The Annals of Statistics
2020, Vol. 48, No. 4, 2347–2380
https://doi.org/10.1214/19-AOS1890

ADAPTIVE DISTRIBUTED METHODS UNDER COMMUNICATION CONSTRAINTS

BY BOTOND SZABÓ[1] AND HARRY VAN ZANTEN[2]

[1]*Mathematical Institute, Leiden University, b.t.szabo@math.leidenuniv.nl*
[2]*Department of Mathematics, Faculty of Sciences, VU Amsterdam, j.h.van.zanten@vu.nl*

We study estimation methods under communication constraints in a distributed version of the nonparametric random design regression model. We derive minimax lower bounds and exhibit methods that attain those bounds. Moreover, we show that adaptive estimation is possible in this setting.

1. Introduction. In this paper we study some aspects of the fundamental possibilities and limitations of distributed methods for high-dimensional or nonparametric problems. The design and study of such methods has attracted substantial attention recently. This is for a large part motivated by the ever increasing size of datasets, leading to the necessity to analyze data while distributed over multiple machines and/or cores. Other reasons to consider distributed methods include privacy considerations or the simple fact that in some situations data are physically collected at multiple locations.

By now a variety of methods are available for estimating nonparametric or high-dimensional models to data in a distributed manner. A (certainly incomplete) list of recent references includes the papers [1, 4, 8, 10, 12, 15–17, 23]. Some of these papers propose new methods, some study theoretical aspects of such methods and some do both. The number of theoretical papers on the fundamental performance of distributed methods is still rather limited however. In the paper [19] we recently introduced a distributed version of the canonical signal-in-white-noise model to serve as a benchmark model to study aspects like convergence rates and optimal tuning of distributed methods. We used it to compare the performance of a number of distributed nonparametric methods recently introduced in the literature. The study illustrated the intuitively obvious fact that in order to achieve an optimal bias-variance trade-off or, equivalently, to find the correct balance between over- and underfitting, distributed methods need to be tuned differently than methods that handle all data at once. Moreover, our comparison showed that some of the proposed methods are more successful at this than others.

A major challenge and fundamental question for nonparametric distributed methods is whether or not it is possible to achieve a form of adaptive inference. In other words, whether we can design methods that do automatic, data-driven tuning in order to achieve the optimal bias-variance trade-off. We illustrated by example in [19] that naively using methods that are known to achieve optimal adaptation in nondistributed settings can lead to suboptimal performance in the distributed case. In the recent paper [26], which considers the same distributed signal-in-white-noise model and was written independently and concurrently as the present paper, it is in fact conjectured that adaptation in the considered particular distributed model is not possible.

In order to study convergence rates and adaptation for distributed methods in a meaningful way, the class of methods should be restricted somehow. Indeed, if there is no limitation on

Received February 2019; revised July 2019.
MSC2020 subject classifications. Primary 62G20, 62G08; secondary 62B10.
Key words and phrases. Distributed computation, minimax rates, adaptation, nonparametric regression, communication constraints.

communication or computation, then we could simply communicate all data from the various local machines to a central machine, aggregate it and use some existing adaptive, rate-optimal procedure. In this paper we consider a setting in which the communication between the local and the global machines is restricted, much in the same way as the communication restrictions imposed in [23] in a parametric framework and recently in the simultaneously written paper [26] in the context of the distributed signal-in-white-noise model we introduced in [19].

In the distributed nonparametric regression model with communication constraints that we consider, we can derive minimax lower bounds for the best possible rate that any distributed procedure can achieve under smoothness conditions on the true regression function. Technically, this essentially relies on an extension of the information theoretic approach of [23] to the infinite-dimensional setting (this is different from the approach taken in [26] which relies on results from [21, 25]). It turns out there are different regimes, depending on how much communication is allowed. On the one extreme end and in accordance with intuition, if enough communication is allowed, then it is possible to achieve the same convergence rates in the distributed setting as in the nondistributed case. The other extreme case is that there is so little communication allowed that combining different machines does not help. Then, the optimal rate under the communication restriction can already be obtained by just using a single local machine and discarding the others. The interesting case is the intermediate regime. For that case we show there exists an optimal strategy that involves grouping the machines in a certain way and letting them work on different parts of the regression function.

These first results on rate-optimal distributed estimators are not adaptive, in the sense that the optimal procedures depend on the regularity of the unknown regression function. The same holds true for the procedure obtained in parallel in [26]. In this paper we go a step further and show that contrary perhaps to intuition and contrary to the conjecture in [26], adaptation is in fact possible. Indeed, we exhibit in this paper an adaptive distributed method which involves a very specific grouping of the local machines in combination with a Lepski-type method that is carried out in the central machine. We prove that the resulting distributed estimator adapts to a range of smoothness levels of the unknown regression function and that, up to logarithmic factors, it attains the minimax lower bound.

Although our analysis is theoretical, we believe it contains interesting messages that are ultimately very relevant for the development of applied distributed methods in high-dimensional settings. First of all, we show that, depending on the communication budget, it might be advantageous to group local machines and let different groups work on different aspects of the high-dimensional object of interest. Second, we show that it is possible to have adaptation in communication restricted distributed settings, that is, to have data-driven tuning that automatically achieves the correct bias-variance trade-off. We note, however, that although our proof of this fact is constructive, the method we exhibit appears to be still somewhat unpractical. We view our adaptation result primarily as a first proof of concept that hopefully invites the development of more practical adaptation techniques for distributed settings.

1.1. *Notation.* For two positive sequences a_n, b_n we use the notation $a_n \lesssim b_n$ if there exists an universal positive constant C such that $a_n \leq C b_n$. Along the lines $a_n \asymp b_n$ denotes that $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. Furthermore, we write $a_n \ll b_n$ if $a_n/b_n = o(1)$. Let us denote by $\lceil a \rceil$ and $\lfloor a \rfloor$ the upper and lower integer value of the real number a, respectively. The sum $\sum_{i=a}^{b} x_j$, for a, b real numbers, denotes the sum $\sum_{i \in \mathbb{N}: a \leq i \leq b} x_j$. For a set A let $|A|$ denote the size of the set. For $f \in L_2[0,1]$ we denote the standard L_2-norm as $\|f\|_2^2 = \int_0^1 f(x)^2\,dx$ while for bounded functions $\|f\|_\infty$ denotes the L_∞-norm. The function sign : $\mathbb{R} \mapsto \{0,1\}$ evaluates to 0 on $(-\infty, 0)$ and 1 on $[0, \infty)$. Furthermore, we use the notation $\text{mean}\{a_1, \ldots, a_n\} = (a_1 + \cdots + a_n)/n$. Throughout the paper c and C denote global constants whose value may change from one line to another.

2. Main results. We work with the distributed version of the random design regression model. We assume that we have m "local" machines and in the ith machine we observe pairs of random variables $(T_\ell^{(i)}, X_\ell^{(i)})$, $\ell = 1, \dots, n/m$, (with $n/m \in \mathbb{N}$) satisfying

$$
\begin{aligned}
&X_\ell^{(i)} = f_0(T_\ell^{(i)}) + \sigma \varepsilon_\ell^{(i)} \quad \text{where} \\
&T_\ell^{(i)} \overset{iid}{\sim} U(0,1), \qquad \varepsilon_\ell^{(i)} \overset{iid}{\sim} N(0,1), \quad \ell = 1, \dots, n/m, i = 1, \dots, m,
\end{aligned} \tag{2.1}
$$

and $f_0 \in L_2[0,1]$ (which is the same for all machines) is the unknown functional parameter of interest. For simplicity we take $\sigma = 1$. We denote the data distribution and expectation corresponding to the ith machine in (2.1) by $\mathbb{P}_{f_0,T}^{(i)}$ and $\mathbb{E}_{f_0,T}^{(i)}$, respectively, and the joint distribution and expectation over all machines $i = 1, \dots, m$, by $\mathbb{P}_{f_0,T}$ and $\mathbb{E}_{f_0,T}$, respectively. We assume that the total sample size n is known to every local machine. For our theoretical results we will assume that the unknown true function f_0 belongs to some regularity class. We work in our analysis with Besov smoothness classes, more specifically we assume that for some degree of smoothness $s > 0$ we have $f_0 \in B_{2,\infty}^s(L)$ or $f_0 \in B_{\infty,\infty}^s(L)$. The first class is of Sobolev type, while the second one is of Hölder type with minimax estimation rates $n^{-s/(1+2s)}$ and $(n/\log n)^{-s/(1+2s)}$, respectively. For precise definitions see Section B in the Supplementary Material [18]. Each local machine carries out (parallel to the others) a local statistical procedure and transmits the results to a central machine which produces an estimator for the signal f_0 by somehow aggregating the messages received from the local machines.

We study these distributed procedures under communication constraints between the local machines and the central machine. We allow each local machine to send at most $B^{(i)}$ bits on average to the central machine. More formally, a distributed estimator $\hat{f}$ is a measurable function of m binary strings, or messages, passed from the local machines to the central machine. We denote by $Y^{(i)}$ the finite binary string transmitted from machine i to the central machine which is a measurable function of the local data $T^{(i)}$, $X^{(i)}$. For a class of potential signals $\mathcal{F} \subset L_2[0,1]$, we restrict the communication between the machines by assuming that for numbers $B^{(1)}, \dots, B^{(m)}$, it holds that $\mathbb{E}_{f_0,T}[l(Y^{(i)})] \le B^{(i)}$ for every $f_0 \in \mathcal{F}$ and $i = 1, \dots, m$, where $l(Y)$ denotes the length of the string Y. We denote the resulting class of communication restricted distributed estimators $\hat{f}$ by $\mathcal{F}_{\text{dist}}(B^{(1)}, \dots, B^{(m)}; \mathcal{F})$. The number of machines $m = m_n$ and the communication constraints $B^{(i)} = B_n^{(i)}$ are allowed to change with the overall sample size n. In fact, that is the interesting situation. However, to alleviate the notational burden somewhat, we do not make this explicit in the notation.

2.1. *Distributed minimax lower bounds for the L_2-risk.* The first theorem we present gives a minimax lower bound for distributed procedures for the L_2-risk, uniformly over Sobolev-type Besov balls; see Section B in the Supplementary Material for rigorous definitions.

THEOREM 2.1. *Consider $s, L > 0$, $\log_2 n \le m = O(n^{\frac{2s}{1+2s}}/\log^2 n)$ and communication constraints $B^{(1)}, \dots, B^{(m)} > 0$. Let the sequence $\delta_n = o(1)$ be defined as the solution to the equation*

$$
\delta_n = L^{-2} \min\left\{ \frac{m}{n \log_2 n}, \frac{m}{n \sum_{i=1}^m [(\log_2(n) \delta_n^{\frac{1}{1+2s}} B^{(i)}) \wedge 1]} \right\}. \tag{2.2}
$$

Then, in distributed random design nonparametric regression model (2.1) we have that

$$
\inf_{\hat{f} \in \mathcal{F}_{\text{dist}}(B^{(1)}, \dots, B^{(m)}; B_{2,\infty}^s(L))} \sup_{f_0 \in B_{2,\infty}^s(L)} \mathbb{E}_{f_0,T} \|\hat{f} - f_0\|_2^2 \gtrsim L^2 \delta_n^{\frac{2s}{1+2s}}.
$$

PROOF. See Section 3.1. □

We briefly comment on the derived result. First of all, note that the quantity δ_n in (2.2) is well defined, since the left-hand side of the equation is increasing, while the right-hand side is decreasing in δ_n. In general, there is no explicit expression for δ_n. See the corollary below, however, for the special case that the communication constraints are the same for each machine, in which case we have explicit lower bounds.

The proof of the theorem is based on an application of a version of Fano's inequality, frequently used to derive minimax lower bounds. More specifically, as a first step we find, as usual, a large enough finite subset of the functional space $B^s_{2,\infty}(L)$ over which the minimax rate is the same as over the whole space. This is done by finding the "effective resolution level" j_n in the wavelet representation of the function of interest and perturbing the corresponding wavelet coefficients while setting the rest of the coefficients to zero. This effective resolution level for s-smooth functions is usually $(1+2s)^{-1}\log_2 n$ in case of the L_2-norm for nondistributed models (e.g., [7]). However, in our distributed setting the effective resolution level changes to $(1+2s)^{-1}\log \delta_n^{-1}$, which can be substantially different from the nondistributed case, as it strongly depends on the number of transmitted bits. The dependence on the expected number of transmitted bits enters the formula by using a variation of Shannon's source coding theorem. Many of the information theoretic manipulations in the proof are an extended and adapted version of the approach introduced in [23], where similar results were derived in context of distributed methods with communication constraints over parametric models.

