]
since \(\mathbb{E}\left[ \left\| \hat{f} - f \right\|_r^r \middle| G \right] \) depends polynomially on \(n2^f\).

IV. MULTI-LEVEL ESTIMATOR: AN ADAPTIVE ESTIMATOR

The single-level estimator described and analyzed Section III suffers two drawbacks: first, it requires prior knowledge of \(H\), which, in turn, requires knowing the parameter \(s\). Second, the estimator is rate-optimal only when \(r \leq p\). In this section, we remedy these deficiencies. Our multi-level estimator requires knowing only an upper bound on \(s\), and is rate-optimal (up to logarithmic factors) for all ranges of \(p, q, r\). As in the centralized setting [5], this adaptivity is achieved using thresholding of estimated wavelet coefficients across multiple scales. The main contribution of this section is making such a thresholding-based scheme work in the communication-constrained setting.

The key observation in designing our multi-level estimator is that different coefficients need to be recovered with different accuracy. We enable this by dividing players into groups for estimating different coefficients, and using a different level of quantization for each group. This is in contrast to simply mimicking the centralized adaptive estimator (9), which would suggest that a player with sample \(X\) should quantize and communicate information about
\( \{ \phi_{L,k}(X) \}_{k}, \{ \psi_{J,k}(X) \}_{k} \) \( J \in [L,H] \). Instead, we do the following: Divide \( n \) players into \( H-L+1 \) groups of equal size (so, each group has \( n' = \frac{n}{H-L+1} \) players). Label the groups \( L, L+1, \ldots, H \). The players in group \( L \) only focus on \( \{ \phi_{L,k}(X) \}_{k} \), \( \{ \psi_{J,k}(X) \}_{k} \). The players in group \( J, J \in [L+1, H] \), only focus on \( \{ \psi_{J,k}(X) \}_{k} \). Moreover, the players in group \( J, J \in [L, H] \), quantize their sample \( X \) using 2\(^J\) uniform bins. As before, by Claim 4, this is tantamount to identifying at most a constant number of indices for which the wavelet function evaluates to a non-zero value (since the players in group \( J \) only consider \( \phi_{L,k} \) or \( \psi_{J,k} \)). The player then quantizes the vector containing these values using Algorithm 1, before using distributed simulation.

1) **Shared Context:** The \( n \) players and the referee agree beforehand on the following: wavelet functions \( \phi, \psi; L, H \in \mathbb{Z}_+ \); division of players into \( H-L+1 \) equally sized groups. Further, for each \( J \in [L, H] \), the \( n' = \frac{n}{H-L+1} \) players in group \( J \) and the referee agree on the following: partition \( [0,1] = \bigcup_{t=0}^{2^J-1} E_t(J) \) as in (10); collection of indices \( A_t(J), B_t(J), t \in \{0,2^J-1\} \), as in (11), (12). For every \( J, t \), the indices in \( A_t(J), B_t(J) \) are arranged in ascending order.

2) **Player's Side (Algorithm 4):** Label players in group \( J \) as \( (i), (J), \ldots, (n', J) \). We denote by \( X_{i,J} \) the sample with player \( (i) \), essentially, the players in group \( J \) run quantization and simulation steps as in the single-level algorithm (Algorithm 2), with \( H \) replaced by \( J \). Note that the players in group \( J \) quantize for both \( \{ \phi_{L,k}(X) \}_{k} \) and \( \{ \psi_{L,k}(X) \}_{k} \), whereas the players in group \( J \) for \( J \geq L+1 \), quantize only for \( \{ \psi_{J,k}(X) \}_{k} \). The resulting alphabet \( \mathcal{Z}(J) \) satisfies \( \mathcal{Z}(J) = O(2^J) \) for every \( J \in [L, H] \). Thus, for a given \( J \), the \( Z_1(J), \ldots, Z_{n',J} \) i.i.d. samples from a \( \mathcal{Z}(J) \)-ary distribution, denoted \( p_{Z(J)} \). The players in group \( J \) resort to \( \text{DistrSim}_k \) to compute their \( \ell \)-bit message if \( |\mathcal{Z}(J)| > 2^\ell \).

3) **Referee's Side (Algorithm 5):** The referee obtains \( m_J \) i.i.d. samples \( Z_{1,J}, \ldots, Z_{m_J,J} \sim p_{Z(J)} \) from players in group \( J \), where \( m_J \) is determined by the distributed simulation protocol \( \text{DistrSim}_k \) if \( 2^\ell < |\mathcal{Z}(J)| \). If there exists a \( J \in [L, H] \) such that \( m_J < n'^2/8 \mathcal{Z}(J) \), the referee disregards all the samples and outputs constant function as the density estimate (step 5 in Algorithm 3). Otherwise, in the high probability event that \( m_J \geq n'^2/8 \mathcal{Z}(H) \) for every \( J \in [L, H] \), the referee computes the density estimate similar to the adaptive centralized estimator (9), with threshold value \( t_J = \kappa \sqrt{J/m_J} \), for a constant \( \kappa \) (more details in Section IV-A). Note that, higher the \( J \), fewer the simulated samples; this dependence on \( J \) of the number of samples available with the referee is one of the major differences between adaptive estimators in the centralized and the distributed setting.

For \( L \) satisfying \( 2^L \leq \min \{ (n'^2)^{\frac{1}{2(1+1/\gamma)}} \), \( n'^2 \mathcal{Z}^{1+1/\gamma} \} \), and \( H \) satisfying \( 2^H \leq \min \{ \sqrt{n'^2}/\log(n'^2), n'/\log^2 n \} \), the multi-level estimator yields the guarantees in Theorem 3 as long as \( s \in (1/p, N+1) \) (recall that \( N \) is the regularity of the wavelet basis, which is in control of the algorithm designer). Since \( L, H \) do not depend on specific Besov parameters, the estimator is adaptive. Moreover, it is noninteractive.

### Algorithm 4 Multi-Level Estimator (Players)

**Input:** Player \( (i, J) \) has input \( X_{i,J}, i \in \{n'\}, J \in [L, H] \) (where \( n' = \frac{n}{H-L+1} \)).

1. for \( J = L, L+1, \ldots, H \) do
2. for \( i = 1, \ldots, n' \) do
3. Player \( (i, J) \) computes \( Z_{i,J} = (B_{i,J}, Q(V_{i,J})) \in \mathcal{Z}(J) \), where: (i) \( B_{i,J} \) is the bin (out of \( 2^J \) bins) in which \( X_{i,J} \) lies; (ii) \( Q(V_{i,J}) \) is an unbiased quantization of the vector \( V_{i,J} \), where

\[
V_{i,J} := \begin{cases} 
(2^{-J/2} \psi_{J,k}(X_{i,J}))_{k \in B_{i,J}} \text{, if } J \in [L+1, H]; \\
(2^{-L/2} \phi_{L,k}(X_{i,J}))_{k \in B_{i,J}} \cup \{2^{-L/2} \psi_{L,k}(X_{i,J})\}_{k \in B_{i,J}} \text{, if } J = L. 
\end{cases}
\]

(\( \oplus \) denotes concatenation of vectors.)

4. Player \( (i, J) \) computes \( \ell \)-bit message \( Y_{i,J} \), where

\[
Y_{i,J} := \begin{cases} 
Z_{i,J}, & \text{if } 2^\ell \geq |\mathcal{Z}(J)|, \\
\text{DistrSim}_k((i, J), Z_{i,J}), & \text{if } 2^\ell < |\mathcal{Z}(J)|. 
\end{cases}
\]

(see Fact 6)

\( \triangleright \text{Quantization} \)

5. Player \( (i, J) \) sends \( Y_{i,J} \) to the referee.

6. \textbf{end for}

7. \textbf{end for}

### A. Setting Parameters for Adaptive Estimation

Since we want our multi-level estimator to be adaptive, the parameters \( L, H, \) and \( \{ t_J \}_{J \in [L, H]} \) should not depend explicitly on Besov parameters. We set \( L, H \) as

\[
2^L := C \left( (n'^2)^{\frac{1}{2(1+1/\gamma)}} \wedge n'^2 \mathcal{Z}^{1+1/\gamma} \right),
\]

\[
2^H := C' \left( \sqrt{n'^2} / \log{n'^2} \wedge \frac{n}{\log^2 n} \right)
\]

where \( C, C' > 0 \) are two constants, sufficiently large and small, respectively. Also, note that, since players in group \( J \) have alphabet size \( O(2^J) \), we have

\[
m_J = \frac{n}{(H-L+1)} \cdot \left( \frac{2^J}{2^\ell} \wedge 1 \right) \leq \frac{n2^J}{H(2^J \sqrt{2^J})}. 
\]

Note that the choice of \( H \) implies both \( H2^H \ll \sqrt{n'^2} \) and \( H2^H \ll n \), hold, and consequently \( m_J \geq 2J^J \).

1) **Threshold Values:** How should we set the threshold values \( \{ t_J \}_{J \in [L, H]} \)? Since we will pay a cost for the coefficients we zero out (increase in bias), we would like to choose \( t_J \) as small as possible. But, in order to have reasonable concentration, we also need \( t_J \) to satisfy, for every (sufficiently large) \( \gamma > 0 \),

\[
\text{Pr}\left( |\hat{\beta}_{J,k} - \beta_{J,k}| \geq \gamma t_J \right) \leq 2^{-\gamma J}.
\]

That is, we want our estimates to concentrate well around their true value, so that we only zero them out wrongly with very small probability. A natural approach to choose \( t_J \) according to this constraint would be to use Hoeffding’s inequality, as \( \beta_{J,k} \) is the empirical mean of \( m_J \) unbiased estimates of \( \beta_{J,k} \), each
Algorithm 5 Multi-Level Algorithm (Referee)