To understand the result, it is illustrative to consider the special case that the communication constraints are the same for all machines, that is, $B^{(1)} = \cdots = B^{(m)} = B$ for some $B > 0$. We can then distinguish three regimes: (i) the case $B \geq (L^2 n)^{1/(1+2s)}/\log_2 n$; (ii) the case $(L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)} \leq B < (L^2 n)^{1/(1+2s)}/\log_2 n$; and (iii) the case $B < (L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)}$.

In regime (i) we have a large communication budget and by elementary computations we get that the minimum in (2.2) is attained in the second fraction and hence that $\delta_n = 1/(L^2 n)$. This means that in this case the derived lower bound corresponds to the usual nondistributed minimax rate $L^{2/(1+2s)} n^{-2s/(1+2s)}$. In the other extreme case, regime (iii), the minimum is taken at the first term in (2.2) and $\delta_n = m/(L^2 n \log_2 n)$, so the lower bound is of the order $L^{2/(1+2s)}(n\log_2(n)/m)^{-2s/(1+2s)}$. This rate is, up to the $\log_2 n$ factor, equal to the minimax rate corresponding to the sample size n/m. Consequently, in this case it does not make sense to consider distributed methods, since by just using a single machine the best rate can already be obtained (up to a logarithmic factor). In the intermediate case (ii) it is straightforward to see that $\delta_n = (L^2 n B \log_2 n)^{(1+2s)/(2+2s)}$. It follows that if $B = o(n^{1/(1+2s)}/\log_2 n)$, that is, if we are only allowed to communicate "strictly" less than in case (i), then the lower bound is strictly worse than the minimax rate corresponding to the nondistributed setting.

The findings above are summarized in the following corollary.

COROLLARY 2.2. *Consider* $s, L > 0$, *communication constraints* $B^{(1)} = \cdots = B^{(m)} = B > 0$ *and assume that* $\log_2 n \leq m = O(n^{\frac{2s}{1+2s}}/\log^2 n)$. *Then*

(i) *if* $B \geq (L^2 n)^{1/(1+2s)}/\log_2 n$,

$$\inf_{\hat f \in \mathcal{F}_{\mathrm{dist}}(B,\ldots,B;B^s_{2,\infty}(L))} \sup_{f_0 \in B^s_{2,\infty}(L)} \mathbb{E}_{f_0,T}\|\hat f - f_0\|_2^2 \gtrsim L^{\frac{2}{1+2s}} n^{-\frac{2s}{1+2s}};$$

(ii) *if* $(L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)} \leq B < (L^2 n)^{1/(1+2s)}/\log_2 n$,

$$\inf_{\hat f \in \mathcal{F}_{\mathrm{dist}}(B,\ldots,B;B^s_{2,\infty}(L))} \sup_{f_0 \in B^s_{2,\infty}(L)} \mathbb{E}_{f_0,T}\|\hat f - f_0\|_2^2 \gtrsim L^{\frac{2}{1+s}} \left(\frac{n^{1/(1+2s)}}{B\log_2 n}\right)^{\frac{s}{1+s}} n^{-\frac{2s}{1+2s}};$$

(iii) *if* $(L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)} > B$,

$$\inf_{\hat{f}\in\mathcal{F}_{\mathrm{dist}}(B,\ldots,B;B^s_{2,\infty}(L))} \sup_{f_0\in B^s_{2,\infty}(L)} \mathbb{E}_{f_0,T}\|\hat{f}-f_0\|_2^2 \gtrsim L^{\frac{2}{1+2s}}\left(\frac{n\log_2 n}{m}\right)^{-\frac{2s}{1+2s}}.$$

2.2. *Nonadaptive rate-optimal distributed procedures for L_2-risk.* Next, we show that the derived lower bounds are (nearly) sharp by presenting distributed procedures that attain the bounds (up to logarithmic factors). We note that it is sufficient to consider only the case $B \geq (L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)}$, since otherwise distributed techniques do not perform better than standard techniques carried out on one of the local servers. Therefore, in case (iii) one would probably prefer to use a single local machine instead of a complicated distributed method with (possibly) worse performance.

As a first step let us consider Daubechies wavelets ψ_{jk}, $j = 0, \ldots$, $k = 0, 1, \ldots, 2^j - 1$ with at least s vanishing moments (for details see Section B in the Supplementary Material). Then, let us estimate the wavelet coefficients of the underlying function f_0 in each local problem, that is, for every $j = 0, \ldots$, and $k = 0, 1, \ldots, 2^j - 1$ let us construct

$$\hat{f}^{(i)}_{jk} = \frac{m}{n}\sum_{\ell=1}^{n/m} X^{(i)}_\ell \psi_{jk}(T^{(i)}_\ell)$$

and note that

$$\mathbb{E}_{f_0,T}\hat{f}^{(i)}_{jk} = \int_0^1 f_0(t)\psi_{jk}(t)\,dt = f_{0,jk}.$$

Since one can only transmit finite amount of bits, we have to approximate the estimators of the wavelet coefficients. Let us take an arbitrary $x \in \mathbb{R}$ and write it in a scientific binary representation, that is, $|x| = \sum_{k=-\infty}^{\log_2|x|} b_k 2^k$, with $b_k \in \{0, 1\}$, $k \in \mathbb{Z}$. Then, let us take y consisting the same digits as x up to the $(D\log_2 n)$th digits, for some $D > 0$, after the binary dot (and truncated there), that is, $|y| = \sum_{k=-D\log_2 n}^{\log_2|x|} b_k 2^k$; see also Algorithm 1.

Observe that the length of y (viewed as a binary string) is bounded from above by $1 + (1 \vee \log_2|x|) + D\log_2 n$ bits. The following lemma asserts that if $\mathbb{E}(1 \vee \log_2|X|) = o(\log_2 n)$, then the expected length $\mathbb{E}[l(Y)]$ of the constructed binary string approximating X is less than constant times $\log_2 n$ (for sufficiently large n) and the approximation is polynomially close to X.

LEMMA 2.3. *Assume that* $\mathbb{E}(1 \vee \log_2|X|) = o(\log_2 n)$. *Then, the approximation Y of X given in Algorithm 1 satisfies that*

$$0 \leq |X - Y| \leq n^{-D} \quad \textit{and} \quad \mathbb{E}\big[l(Y)\big] \leq \big(D + o(1)\big)\log_2(n).$$

PROOF. See Section 3.4. □

After these preparations we can exhibit procedures attaining (nearly) the theoretical limits obtained in Corollary 2.2.

Algorithm 1 Transmitting a finite-bit approximation of a number

1: **procedure** TRANSAPPROX(x)
2: *Transmit*: $\mathrm{sign}(x)$, $b_{-\lfloor D\log_2 n\rfloor}, \ldots, b_{\lfloor\log_2|x|\rfloor}$.
3: *Construct*: $y = (2\mathrm{sign}(x) - 1)\sum_{k=-D\log_2 n}^{\log_2|x|} b_k 2^k$.

Algorithm 2 Nonadaptive L_2-method, case (i)

In the local machines:
for $i = 1$ to m **do**:
  for $2^j + k = 1$ to $(L^2 n)^{1/(1+2s)} \wedge (B/\log_2 n)$ **do**
    $Y_{jk}^{(i)} := \text{TransApprox}(\hat{f}_{jk}^{(i)})$
In the central machine:
for $2^j + k = 1$ to $(L^2 n)^{1/(1+2s)} \wedge (B/\log_2 n)$ **do**
  $\hat{f}_{jk} := \text{mean}\{Y_{jk}^{(i)} : 1 \le i \le m\}$.
Construct: $\hat{f} = \sum \hat{f}_{jk}\psi_{jk}$.

We first consider the case (i) that $B \ge (L^2 n)^{1/(1+2s)}/\log_2 n$. In this case each local machine $i = 1, \dots, m$ transmits the approximations $Y_{jk}^{(i)}$ (given in Algorithm 1 with $D = 1/2$) of the first $(L^2 n)^{1/(1+2s)} \wedge (B/\log_2 n)$ wavelet coefficients $\hat{f}_{jk}^{(i)}$, that is, for $2^j + k \le (L^2 n)^{1/(1+2s)} \wedge (B/\log_2 n)$. Then, in the central machine we simply average the transmitted approximations to obtain the estimated wavelet coefficients

$$\hat{f}_{jk} = \begin{cases} \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} Y_{jk}^{(i)} & \text{if } 2^j + k \le \big(L^2 n\big)^{1/(1+2s)} \wedge (B/\log_2 n), \\ 0 & \text{else.} \end{cases}$$

The final estimator $\hat{f}$ for f_0 is the function in $L_2[0,1]$ with these wavelet coefficients, that is, $\hat{f} = \sum \hat{f}_{jk}\psi_{jk}$. The method is summarized as Algorithm 2.

We note again that the procedure outlined in Algorithm 2 is just a simple averaging, sometimes called "divide and conquer" or "embarrassingly parallel" in the learning literature (e.g., [14, 24]).

The following theorem asserts that the constructed estimator indeed attains the lower bound in case (i) (up to a logarithmic factor for B close to the threshold). Note that in this upper bound result (and in the ones ahead) we do not have to assume the technical condition on the number of machines as in the lower bounds.

THEOREM 2.4. *Let $s, L > 0$, $m \le n$, and suppose that $B \ge (L^2 n)^{1/(1+2s)}/\log_2 n$. Then, the distributed estimator $\hat{f}$ described in Algorithm 2 belongs to $\mathcal{F}_{\text{dist}}(B, \dots, B; B_{2,\infty}^s(L))$ and satisfies*

$$\sup_{f_0 \in B_{2,\infty}^s(L), \|f_0\|_\infty \le M} \mathbb{E}_{f_0,T}\|\hat{f} - f_0\|_2^2 \lesssim \big(L^{\frac{2}{1+2s}} n^{-\frac{2s}{1+2s}}\big) \vee \big(L^2 (B/\log_2 n)^{-2s}\big).$$

PROOF. See Section 3.2. □

Next we consider the case (ii) of Corollary 2.2, that is, the case that the communication restriction satisfies $(L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)} \le B < (L^2 n)^{1/(1+2s)}/\log_2 n$. For technical reasons we also assume that $B \ge \log_2 n$. Using Algorithm 2 in this case would result in a highly suboptimal procedure, as we prove at the end of Section 3.3. It turns out that under this more severe communication restriction we can do much better if we form different groups of machines that work on different parts of the signal.