**Input:** \( \{Y_{i,j}\}_{i \in [n], j \in [L, H]} \) (\( L \)-bit messages from \( n \) players), \( L, H, \kappa \)

1. **for** \( J = L, L+1, \ldots, H \) **do**
   2. From \( Y_{1,j}, \ldots, Y_{n,j} \), referee obtains \( m_j \) i.i.d. samples \( Z_{1,j}', \ldots, Z_{m_j,j}' \sim \mathcal{P}_{Z(j)} \), where \( Z_{i,j}' = (B_{i,j}', Q_{i,j}') \in \mathcal{Z}(j) \), and
      \[
      m_j = n', \quad \text{if } 2^t \geq |\mathcal{Z}(j)|,
      \]
      \[
      m_j \text{ is determined by } \text{DISTRSIM}_f, \quad \text{if } 2^t < |\mathcal{Z}(j)|.
      \]
3. **end for**
4. **if** there exists \( J \) s.t. \( 2^t < |\mathcal{Z}(j)| \) and \( m_j < n'2^t \) **then**
5. Referee outputs density estimate \( \hat{f} \), where
   \[
   \hat{f}(x) = 1 \quad \text{for every } x \in [0, 1].
   \]
6. **else**
7. **for** \( J = L, L+1, \ldots, H \) **do**
8. **for** \( i = 1, \ldots, m_j \) **do**
9. **if** \( J = L \) **then**
10. Referee computes \( \{\hat{\phi}^{(i)}_{L,k}\}_{k \in \mathbb{Z}} \) as
    \[
    \hat{\phi}^{(i)}_{L,k} := \begin{cases} 2^{2L/2} Q'_{i,L}(k) & \text{if } k \in \mathcal{A}_B^{(i)} \\ 0 & \text{otherwise}. \end{cases}
    \]
11. **end if**
12. Referee computes \( \{\hat{\psi}^{(i)}_{j,k}\}_{k \in \mathbb{Z}} \) as
    \[
    \hat{\psi}^{(i)}_{j,k} := \begin{cases} 2^{1/2} Q'_{i,j}(k) & \text{if } k \in \mathcal{B}_B^{(i)} \\ 0 & \text{otherwise}. \end{cases}
    \]
13. **end for**
14. **end for**
15. Referee outputs the density estimate
    \[
    \hat{f} = \sum_k \hat{\alpha}_{L,k} \hat{\phi}_{L,k} + \sum_{J=L}^{H} \sum_k \hat{\beta}_{j,k} \hat{\psi}_{j,k},
    \]
    where \( \hat{\alpha}_{L,k} = \frac{1}{m_L} \sum_{i=1}^{m_L} \hat{\phi}^{(i)}_{L,k}, \quad \hat{\beta}_{j,k} = \frac{1}{m_j} \sum_{i=1}^{m_j} \hat{\psi}^{(i)}_{j,k} \), and
    \[
    \hat{\beta}_{j,k} = \hat{\beta}_{j,k} \mathbb{1}_{\{\beta_{j,k} \leq 2^t\}}.
    \]
16. **end if**

with magnitude \( \leq 2^{2t/2} \). One can check that this would lead to the setting of \( t_J \gg \sqrt{J/2^t/m_J} \), which, unfortunately, is too big (by a factor of \( 2^{2t/2} \)) to give optimal rates.

However, recall that the \( m_j \) unbiased estimates, \( \hat{\psi}_{j,k}(X_{i,j}) \), are not only such that \( |\hat{\psi}_{j,k}(X_{i,j})| \leq 2^{1/2} \); in many cases, they are actually zero, since \( |\hat{\psi}_{j,k}(X_{i,j})| \sim 2^{1/2} \mathbb{1}_{\{k \in \mathcal{B}_B^{(i)}\}} \).

This allows us to derive the following, improving upon the naïve use of Hoeffding’s inequality.

**Lemma 7:** For \( J \in [L, H] \), setting \( t_J := \sqrt{J/m_J} \), we have
\[
\Pr\left(|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \gamma t_J\right) \leq 2^{-\gamma J}
\]
for every \( \gamma \geq 6A\|f\|_{\infty} \).

**Proof:** Fix \( J \) and \( k \), and consider any \( i \in [m_J] \). Since
\[
|\hat{\psi}_{j,k}(X_{i,j})| \leq b := 2^{1/2}, \quad \text{we have}
\]
\[
\mathbb{E}\left[|\hat{\psi}_{j,k}(X_{i,j})|^2\right] = 2^t \Pr(k \in \mathcal{B}_B^{(i)}) \leq 2^t \cdot \|f\|_{\infty} \cdot \frac{2A}{2^t} = A\|f\|_{\infty} := v,
\]
where the inequality follows from our assumption that \( \text{supp}(\psi) \subseteq [-A, A] \). In particular, we have \( v \gg 1 \). Recalling the definition of \( \hat{\beta}_{j,k} \) from (17), we can apply Bernstein’s inequality (Theorem 20) to obtain, for \( t \geq 0 \) and \( \gamma \geq 3v \), that
\[
\Pr\left(|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \gamma t\right) \leq e^{-\frac{3\gamma^2 v}{6(1+2\gamma^2/v)}}.
\]
Setting \( t_J := \sqrt{\frac{J}{m_J}} \), we get
\[
\Pr\left(|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \gamma t_J\right) \leq 2^{-\gamma J}.
\]
Finally, our choice of \( H \) and \( m_j \) implies that
\[
\sqrt{\frac{J}{m_J}} \cdot \sqrt{J/2^t} = \sqrt{J/2^t}, \quad \text{as } m_j \geq J/2^t \text{ for all } J \leq H.
\]

2) **Conclusion:** For constants \( C, C', \kappa > 0 \), the values of parameters are summarized below.
\[
2^L := C\left(\frac{n2^L}{n2^t} \wedge \frac{n}{n2^t}\right) \quad (18)
\]
\[
2^H := C'\left(\sqrt{n2^t} \wedge \frac{n}{\log^2 n}\right) \quad (19)
\]
\[
m_j := \frac{n}{(H-L+1)\left(\frac{2^t}{2^t} \wedge 1\right)} = \frac{n2^t}{H(2^t \vee 2^t)} \quad (20)
\]
\[
t_J := \kappa \sqrt{\frac{J}{m_J}} \quad (21)
\]
As previously mentioned, the choices imply both \( H_2^H \ll \sqrt{n2^t} \) and \( H^22^H \ll n \), and consequently \( m_j \geq J/2^t \).

**B. Analysis of Multi-Level Estimator**

Our goal is to upper bound the worst-case \( L' \) loss
\[
\mathbb{E}\left[\left\|\hat{f} - f\right\|_{L'}\right], \quad \text{where } \hat{f} \text{ is the estimate output by the referee in Algorithm 5.}
\]
As in Section III-B, assume for the time being that the referee can always simulate, for every \( J \in [L, H] \), \( m_j = \Theta\left(n2^t/|\mathcal{Z}(j)|\right) \) samples using \( \text{DISTRSIM}_f \).

At the end of this section, we will show that this assumption does not affect the rate since the event that referee simulates fewer than \( n2^t/|\mathcal{Z}(j)| \) samples happens with exponentially small probability (Fact 6).

We present an outline of the analysis here, leaving the supporting facts, claims, and detailed calculations to Appendices A, B and C. Following the outline of [5, Theorem 3] and [6, Theorem 5.1], we will bound \( L' \) loss as
\[
\mathbb{E}\left[\left\|f - \hat{f}\right\|_{L'}\right] \leq 3^{-1}(\text{bias}(f) + \text{linear}(f) + \text{details}(f))
\]
(22)
where
\[
\text{bias}(f) = \mathbb{E} \left[ \left\| f - \sum_{k \in \mathbb{Z}} \alpha_{H,k} \phi_{H,k} \right\|_r \right]
\]
\[
\text{linear}(f) = \mathbb{E} \left[ \left\| \sum_{k \in \mathbb{Z}} (\bar{\alpha}_{L,k} - \alpha_{L,k}) \phi_{L,k} \right\|_r \right]
\]
\[
\text{details}(f) = \mathbb{E} \left[ \left\| \sum_{j=L}^{H} \sum_{k \in \mathbb{Z}} (\bar{\beta}_{J,k} - \beta_{J,k}) \psi_{J,k} \right\|_r \right]
\]
and handle each of the three terms separately. Note that only the third term is affected by thresholding.

1) Linear Term: To bound \( \text{linear}(f) \), we invoke Fact 18 and Corollary 23 as in the analysis of single-level estimator. This gives
\[
\mathbb{E} \left[ \sum_{k \in \mathbb{Z}} (\bar{\alpha}_{L,k} - \alpha_{L,k}) \phi_{L,k} \right]_r \leq 2^{L/2} - 1 \sum_{k \in \mathbb{Z}} \mathbb{E} [\bar{\alpha}_{L,k} - \alpha_{L,k}]_r^s \leq 2^{L/2} \cdot \frac{2L}{m^2_L}
\]
\[
= H^\frac{r}{2} \left( \frac{2L}{n^{2/2}} \right)^\frac{r/2}{2} \left( \frac{2L}{n} \right)^{r/2}
\]
\[
\leq H^\frac{r}{2} \left( n^{2/2} \right)^{r/2} \left( \frac{2L}{n} \right)^{r/2}
\]
\[
\leq H^\frac{r}{2} \left( (n2^e)^{- \frac{2}{s+1}(s-1/p+1/r)} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
\[
\leq H^\frac{r}{2} \left( (n2^e)^{- \frac{2}{s+1}(s-1/p+1/r)} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
\[
\leq H^\frac{r}{2} \left( (n2^e)^{- \frac{2}{s+1}(s-1/p+1/r)} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
\[
\leq H^\frac{r}{2} \left( (n2^e)^{- \frac{2}{s+1}(s-1/p+1/r)} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
where the second-to-last inequality relies on our choice of \( L \).

2) Bias Term: To bound \( \text{bias}(f) \), we use Fact 15 to get, for \( s' = s - 1/p + 1/r \),
\[
\text{bias}(f) \leq C \cdot 2^{-H^s' r}
\]
\[
\leq C' \cdot \left( \sqrt{\log^2 \left( \frac{2L}{n} \right)} + \sqrt{\log n} \right)^{r(\frac{s-1}{p}+\frac{1}{r})}
\]
(25)

3) Details Term: To bound the term \( \text{details}(f) \), we define, for \( J \in [L, H] \), three sets of indices:
\[
\tilde{J}_J := \{ k \in \mathbb{Z} : |\beta_{J,k}| > \kappa_J \}
\]
(estimate big: not thresholded)
\[
J^s_J := \{ k \in \mathbb{Z} : |\beta_{J,k}| \leq \frac{1}{4r} \kappa_J \}
\]
(small coefficients)
\[
J^b_J := \{ k \in \mathbb{Z} : |\beta_{J,k}| > 2r \kappa_J \}
\]
(big coefficients)

We will partition the error according to these sets of indices, and argue about them separately. Specifically, we write
\[
\text{details}(f) = \text{bias}(f) + \text{linear}(f) + \text{error}(f)
\]
\[
= \mathbb{E} \left[ \left\| \sum_{j=L}^{H} \sum_{k \in \tilde{J}_j} (\beta_{J,k} - \beta_{J,k}) \psi_{J,k} \right\|_r \right]
\]
\[
+ \mathbb{E} \left[ \left\| \sum_{j=L}^{H} \sum_{k \in J^s_J} (\beta_{J,k} - \beta_{J,k}) \psi_{J,k} \right\|_r \right]
\]
\[
+ \mathbb{E} \left[ \left\| \sum_{j=L}^{H} \sum_{k \in J^b_J} \beta_{J,k} \psi_{J,k} \right\|_r \right]
\]
The four errors coming from the “big-big,” “big-small,” “small-big,” and “small-small” indices, respectively. We defer the details of the analysis of these errors to Appendix C and only record the final results here:

\[
E_{bs} \lesssim H^{r/2} \left( \left( \frac{H}{n} \right)^{r/2} \right) \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
\[
E_{bb} \lesssim H^{r/2} \left( \left( \frac{H}{n} \right)^{r/2} \right) \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
\[
E_{ss} \lesssim H^{r/2} \left( \left( \frac{H}{n} \right)^{r/2} \right) \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
(29)