We introduce the notation $\eta = \lfloor (L^2 n)^{\frac{1}{2+2s}} ((\log_2 n)/B)^{\frac{1+2s}{2+2s}} \rfloor \wedge m$. Then, we group the local machines into η groups and let the different groups work on different parts of wavelet domain as follows: the machines with numbers $1 \le i \le m/\eta$ each transmit the approximations $Y_{jk}^{(i)}$ of

Algorithm 3 Nonadaptive L_2-method, case (ii)

In the local machines:
for $\ell = 1$ to η **do**
  for $i = \lfloor (\ell-1)m/\eta \rfloor + 1$ to $\lfloor \ell m/\eta \rfloor$ **do**
    for $2^j + k = (\ell-1)\lfloor B/\log_2 n \rfloor + 1$ to $\ell \lfloor B/\log_2 n \rfloor$ **do**
      $Y_{jk}^{(i)} := \mathrm{TransApprox}(\hat{f}_{jk}^{(i)})$.
In the central machine:
for $2^j + k = 1$ to $\eta \lfloor B/\log_2 n \rfloor$ **do**
  $\hat{f}_{jk} := \mathrm{mean}\{Y_{jk}^{(i)} : \mu_{jk} m/\eta < i \le (\mu_{jk}+1)m/\eta\}$.
Construct: $\hat{f} = \sum \hat{f}_{jk}\psi_{jk}$.

the estimated wavelet coefficients $\hat{f}_{jk}^{(i)}$ for $1 \le 2^j + k \le \lfloor B/\log_2 n \rfloor$; the next machines, with numbers $m/\eta < i \le 2m/\eta$, each transmit the approximations $Y_{jk}^{(i)}$ for $\lfloor B/\log_2 n \rfloor < 2^j + k \le 2\lfloor B/\log_2 n \rfloor$, and so on. The last machines with numbers $(\eta-1)m/\eta < i \le m$ transmit the $Y_{jk}^{(i)}$ for $(\eta-1)\lfloor B/\log_2 n \rfloor < 2^j + k \le \eta \lfloor B/\log_2 n \rfloor$. Then, in the central machine we average the corresponding transmitted noisy coefficients in the obvious way. Formally, using the notation $\mu_{jk} = \lceil (2^j + k)\lfloor B/\log_2 n \rfloor^{-1} \rceil - 1$, the aggregated estimator $\hat{f}$ is the function with wavelet coefficients given by

$$\hat{f}_{jk} = \begin{cases} \mathrm{mean}\left\{ Y_{jk}^{(i)} : \dfrac{\mu_{jk} m}{\eta} < i \le \dfrac{(\mu_{jk}+1)m}{\eta} \right\} & \text{if } 2^j + k \le \eta \lfloor B/\log_2 n \rfloor, \\ 0 & \text{else.} \end{cases}$$

The procedure is summarized as Algorithm 3.

The following theorem asserts that this estimator attains the lower bound in case (ii) (up to a logarithmic factor). We also prove in Section 3.3 that Algorithm 2 is suboptimal in this case.

THEOREM 2.5. *Let* $s, L > 0$, $m \le n$ *and suppose that* $(L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)} \vee \log_2 n \le B < (L^2 n)^{1/(1+2s)}/\log_2 n$. *Then, the distributed estimator* $\hat{f}$ *described in Algorithm 3 belongs to* $\mathcal{F}_{\mathrm{dist}}(B, \dots, B; B^s_{2,\infty}(L))$ *and satisfies*

$$\sup_{f_0 \in B^s_{2,\infty}(L), \|f_0\|_\infty \le M} \mathbb{E}_{f_0,T} \|\hat{f}_n - f_0\|_2^2 \lesssim M_n L^{\frac{2}{1+s}} \left(\frac{n^{1/(1+2s)}}{B \log_2 n} \right)^{\frac{s}{1+s}} n^{-\frac{2s}{1+2s}},$$

with $M_n = (\log_2 n)^{2s}$.

PROOF. See Section 3.3. □

REMARK 2.6. Instead of an upper bound on the expected number of transmitted bits $\mathbb{E}_{f_0,T}[l(Y^{(i)})] \le B^{(i)}$, one could consider a stronger, almost sure restriction, that is, $l(Y^{(i)}) \le B^{(i)}$ holds $\mathbb{P}^{(i)}_{f_0,T}$-a.s., for all $i = 1, \dots, m$. It is straightforward to see that the minimax lower bounds derived in Theorem 2.1 and Corollary 2.2 still hold under this assumption. Furthermore, we can show that the lower bounds are tight, that is, by slightly modifying the Algorithms 2 and 3 we get the same convergence rate as in Theorems 2.4 and 2.5 under the more restrictive almost sure upper bound on the number of communicated bits. The proof of the remark is deferred to Section 3.5.

REMARK 2.7. The computational complexity of the estimator $\hat{f}_{jk}^{(i)}$, for any j, k, i is $O(n/m)$. Since each local machine transmits at most $(B/\log_2 n) \vee n^{1/(1+2s)}$ wavelet coefficients, the total computational complexity is $O(((B/\log_2 n) \vee n^{1/(1+2s)})n/m)$. In the central machine we average out the local estimators. The computational cost of each estimator $\hat{f}_{jk}$ is $O(m/\eta)$ and, since we compute $\eta B/\log_2 n$ coefficients, the total computational cost in the central machine is $O(mB/\log_2 n)$. As a benchmark the computational complexity of a nondistributed wavelet thresholding estimator is $O(n^{1/(1+2s)}n)$.

2.3. *Distributed minimax results for L_∞-risk.* When we replace the L_2-norm by the L_∞-norm and correspondingly change the type of Besov balls we consider, we can derive a lower bound similar to Theorem 2.1 (see Section B in the Supplementary Material for the rigorous definition of Besov balls).

THEOREM 2.8. *Consider $s, L > 0$, communication constraints $B^{(1)}, \dots, B^{(m)} > 0$, and assume that $\log_2 n \le m = O(n^{\frac{2s}{1+2s}}/\log^2 n)$. Let the sequence $\delta_n = o(1)$ be defined as the solution to the equation (2.2). Then, in the distributed random design regression model (2.1) we have that*

$$\inf_{\hat{f} \in \mathcal{F}_{\text{dist}}(B^{(1)},\dots,B^{(m)};\, B^s_{\infty,\infty}(L))} \sup_{f_0 \in B^s_{\infty,\infty}(L)} \mathbb{E}_{f_0,T} \|\hat{f} - f_0\|_\infty$$
$$\gtrsim L^{\frac{1}{1+2s}} \left(\frac{n}{\log n}\right)^{-\frac{s}{1+2s}} \vee L\delta_n^{\frac{s}{1+2s}}.$$

PROOF. See Section 4.1. □

The proof of the theorem is very similar to the proof of Theorem 2.8. The first term on the right-hand side follows from the usual nondistributed minimax lower bound. For the second term we use the standard version of Fano's inequality. We again consider a large enough finite subset of $B^s_{\infty,\infty}(L)$. The effective resolution level for the L_∞-norm in the nondistributed case is $(1+2s)^{-1}\log_2(n/\log_2 n)$. Similar to the L_2 case, the effective resolution level changes to $(1+2s)^{-1}\log \delta_n^{-1}$ in the distributed setting which can be again substantially different from the nondistributed case. The rest of the proof follows the same line of reasoning as the proof of Theorem 2.8.

Similarly to the L_2-norm, we consider again the specific case where all communication budgets are taken to be equal, that is, $B^{(1)} = B^{(2)} = \cdots = B^{(m)} = B$. One can easily see that there are again three regimes of B (slightly different compared to the L_2-case).

COROLLARY 2.9. *Consider $s, L > 0$, communication constraint $B^{(1)} = \cdots = B^{(m)} = B > 0$, and assume that $\log_2 n \le m = O(n^{\frac{2s}{1+2s}}/\log^2 n)$.*

(ib) *If $B \ge (L^2 n/(\log_2 n)^{3+4s})^{1/(1+2s)}$, then*

$$\inf_{\hat{f} \in \mathcal{F}_{\text{dist}}(B,\dots,B;\, B^s_{\infty,\infty}(L))} \sup_{f_0 \in B^s_{\infty,\infty}(L)} \mathbb{E}_{f_0,T} \|\hat{f} - f_0\|_\infty \gtrsim L^{\frac{1}{1+2s}} (n/\log_2 n)^{-\frac{s}{1+2s}}.$$

(iib) *If $(L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)} \le B < (L^2 n/(\log_2 n)^{3+4s})^{1/(1+2s)}$, then*

$$\inf_{\hat{f} \in \mathcal{F}_{\text{dist}}(B,\dots,B;\, B^s_{\infty,\infty}(L))} \sup_{f_0 \in B^s_{\infty,\infty}(L)} \mathbb{E}_{f_0,T} \|\hat{f} - f_0\|_\infty$$
$$\gtrsim L^{\frac{1}{1+s}} \left(\frac{n^{\frac{1}{1+2s}}}{B(\log_2 n)^{\frac{3+4s}{1+2s}}}\right)^{\frac{s}{2+2s}} \left(\frac{n}{\log_2 n}\right)^{-\frac{s}{1+2s}}.$$

(iiib) *If* $(L^2 n \log_2(n)/m^{2+2s})^{1/(1+2s)} > B$, *then*

$$\inf_{\hat{f}\in\mathcal{F}_{\mathrm{dist}}(B,\dots,B;B^s_{\infty,\infty}(L))} \sup_{f_0\in B^s_{\infty,\infty}(L)} \mathbb{E}_{f_0,T}\|\hat{f}-f_0\|_\infty \gtrsim L^{\frac{1}{1+2s}}\left(\frac{n\log_2 n}{m}\right)^{-\frac{s}{1+2s}}.$$

Next we provide matching upper bounds (up to a $\log n$ factor) in the first two cases, that is, (ib) and (iib). In the third case the lower bound matches (up to a logarithmic factor) the minimax rate corresponding to a single local machine, hence it is not advantageous at all to develop complicated distributed techniques as a single server with only fraction of the total information performs at least as well. In the previous section dealing with L_2 estimation, we have provided two algorithms (one where the machines had the same tasks and one where the machines were divided into groups and were assigned different tasks) to highlight the differences between the cases. Here, for simplicity we combine the algorithms to a single one, but essentially the same techniques are used as before.

In each local machine we compute the local estimators of the wavelet coefficients $\hat{f}^{(i)}_{jk}$ and transmit a finite digit approximation of them $Y^{(i)}_{jk}$, as in the L_2-case. Then, let us divide the machines into $\eta = (\lfloor (L^2 n(\log_2 n)^{2s}/B^{1+2s})^{\frac{1}{2+2s}}\rfloor \wedge m) \vee 1$ equal sized groups ($\eta = 1$ corresponds to case (ib), while $\eta > 1$ corresponds to case (iib)). Similarly to before machines with numbers $1 \le i \le m/\eta$ transmit the approximations $Y^{(i)}_{jk}$ of the estimated wavelet coefficients $\hat{f}^{(i)}_{jk}$ for $1 \le 2^j + k \le \lfloor B/\log_2 n\rfloor \wedge (n/\log_2 n)^{\frac{1}{1+2s}}$, and so on, the last machines with numbers $(\eta-1)m/\eta < i \le m$ transmit the approximations $Y^{(i)}_{jk}$ for $((\eta-1)\lfloor B/\log_2 n\rfloor) \wedge (n/\log_2 n)^{\frac{1}{1+2s}} < 2^j + k \le (\eta\lfloor B/\log_2 n\rfloor) \wedge (n/\log_2 n)^{\frac{1}{1+2s}}$. In the central machine we average the corresponding transmitted coefficients in the obvious way, that is, the aggregated estimator $\hat{f}$ is the function with wavelet coefficients given by

$$\hat{f}_{jk} = \begin{cases} \mathrm{mean}\left\{ Y^{(i)}_{jk} : \dfrac{\mu_{jk} m}{\eta} < i \le \dfrac{(\mu_{jk}+1)m}{\eta} \right\} \\ \qquad \text{if } 2^j + k \le \eta \left\lfloor \dfrac{B}{\log_2 n} \right\rfloor \wedge \left(\dfrac{n}{\log n}\right)^{\frac{1}{1+2s}}, \\ 0 \quad \text{else,} \end{cases}$$

where $\mu_{jk} = \lceil (2^j + k)\lfloor B/\log_2 n\rfloor^{-1}\rceil - 1$. The procedure is summarized as Algorithm 4 and the (up to a logarithmic factor) optimal behaviour is given in Theorem 2.10 below.

Algorithm 4 Nonadaptive L_∞-method, combined

In the local machines:
for $\ell = 1$ to η **do**
  for $i = \lfloor(\ell-1)m/\eta\rfloor + 1$ to $\lfloor \ell m/\eta\rfloor$ **do**
    for $2^j + k = (\ell-1)\lfloor B/\log_2 n\rfloor + 1$ to $\ell\lfloor B/\log_2 n\rfloor$ **do**
      $Y^{(i)}_{jk} := \mathrm{TransApprox}(\hat{f}^{(i)}_{jk})$.
In the central machine:
for $2^j + k = 1$ to $\eta\lfloor B/\log_2 n\rfloor$ **do**
  $\hat{f}_{jk} := \mathrm{mean}\{Y^{(i)}_{jk} : \mu_{jk} m/\eta < i \le (\mu_{jk}+1)m/\eta\}$.
Construct: $\hat{f} = \sum \hat{f}_{jk}\psi_{jk}$.