4) Total Error: Defining, for \( r \geq 1, s \geq 0, \) and \( p \geq 1, \) the quantities
\[
\nu(r, p, s) := \frac{r s}{2s + 2} \frac{1}{p_n} \frac{1}{p_n} + \frac{r}{2(s - 1/p + 1/r)} + 2 \frac{1}{p_n} \frac{1}{p_n} \frac{1}{p_n}
\]
\[
\mu(r, p, s) := \frac{r s}{2s + 1} \frac{1}{p_n} \frac{1}{p_n} + \frac{r}{2(s - 1/p + 1/r)} + 2 \frac{1}{p_n} \frac{1}{p_n} \frac{1}{p_n}
\]
we can gather all the error terms from Eqs. (24), (25), (26), (27), (28) and (29), to get
\[
\mathbb{E} \left[ \left\| f - \tilde{f} \right\|_r \right] \lesssim H^{r/2} \left( \left( \frac{H}{n} \right)^{r/2} \right) \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
\[
\left( \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2} \right) \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
\[
\left( \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2} \right) \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
\[
\left( \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2} \right) \left( \left( \frac{n2^e}{r} \right)^{r/2} \right) \left( \frac{2L}{n} \right)^{r/2}
\]
where \( \kappa = \kappa(s, r, p) \) is a constant obtained for simplicity by taking the maximum of the exponent of \( H \) in the previous bounds. To simplify this expression, we observe that the following holds for all \( p, r, s \geq 1 \):
\begin{itemize}
  \item \( \nu(r, p, s) \leq \frac{r}{2(s - 1/p + 1/r)} \)
  \item \( \nu(r, p, s) \leq \frac{r}{2(s - 1/p + 1/r)} \)
  \item \( \mu(r, p, s) \leq \frac{r}{2(s - 1/p + 1/r)} \)
  \item \( \mu(r, p, s) \geq \frac{r}{2(s - 1/p + 1/r)} \)
\end{itemize}
and, of course, \( \mu(r, p, s) \geq \frac{r}{2(s - 1/p + 1/r)} \) if and only if \( r \in (s + 1)p, (2s + 1)p \) and, of course, \( \mu(r, p, s) = \frac{r}{2(s - 1/p + 1/r)} \) if \( r \leq (s + 1)p \)
(this follows from algebraic manipulations and distinctions of cases). Given the above, we finally get the following bound

$$\mathbb{E} \left[ \left\| \hat{f} - \tilde{f} \right\|_r \right] \leq H^c \left( (n^2)^{-\nu(r,p,s)} \vee n^{-\nu(r,p,s)} \vee n^{-\frac{1}{2(r+1)}} \right)$$

$$= \log^c n,$$

where we loosened the bound on the exponent of $H$ to make the result simpler to state.

5) Error From Distributed Simulation: We now show that the small probability of failure of DISTRSIM$_1$ to simulate more than $n^{2d}/8|Z(J)|$ samples does not affect the rate. Let $\mathcal{G}$ be the “good” event that, for every $J \in \{L, H\}$, the number of simulated samples $m_J$ satisfies $m_J \geq n^{2d}/8|Z(J)|$. Then, we have the following:

$$\mathbb{E} \left[ \left\| \hat{f} - \tilde{f} \right\|_r \right] = \mathbb{E} \left[ \left\| \hat{f} - \tilde{f} \right\|_r \mid \mathcal{G} \right] \Pr \{ \mathcal{G} \} + C \cdot \Pr \{ \mathcal{G}^c \} \right.$$  

(for a constant $C$ depending on $p, q, s$)

$$\leq \mathbb{E} \left[ \left\| \hat{f} - \tilde{f} \right\|_r \mid \mathcal{G} \right] + C \cdot \sum_{j = L}^{H} e^{-n^{2d}/32|Z(J)|}$$

$$\leq \mathbb{E} \left[ \left\| \hat{f} - \tilde{f} \right\|_r \mid \mathcal{G} \right] + C \cdot H \cdot e^{-n^{2d}/32|Z(L)|}$$

where $|Z(J)| = \frac{1}{2}$, see the end of Section IV-A)

$$\leq \mathbb{E} \left[ \left\| \hat{f} - \tilde{f} \right\|_r \mid \mathcal{G} \right] + C \cdot \sqrt{n^{2d}} \log n^{2d} \nonumber$$

$$\cdot e^{-n^{2d}/2^{r+1} + 2} = O \left( \mathbb{E} \left[ \left\| \hat{f} - \tilde{f} \right\|_r \mid \mathcal{G} \right] \right),$$

since $\mathbb{E} \left[ \left\| \hat{f} - \tilde{f} \right\|_r \mid \mathcal{G} \right]$ decays polynomially in $n^{2d}$ (see (30)).

V. LOWER BOUNDS

We conclude by proving information-theoretic lower bounds (Theorem 1) for the minimax loss $L^*_i(n, \ell, p, q, s)$, which applies to the broader class of interactive protocols (recall that our matching upper bounds are obtained by noninteractive ones). To derive lower bounds, we consider a family of probability distributions $\mathcal{P}$ parameterized by $\{-1, 1\}^d$ for some $d \in \mathbb{Z}_+$; that is, $\mathcal{P} = \{ \mathbf{p}_z : z \in \{-1, 1\}^d \}$, where $p_z$ has density $f_z$. Moreover, we specify a prior $\pi$ on $Z = (Z_1, \ldots, Z_d) \in \{-1, 1\}^d$, defined as $Z_i \sim \text{Rademacher}(\tau)$ independently for each $i \in [d]$, for some $\tau \in (0, 1/2)$. We then consider the following scenario:

For $Z \sim \pi$, let $X_1, \ldots, X_n$ be i.i.d. samples from $\mathbf{p}_Z$ distributed across $n$ players. Let $Y_1, \ldots, Y_n$ be $\ell$-bit messages sent by the players (possibly interactively) to the referee. Denote by $\mathbf{p}^Y_{+i}$ (resp. $\mathbf{p}^Y_{-i}$) the joint distribution of $Y_1, \ldots, Y_n$ given $Z_i = 1$ (resp. $Z_i = -1$). That is,

$$\mathbf{p}^Y_{+i} = \frac{1}{\tau} \sum_{z_i = 1} \pi(z) \mathbf{p}^z_{+i}, \quad \mathbf{p}^Y_{-i} = \frac{1}{\tau} \sum_{z_i = -1} \pi(z) \mathbf{p}^z_{-i},$$

(31)

where $\mathbf{p}^z_{+i}$ is the joint distribution of $Y_1, \ldots, Y_n$, given $Z = z$.

In this scenario, we analyze the “average discrepancy” $\frac{1}{d} \sum_{z \in [d]} d_{TV}(\mathbf{p}^z_{+i}, \mathbf{p}^z_{-i})$ where $d_{TV}(\mathbf{p}, \mathbf{q})$ denotes the total variation distance between $\mathbf{p}$ and $\mathbf{q}$. On the one hand, a result from [4] will give us an upper bound on this average discrepancy as a function of $n$ and $\ell$ which holds for any interactive protocol generating $Y_1, \ldots, Y_n$ (Theorem 11). On the other hand, we will derive a lower bound on average discrepancy (as a function of the error rate $\varepsilon$) as follows: Consider a communication-constrained density estimation algorithm (possibly interactive) which outputs $\hat{f}$ satisfying $\mathbb{E} \left[ \left\| \hat{f} - f \right\|_r \right] \leq \varepsilon r$. We will show that one can use the messages $Y_1, \ldots, Y_n$ generated by this algorithm to solve, for each $i \in [d]$, the binary hypothesis testing problem of deciding whether $Z_i = 1$ or $Z_i = -1$. This, in turn, will imply a lower bound on $\frac{1}{d} \sum_{z \in [d]} d_{TV}(\mathbf{p}^z_{+i}, \mathbf{p}^z_{-i})$. Putting together the upper and lower bounds on average discrepancy will give us a lower bound on $\varepsilon$.

The parameterized family of distributions $\mathcal{P}$ is constructed as follows: Let $f_0$ be a function supported on $[0, 1]$. Let $I_1, \ldots, I_d \subseteq \{0, 1\}$ be mutually disjoint intervals of equal length. Let $\psi_i$ be a “bump” function supported on interval $I_i$, where $\psi_i$’s are all translations of the same bump function. For, then $z = (z_1, \ldots, z_d) \in \{-1, 1\}^d$, we define $p_z$ to be a probability distribution with density $f_z$, defined as the “baseline” $f_0$ perturbed by adding (a rescaling of) the bump $\psi_i$ according to the value of $z_i$. In more detail, to get the desired lower bounds, we distinguish two cases depending on whether $r < (s+1)p$ or not, and construct two families of distributions: $\mathcal{P}_1$ (when $r < (s+1)p$) and $\mathcal{P}_2$ (when $r \geq (s+1)p$).

- For $\mathcal{P}_1$, we use a uniform prior on $Z = (Z_1, \ldots, Z_d)$, i.e., $Z$ has independent Rademacher coordinates, and set

$$f_z = f_0 + \gamma \sum_{i=1}^{d} z_i \psi_i$$

for some suitably small parameter $\gamma > 0$. That is, the baseline density $f_0$ has disjoint bumps, which are either $\psi_i$ or $-\psi_i$, depending on the value of $z_i$. See Fig. 1 for an illustration, and Section V-C for the details.

- For $\mathcal{P}_2$, we use a non-uniform (‘sparse’) prior on $Z$, where $Z_1, \ldots, Z_d$ are independent with parameter $1/d$, and we set

$$f_z = f_0 + \gamma \sum_{i=1}^{d} (1 + z_i) \psi_i$$

(see Fig. 1 for an illustration, and Section V-C for details).

Applying to these constructions the method described above allows us to derive the lower bounds of Theorem 1.
Remark 8: We note that a similar proof would enable us to derive an analogous result for local privacy, thus extending the lower bounds of [6] (which are restricted to noninteractive protocols) to the interactive setting.

Before delving into the case-wise details, we first recall the result from [4] that we will use to upperbound average discrepancy, and then discuss how the consideration of binary hypothesis testing problem gives a lower bound on average discrepancy.

A. Upper Bound on Average Discrepancy

Consider the following assumptions on $\mathcal{P} = \{p_z : z \in \{-1, 1\}^d\}$, where $p_z$’s are probability distributions on $[0, 1]$.

Assumption 9 (Densities Exist): For any $z \in \{-1, 1\}^d$ and $i \in [d]$, there exist functions $\phi_{z,i} : [0, 1] \rightarrow \mathbb{R}$ such that $E_{p_z}[\phi_{z,i}^2] = 1$ and

$$\frac{dp_{z,i}}{dp_z} = 1 + \alpha \phi_{z,i},$$

where $\alpha \in \mathbb{R}$ is a fixed constant independent of $z, i$.