THEOREM 2.10. *Let $s, L > 0$. Then, the distributed estimator $\hat{f}$ described in Algorithm 4 belongs to $\mathcal{F}_{\mathrm{dist}}(B, \ldots, B; B^s_{\infty,\infty}(L))$ and satisfies*

- *for $B \geq n^{1/(1+2s)}(\log_2 n)^{2s/(1+2s)}$,*

$$\sup_{f_0 \in B^s_{\infty,\infty}(L)} \mathbb{E}_{f_0,T} \|\hat{f}_n - f_0\|_\infty \lesssim L^{\frac{1}{1+2s}} (n/\log_2 n)^{-\frac{s}{1+2s}};$$

- *for $(n(\log_2 n)/m^{2+2s})^{1/(1+2s)} \vee \log_2 n \leq B < n^{1/(1+2s)}(\log_2 n)^{2s/(1+2s)}$,*

$$\sup_{f_0 \in B^s_{\infty,\infty}(L)} \mathbb{E}_{f_0,T} \|\hat{f}_n - f_0\|_\infty \lesssim M_n L^{\frac{1}{1+s}} \left(\frac{n^{\frac{1}{1+2s}}}{B(\log_2 n)^{\frac{3+4s}{1+2s}}} \right)^{\frac{s}{2+2s}} (n/\log_2 n)^{-\frac{s}{1+2s}},$$

with $M_n = (\log_2 n)^{s \vee \frac{3s}{2+2s}}$.

PROOF. See Section 4.2. □

We can draw similar conclusions for the L_∞-norm as for the L_2-norm. If we do not transmit a sufficient amount of bits (at least $n^{1/(1+2s)}$ up to a $\log n$ factor) from the local machines to the central one, then the lower bound from the theorem exceeds the minimax risk corresponding to the nondistributed case. Furthermore, by transmitting the sufficient amount of bits (i.e., $n^{1/(1+2s)}$ up to a $\log n$ factor) corresponding to the class $B^s_{\infty,\infty}(L)$, the lower bound will coincide with the nondistributed minimax estimation rate.

REMARK 2.11. We have restricted the analysis to $B^s_{2,\infty}$ and $B^s_{\infty,\infty}$ Besov balls. Obviously, a more complete picture over a broader scale of Besov spaces would be desirable. We note, however, that $B^s_{p,q}$ spaces can be handled similarly to the $B^s_{2,\infty}$ case with the cost of some additional technical and notational complexity; see for instance [2, 7, 9] for extension of results in $B^s_{2,\infty}$, $B^s_{\infty,\infty}$ to general B^s_{pq}.

2.4. *Adaptive distributed estimation.* The (almost) rate-optimal procedures considered so far have in common that they are nonadaptive, in the sense that they all use the knowledge of the regularity level s of the unknown functional parameter of interest. In this section we exhibit a distributed algorithm attaining the lower bounds (up to a logarithmic factor) across a whole range of regularities s simultaneously. In the nondistributed setting it is well known that this is possible, and many adaptation methods exist, including, for instance, the block Stein method, Lepski's method, wavelet thresholding and Bayesian adaptation methods, just to mention but a few (e.g., [7, 20]). In the distributed case the matter is more complicated. Using the usual adaptive tuning methods in the local machines will typically not work (see [19]), and in fact it was recently conjectured that adaptation, if at all possible, would require more communication than is allowed in our model (see [26]).

We will show, however, that in our setting, if all machines have the same communication restriction given by $B \geq \log_2 n$, it is possible to adapt to regularities s ranging in the interval $(s_{\min}, s_{\max})$, where

$$s_{\min} = \limsup_n \left(\inf\left\{ s > 0 : \left(n(\log_2 n)^2 / m^{2+2s} \right)^{\frac{1}{1+2s}} \leq B \right\} \right) \tag{2.3}$$

and $s_{\max}$ is the regularity of the considered Daubechies wavelet and can be chosen arbitrarily large. Note that $s_{\min}$ is well defined. If $s \in (s_{\min}, s_{\max})$, then we are in one of the nontrivial cases (i) or (ii) of Corollary 2.2. We will construct a distributed method which, up to logarithmic factors, attains the corresponding lower bounds, without using knowledge about the regularity level s.

REMARK 2.12. We provide some examples for the value of $s_{\min}$ for different choices of B and m. Taking $m = \sqrt{n}$, we have for all $B \geq \log_2 n$ that $s_{\min} = 0$. For $m = \log n$ and $B = \sqrt{n}$, we get $s_{\min} = 1/2$. For $m = \log n$ and $B = \log_2 n$, we have that $s_{\min} = \infty$. Note that it is intuitively clear that in case the number of machines is large, then it is typically advantageous to use a distributed method compared to a single local machine as we would lose too much information in the later case. However, if we have a small number of machines and can transmit only a very limited amount of information, then it might be more advantageous to use only a single machine to make inference.

In the non-adaptive case we saw that different strategies were required to attain the optimal rate, case (ii) requiring a particular grouping of the local machines. The cut-off between cases (i) and (ii) depends, however, on the value of s, so in the present adaptive setting we do not know beforehand in which of the two cases we are. In order to tackle this problem, we introduce a somewhat more involved grouping of the machines which basically gives us the possibility to carry out both strategies simultaneously. This is combined with a modified version of Lepski's method, carried out in the central machine, ultimately leading to (nearly) optimal distributed concentration rates for every regularity class $s \in (s_{\min}, s_{\max})$, simultaneously. We note that in our distributed regression setting, deriving an appropriate version of Lepski's method requires some nonstandard technical work; see Section 3.6. For a treatment and discussion of Lepski's method in the usual signal-in-white-noise model, see, for instance, Chapter 8 of [7].

Loosely speaking, the grouping of the machines can be described as follows. As a first step we divide the machines into two equal size clusters. Machines in the first cluster are all assigned the same task; each of them transmits the wavelet coefficients up to resolution level $j_{B,n}$, depending on the communication budget. The machines in the second cluster are then responsible for transmitting the remaining wavelet coefficients (up to some large enough resolution level $j_{\max} = c \log_2 n$, for some constant $c > 0$). Since the number of these wavelet coefficients (typically) exceeds the communication budget of a single machine, it is not possible to assign the same protocol to each of the machines in the second cluster. We therefore further divide the machines in the second cluster into $j_{\max} - j_{B,n}$ equally sized subclusters. Then, the machines in each subcluster are assigned to transmit the wavelet coefficients at a given resolution level between $j_{B,n} + 1$ and $j_{\max}$. Since the numbers of coefficients at these resolution levels still exceed the communication budget, we further divide the subclusters into equally large subgroups, such that any coefficient will be transmitted by the machines belonging to exactly one subgroup. We proceed by making this strategy precise.

As a first step in our adaptive procedure, we divide the machines into groups. To simplify the notation somewhat, we assume that m is even (otherwise, replace $m/2$ by $\lfloor m/2 \rfloor$ or $\lceil m/2 \rceil$ where appropriate). We first take $m/2$ machines and denote the set of their index numbers by I. Then, the remaining machines are split into $\tilde{\eta} = \tilde{\eta}_n = j_{\max} - j_{B,n}$ equally sized groups (for simplicity each group has $\lfloor (m/2)/\tilde{\eta} \rfloor$ machines and the leftovers are discarded), where

$$j_{B,n} := \lfloor \log_2 \lfloor B / \log_2 n \rfloor \rfloor,$$

$$j_{\max} := \lceil (2 + 2s_{\min})^{-1} \log_2 (nB) \rceil \wedge \lceil (1 + 2s_{\min})^{-1} \log_2 n \rceil.$$

The corresponding sets of indexes are denoted by $I_0, I_1, \ldots, I_{\tilde{\eta}-1}$. Note that $|I_t| \asymp m / \log_2 n$, for $t \in \{0, \ldots, \tilde{\eta} - 1\}$. Then, the machines in the group I transmit the approximations $Y_{jk}^{(i)}$ (with $D = 1/2$ in Algorithm 1) of the local estimators of the wavelet coefficients $\hat{f}_{jk}^{(i)}$, for $0 \leq j \leq j_{B,n} - 1$, $k = 0, \ldots, 2^j - 1$ to the central machine. The ma-

chines in group I_t, $t \in \{0, \dots, \tilde{\eta} - 1\}$, will be responsible for transmitting the coefficients at resolution level $j = j_{B,n} + t$. First, for every $t \in \{0, \dots, \tilde{\eta} - 1\}$, the machines in group I_t are split again into 2^t equal size groups (for simplicity each group has $\lfloor 2^{-t} \lfloor (m/2)/\tilde{\eta} \rfloor \rfloor \geq 1$ machines and the leftovers are discarded again), denoted by $I_{t,1}, I_{t,2}, \dots, I_{t,2^t}$. A machine i in one of the groups $I_{t,\ell}$ for $\ell \in \{1, \dots, 2^t\}$ transmits the approximations $Y_{jk}^{(i)}$ (again with $D = 1/2$ in Algorithm 1) of the local estimators of the wavelet coefficients $\hat{f}_{jk}^{(i)}$, for $j = j_{B,n} + t$ and $(\ell - 1)2^{j_{B,n}} \leq k < \ell 2^{j_{B,n}}$ to the central machine.

In the central machine we first average the transmitted approximations of the corresponding coefficients. We define

$$\hat{f}_{jk} = \begin{cases} |I|^{-1} \sum_{i \in I} Y_{jk}^{(i)} & \text{if } j < j_{B,n}, k = 0, \dots, 2^j - 1, \\ |I_{t,\ell}|^{-1} \sum_{i \in I_{t,\ell}} Y_{jk}^{(i)} & \text{if } j_{B,n} \leq j \leq j_{B,n} + \tilde{\eta}, k = 0, \dots, 2^j - 1. \end{cases} \tag{2.4}$$

Using these coefficients we can construct for every j the preliminary estimator

$$\tilde{f}(j) = \sum_{l \leq j-1} \sum_{k=0}^{2^l - 1} \hat{f}_{lk} \psi_{lk}. \tag{2.5}$$

This gives us a sequence of estimators from which we select the appropriate one using a modified version of Lepski's method. We consider $\mathcal{J} = \{0, \dots, j_{\max}\}$ and define $\hat{j}$ as

$$\hat{j} = \min\{ j \in \mathcal{J} : \| \tilde{f}(j) - \tilde{f}(l) \|_2^2 \leq \tau 2^l / n_l, \forall l > j, l \in \mathcal{J} \}, \tag{2.6}$$

for some sufficiently large parameter $\tau > 1$ (defined later) and $n_j = |I_{j - j_{B,n},1}| n/m \asymp \frac{nB}{2^j (\log_2 n)^2}$, for $j \geq j_{B,n}$ and $n_j = |I| n/m \asymp n$ for $j < j_{B,n}$. Then, we construct our final estimator $\hat{f}$ simply by taking $\hat{f} = \tilde{f}(\hat{j})$.

We summarize the above procedure (without discarding servers for achieving exactly equal size subgroups) in Algorithm 5, below.

THEOREM 2.13. *For every $s, L > 0$ the distributed method $\hat{f}$ described above belongs to $\mathcal{F}_{\text{dist}}(B, \dots, B; B_{2,\infty}^s(L))$, and for all $s \in (s_{\min}, s_{\max})$*

$$\sup_{f_0 \in B_{2,\infty}^s(L)} \mathbb{E}_{f_0,T} \| \hat{f} - f_0 \|_2 \lesssim \begin{cases} L^{1/(1+2s)} n^{-s/(1+2s)} \\ \quad \text{if } B \geq 4(L^2 n)^{1/(1+2s)} \log_2 n, \\ M_n L^{\frac{1}{1+s}} \left(\dfrac{n^{1/(1+2s)}}{B \log_2 n} \right)^{\frac{s}{2+2s}} n^{-\frac{s}{1+2s}} \\ \quad \text{if } B < 4(L^2 n)^{1/(1+2s)} \log_2 n, \end{cases}$$

with $M_n = (\log_2 n)^{s/(1+2s)}$.

PROOF. See Section 3.6. □

REMARK 2.14. Compared to the lower bound in Corollary 2.2, one can observe that in case $B \geq n^{1/(1+2s)} \log_2 n$ the upper bound is sharp. For $B < 4(L^2 n)^{1/(1+2s)} \log_2 n$, we might get an extra slowly varying term of order at most $O((\log n)^{s/(1+2s)})$. Also, note that our method is sharp in the radius of the Besov ball L.

Algorithm 5 Adaptive L_2-method

In the local machines:
for $i = 1$ to $m/2$ **do**
  for $j = 0$ to $j_{B,n} - 1$ **do**
    for $k = 0$ to $2^j - 1$ **do**
      $Y_{jk}^{(i)} := \text{TransApprox}(\hat{f}_{jk}^{(i)})$.
for $t = 0$ to $\tilde{\eta} - 1$ **do**
  Let $j := j_{B,n} + t$.
  for $\ell = 1$ to 2^t **do**
    for $i = m/2 + t\lfloor \frac{m/2}{\tilde{\eta}} \rfloor + (\ell - 1)\lfloor 2^{-t} \lfloor \frac{m/2}{\tilde{\eta}} \rfloor \rfloor + 1$ to $m/2 + t\lfloor \frac{m/2}{\tilde{\eta}} \rfloor + \ell \lfloor 2^{-t} \lfloor \frac{m/2}{\tilde{\eta}} \rfloor \rfloor$ **do**
      for $k = (\ell - 1)2^{j_{B,n}}$ to $\ell 2^{j_{B,n}} - 1$ **do**
        $Y_{jk}^{(i)} := \text{TransApprox}(\hat{f}_{jk}^{(i)})$.
In the central machine:
(1) Averaging the local observations:
for $j = 0$ to $j_{B,n} - 1$ **do**
  for $k = 0$ to $2^j - 1$ **do**
    $\hat{f}_{jk} := \text{mean}\{Y_{jk}^{(i)} : i \le m/2\}$.
for $t = 0$ to $\tilde{\eta} - 1$ **do**
  Let $j := j_{B,n} + t$.
  for $\ell = 1$ to 2^t **do**
    for $k = (\ell - 1)2^{j_{B,n}}$ to $\ell 2^{j_{B,n}} - 1$ **do**
      $\hat{f}_{jk} := \text{mean}\{Y_{jk}^{(i)} : m/2 + t\lfloor \frac{m/2}{\tilde{\eta}} \rfloor + (\ell - 1)\lfloor 2^{-t} \lfloor \frac{m/2}{\tilde{\eta}} \rfloor \rfloor < i \le m/2 + t\lfloor \frac{m/2}{\tilde{\eta}} \rfloor + \ell \lfloor 2^{-t} \lfloor \frac{m/2}{\tilde{\eta}} \rfloor \rfloor\}$.
(2) Lepski's method:
for $j = 0$ to $j_{\max}$ **do**
  $\tilde{f}(j) := \sum_{l \le j-1} \sum_{k=0}^{2^j - 1} \hat{f}_{jk} \psi_{jk}$.
Let $\hat{j} := j_{\max}$, *stop* := *FALSE*.
while *stop* == *FALSE* and $\hat{j} \ge 0$ **do**
  Let $l := \hat{j} + 1$.
  while *stop* == *FALSE* and $l \le j_{\max}$ **do**
    if $\|\tilde{f}(j) - \tilde{f}(l)\|_2^2 \le \tau 2^l / n_l$ **then**
      $l := l + 1$.
    else *stop* := *TRUE*.
  if *stop* == *FALSE* **then**
    $\hat{j} := \hat{j} - 1$.
Construct: $\hat{f} = \tilde{f}(\hat{j})$.

REMARK 2.15. The computational complexity of the adaptive algorithm in each local machine is $O(Bn/(m \log_2 n))$, since $B/\log_2 n$ empirical wavelet coefficients $\hat{f}_{jk}^{(i)}$ are computed and each of them requires $O(n/m)$ operations. In the central machine the computational complexity of the estimators $\hat{f}_{jk}$ for $j < j_{B,n}$ is $O(m)$, while for $j_{B,n} \le j \le j_{\max}$ is $O(m2^{j_{B,n}-j}/\log_2 n)$, hence the total computational complexity of the estimators $\hat{f}_{jk}$, $j \le j_{\max}$, $0 \le k \le 2^{j-1}$ is $O(mB/\log_2 n)$. Then, to compute $\|\tilde{f}(j) - \tilde{f}(l)\|_2^2$, $j < l < j_{\max}$, requires $O(2^l) = O(n^{1/(1+2s_{\min})})$ operations; hence, a conservative upper bound for the com-

putational complexity of $\hat{j}$ is $O(n^{1/(1+2s_{\min})}\log_2^2 n)$, but this could be further reduced by saving the values $\|\tilde{f}(j)-\tilde{f}(j+1)\|_2^2$ and reusing them multiple times.

A slight modification of the above algorithm also leads to a (up to a logarithmic factor) minimax adaptive estimation rate in the L_∞-norm. We construct the truncation estimator $\tilde{f}(j)$ as in Algorithm 5; see (2.5). The only difference to the L_2-case is that we introduce an extra l term in the definition of $\hat{j}$, that is,

$$\hat{j}=\min\{j\in\mathcal{J}:\|\tilde{f}(j)-\tilde{f}(l)\|_\infty\le\tau\sqrt{l2^l/n_l},\forall l>j,l\in\mathcal{J}\}.$$

Finally, we define $\hat{f}=\tilde{f}(\hat{j})$ and show below that it attains the nearly optimal minimax rate adaptively.

THEOREM 2.16. *For every $L,s>0$, the distributed method $\hat{f}$ described above belongs to $\mathcal{F}_{\mathrm{dist}}(B,\dots,B;B^s_{\infty,\infty}(L))$. Furthermore, for all $s\in(s_{\min},s_{\max})$,*

$$\sup_{f_0\in B^s_{\infty,\infty}(L)}\mathbb{E}_{f_0,T}\|\hat{f}-f_0\|_\infty\lesssim\begin{cases}L^{1/(1+2s)}(n/\log_2 n)^{-s/(1+2s)}\\ \quad \text{if } B\ge\overline{B},\\ M_nL^{1/(1+s)}\Big(\dfrac{n^{1/(1+2s)}}{B(\log_2 n)^{\frac{3+4s}{1+2s}}}\Big)^{\frac{s}{2+2s}}\Big(\dfrac{n}{\log_2 n}\Big)^{-\frac{s}{1+2s}}\\ \quad \text{if } B<\overline{B},\end{cases}$$

with $\overline{B}=4(L^2n(\log_2 n)^{2s})^{\frac{1}{1+2s}}$ and $M_n=(\log_2 n)^{\frac{1+2s}{1+s}}$.

PROOF. See Section 4.3. □

3. Proofs for the L_2-norm.

3.1. *Proof of Theorem 2.1.* Note that without loss of generality we can multiply δ_n with an arbitrary constant. In the proof we define δ_n as the solution to

$$\delta_n=C_1^{-1}L^{-2}\min\Big\{\frac{m}{n\log_2 n},\frac{m}{n\sum_{i=1}^m[(\delta_n^{\frac{1}{1+2s}}\log_2(n)B^{(i)})\wedge 1]}\Big\},\tag{3.1}$$

for some sufficiently large $C_1\ge 1$ to be specified later. We prove the desired lower bound for the minimax risk using a modified version of Fano's inequality, given ahead. As a first step we construct a finite subset $\mathcal{F}_0\subset B^s_{2,\infty}(L)$. We use the wavelet notation outlined in Section B of the Supplementary Material and consider Daubechies wavelets with at least s vanishing moments. Define $j_n=\lfloor(\log_2\delta_n^{-1})/(1+2s)\rfloor$. Next, we divide the interval $[0,1]$ into a partition of $2^{j_n}/C_2$ disjoint intervals $I_1,\dots,I_{2^{j_n}/C_2}$, for some large enough $C_2>0$ (without loss of generality we assume that $2^{j_n}/C_2\in\mathbb{N}$), such that each interval I_k contains the full support of a wavelet basis function $\psi_{j_n,\ell}$, $\ell\in\{0,\dots,2^{j_n}-1\}$ (for Daubechies wavelets with s vanishing moments this is possible for $C_2\ge 2s+2$). Slightly abusing our notation, let us denote a basis function corresponding to the kth interval I_k by $\psi_{j_n,k}$ and by $K_{j_n}=\{1,2,\dots,2^{j_n}/C_2\}$ the index set of the intervals (and basis functions). Note that the basis functions $\psi_{j_n,k}$, $k\in K_{j_n}$, have disjoint supports.

For $\beta\in\{-1,1\}^{|K_{j_n}|}$, let $f_\beta\in L_2[0,1]$ be the function with wavelet coefficients

$$f_{\beta,jk}=\begin{cases}L\beta_k\delta_n^{1/2} & \text{if } j=j_n,k\in K_{j_n},\\ 0, & \text{else,}\end{cases}\tag{3.2}$$

and take $C_1 = 2^{17} C_2 \|\psi\|_\infty^2$. Now, define $\mathcal{F}_0 = \{f_\beta : \beta \in \{-1,1\}^{|K_{j_n}|}\}$. Note that $\mathcal{F}_0 \subset B^s_{2,\infty}(L)$, since

$$\|f_\beta\|^2_{B^s_{2,\infty}} = \sup_j 2^{2sj} \sum_{k=0}^{2^j-1} f^2_{\beta,jk} \le L^2 2^{2sj_n} |K_{j_n}| \delta_n \le L^2.$$

For this set of functions $\mathcal{F}_0$, the maximum and minimum number of elements in balls of radius $t > 0$, given by

$$N_t^{\max} = \max_{f_\beta \in \mathcal{F}_0} \big|\{f_{\beta'} \in \mathcal{F}_0 : \|f_\beta - f_{\beta'}\|_2 \le t\}\big|,$$

$$N_t^{\min} = \min_{f_\beta \in \mathcal{F}_0} \big|\{f_{\beta'} \in \mathcal{F}_0 : \|f_\beta - f_{\beta'}\|_2 \le t\}\big|,$$

satisfy $N_t^{\max} = N_t^{\min} = \sum_{i=0}^{\tilde{t}} \binom{|K_{j_n}|}{i} < |\mathcal{F}_0|/2$ for $\tilde{t} = t^2/(4\delta_n L^2) < |K_{j_n}|/2$ (and therefore $N_t^{\max} < |\mathcal{F}_0| - N_t^{\min}$).

Let F be a uniform random variable over the set $\{-1,1\}^{|K_{j_n}|}$, which we identify with the set $\mathcal{F}_0$. Note that the design T is independent of F, while the data X depends on F. In each local machine i we observe the pair of random variables $(T^{(i)}, X^{(i)})$, and we transmit a measurable function $Y^{(i)}$ of this local data to the central machine. This provides us the Markov chains $F \to (T^{(i)}, X^{(i)}) \to Y^{(i)}$, $i = 1, \dots, m$ or by jointly writing them in the form

$$F \to (T, X) \to Y. \tag{3.3}$$

In this setting the following general theorem applies. It is a slight extension of Corollary 1 of [5]; see also Theorem A.6 with the corresponding proof in the Supplementary Material.

THEOREM 3.1. *If the semimetric space $(\mathcal{F}, d)$ contains a finite set $\mathcal{F}_0$ and $|\mathcal{F}_0| - N_t^{\min} > N_t^{\max}$, then for all $p, t > 0$,*

$$\inf_{\hat{f} \in \mathcal{E}(Y)} \sup_{f \in \mathcal{F}} \mathbb{E}_f d^p(\hat{f}, f) \ge t^p \left(1 - \frac{I(F;Y) + \log 2}{\log(|\mathcal{F}_0|/N_t^{\max})}\right),$$

where $\mathcal{E}(Y)$ denotes the set of measurable functions of Y, $I(F;Y)$ is the mutual information between the uniform random variable F (on $\mathcal{F}_0$) and Y in the Markov chain $F \to X \to Y$, and $\mathbb{E}_f$ is the expectation with respect to the distribution of Y given $F = f$.