Assumption 10 (Orthonormality): For all $z \in \{-1, 1\}^d$ and $i, j \in [d]$, $E_{p_z}[\phi_{z,i}\phi_{z,j}] = 1_{\{i = j\}}$.

Under the two assumptions, the following theorem gives an upper bound on the average discrepancy:

Theorem 11 (Corollary 2 in [4]): Suppose $\mathcal{P}$ satisfies assumptions 9 and 10. For some $\tau \in (0, 1/2]$, let $p$ be a prior on $Z \in \{-1, 1\}^d$ defined as $Z_i \sim \text{Rademacher}(\tau)$ independently for each $i \in [d]$. For $Z \sim p$, let $X_1, \ldots, X_n$ be i.i.d. samples from $p_Z$. Then, for any interactive protocol generating $\ell$-bit messages $Y_1, \ldots, Y_n$, we have

$$\left(\frac{1}{d} \sum_{i=1}^{d} d_{\text{TV}}(p_{Y_i}^{n}, p_{Y_i}^{\tau})\right)^2 \leq \frac{7}{d} n\alpha^2 2^\ell.$$

B. Lower Bound on Average Discrepancy

For $Z \sim \pi$, let $X_1, \ldots, X_n$ be i.i.d. samples from $p_Z$ distributed across $n$ players, and let $Y_1, \ldots, Y_n$ be $\ell$-bit messages sent by the players (possibly interactively) to the referee. Based on the $\ell$-bit messages $Y_1, \ldots, Y_n$, suppose the referee outputs an estimate $\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_d)$ of $Z = (Z_1, \ldots, Z_d)$.

Then, an upper bound on $\sum_{i=1}^{d} \Pr \{\hat{Z}_i \neq Z_i\}$ gives a lower bound on $\sum_{i=1}^{d} D(p_{Y_i}^{\tau} || p_{Y_i}^{n})$. To see this, note that, for a given $i \in [d]$,

$$\Pr \{\hat{Z}_i \neq Z_i\} = \Pr \{\hat{Z}_i = -1|Z_i = 1\} \Pr \{Z_i = 1\}$$

$$+ \Pr \{\hat{Z}_i = 1|Z_i = -1\} \Pr \{Z_i = -1\}$$

$$= \tau (1 - \Pr \{\hat{Z}_i = 1|Z_i = 1\})$$

$$+ (1 - \tau) \Pr \{\hat{Z}_i = 1|Z_i = -1\}$$

$$\geq \tau (1 - \Pr \{\hat{Z}_i = 1|Z_i = 1\}) + \tau \Pr \{\hat{Z}_i = 1|Z_i = -1\}$$

(since $(1 - \tau) \geq \tau$ for $\tau \leq 1/2$)

$$= \tau (1 - \Pr \{\hat{Z}_i = 1|Z_i = 1\}) - \Pr \{\hat{Z}_i = 1|Z_i = -1\})$$

$$\geq \tau (1 - d_{\text{TV}}(p_{Y_i}^{\tau}, p_{Y_i}^{n}))\).$$

Thus,

$$\sum_{i=1}^{d} \Pr \{\hat{Z}_i \neq Z_i\} \geq \tau \left(d - \sum_{i=1}^{d} d_{\text{TV}}(p_{Y_i}^{n}, p_{Y_i}^{\tau})\right),$$

which gives

$$\frac{1}{d} \sum_{i=1}^{d} d_{\text{TV}}(p_{Y_i}^{n}, p_{Y_i}^{\tau}) \geq 1 - \frac{1}{d\tau} \sum_{i=1}^{d} \Pr \{\hat{Z}_i \neq Z_i\}.$$  (32)
In conclusion, to get a lower bound on average discrepancy, it suffices to upperbound \(\sum_{i=1}^{d} \Pr \{ \hat{Z}_i \neq Z_i \} \) for an estimator \( \hat{Z} \) of \( Z \).

C. Lower Bound on \( L^2_r(n, \ell, p, q, s) \) for \( r < (s+1)p \)

**Construction.** The family of distributions \( \mathcal{P}_1 \) that we will use to derive lower bound when \( r < (s+1)p \) has also been used in deriving lower bounds in the unconstrained setting [5], [20] and in the LDP setting [6].

Let \( g_0 \) be a density function (see [20, p.157]) such that

1. \( \operatorname{supp}(g_0) \subseteq [0,1] \);
2. \( \|g_0\|_{p\ell,2} \leq 1/2 \);
3. \( g_0 \equiv c_0 > 0 \) on some interval \([a,b]\) \subseteq [0,1].

In what follows, \( j \) is a free parameter that will be suitably chosen later in the proof. Let \( \psi_{j,k} \) be defined as \( \psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k) \), where \( \psi \) is the mother wavelet used to define \( \| \cdot \|_{p\ell,2} \) (see Section II). It is a fact that \( \int \psi_{j,k}(x) dx = 0 \) for every \( j, k \) [20].

For a given \( z \in \{-1,1\}^d \), define

\[
\begin{align*}
f_z := g_0 + \gamma \sum_{k \in I_j} z_k \psi_{j,k} & \quad \text{for every } j, k. \tag{33}
\end{align*}
\]

where

- \( I_j \) is the set of indices \( k \in \mathbb{Z} \) such that
  - i. \( \operatorname{supp}(\psi_{j,k}) \subseteq [a,b] \) for every \( k \in I_j \);
  - ii. for \( k, k' \in I_j \), \( k \neq k' \), \( \psi_{j,k} \) and \( \psi_{j,k'} \) have disjoint support;
  - iii. \( d := |I_j| = C 2^j \), for a constant \( C \). Here on, we will assume for simplicity that \( d = 2^j \).

- \( \gamma \) is chosen such that
  - i. \( f_z(x) \geq c_0/2 \) for every \( x \in [a,b] \); this condition is satisfied if \( c_0 - \gamma 2^{j/2} \| \psi \|_{\infty} \geq c_0/2 \), i.e., \( \gamma \leq (c_0/2 \| \psi \|_{\infty})^{2j/2} \).
  - ii. \( \|f_z\|_{p\ell,2} \leq 1 \); since \( \|f_z\|_{p\ell,2} \leq \|g_0\|_{p\ell,2} + \gamma \|\psi_{j,k}\|_{p\ell,2} \leq 1/2 + \gamma C 2^j/2 2^{(s+1)/2} \) (see pg. 160 in [20]), we get that \( \|f_z\|_{p\ell,2} \leq 1 \) if \( \gamma \leq (1/2C)^{2j/(s+1)} \).

Since \( s > 1/p > 0 \), we get that for \( j \) large enough, if \( \gamma \) satisfies condition (ii), it automatically satisfies condition (i). Thus, we choose \( \gamma = C 2^{-j(s+1)/2} \) for some constant \( C \).

Finally, we define the family of distributions as

\[
\mathcal{P}_1 = \left\{ p_z : p_z \text{ has density } f_z = g_0 + \gamma \sum_{k \in I_j} z_k \psi_{j,k}, \ z \in \{-1,1\}^d \right\}. \tag{34}
\]

Figure 1 illustrates how \( p_z \in \mathcal{P}_1 \) might look.

1) **Prior on \( Z \):** We assume a uniform prior on \( Z \in \{-1,1\}^d \), i.e., \( Z_i \sim \text{Rademacher}(1/2) \) independently for each \( i \in [d] \).

2) **Upper Bound on Average Discrepancy:** To upperbound average discrepancy, we verify that \( \mathcal{P}_1 \) satisfies the three assumptions described in Section V-A, and then use Theorem 11. For any \( z \in \{-1,1\}^d \), we have

\[
\frac{d p_{z \oplus k}}{d p_z}(x) = 1 - \frac{2 \gamma z_k \psi_{j,k}(x)}{c_0 + \gamma z_k \psi_{j,k}(x)}.
\]

Since \( \operatorname{supp}(\psi_{j,k}) \cap \operatorname{supp}(\psi_{j,k'}) \) is empty for \( k \neq k' \), it follows that assumptions 9 and 10 hold. We now compute an upper bound on \( \alpha' := \mathbb{E}_{p_z} \left[ \left( \frac{\gamma z_k \psi_{j,k}(X)}{c_0 + \gamma z_k \psi_{j,k}(X)} \right)^2 \right] \).

\[
\mathbb{E}_{p_z} \left[ \left( \frac{2 \gamma z_k \psi_{j,k}(X)}{c_0 + \gamma z_k \psi_{j,k}(X)} \right)^2 \right] = 4 \gamma^2 \int_{\operatorname{supp}(\psi_{j,k})} \psi_{j,k}(x)^2 \left( \frac{c_0 + \gamma z_k \psi_{j,k}(x)}{c_0 + \gamma z_k \psi_{j,k}(x)} \right)^2 dx
\]

\[
\leq 2 \gamma^2 c_0 \int_{\operatorname{supp}(\psi_{j,k})} \psi_{j,k}(x) dx
\]

\[
\leq 2 \gamma^2 c_0 \times (2^{j+1} \| \psi \|_{\infty})^2 \times \text{length}(\operatorname{supp}(\psi_{j,k}))
\]

\[
\leq 2 \gamma^2 c_0 \times (2^{j+1} \| \psi \|_{\infty})^2 \times C'' \frac{2^j}{2^j} \quad \text{(for a constant } C'' > 0)\]

\[
= C' \gamma^2
\]

\[
= C' 2^{j(s+1)/2} \quad \text{(for a constant } C' > 0)\]

Thus, using Theorem 11, we get

\[
\left( \frac{1}{2j} \sum_{k \in I_j} d_{TV}(P^n_{\mathcal{Z},k}, P^n_{\mathcal{Z},k'}) \right)^2 \lesssim (n2^j)^{2j(s+1)/2}. \tag{35}
\]

3) **Lower Bound on Average Discrepancy:** To lower bound the average discrepancy, we will use the idea described in Section V-B. Consider a communication-constrained density estimation algorithm (possibly interactive) that outputs \( \hat{f} \) satisfying \( \operatorname{supp}_{f \in B(p,q,s)} \mathbb{E}_{f} \left[ \left\| f - \hat{f} \right\|_r \right] \leq \varepsilon' \). Using this density estimator, we estimate \( \hat{Z} \) as

\[
\hat{Z} = \arg\min_{z} \left\| f_z - \hat{f} \right\|_r.
\]

Then

\[
\mathbb{E}_{p_z} \left[ \left\| f_z - f_{\hat{Z}} \right\|_r \right] \leq 2^{r-1} \left( \mathbb{E}_{p_z} \left[ \left\| f_z - \hat{f} \right\|_r \right] + \mathbb{E}_{p_z} \left[ \left\| f_z - f_{\hat{Z}} \right\|_r \right] \right) \leq 2^r \varepsilon'. \tag{36}
\]

Now, for \( z \neq z' \), we have

\[
\left\| f_z - f_{z'} \right\|_r = \int_0^1 \left\| f_z(x) - f_{z'}(x) \right\|_r dx
\]

\[
= \gamma^r \int_0^1 \left\| \sum_{k \in I_j} \psi_{j,k}(x) \mathbb{I}_{\{z_k \neq z'_k \}} \right\|_r dx
\]

\[
= \gamma^r \int_0^1 \left\| \sum_{k \in I_j} \psi_{j,k}(x) \mathbb{I}_{\{z_k \neq z'_k \}} \right\|_r dx
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
acharya et al.: optimal rates for nonparametric density estimation 1953

D. Lower Bound on \( L_2^*(n, \epsilon, p, q, s) \) for \( r \geq (s + 1)p \)

1) Construction: The family of distributions \( \mathcal{P}_2 \) that we will use to derive lower bound when \( r \geq (s + 1)p \) is not exactly the same as that in the unconstrained and in the LDP setting \([5], [6], [20]\); but, combined with the prior that we will choose on \( Z \), it will essentially mimic that.