We apply this theorem with $p = 2$, $t^2 = 2L^2\delta_n |K_{j_n}|/3$, and $d(f,g) = \|f - g\|_2$ to obtain

$$\begin{aligned} &\inf_{\hat{f} \in \mathcal{F}_{\mathrm{dist}}(B^{(1)}, \dots, B^{(m)}; B^s_{2,\infty}(L))} \sup_{f_0 \in B^s_{2,\infty}(L)} \mathbb{E}_{f_0,T} \|\hat{f} - f_0\|_2^2 \\ &\gtrsim L^2 \delta_n |K_{j_n}| \left(1 - \frac{I(F;Y) + \log 2}{\log(|\mathcal{F}_0|/N_t^{\max})}\right), \end{aligned} \tag{3.4}$$

where $I(F;Y)$ is the mutual information between the random variables F and Y.

To lower bound the right-hand side, first note that $N_t^{\max} = \sum_{i=0}^{\tilde{t}} \binom{|K_{j_n}|}{i} < 2\binom{|K_{j_n}|}{\tilde{t}} \le 2(e|K_{j_n}|/\tilde{t})^{\tilde{t}}$ and, therefore, for $\tilde{t} = |K_{j_n}|/6$ (i.e., $t^2 = 2L^2\delta_n |K_{j_n}|/3$),

$$\log(|\mathcal{F}_0|/N_t^{\max}) \ge |K_{j_n}| \log\big(2(6e)^{-1/6} 2^{-1/|K_{j_n}|}\big) \ge |K_{j_n}|/6.$$

Hence, to derive the statement of the theorem from (3.4) it is sufficient to show that

$$I(F;Y) \le |K_{j_n}|/8 + O(1). \tag{3.5}$$

The proof of the next lemma is deferred to Section 5.1.

LEMMA 3.2. *For the Markov chain $F \to (T, X) \to Y$ introduced in (3.3) we have for $m = O(n^{\frac{2s}{1+2s}} / \log_2^2 n)$ that*

$$I(F;Y) \le \frac{4L^2 C_2 \|\psi\|_\infty^2 \delta_n |K_{j_n}| n}{m} \sum_{i=1}^{m} \big((2^{12} \log_2(n) |K_{j_n}|^{-1} B^{(i)}) \wedge 1 \big) + O(1). \tag{3.6}$$

Since in view of the definition of δ_n, we have that

$$\delta_n \le \frac{2^{12} C_1^{-1} L^{-2} m}{n \sum_{i=1}^{m} [(2^{12} \log_2(n) \delta_n^{\frac{1}{1+2s}} B^{(i)}) \wedge 1]};$$

the right-hand side of (3.6) is further bounded by $2^{-3}|K_{j_n}| + O(1)$, finishing the proof of assertion (3.5) and concluding the proof of the theorem.

3.2. *Proof of Theorem 2.4.* First, note that by using Cauchy–Schwarz inequality we get that

$$\begin{aligned} \mathbb{E}_{f_0,T} \big(\log_2 |\hat{f}_{jk}^{(i)}| \vee 1 \big) &\le 1 + \mathbb{E}_{f_0,T} |\hat{f}_{jk}^{(i)}| = 1 + \mathbb{E}_{f_0,T} |X_1^{(i)} \psi_{jk}(T_1^{(i)})| \\ &\le 1 + \|f_0\|_2 \|\psi_{jk}\|_2 + \|\psi_{jk}\|_2 \mathbb{E}_{f_0} |\varepsilon_1^{(i)}| = O(1). \end{aligned}$$

Hence, in view of Lemma 2.3 (with $D = 1/2$) the approximation satisfies

$$0 \le |\hat{f}_{jk}^{(i)} - Y_{jk}^{(i)}| \le 1/\sqrt{n} \quad \text{and} \quad \mathbb{E}_{f_0,T} \big[l(Y_{jk}^{(i)}) \big] \le \big(1/2 + o(1) \big) \log_2 n.$$

Therefore, we need at most $(1/2 + o(1))B$ bits in expected value to transmit $\{Y_{jk}^{(i)} : 2^j + k \le (L^2 n)^{1/(1+2s)} \wedge \lfloor B / \log_2 n \rfloor\}$; hence, $\hat{f}_n \in \mathcal{F}_{\text{dist}}(B, \dots, B; B_{2,\infty}^s(L))$.

Next, for convenience we introduce the notation for the approximation error $W_{jk}^{(i)} = Y_{jk}^{(i)} - \hat{f}_{jk}^{(i)}$, satisfying $|W_{jk}^{(i)}| \le n^{-1/2}$. The estimator $\hat{f}$ is given by its wavelet coefficients $\hat{f}_{jk}$, $j \in \mathbb{N}$, $k \in \{0, 1, \dots, 2^j - 1\}$. For $2^j + k > (L^2 n)^{1/(1+2s)} \wedge \lfloor B / \log_2 n \rfloor$, we have $\hat{f}_{jk} = 0$, while for $2^j + k \le (L^2 n)^{1/(1+2s)} \wedge \lfloor B / \log_2 n \rfloor$,

$$\hat{f}_{jk} = \frac{1}{m} \sum_{i=1}^{m} Y_{jk}^{(i)} = \frac{1}{m} \sum_{i=1}^{m} (\hat{f}_{jk}^{(i)} + W_{jk}^{(i)}) = f_{0,jk} + Z_{jk} + W_{jk},$$

where $Z_{jk} = m^{-1} \sum_{i=1}^{m} (\hat{f}_{jk}^{(i)} - \mathbb{E}_{f_0,T} \hat{f}_{jk}^{(i)})$ and $|W_{jk}| = |m^{-1} \sum_{i=1}^{m} W_{jk}^{(i)}| \le n^{-1/2}$. Note that in view of assumption $\|f_0\|_\infty \le M$

$$\begin{aligned} \mathbb{E}_{f_0,T} Z_{jk}^2 &\le 2\mathbb{E}_{f_0,T} \left(\frac{1}{n} \sum_{i=1}^{m} \sum_{\ell=1}^{n/m} f_0(T_\ell^{(i)}) \psi_{jk}(T_\ell^{(i)}) - \mathbb{E}_{f_0,T} f_0(T_\ell^{(i)}) \psi_{jk}(T_\ell^{(i)}) \right)^2 \\ &\quad + 2\mathbb{E}_{f_0,T} \left(\frac{1}{n} \sum_{i=1}^{m} \sum_{\ell=1}^{n/m} \varepsilon_\ell^{(i)} \psi_{jk}(T_\ell^{(i)}) \right)^2 \\ &\le 2n^{-1} \mathbb{E}_T \big(f_0(T_1^{(1)}) \psi_{jk}(T_1^{(1)}) - \mathbb{E}_T f_0(T_1^{(1)}) \psi_{jk}(T_1^{(1)}) \big)^2 \\ &\quad + 2n^{-1} \mathbb{E}_{f_0} (\varepsilon_1^{(1)})^2 \mathbb{E}_T \psi_{jk}^2(T_1^{(1)}) \\ &\le 2n^{-1} \int_0^1 f_0^2(t) \psi_{jk}^2(t)\, dt + 2n^{-1} \le 2(M^2 + 1)/n. \end{aligned}$$

For convenience we also introduce the notation $j_n = \lfloor \log_2((L^{\frac{2}{1+2s}} n^{\frac{1}{1+2s}}) \wedge \lfloor B/\log_2 n \rfloor) \rfloor$. Then, by combining the above inequalities we get that the risk is bounded from above by

$$
\begin{aligned}
\mathbb{E}_{f_0,T} \|\hat{f} - f_0\|_2^2 &\le \sum_{j \ge j_n} \sum_{k=0}^{2^j-1} f_{0,jk}^2 + 2 \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} \mathbb{E}_{f_0,T}(Z_{jk}^2 + W_{jk}^2) \\
&\lesssim \sum_{j \ge j_n} 2^{-2js} \sup_{j \ge j_n} 2^{2js} \sum_{k=0}^{2^j-1} f_{0,jk}^2 + \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} n^{-1} \\
&\lesssim L^2 2^{-2j_n s} + 2^{j_n}/n \\
&\lesssim \big(L^{\frac{2}{1+2s}} n^{-2s/(1+2s)}\big) \vee \big(L^2 (B/\log_2 n)^{-2s}\big).
\end{aligned} \tag{3.7}
$$

3.3. *Proof of Theorem 2.5.* Similarly to the proof of Theorem 2.4 we get that $\mathbb{E}_{f_0,T}[l(Y_{jk}^{(i)})] \le (1/2 + o(1)) \log_2 n$ and since each machine transmits at most $\lfloor B/\log_2 n \rfloor$ coefficients, the total amount of transmitted bits per machine is bounded from above by B (for large enough n); hence, $\hat{f} \in \mathcal{F}_{\mathrm{dist}}(B, \dots, B; B^s_{2,\infty}(L))$.

Next, let $A_{jk} = \{\lfloor \mu_{jk} m/\eta \rfloor + 1, \dots, \lfloor (\mu_{jk}+1) m/\eta \rfloor\}$ be the collection of machines transmitting the (j,k)th approximated wavelet coefficient $Y_{jk}^{(i)}$, and note that the size of the set satisfies $|A_{jk}| \asymp m/\eta$. Then, our aggregated estimator $\hat{f}$ satisfies for $2^j + k \le \eta \lfloor B/\log_2 n \rfloor$ (i.e., the total number of different coefficients transmitted) that

$$
\hat{f}_{jk} = \frac{1}{|A_{jk}|} \sum_{i \in A_{jk}} Y_{jk}^{(i)} = f_{0,jk} + Z_{jk} + W_{jk},
$$

where $|W_{jk}| = \frac{1}{|A_{jk}|} |\sum_{i \in A_{jk}} W_{jk}^{(i)}| \le n^{-1/2}$ and $Z_{jk} = \frac{1}{|A_{jk}|} \sum_{i \in A_{jk}} (\hat{f}_{jk}^{(i)} - \mathbb{E}_{f_0,T} \hat{f}_{jk}^{(i)})$. Note that similarly to above

$$
\begin{aligned}
\mathbb{E}_{f_0,T} Z_{jk}^2 &\le 2\mathbb{E}_{f_0,T} \bigg(\frac{m}{n|A_{jk}|} \sum_{i \in A_{jk}} \sum_{\ell=1}^{n/m} f_0(T_\ell^{(i)}) \psi_{jk}(T_\ell^{(i)}) \\
&\qquad - \mathbb{E}_{f_0,T} f_0(T_\ell^{(i)}) \psi_{jk}(T_\ell^{(i)}) \bigg)^2 \\
&\qquad + 2\mathbb{E}_{f_0,T} \bigg(\frac{m}{n|A_{jk}|} \sum_{i \in A_{jk}} \sum_{\ell=1}^{n/m} \varepsilon_\ell^{(i)} \psi_{jk}(T_\ell^{(i)}) \bigg)^2 \\
&\le \frac{2m}{n|A_{jk}|} \mathbb{E}_T \big(f_0(T_1^{(1)}) \psi_{jk}(T_1^{(1)}) - \mathbb{E}_T f_0(T_1^{(1)}) \psi_{jk}(T_1^{(1)}) \big)^2 \\
&\qquad + \frac{2m}{n|A_{jk}|} \mathbb{E}_{f_0}(\varepsilon_1^{(1)})^2 \mathbb{E}_T \psi_{jk}^2(T_1^{(1)}) \\
&\le \frac{2(M^2+1)m}{n|A_{jk}|}.
\end{aligned} \tag{3.8}
$$

Let $j_n = \lfloor \log_2(\eta \lfloor B/\log_2 n \rfloor) \rfloor$. Then, similarly to (3.7) the risk of the aggregated estimator is bounded as

$$\begin{aligned}
&\mathbb{E}_{f_0,T} \|\hat{f} - f_0\|_2^2 \\
&\quad \le \sum_{j \ge j_n} \sum_{k=0}^{2^j-1} f_{0,jk}^2 + 2 \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} \mathbb{E}_{f_0,T}(Z_{jk}^2 + W_{jk}^2) \\
&\quad \lesssim \sum_{j \ge j_n} 2^{-2js} \sup_{j \ge j_n} 2^{2js} \sum_{k=0}^{2^j-1} f_{0,jk}^2 + \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} \eta/n \\
&\quad \lesssim L^2 \Big(\frac{B\eta}{\log_2 n} \Big)^{-2s} + \frac{B\eta^2}{n \log_2 n} \qquad (3.9) \\
&\quad \asymp \big(L^{\frac{2}{1+s}} (nB/\log_2 n)^{-\frac{s}{1+s}} \big) \vee \Big(L^2 \Big(\frac{Bm}{\log_2 n} \Big)^{-2s} \Big) \\
&\quad = \Big(L^{\frac{2}{1+s}} (\log_2 n)^{\frac{2s}{1+s}} \Big(\frac{n^{1/(1+2s)}}{B \log_2 n} \Big)^{\frac{s}{1+s}} n^{-\frac{2s}{1+2s}} \Big) \vee \Big(L^2 \Big(\frac{Bm}{\log_2 n} \Big)^{-2s} \Big), \qquad (3.10)
\end{aligned}$$

concluding the proof of the theorem.