Let \( g_0, \psi_{j,k}, I_j \) be as in Section V-C. For a given \( z \in \{-1, 1\}^d \) (where \( d := |I_j| \), define

\[
f_z := g_0 + \gamma \sum_{k \in I_j} (1 + z_k) \psi_{j,k}.
\]

(40)

where we will choose \( \gamma \) after we describe the prior on \( Z \).

Finally, we define the family of distributions as

\[
\mathcal{P}_2 = \left\{ \mathbf{p}_z : \text{has density } f_z = g_0 + \gamma \sum_{k \in I_j} (1 + z_k) \psi_{j,k}, \quad z \in \{-1, 1\}^d \right\}.
\]

(41)

Figure 2 illustrates how \( \mathbf{p}_z \in \mathcal{P}_2 \) might look.

2) Prior on \( Z \): We assume a “sparse” prior on \( Z \in \{-1, 1\}^d \), defined as \( Z_k \sim \text{Rademacher}(1/d) \) independently for each \( k \in [d] \). We call it “sparse” because, with high probability, for \( Z = (Z_1, \ldots, Z_d) \) sampled from this prior, the number of indices \( k \) with \( Z_k = 1 \) will be small (we will quantify this soon). Now, since \( f_z = g_0 + \gamma \sum_{k \in I_j} (1 + Z_k) \psi_{j,k} \), this means that with high probability \( 1 + Z_k = 0 \) for a large number of \( k \)'s, and thus there will be only a few “bumps” in \( f_z \).

3) Choosing \( \gamma \): Define \( \mathcal{G} \subset \{-1, 1\}^d \) as

\[
\mathcal{G} := \left\{ z \in \{-1, 1\}^d : \sum_{k=1}^d \mathbb{1}_{\{z_k = 1\}} \leq 2j \right\}.
\]

Then, by Bernstein’s inequality

\[
\Pr \{ Z \in \mathcal{G} \} \geq 1 - 4 \cdot 2^{-2j}.
\]

(42)

We will choose \( \gamma \) such that

i. \( f_z(x) \geq c_0/2 \) for every \( x \in [a, b] \); as seen in Section V-C, this condition is satisfied if \( \gamma \leq (c_0/2||\psi||_\infty)^2 \).

ii. \( \|f_z\|_{pq} \leq 1 \) for every \( z \in \mathcal{G} \); argument similar to that in Section V-C gives that \( \|f_z\|_{pq} \leq 1 \) for \( z \in \mathcal{G} \) if \( \gamma \leq 2^{-j(s+1/2-1/p)}j^{-1/p} \).

Since \( s > 1/p \), we get that for \( j \) large enough (\( j \) is a free parameter that we choose later), if \( \gamma \) satisfies condition (ii), it automatically satisfies condition (i). Thus, we choose \( \gamma = C2^{-j(s+1/2-1/p)}j^{-1/p} \) for some constant \( C \).

Note that, for \( z \notin \mathcal{G} \), this choice of \( \gamma \) still results in \( f_z \) being a density function (since \( \int \psi_{j,k}(x)dx = 0 \), but it may be the case that \( \|f_z\|_{pq} > 1 \).

4) Upper Bound on Average Discrepancy: To upperbound average discrepancy, we verify that \( \mathcal{P}_2 \) satisfies the three assumptions described in Section V-A. For any \( z \in \{-1, 1\}^d \),
Thus, using Theorem 11, we get
\[
\frac{dp_z \otimes k}{dp_z}(x) = \frac{g_0 + \gamma \sum_{i \in I_j} (1 - z_k) \psi_j, k(x)}{g_0 + \gamma \sum_{i \in I_j} (1 + z_k) \psi_j, k(x)} = 1 - \frac{2\gamma z_k \psi^j, k(x)}{c_0 + \gamma \sum_{i \in I_j} z_k \psi_j, k(x)}
\]
which is same as what we had in Section V-C. Similar arguments lead to the conclusion that assumptions 9 and 10 are satisfied. Moreover, an upper bound on \(\alpha^2 := \mathbb{E}_{p_z} \left[ \frac{\psi^j, k(X)}{\psi^j, k(X) + \psi^j, k(Z)} \right] \) follows similarly (with different value of \(\gamma\)), and we get that
\[
\alpha^2 \leq C2^{-2j(s+1/2-1/p)} j^{-2/p}.
\]
Thus, using Theorem 11, we get
\[
\left( \frac{1}{2} \sum_{k \in I_j} d_{TV} \left( p_{Y_k \leftarrow k}^{\bowtie}, p_{Y_k \rightarrow k}^n \right) \right)^2 \lesssim (n2^j)^{-2j(s+1-1/p)} j^{-2/p}.
\]

5) Lower Bound on Average Discrepancy: To lower-bound average discrepancy, we proceed as in Section V-C. Consider a communication-constrained density estimation algorithm (possibly interactive) that outputs \(\hat{f}\) satisfying
\[
\sup_{f \in B(p, q, s)} \mathbb{E}_f \left[ \| f - \hat{f} \|_r \right] \leq \varepsilon^r.
\]
Using this density estimator, we estimate \(\hat{z}\) as
\[
\hat{z} = \arg \min_z \| f_z - \hat{f} \|_r.
\]
Then, for \(z \in G\),
\[
\mathbb{E}_{p_z} \left[ \| f_z - f_{\hat{z}} \|_r \right] \leq 2^r \varepsilon^r.
\]
This only holds for \(z \in G\) because the estimator’s guarantee only holds if samples come from a density \(f\) satisfying \(\| f \|_{spg} \leq 1\). Now, for \(z \neq z'\), plugging in the value of \(\gamma\) in the calculation done in Section V-C, we get
\[
\| f_z - f_{z'} \|_r = CJ^{-r/p} 2^{-j(r(s-1/p)+1)} \sum_{k \in I_j} \mathbb{I} \{ z_k \neq z'_k \}
\]
(for a constant \(C > 0\))
which gives that, for any estimator \(\hat{z}\),
\[
\mathbb{E}_{p_z} \left[ \| f_z - f_{\hat{z}} \|_r \right] = CJ^{-r/p} 2^{-j(r(s-1/p)+1)} \sum_{k \in I_j} \Pr \{ \hat{z}_k \neq Z_k \}.
\]
Combining this with (44), we get that
\[
\sum_{k \in I_j} \Pr \{ Z_k \neq \hat{z}_k, Z \in G \} \lesssim \varepsilon^r 2^{j(r(s-1/p)+1)}.
\]
Thus,\[
\sum_{k \in I_j} \Pr \{ \hat{z}_k \neq Z_k \} = \sum_{k \in I_j} \Pr \{ Z_k \neq \hat{z}_k, Z \in G \} + \sum_{k \in I_j} \Pr \{ Z_k \neq \hat{z}_k, Z \notin G \} \leq \sum_{k \in I_j} \Pr \{ Z_k \neq \hat{z}_k, Z \in G \} + \left( \sum_{k \in I_j} \Pr \{ \hat{z}_k \neq Z_k | Z \notin G \} \right) \Pr \{ Z \notin G \}
\]
\[
\lesssim \varepsilon^r 2^j 2^{-j/2}.
\]
(43) (44)
Thus, substituting $d = 2^j$ and $τ = 1/d = 2^{-j}$ in (32) and ignoring multiplicative constants, we get
\[
\frac{1}{2^j} \sum_{k ∈ I_j} d_{TV}(P_{Y,k}^n, P_{+k}^n) \lesssim 1 - ε^r j^r/p 2^{j(−(1/p)+1/r)} - 2^{-j}
\]
Choosing $j$ such that
\[
ε^r 2^{jr(s−1/p+1/r)} j^r/p \lesssim 1
\]
gives
\[
\frac{1}{2^j} \sum_{k ∈ I_j} d_{TV}(P_{Y,k}^n, P_{+k}^n) \gtrsim 1.
\]

6) Putting Things Together: From (43) and (47), we get, for any $j$ satisfying $ε^r 2^{jr(s−1/p+1/r)} j^r/p \lesssim 1$, that $1 \lesssim (n^2)^{(s−1/p+1/r)/2} j^r/p$. This then yields
\[
2^{j(s+1−1/p)} j^r/p \lesssim n^2.
\]

To get a rough idea of the bound this will give, let us ignore $j^2/p$ to get,
\[
2^j \lesssim (n^2)^{(r−1/p+1/r)/2}.
\]

Now, since $ε^r 2^{jr(s−1/p+1/r)} j^r/p \lesssim 1$, we get, roughly, (ignoring $j^r/p$)
\[
2^j \sim (1/ε)^{r−1/p+1/r}.
\]

Combining (49), (50), we get that (up to logarithmic factors)
\[
ε^r \gtrsim (n^2)^{−r−1/p+1/r}.
\]

This implies that
\[
\frac{1}{ε} \lesssim \left( \frac{n^2}{2^{(r−1/p+1/r)j^r/p}} \right)^{−2/p} \left( \frac{\log(n)^2}{2^{(r−1/p+1/r)j^r/p}} \right)^{−2/p}.
\]

Thus,
\[
ε^r \gtrsim (n^2)^{−r−1/p+1/r} \left( \frac{\log(n)^2}{2^{(r−1/p+1/r)j^r/p}} \right)^{−2/p}.
\]

E. Concluding the Proof of Theorem 1.1

Combining lower bounds from Sections V-C and V-D with lower bounds in the classical setting [5] (where the rate transition happens at $r = (2s + 1)p$), we get Theorem 1.1.

APPENDIX A

USEFUL FACTS

We recall some results that will be used in our analysis.

A. Useful Facts About Besov Spaces

We record a few facts about Besov spaces that will be used in our analysis. Throughout, we assume that $f ∈ B(p, q, s)$ with $∥f∥_{Bpq} ≤ 1$ and $supp(f) ⊆ [0, 1]$. The following facts are essential to our discussion on wavelets.

Fact 12: Let the wavelet expansion of $f$ be as in (4). For $H ≥ L$, define $f^{(H)} := \sum_{k ∈ Z} \theta_{L,k} \phi_{L,k} + \sum_{j=L}^{H−1} \sum_{k ∈ Z} β_{j,k} \psi_{j,k}$. Then, $f^{(H)} : \sum_{k ∈ Z} α_{H,k} \phi_{H,k}$. Since $supp(f) ⊆ [0, 1]$, and $supp(φ_{j,0}) = [−A2^{−j}, A2^{−j}]$, there is no overlap between $supp(f)$ and $supp(φ_{j,0})$ for all but a finite number of indices $k$. In particular, we have the following.

Fact 13: Let the wavelet expansion of $f$ be as in (4). Then, for any given $j ∈ Z_+$, there are $O(2^j)$ translation indices $k$ such that $φ_{j,k}(x)$ or $ψ_{j,k}(x)$ is possibly non-zero, where $x ∈ supp(f)$.

For $f ∈ B(p, q, s)$, it is clear from the definition of Besov norm the wavelet coefficients must decay sufficiently fast. More precisely, we have the following.

Fact 14: If $f ∈ B(p, q, s)$, then
\[
lim_{j → −∞} 2^{j(p+1−1/p)} \sum_{k ∈ Z} |β_{j,k}|^p = 0
\]

and in particular there exists $C > 0$ such that
\[
∥φ_j∥_p = \sum_{k ∈ Z} |φ_{j,k}| ≤ C \cdot 2^{−j(p+1−1/p)}.
\]

The next fact quantifies the approximation error when the wavelet expansion of $f$ is truncated.

Fact 15: Let $f^{(H)}$ be as in Fact 12. Then, for $r ≥ 1$,
\[
∥f^{(H)} − f∥_r ≤ C \begin{cases} 2^{−H} & \text{if } r ≤ p, \\ 2^{−H−(s−1/p+1/r)} & \text{if } r > p. \end{cases}
\]

The next fact (from equation (15) in [5]) gives a bound on $∥f∥_∞$ when $∥f∥_{Bpq} ≤ 1$.
Fact 16: Let \( s > 1/p \). Then
\[
\|f\|_{\infty} \leq \left(1 - 2^{-(s-1/p)q^p}\right)^{1/q}
\]
where \( 1/q + 1/q' = 1 \).

Now, let \( X_1, \ldots, X_n \) be independent samples from distribution with density \( f \). For \( j, k \in \mathbb{Z} \), define
\[
\hat{\alpha}_{j,k} := \frac{1}{n} \sum_{i=1}^{n} \phi_{j,k}(X_i), \quad \hat{\beta}_{j,k} := \frac{1}{n} \sum_{i=1}^{n} \psi_{j,k}(X_i).
\]
Observe that \( \hat{\alpha}_{j,k} \) (resp., \( \hat{\beta}_{j,k} \)) is an unbiased estimate of \( \alpha_{j,k} \) (resp., \( \beta_{j,k} \)). The following fact is from equation (16) in [5].

Fact 17: Let \( n \geq 2^r \). Then, for \( r \geq 1 \),
\[
\mathbb{E}[|\hat{\alpha}_{j,k} - \alpha_{j,k}|^r] \leq C n^{-r/2}, \quad \mathbb{E}[|\hat{\beta}_{j,k} - \beta_{j,k}|^r] \leq C n^{-r/2}
\]
where \( C \) is a constant that depends on \( p, q, s, r, \phi, \psi \). We note another useful fact (obtained after setting \( \beta = 0 \) in equation (21) in [5]).

Fact 18: Let \( g = \sum_{j=L}^{H} \sum_{k \in \mathbb{Z}} \hat{\theta}_{j,k} \psi_{j,k} \), where \( \hat{\theta}_{j,k} \) is random. Then, for \( r \geq 1 \),
\[
\mathbb{E}[\|g\|_p^r] \leq C (H-L)^{(r/2)-1} + H \sum_{j=L}^{H} 2^{r(r/2)-1} \sum_{k \in \mathbb{Z}} \mathbb{E}[|\hat{\theta}_{j,k}|^r]
\]
where \( C \) is a constant that depends on \( r \).

B. Useful Probabilistic Inequalities

Theorem 19 (Rosenthal’s inequality [22]): Let \( X_1, \ldots, X_n \) be independent random variables such that \( \mathbb{E}[X_i] = 0 \) and \( \mathbb{E}[|X_i|^r] < \infty \) for every \( i \).

1) Suppose \( \mathbb{E}[X_i^2] < \infty \) for every \( i \). Then, for \( 1 \leq r \leq 2 \),
\[
\mathbb{E}\left[\sum_{i=1}^{n} |X_i|^r\right] \leq \left( \mathbb{E}[X_i^2] \right)^{r/2}.
\]
(This just follows from concavity of \( f(x) = x^{r/2} \) for \( r \leq 2 \).)

2) Suppose \( \mathbb{E}[|X_i|^r] < \infty \) for every \( i \). Then, for \( r > 2 \), there exists a constant \( K_r \) depending only on \( r \) such that
\[
\mathbb{E}\left[\sum_{i=1}^{n} |X_i|^r\right] \leq K_r \left( \mathbb{E}[X_i^2] \right)^{r/2} + \left( \mathbb{E}[X_i^2] \right)^{r/2}.
\]

Theorem 20 (Bernstein’s inequality): Let \( X_1, \ldots, X_n \) be independent random variables such that \( |X_i| \leq b \) almost surely, and \( \mathbb{E}[X_i^2] \leq \nu_i \) for every \( i \). Let \( X := \sum_{i=1}^{n} X_i \) and \( V := \sum_{i=1}^{n} \nu_i \). Then, for every \( u \geq 0 \),
\[
\Pr(|X - \mathbb{E}[X]| \geq u) \leq \exp \left( -\frac{u^2}{2(V + \frac{b^2}{3})} \right).
\]

APPENDIX B

QUANTIZATION ERROR FOR WAVELET COEFFICIENTS

Our goal is to prove Claim 22 and Corollary 23, which bound the error between quantized estimate of wavelet coefficients and true wavelet coefficients.

Denote by \( p \) the probability distribution corresponding to the unknown density \( f \). Then, we want to bound the error between quantized estimate of wavelet coefficients and true wavelet coefficients in the following setting:

- There are \( m \) i.i.d. samples \( X_1, \ldots, X_m \sim p \).
- For each \( i \in [m] \), let
\[
\hat{\phi}_{j,k}(X_i) = \begin{cases} 0 & \text{if } k \notin A_{(j)}^i, \\ 2^{j/2}Q(V_i)(k) & \text{if } k \in A_{(j)}^i, \end{cases}
\]
where \( B_i \) is the bin \( X_i \) lies in, \( Q(V_i) \) is obtained by quantizing \( V_i := \{2^{-j/2}\phi_{H,k}(X_i)\}_{k \in A_{B_i}^j} \) using Algorithm 1, and \( Q(V_i)(k) \) is the entry in \( Q(V_i) \) corresponding to \( k \in A_{B_i}^j \). In other words, \( \{\hat{\phi}_{j,k}(X_i)\}_{k \in \mathbb{Z}} \) is the quantized version of \( \{\phi_{j,k}(X_i)\}_{k \in \mathbb{Z}} \).
- Similarly, for each \( i \in [m] \), let
\[
\hat{\psi}_{j,k}(X_i) = \begin{cases} 0 & \text{if } k \notin B_{(j)}^i, \\ 2^{j/2}Q(W_i)(k) & \text{if } k \in B_{(j)}^i, \end{cases}
\]
where \( B_i \) is the bin \( X_i \) lies in, \( Q(W_i) \) is obtained by quantizing \( W_i := \{2^{-j/2}\psi_{H,k}(X_i)\}_{k \in B_{(j)}^B} \) using Algorithm 1, and \( Q(W_i)(k) \) is the entry in \( Q(W_i) \) corresponding to \( k \in B_{(j)}^B \). In other words, \( \{\hat{\psi}_{j,k}(X_i)\}_{k \in \mathbb{Z}} \) is the quantized version of \( \{\psi_{j,k}(X_i)\}_{k \in \mathbb{Z}} \).
- For \( j, k \in \mathbb{Z} \), define
\[
\tilde{\alpha}_{j,k} := \frac{1}{m} \sum_{i=1}^{m} \hat{\phi}_{j,k}(X_i), \quad \tilde{\beta}_{j,k} := \frac{1}{m} \sum_{i=1}^{m} \hat{\psi}_{j,k}(X_i),
\]
\[
\hat{\alpha}_{j,k} := m \sum_{i=1}^{m} \phi_{j,k}(X_i), \quad \hat{\beta}_{j,k} := m \sum_{i=1}^{m} \psi_{j,k}(X_i).
\]
We call \( \hat{\alpha}_{j,k}, \tilde{\beta}_{j,k} \) the quantized estimates and \( \hat{\alpha}_{j,k}, \tilde{\beta}_{j,k} \) the unquantized estimates of the true wavelet coefficients \( \alpha_{j,k}, \beta_{j,k} \), respectively.

The following claim bounds the error between quantized and unquantized (centralized) estimates of wavelet coefficients.

Claim 21 (Error Between Quantized and Unquantized Estimates): For \( r \geq 1 \), we have
\[
\mathbb{E}[|\hat{\alpha}_{j,k} - \tilde{\alpha}_{j,k}|^r] \leq C \left( \frac{1}{2^{j(r/2-1)}} + \frac{1}{m^{r/2}} \right), \quad \text{if } r \in [1, 2],
\]
\[
\mathbb{E}[|\hat{\alpha}_{j,k} - \tilde{\alpha}_{j,k}|^r] \leq C \left( \frac{1}{2^{j(r/2-1)}} + \frac{1}{m^{r/2}} \right), \quad \text{if } r > 2,
\]
for a constant \( C \). The same bound holds for \( \mathbb{E}[|\hat{\alpha}_{j,k} - \tilde{\alpha}_{j,k}|^r] \) as well.
Proof: For a given $j, k,$
\[
\begin{align*}
E[|\tilde{\alpha}_{j,k} - \alpha_{j,k}|^r] &= E\left[\frac{1}{m} \sum_{i=1}^{m} \left(\tilde{\phi}_{j,k}(X_i) - \phi_{j,k}(X_i)\right) \mathbf{1}\{A_{j,k}^i \geq k\}\right] \\
&= \frac{1}{m^r} E\left[\sum_{i=1}^{m} \tilde{Y}_{ik}\right]
\end{align*}
\]
where
\[
Y_{ik} := \left(\tilde{\phi}_{j,k}(X_i) - \phi_{j,k}(X_i)\right) \mathbf{1}\{k \in A_{j,k}^i\}.
\]
Note that, since the quantization is unbiased, we have $E[Y_{ik}] = 0$. Moreover, $|Y_{ik}| \lesssim 2^{r/2}$ almost surely. We first consider the case $r > 2$. Then, by Rosenthal’s inequality (Theorem 19),
\[
\begin{align*}
E\left[\sum_{i=1}^{m} Y_{ik}\right] &\lesssim (2^{r/2})^2 \sum_{i=1}^{m} E[Y_{ik}^2] + \left(\sum_{i=1}^{m} E[Y_{ik}^2]\right)^{5/2} \\
&= 2^{r(r-1)/2} m E[Y_{ik}^2] + m^{3/2} E[Y_{ik}^2]^{7/2}.
\end{align*}
\]
Moreover,
\[
E[Y_{ik}^2] = E\left[\left(\tilde{\phi}_{j,k}(X_i) - \phi_{j,k}(X_i)\right)^2 \mathbf{1}\{k \in A_{j,k}^i\}\right] \\
&\lesssim 2^j \text{Pr}\left(k \in A_{j,k}^i\right).
\]
Now, note that
\[
\text{Pr}\left(k \in A_{j,k}^i\right) = \text{Pr}(X_i \in \text{supp}(\phi_{j,k})) \\
\leq \frac{2A}{2^r} \|f\|_{\infty} \lesssim \frac{1}{2^j} \quad \text{(using Fact 16)}
\]
which gives
\[
E[Y_{ik}^2] \lesssim 1.
\]
Substituting this in (58), we get the desired result when $r > 2$. For $r \in [1, 2]$, using part (1) of Theorem 19, only the second term in (58) remains. This gives the result for $r \in [1, 2]$. The proof for $E[|\hat{\beta}_{j,k} - \beta_{j,k}|^r]$ is analogous. ■

The claim above, combined with Fact 17, lets us bound the error between quantized estimates and true coefficients as follows.