Finally, we show that Algorithm 2 is in general suboptimal in this case. Consider the function $f_0 \in B_{2,\infty}^s(1)$ with wavelet coefficients $f_{0,jk} = 2^{-j(s+1/2)}$, $j \in \mathbb{N}$, $k = 0, \dots, 2^j - 1$, and take $j_n = \lfloor \log_2 \lfloor B/\log_2 n \rfloor \rfloor$, then

$$\begin{aligned}
\mathbb{E}_{f_0,T} \|\hat{f} - f_0\|_2^2 &\ge \sum_{j \ge j_n} \sum_{k=0}^{2^j-1} f_{0,jk}^2 \ge \sum_{k=0}^{2^j-1} 2^{-j_n(2s+1)} \\
&\gtrsim \Big(\frac{B}{\log_2 n} \Big)^{-2s} = \tilde{M}_n \Big(\frac{n^{1/(1+2s)}}{B \log_2 n} \Big)^{\frac{s}{1+s}} n^{-\frac{2s}{1+2s}},
\end{aligned}$$

where the multiplication factor $\tilde{M}_n = (\frac{n(\log_2 n)^{3+2s}}{B^{1+2s}})^{\frac{s}{1+s}}$ tends to infinity and can be of polynomial order, yielding a highly suboptimal rate.

3.4. *Proof of Lemma 2.3.* One can easily see by construction that

$$0 \le |X - Y| \le n^{-D}. \qquad (3.11)$$

Next, note that the expected number of transmitted bits is bounded from above by

$$\begin{aligned}
\mathbb{E}\big(1 + (1 \vee \log_2 |X|) + D \log_2 n\big) &= 1 + D \log_2(n) + \mathbb{E}(1 \vee \log_2 |X|) \\
&= \big(D + o(1)\big) \log_2 n.
\end{aligned}$$

3.5. *Proof of Remark 2.6.* Let us assume for simplicity that $B \ge 2 \log n$. We propose a simple modification of Algorithms 2 and 3 such that the resulting estimator $\tilde{f}$ satisfies the stronger, almost sure communication constraints and achieves the same convergence rate, as $\hat{f}$. In the data transmission subroutine (i.e., Algorithm 1) we distinguish two cases; if $\log_2 |x| < \log n$, then we follow the protocol of Algorithm 1 (with $D = 1/2$) and transmit $Y_{jk}^{(i)}$, else, we transmit a single 0 digit (to note that the number we want to transmit is larger than n). We also reduce the number of transmitted coefficients per local machines from $B/\log_2 n$ to $B/(2 \log_2 n)$. Then, in the central machine for the coordinate (j,k) we follow the routine

of Algorithms 2 and 3 (i.e., $\tilde{f}_{jk} = \hat{f}_{jk}$), if $\hat{f}_{jk}^{(i)} \le n$, for all $i = 1, \dots, m$, else we simply set $\tilde{f}_{jk} = 0$.

It is straightforward to see that the proposed algorithm satisfies the stronger, almost sure communication constraints. Next, let us denote by E_{jk}, $j = 1, \dots, \log_2 n$, $k = 0, \dots, 2^j - 1$ the event that $\tilde{f}_{jk} \neq 0$ and note that

$$\begin{aligned} P_{\theta_0}(E_{jk}^c) &\le \sum_{i=1}^{m} P_{\theta_0}\big(|\hat{f}_{jk}^{(i)}| \ge n\big) \le m P_{\theta_0}\left(\frac{m}{n}\sum_{\ell=1}^{n/m} X_\ell^{(i)} \psi_{jk}(T_\ell^{(i)}) \ge n\right) \\ &\le m P_{\theta_0}\Big(2^{j/2}\|\psi\|_\infty \max_\ell \big(|Z_\ell^{(i)}| + M\big) \ge n\Big) \\ &\le n P_{\theta_0}\big(|Z_\ell^{(1)}| \ge c_1\sqrt{n}\big) \le n e^{-c_2 n} = o\big(n^{-2}\big), \end{aligned}$$

for some small enough constants $c_1, c_2 > 0$ and large enough n. Then, note that for arbitrary $j_n \le \log_2 n$,

$$\begin{aligned} \mathbb{E}_{f_0,T}\|\tilde{f} - f_0\|_2^2 &= \sum_{j \ge j_n}\sum_{k=0}^{2^j-1} f_{0,jk}^2 + \sum_{j=0}^{j_n}\sum_{k=0}^{2^j-1} \mathbb{E}_{f_0,T}(f_{0,jk} - \tilde{f}_{jk})^2 1_{E_{jk}} \\ &\quad + \sum_{j=0}^{j_n}\sum_{k=0}^{2^j-1} \mathbb{E}_{f_0,T}(f_{0,jk} - \tilde{f}_{jk})^2 1_{E_{jk}^c} \\ &\le \sum_{j \ge j_n} 2^{-2js} \sup_{l \ge j_n} 2^{2ls} \sum_{k=0}^{2^l-1} f_{0,lk}^2 + \sum_{j=0}^{j_n}\sum_{k=0}^{2^j-1} \mathbb{E}_{f_0,T}(f_{0,jk} - \hat{f}_{jk})^2 \\ &\quad + \sum_{j=0}^{j_n}\sum_{k=0}^{2^j-1} f_{0,jk}^2 P_{\theta_0}(E_{jk}^c) \\ &\lesssim L^2 2^{-2j_n s} + \sum_{j=0}^{j_n}\sum_{k=0}^{2^j-1} \mathbb{E}_{f_0,T}\big(Z_{jk}^2 + W_{jk}^2\big) + o\big(n^{-1}\big). \end{aligned}$$

We conclude the proof by noting that the first two terms on the right-hand side have the required upper bounds (see the proofs of Theorems 2.4 and 2.5), while the third term is of smaller order than the previous ones.

3.6. *Proof of Theorem 2.13.* First, recall that for every $s, L > 0$ and $f_0 \in B_{2,\infty}^s(L)$ we have $f_{0,jk}^2 \le L^2$, $j \ge 0$, $k \in \{0, 1, \dots, 2^j - 1\}$. Therefore, in view of Lemma 2.3 (with $D = 1/2$) we have $\mathbb{E}_{f_0,T}[l(Y_{jk}^{(i)})] \le (1/2 + o(1)) \log_2 n$. Since the machines in group I and the machines in $I_{t,\ell}$, $t \in \{0, \dots, \tilde{\eta} - 1\}$, $\ell \in \{1, \dots, 2^t\}$ transmit at most $\lfloor B/\log_2 n \rfloor$ coefficients, we have that in expected value at most

$$\lfloor B/\log_2 n \rfloor \big(1/2 + o(1)\big) \log_2 n \le B$$

bits are transmitted per machine (for n large enough). Therefore, the estimator indeed belongs to $\mathcal{F}_{\text{dist}}(B, \dots, B; B_{2,\infty}^s(L))$.

Next, we show that the estimator $\hat{f}$ achieves the minimax rate. First, let us introduce the notation $|W_{jk}^{(i)}| = |Y_{jk}^{(i)} - \hat{f}_{jk}^{(i)}| \le n^{-1/2}$. Then, note that for $j \le j_{\max}$ and $k \in \{0, 1, \dots, 2^j - 1\}$

the aggregated quantities $\hat{f}_{jk}$ defined in (2.4) are equal to

$$\hat{f}_{jk} = \frac{1}{|A_{jk}|}\sum_{i\in A_{jk}} Y_{jk}^{(i)} = f_{0,jk} + Z_{jk} + W_{jk}, \tag{3.12}$$

where

$$A_{jk} = \begin{cases} I & \text{if } j < j_{B,n}, k = 0, 1, \dots, 2^j - 1, \\ I_{j-j_{B,n},\ell} & \text{if } j \geq j_{B,n}, (\ell-1)2^{j_{B,n}} \leq k < \ell 2^{j_{B,n}}, \end{cases}$$

where $|W_{jk}| = n_j^{-1}|\sum_{i\in A_{jk}} W_{jk}^{(i)}| \leq n^{-1/2}$, $Z_{jk} = |A_{jk}|^{-1}\sum_{i\in A_{jk}}(\hat{f}_{jk}^{(i)} - \mathbb{E}_{f_0,T}\hat{f}_{jk}^{(i)})$, and recall that $n_j = n|A_{jk}|/m$ for every $j \leq j_{\max}$, $k \in \{0, .., 2^j - 1\}$. Recall also that $n_j \asymp nB/(2^j(\log_2 n)^2)$ for $j \geq j_{B,n}$ and $n_j \asymp n$ for $j < j_{B,n}$.

Note that the squared bias satisfies

$$\|\mathbb{E}_{f_0,T}\tilde{f}(j) - f_0\|_2^2 \lesssim \|K(f_0, j) - f_0\|_2^2 + 2^j/n \lesssim 2^{-2js}\|f_0\|_{B_{2,\infty}^s}^2 + 2^j/n,$$

where $K(f_0, j) = \sum_{l=0}^{j-1}\sum_{k=0}^{2^l-1} f_{0,lk}\psi_{lk}$. Furthermore, also note that for $\ell \leq j$ we have $n_\ell \geq n_j$ and, hence, in view of (3.8)

$$\begin{aligned}\mathbb{E}_{f_0,T}\|\tilde{f}(j) - \mathbb{E}_{f_0,T}\tilde{f}(j)\|_2^2 &\lesssim \sum_{\ell\leq j-1}\sum_{k=0}^{2^\ell-1}\left(\mathbb{E}_{f_0,T}Z_{\ell k}^2 + \mathbb{E}_{f_0,T}W_{\ell k}^2\right) \\ &\lesssim \sum_{\ell\leq j-1}\sum_{k=0}^{2^\ell-1} n_\ell^{-1} \leq 2^j/n_j.\end{aligned}$$

Let us introduce the notation $B(j, f_0) = 2^{-2js}\|f_0\|_{B_{2,\infty}^s}^2$ and define the optimal choice of the parameter j (the optimal resolution level) as

$$j^* = \min\{j \in \mathcal{J} : B(j, f_0) \leq 2^j/n_j\},$$

balancing out the squared bias and variance terms. Note that since the right-hand side is monotone increasing and the left-hand side is monotone decreasing in j, we have that

$$\begin{aligned} &B(j, f_0) \leq 2^j/n_j \quad \text{for } j \geq j^* \quad \text{and} \\ &B(j, f_0) > 2^j/n_j \quad \text{for } j < j^*. \end{aligned} \tag{3.13}$$

Therefore,

$$2^{j^*-1}/n_{j^*-1} < B(j^* - 1, f_0) = 2^{2s}B(j^*, f_0) \leq 2^{2s}2^{j^*}/n_{j^*}.$$