Claim 22 (Error Between Quantized Estimates and True Coefficients): Let $m \geq 2^j$. Then, for $r \geq 1$, we have
\[
E[|\tilde{\alpha}_{j,k} - \alpha_{j,k}|^r] \leq C \frac{2^{j(r/2-1)}}{m^{r/2}}, \quad E[|\hat{\beta}_{j,k} - \beta_{j,k}|^r] \leq C \frac{2^{j(r/2-1)}}{m^{r/2}}
\]
for a constant $C$.

Proof: Note that
\[
E[|\tilde{\alpha}_{j,k} - \alpha_{j,k}|^r] \leq 2^r - 1 (E[|\tilde{\alpha}_{j,k} - \alpha_{j,k}|^r] + E[|\tilde{\alpha}_{j,k} - \alpha_{j,k}|^r]).
\]
The first term can be handled with Claim 21. The second term can be bound using Fact 17. Overall, for $r > 2$, we get
\[
E[|\tilde{\alpha}_{j,k} - \alpha_{j,k}|^r] \lesssim \frac{2^{j(r/2-1)}}{m^{r/2}} + \frac{1}{m^{r/2}} \leq \frac{2^{j(r/2-1)}}{m^{r/2}},
\]
since $m \geq 2^j$. We get the same bound for $r \in [1, 2]$. The result follows. The bound on $E[|\hat{\beta}_{j,k} - \beta_{j,k}|^r]$ is obtained in the same way.

Since, for any $j$, there are $O(2^j)$ translations $k$ for which the coefficients are non-zero (Fact 13), Claim 22 readily implies the corollary below.

Corollary 23: Let $m \geq 2^j$. Then, for $r \geq 1$ and a constant $C$, we have
\[
\begin{align*}
\sum_{k \in Z} E[|\tilde{\alpha}_{j,k} - \alpha_{j,k}|^r] &\leq C \frac{2^j}{m^{r/2}}, \\
\sum_{k \in Z} E[|\tilde{\beta}_{j,k} - \beta_{j,k}|^r] &\leq C \frac{2^j}{m^{r/2}}.
\end{align*}
\]

APPENDIX C
ANALYSIS OF MULTI-LEVEL ESTIMATOR

Our goal is to bound the terms $E_{bs}, E_{bb}, E_{sb}, E_{ss}$ in the analysis of multi-level estimator. Recall that these terms appear while analyzing
\[
details(f) := E\left[\left\|\sum_{J=L}^{H} \sum_{k \in Z} (\tilde{\beta}_{J,k} - \beta_{J,k}) \psi_{J,k}\right\|^r\right].
\]
We define three sets of indices:
\[
\begin{align*}
\tilde{I}_J &:= \{k \in Z : |\tilde{\beta}_{J,k}| > \kappa t_J\} \\
(I_J) &:= \{k \in Z : |\beta_{J,k}| \leq \frac{1}{2} \kappa t_J\} \quad \text{(small coefficients)} \\
I_J^b &:= \{k \in Z : |\beta_{J,k}| > 2 \kappa t_J\} \quad \text{(big coefficients)}
\end{align*}
\]
and write
\[
\begin{align*}
details(f) &= E\left[\left\|\sum_{J=L}^{H} \sum_{k \in \tilde{I}_J} (\tilde{\beta}_{J,k} - \beta_{J,k}) \psi_{J,k}\right\|^r\right] \\
&+ E\left[\left\|\sum_{J=L}^{H} \sum_{k \in \tilde{I}_J^b} (\tilde{\beta}_{J,k} - \beta_{J,k}) \psi_{J,k}\right\|^r\right] \\
&+ E\left[\left\|\sum_{J=L}^{H} \sum_{k \in \tilde{I}_J} \beta_{J,k} \psi_{J,k}\right\|^r\right] \\
&+ E\left[\left\|\sum_{J=L}^{H} \sum_{k \in \tilde{I}_J^b} \beta_{J,k} \psi_{J,k}\right\|^r\right] \\
&= E_{bs} + E_{bb} + E_{sb} + E_{ss},
\end{align*}
\]
where the four terms correspond to “big-small,” “big-big,” “small-big,” and “small-small” indices. Our analysis is along the lines of that in [6].
1) The Term $E_{bs}$: We can write

$$E_{bs} \lesssim H'^2/2 \sum_{j=L}^H \sum_{k \in \mathbb{Z}} 2^{j(\alpha - 1)} \mathbb{E} \left[ |\hat{\beta}_{j,k} - \beta_{j,k}| \right]$$

where the last inequality follows from Claim 22 and the $O(2^L)$-sparsity of coefficients (Fact 13). Going forward, recalling our setting of $m_j$ we get

$$E_{bs} \lesssim H'^2/2 \sum_{j=L}^H \frac{2^{j \alpha - 1}}{m_j^{r/2}}$$

where the last inequality holds for $\kappa > 2r$.

2) The Term $E_{bb}$: Turning to the term $E_{bb}$, we have

$$E_{bb} \lesssim \frac{H}{n^2} \sum_{j=L}^H \sum_{k \in \mathbb{Z}} 2^{j(r-\xi)/2} \left( 2^{\ell} \sqrt{2} \right)^{r/2}$$

further bound this as

$$E_{bb} \lesssim \frac{H}{n^2} \sum_{j=L}^H \sum_{k \in \mathbb{Z}} 2^{j(r-\xi)/2} \left( 2^{\ell} \sqrt{2} \right)^{r/2}$$

the last inequality using $\kappa/2 \geq 1$ and $J \geq 1$ to simplify the expression a little. For now, we ignore the factor $H'^2$ (we will bring it back at the end), and look at two cases:

- If $p > \frac{r}{2 + \ell}$, we continue by writing

$$E_{bb} \lesssim \frac{H}{n^2} \sum_{j=L}^H \sum_{k \in \mathbb{Z}} 2^{j(r-\xi)/2} \left( 2^{\ell} \sqrt{2} \right)^{r/2}$$

where the first inequality, which holds for any $\alpha_j \in [0,p]$, uses the following bound on $\sum_k |\beta_{j,k}|^p$: For any $\alpha \in [0,p]$, we have from Fact 14 and Hölder’s inequality along with the sparsity of coefficients (Fact 13), that

$$\sum_{k \in \mathbb{Z}} |\beta_{j,k}|^\alpha \leq \left( \sum_{k \in \mathbb{Z}} |\beta_{j,k}|^p \right)^{\frac{\alpha}{p}} \lesssim B J^{-1 - \frac{\alpha}{p}}$$

so that $\sum_{k \in \mathbb{Z}} |\beta_{j,k}|^\alpha \lesssim 2^{-J \alpha (s+\frac{1}{2} - \frac{1}{p})}$, as in [6, Section C.2.3]. To bound the resulting sum, we need to choose $\alpha_j \in [0,p]$ for all $J$ in order to minimize the result. Since $m_j \propto 2^{\ell \cdot 2}$ and $\alpha \leq \alpha_j \leq \alpha$, we have

$$\log m_j - J(2s + 1)$$

is decreasing in $J$, and thus becomes negative at some value $M$ (for simplicity, assumed to be an integer), such that

$$2^M \approx \left( \frac{n^2}{H} \right)^{\frac{r}{2 + \ell}} \wedge \left( \frac{n}{H} \right)^{\frac{r}{2 + \ell}}$$

we see that we should set $\alpha_j := 0$ for $J \leq M$, and for $J > M$ set all $\alpha_j$ to some value $\alpha = \alpha(r,s)$ which will

2To see why, recall that

$$\log m_j - J(2s + 1) = \log \frac{n}{H} - (J - \ell)_+ - J(2s + 1) + O(1)$$

from our setting of $m_j$. Finding the value of $J$ for which $\log \frac{n}{H} - (J - \ell)_+ - J(2s + 1)$ cancels gives the claimed relation.
balance the remaining terms. With this choice, we can write
\[
\sum_{J=L}^{H} m_J \frac{r-\alpha}{J} 2^{J(r-(2s+1)\alpha)} \leq \sum_{J=M}^{H} m_J \frac{r-\alpha}{J} 2^{J(r-(2s+1)\alpha)} \\
\leq \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=M}^{H} 2^{J}\frac{r-\alpha}{J} + \sum_{J=M}^{H} m_J \frac{r-\alpha}{J} 2^{J(r-(2s+1)\alpha)} \\
\leq \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J} + \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J}\frac{r-\alpha}{J} \\
+ \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=M}^{H} 2^{J(r-(s+1)\alpha)} \\
+ \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} \sum_{J=M}^{H} 2^{J(r-(2s+1)\alpha)} ,
\]
recalling for the second inequality that \( m_J^{-1} \sim \frac{H}{n^{2s}} \triangleq 1 \).

We can bound the first and second terms as
\[
\left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J} \leq 2^{2r-2r/2 - 1} \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} 2^{M} ,
\]
and from (61) we get that their sum is then
\[
\left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J} + \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J} \\
\leq \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \frac{2^{rM}}{2^{r/2 - 1}} \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} 2^{M} \\
\leq \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \frac{2^{rM}}{2^{r/2 - 1}} \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} 2^{M} \\
\lesssim \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \frac{2^{rM}}{2^{r/2 - 1}} \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} 2^{M} ,
\]
Thus, it only remains to handle the third and fourth terms by choosing a suitable value for \( \alpha \). Recalling that we are in the case \( p > \frac{r}{s+1} \), we pick any \( \frac{r}{s+1} < \alpha \leq p \); for instance, \( \alpha := p \). Since \( r-(s+1)p < 0 \) we then have
\[
\left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=M}^{H} 2^{J(r-(s+1)p)} \\
\leq \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} 2^{M(r-(s+1)p)} \\
\leq \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} 2^{M(r-(s+1)p)} ;
\]

note that \( \frac{1}{1-2^{(2s)(r-(s+1)p)}} > 0 \) is a constant, depending only on \( r, s, p \). Similarly, \( r-(2s+1)p < 0 \), and so
\[
\left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} \sum_{J=M}^{H} 2^{J(r-(2s+1)p)} \\
\leq \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} 2^{M(r-(2s+1)p)} \\
\leq \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} 2^{M(r-(2s+1)p)} ;
\]
From the setting of \( M \) from (61), by a distinction of cases we again can bound their sum as
\[
\left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=M}^{H} 2^{M(r-(s+1)p)} \\
+ \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} \sum_{J=M}^{H} 2^{M(r-(2s+1)p)} \\
\lesssim \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \frac{2^{rM}}{2^{r/2 - 1}} \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} 2^{M} .
\]
Therefore, overall, in the case \( p > \frac{r}{s+1} \) we have (bringing back the factor \( H^r/2 \) we had ignored earlier)
\[
E_{bb} \lesssim H^r/2 \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \frac{2^{rM}}{2^{r/2 - 1}} \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} 2^{M} . \quad (62)
\]
- If \( p \leq \frac{r}{s+1} \), we will choose \( \alpha_J \geq p \) for all \( J \). Under this constraint, we can use the monotonicity of \( \ell_p \) norms (for every \( x, ||x||_p \leq ||x||_q \) if \( p \geq q \)) to write
\[
E_{bb} \lesssim \sum_{J=1}^{H} m_J \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} 2^{J(r-(s+1)p)} \\
\leq \sum_{J=1}^{H} m_J \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} 2^{J(r-(s+1)p)} \\
\lesssim \sum_{J=1}^{H} m_J \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} 2^{J(r-(s+1)p)} \\
= \sum_{J=1}^{H} m_J \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} 2^{J(r-(s+1)p)} . \quad (Fact 14)
\]
As before, one can see that for there exists some \( M \) such that the best choice is to set \( \alpha_J = p \) for \( J \leq M \) (as small as possible given our constraint \( \alpha_J \geq p \)). Moreover, proceeding as in the previous case, we can see that this \( M \) is such that
\[
2^{M} \leq \frac{\left( \frac{n^2}{H} \right)^{\frac{1}{2}r-(s+1)p}}{2^{r/2 - 1}} \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} 2^{M} . \quad (63)
\]
This part of the sum will then contribute
\[
\sum_{J=1}^{M} m_J \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} 2^{J(r-(s+1)p)} (J-s-p(s+\frac{1}{2} - \frac{1}{2})) \\
\times \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J(r-(s+1)p)} \\
\times \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J(r-(s+1)p)} \\
\times \left( \frac{H}{n^{2s}} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J(r-(s+1)p)} \\
\times \left( \frac{H}{n} \right)^{\frac{r-\alpha}{J}} \sum_{J=1}^{M} 2^{J(r-(s+1)p)} .
\]
\[\text{3That is, find the value } J \text{ solving (approximately) the equation } \log \left( \frac{n}{H} \right) - (J-\ell) + J(2s+1-2/p) = 0 \text{ (note that the LHS is again decreasing in } J).\]
For \( J > M \), we choose an arbitrary constant \( \alpha \geq p \) such that \( \alpha > \frac{r-1}{s+1-1/p} \) (so that \( r - 1 - \alpha (s + 1 - 1/p) < 0 \)), and set \( \alpha_j = \alpha \) for all \( J > M \).\(^4\) Observe that this implies \( \alpha > \frac{r^2}{s+1-2+1/p} \). This part of the sum will then contribute at most \[
\sum_{J=M}^H \frac{r-a}{n^2} 2^J(\frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{s+1-1/p}) \]
\[
\leq \left( \frac{H}{n^2} \right)^{r-a} \sum_{J=M}^H 2^{J(r-1-\alpha(s+\frac{1}{2}-1/p))} + \left( \frac{H}{n} \right)^{r-a} \sum_{J=M}^H 2^{J(\frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{s+1-1/p})} \]
\[
\leq \left( \frac{H}{n^2} \right)^{r-a} 2^M(\frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{s+1-1/p}) + \left( \frac{H}{n} \right)^{r-a} 2^M(\frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{s+1-1/p}) \]
\[
\leq \left( \frac{H}{n^2} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} \right) + \left( \frac{H}{n} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} \right) \]
as well. Thus, overall, in the case \( p \leq \frac{r}{s+1} \), we have (bringing back the factor \( H^{r/2} \) we had ignored earlier)
\[
E_{bb} \leq H^{r/2} \left( \frac{H}{n^2} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} + \frac{H}{n} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} \right).
\]

3) The Term \( E_{sb} \): To handle the term \( E_{sb} \), we will rely on the fact that, for any \( r \geq p \), we have the inclusion \( B(p, q, s') \subseteq B(r, q, s') \), for \( s' = s - \left( \frac{n}{p} - \frac{1}{r} \right) \). This will let us use Fact 14 on \( \sum_{k \in \mathbb{Z}} |\beta_{J,k}|^2 \).
\[
E_{sb} \leq H^{r/2} \sum_{J=L}^H 2^{J(\frac{r-1}{s+1}) - 1} \left( \sum_{k \in \mathbb{Z}} |\beta_{J,k}|^2 \right) \left( \sum_{k \in \mathbb{Z}} \right) \left( \frac{h}{a} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} + \frac{H}{n} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} \right).
\]

where for the third-to-last inequality we relied on our choice of \( \kappa \geq r(N+1) \), and for the second-to-last, on our setting of \( L \).

4) The Term \( E_{ss} \): Finally, we bound the last error term for details\((f)\), \( E_{ss} \). In view of proceeding as for \( E_{bb} \), for any nonnegative sequence \( (\alpha_J)_J \) with \( 0 \leq \alpha_J \leq r \), we can write \( E_{ss} \)
\[
E_{ss} \leq H^{r/2} \sum_{J=L}^H 2^{J(\frac{r-1}{s+1}) - 1} \sum_{k \in \mathbb{Z}} |\beta_{J,k}|^{2} \left( \sum_{k \in \mathbb{Z}} \right) \left( \frac{h}{a} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} + \frac{H}{n} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} \right).
\]

which is, except for the extra factor of \( \left( \frac{r}{s+1} \right)^{r} H^{2} \), exactly the same expression as (60). We can thus continue the analysis of \( E_{ss} \) the same way as we did \( E_{bb} \), noting that since \( r \geq p \) all the choices for \( \alpha_J \) in that analysis are still possible; leading to the bound:
\[
E_{ss} \leq \begin{cases} 
H^{r/2} \cdot \left( \left( \frac{H}{n^2} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} + \frac{H}{n} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} \right) \right), & \text{if } p > \frac{r}{s+1}; \\
H^{r/2} \cdot \left( \left( \frac{H}{n^2} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} + \frac{H}{n} \right)^{r-a} \left( \frac{r-1-\alpha(s+\frac{1}{2}-1/p)}{2(s+1-1/p)} \right) \right), & \text{if } p \leq \frac{r}{s+1}.
\end{cases}
\]

REFERENCES

[1] J. Acharya, C. Canonne, A. V. Singh, and H. Tyagi, "Optimal rates for nonparametric density estimation under communication constraints," in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 26754–26766.

[2] L. P. Barnes, Y. Han, and A. Ozgur, "Lower bounds for learning distributions under communication constraints via Fisher information," J. Mach. Learn. Res., vol. 21, no. 236, pp. 1–30, 2020.

[3] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints II: Communication constraints and shared randomness," IEEE Trans. Inf. Theory, vol. 66, no. 12, pp. 7856–7877, Dec. 2020.

[4] J. Acharya, C. L. Canonne, Z. Sun, and H. Tyagi, "Unified lower bounds for interactive high-dimensional estimation under information constraints," 2020, arXiv:2010.06562.

[5] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," Ann. Statist., vol. 24, no. 2, pp. 508–539, Apr. 1996.

[6] C. Butucea, A. Dubois, M. Kroll, and A. Samurad, "Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids," Bernoulli, vol. 26, no. 3, pp. 1727–1764, Aug. 2020.

[7] Y. Zhu and J. Lafferty, "Quantized minimax estimation over Sobolev ellipsoids," Inf. Inference, A J. IMA, vol. 7, no. 1, pp. 31–82, Mar. 2018.
[8] Y. Zhu and J. Lafferty, “Distributed nonparametric regression under communication constraints,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 6009–6017.

[9] B. Szabó and H. van Zanten, “Adaptive distributed methods under communication constraints,” Ann. Statist., vol. 48, no. 4, pp. 2347–2380, Aug. 2020.

[10] B. Szabó and H. van Zanten, “Distributed function estimation: Adaptation using minimal communication,” Math. Statist. Learn., vol. 5, no. 3, pp. 159–199, 2020.

[11] T. T. Cai and H. Wei, “Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation,” Ann. Statist., vol. 50, no. 2, pp. 698–725, Apr. 2022.

[12] J. Liu, “A few interactions improve distributed nonparametric estimation, optimally,” IEEE Trans. Inf. Theory, early access, Aug. 30, 2023, doi: 10.1109/TIT.2023.3309920.

[13] Y. Han, A. Ozgur, and T. Weissman, “Geometric lower bounds for distributed parameter estimation under communication constraints,” in Proc. Conf. Learn. Theory, 2018, pp. 3163–3188.

[14] J. Acharya, C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi, “Interactive inference under information constraints,” IEEE Trans. Inf. Theory, vol. 68, no. 1, pp. 502–516, Jan. 2022.

[15] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, “Information-theoretic lower bounds for distributed statistical estimation with communication constraints,” in Proc. Adv. Neural Inf. Process. Syst., vol. 26, 2013, pp. 1–9.

[16] O. Shamir, “Fundamental limits of online and distributed algorithms for statistical learning and estimation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–9.

[17] A. Garg, T. Ma, and H. Nguyen, “On communication cost of distributed statistical estimation and dimensionality,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–9.

[18] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, “Communication lower bounds for statistical estimation problems via a distributed data processing inequality,” in Proc. 48th Annu. ACM Symp. Theory Comput., Jun. 2016, pp. 1011–1020.

[19] L. P. Barnes, Y. Han, and A. Özgür, “Fisher information for distributed estimation under a blackboard communication protocol,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jul. 2019, pp. 2704–2708.

[20] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov, Wavelets, Approximation, and Statistical Applications, vol. 129. Berlin, Germany: Springer, 2012.

[21] I. Daubechies, Ten Lectures on Wavelets. Philadelphia, PA, USA: SIAM, 1992.

[22] H. P. Rosenthal, “On the subspaces of $L^p(p>2)$ spanned by sequences of independent random variables,” Isr. J. Math., vol. 8, no. 3, pp. 273–303, Sep. 1970.