Let us distinguish three cases according to the value of j^*. If $j^* < j_{B,n}$, then $n_{j^*-1} = n_{j^*} \asymp n$ and therefore $2^{j^*} \asymp (\|f_0\|_{B_{2,\infty}^s}^2 n)^{1/(1+2s)}$ (using the definition $B(j^*, f_0) = 2^{-2j^*s}\|f_0\|_{B_{2,\infty}^s}^2$). Note that the inequality $j^* < j_{B,n}$ is implied by $B(j_{B,n} - 1, f_0) \leq 2^{j_{B,n}-1}/n_{j_{B,n}-1}$ which in turn holds if $2^{j_{B,n}-1} \geq (\|f_0\|_{B_{2,\infty}^s}^2 n)^{1/(1+2s)}$. Therefore, we can conclude that $B \geq 4(\|f_0\|_{B_{2,\infty}^s}^2 n)^{1/(1+2s)}\log_2 n$ implies the inequality $j^* < j_{B,n}$ (by recalling that $2^{j_{B,n}} \geq B/(2\log_2 n)$). If $j^* = j_{B,n}$ (i.e., $2^{j^*} \asymp B/\log_2 n$), then $n_{j^*} \asymp n/\log_2 n$, $n_{j^*-1} \asymp n$ which together with (3.13) (for $j = j^* - 1$) implies $(\|f_0\|_{B_{2,\infty}^s}^2 n/\log_2 n)^{1/(1+2s)} \lesssim 2^{j^*} \lesssim (\|f_0\|_{B_{2,\infty}^s}^2 n)^{1/(1+2s)}$. Finally, if $j^* > j_{B,n}$, then $n_{j^*-1} \asymp n_{j^*} \asymp nB/(2^{j^*}\log_2^2 n)$ which

together with (3.13) (for $j = j^* - 1$) implies that $2^{j^*} \asymp (\|f_0\|^2_{B^s_{2,\infty}} nB/\log_2^2 n)^{1/(2+2s)}$ and $n_{j^*} \gtrsim \|f_0\|^{-\frac{1}{1+s}}_{B^s_{2,\infty}} (nB/\log_2^2 n)^{\frac{1+2s}{2+2s}}$. We summarize these findings in the following displays:

$$2^{j^*} \asymp \begin{cases} \big(\|f_0\|^2_{B^s_{2,\infty}} n\big)^{1/(1+2s)} & \text{if } B \geq \overline{B}, \\ B/\log_2 n & \text{if } \underline{B} \leq B < \overline{B}, \\ \big(\|f_0\|^2_{B^s_{2,\infty}} nB/\log_2^2 n\big)^{1/(2+2s)} & \text{if } B < \underline{B}, \end{cases} \tag{3.14}$$

and

$$n_{j^*} \gtrsim \begin{cases} n & \text{if } B \geq \overline{B}, \\ n/\log_2 n, & \text{if } \underline{B} \leq B < \overline{B}, \\ \|f_0\|^{-1/(1+s)}_{B^s_{2,\infty}} \big(nB/\log_2^2 n\big)^{(1+2s)/(2+2s)} & \text{if } B < \underline{B}, \end{cases} \tag{3.15}$$

where $\overline{B} = 4(\|f_0\|^2_{B^s_{2,\infty}} n)^{1/(1+2s)} \log_2 n$ and $\underline{B} = (\|f_0\|^2_{B^s_{2,\infty}} n)^{\frac{1}{1+2s}} (\log_2 n)^{\frac{2s}{1+2s}}$. Note that in all cases $j^* \leq j_{\max}$ holds.

Let us split the risk into two parts,

$$\begin{aligned} &\mathbb{E}_{f_0,T} \|f_0 - \hat{f}\|_2 \\ &\quad = \mathbb{E}_{f_0,T} \|f_0 - \tilde{f}(\hat{j})\|_2 1_{\hat{j} > j^*} + \mathbb{E}_{f_0,T} \|f_0 - \tilde{f}(\hat{j})\|_2 1_{\hat{j} \leq j^*}, \end{aligned} \tag{3.16}$$

and deal with each term on the right-hand side separately. First, note that

$$\begin{aligned} &\mathbb{E}_{f_0,T} \|f_0 - \tilde{f}(\hat{j})\|_2^2 1_{\hat{j} \leq j^*} \\ &\quad \leq 2\mathbb{E}_{f_0,T} \|\tilde{f}(j^*) - \tilde{f}(\hat{j})\|_2^2 1_{\hat{j} \leq j^*} + 2\mathbb{E}_{f_0,T} \|\tilde{f}(j^*) - f_0\|_2^2 \\ &\quad \lesssim \tau 2^{j^*}/n_{j^*} + \|\mathbb{E}_{f_0,T} \tilde{f}(j^*) - f_0\|_2^2 + \mathbb{E}_{f_0,T} \|\tilde{f}(j^*) - \mathbb{E}_{f_0,T} \tilde{f}(j^*)\|_2^2 \\ &\quad \lesssim 2^{j^*}/n_{j^*} + \|f_0\|^2_{B^s_{2,\infty}} 2^{-2j^* s}, \end{aligned}$$

which implies together with (3.14) and (3.15), that

$$\begin{aligned} &\mathbb{E}_{f_0,T} \|f_0 - \hat{f}\|_2^2 1_{\hat{j} \leq j^*} \\ &\quad \lesssim \begin{cases} \|f_0\|^{2/(1+2s)}_{B^s_{2,\infty}} n^{-2s/(1+2s)} & \text{if } B \geq \overline{B}, \\ B/n & \text{if } \underline{B} \leq B \leq \overline{B}, \\ \|f_0\|^{2/(1+s)}_{B^s_{2,\infty}} \big(nB/\log_2^2 n\big)^{-2s/(2+2s)} & \text{if } B \leq \underline{B}. \end{cases} \end{aligned} \tag{3.17}$$

Since $|f_0\|_{B^s_{2,\infty}} \leq L$, the preceding upper bounds are bounded from above by the ones stated in the theorem.

Next, we deal with the first term on the right hand side of (3.16). By Cauchy–Schwarz inequality and Lemma 3.3 we get that

$$\begin{aligned} &\mathbb{E}_{f_0,T} \|f_0 - \hat{f}\|_2 1_{\hat{j} > j^*} \\ &\quad \leq \sum_{j=j^*+1}^{j_{\max}} \mathbb{E}^{1/2}_{f_0,T} \|f_0 - \tilde{f}(j)\|_2^2 \mathbb{P}^{1/2}_{f_0,T}(\hat{j} = j) \\ &\quad \lesssim \sum_{j=j^*+1}^{j_{\max}} \mathbb{P}^{1/2}_{f_0,T}(\hat{j} = j) \lesssim j_{\max} e^{-(cn^\delta \wedge \sqrt{n_r})} + \sum_{k=1}^{\infty} e^{-(c/2)2^{j^*} k} \\ &\quad = o\big(n^{-1}\big) + o\big(2^{-j^* s}\big), \end{aligned}$$

resulting in the required upper bound in view of (3.17), concluding the proof of our statement.

LEMMA 3.3. *Assume that $f_0 \in B^s_{2,\infty}(L)$, for some $s, L > 0$. Then, there exists a universal constants $c, \delta > 0$ such that for every $j > j^*$ we have*

$$\mathbb{P}_{f_0,T}(\hat{j} = j) \lesssim e^{-(c2^j \wedge n^\delta \wedge \sqrt{n_r})}.$$

PROOF. Let us introduce the notation $j^- = j - 1$ and note that for every $j > j^*$ we have $j^- \geq j^*$. Then, by the definition of $\hat{j}$

$$\mathbb{P}_{f_0,T}(\hat{j} = j) \leq \sum_{l=j}^{j_{\max}} \mathbb{P}_{f_0,T}\big(\|\tilde{f}(j^-) - \tilde{f}(l)\|_2^2 > \tau 2^l/n_l\big).$$

Note that the left-hand side term in the probability in view of Parseval's inequality can be given in the form

$$\begin{aligned}\|\tilde{f}(j^-) - \tilde{f}(l)\|_2^2 &= \sum_{r=j^-}^{l-1} \sum_{k=0}^{2^r-1} (f_{0,rk} + Z_{rk} + W_{rk})^2 \\ &\leq 3 \sum_{r=j^-}^{l-1} \sum_{k=0}^{2^r-1} \big(f_{0,rk}^2 + Z_{rk}^2 + W_{rk}^2\big).\end{aligned}$$

We deal with the three terms on the right-hand side separately. Note that the functions $j \mapsto B(j, f_0)$ and $j \mapsto n_j$ are monotone decreasing; hence, by the definition of j^* we get for $l \geq j^- \geq j^*$

$$\sum_{r=j^-}^{l-1} \sum_{k=0}^{2^r-1} f_{0,rk}^2 \leq B\big(j^-, f_0\big) \leq B\big(j^*, f_0\big) \leq 2^{j^*}/n_{j^*} \leq 2^l/n_l.$$

Furthermore, $\sum_{r=j^-}^{l-1} \sum_{k=0}^{2^r-1} W_{rk}^2 \leq 2^l/n \leq 2^l/n_l$.

Let $S(r) = \{\sum_{l=0}^{r} \sum_{k=0}^{2^l-1} b_{lk}\psi_{lk} : \sum_{l=0}^{r} \sum_{k=0}^{2^l-1} b_{lk}^2 = 1\}$ denote the unite sphere in the linear subspace spanned by the basis functions ψ_{lk}, $l \leq r$, $0 \leq k \leq 2^l - 1$. Then, in view of Lemma 5.3 of [3] (see also Lemma C.4 in the Supplementary Material) and the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ we get that

$$\begin{aligned}\sum_{k=0}^{2^r-1} Z_{rk}^2 &= \sum_{k=0}^{2^r-1} \left(\frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \big(Y_\ell^{(i)} \psi_{jk}(T_\ell^{(i)}) - \mathbb{E}_{f_0,T} Y_\ell^{(i)} \psi_{jk}(T_\ell^{(i)})\big)\right)^2 \\ &\leq 2 \sup_{g \in S(r)} \left(\frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \big(f_0(T_\ell^{(i)}) g(T_\ell^{(i)}) - \mathbb{E}_T f_0(T_\ell^{(i)}) g(T_\ell^{(i)})\big)\right)^2 \\ &\quad + 2 \sum_{k=0}^{2^r-1} \left(\frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \big(\varepsilon_\ell^{(i)} \psi_{jk}(T_\ell^{(i)})\big)\right)^2.\end{aligned} \tag{3.18}$$

We deal with the two terms on the right-hand side separately, starting with the first one. Note that for every $g \in S(r)$ the inequality $\|g\|_\infty \leq C 2^{r/2}$ holds, for some universal constant $C > 0$ and

$$\sup_{g \in S(r)} V_T\big(f_0(T_1^{(1)}) g(T_1^{(1)})\big) \leq \|f_0\|_\infty^2.$$

Next, for convenience let us introduce the notation

$$\nu(g) = \frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \bigl(f_0(T_\ell^{(i)}) g(T_\ell^{(i)}) - \mathbb{E}_T f_0(T_\ell^{(i)}) g(T_\ell^{(i)})\bigr).$$

Then, by the definition of $S(r)$ and Cauchy–Schwarz inequality

$$\begin{aligned}\mathbb{E}_T \sup_{g \in S(r)} \bigl|\nu(g)\bigr| &\le \sum_{k=0}^{2^r-1} \mathbb{E}_T\bigl(\nu(\psi_{r,k})^2\bigr) = \sum_{k=0}^{2^r-1} \frac{1}{n_r} V_T\bigl(f_0(T_1^{(1)}) \psi_{rk}(T_1^{(1)})\bigr) \\ &\le \frac{\|f_0\|_\infty^2 2^r}{n_r}.\end{aligned}$$

Therefore, in view of Lemma 5 of [11] (see also Lemma C.1 in the Supplementary Material) there exist constants $c_1, c_2, c_2 > 0$ such that

$$\begin{aligned}&\mathbb{E}_T \sup_{g \in S(r)} \Biggl[\Biggl(\frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \bigl(f_0(T_\ell^{(i)}) g(T_\ell^{(i)}) - \mathbb{E}_T f_0(T_\ell^{(i)}) g(T_\ell^{(i)})\bigr)\Biggr)^2 \\ &\qquad - c_1 2^r / n_r \Biggr]_+ \\ &\le c_2 \frac{1}{n_r} e^{-c_3 2^r} + c_4 \frac{2^r}{n_r^2} e^{-\sqrt{n_r}} \lesssim \frac{1}{n_r} e^{-(c_3 2^r \wedge \sqrt{n_r})}.\end{aligned} \tag{3.19}$$

Therefore, by Markov's inequality we get that

$$\begin{aligned}&\mathbb{P}_T\Biggl(\sup_{g \in S(r)} \Biggl(\frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \bigl(f_0(T_\ell^{(i)}) g(T_\ell^{(i)}) \\ &\qquad - \mathbb{E}_T f_0(T_\ell^{(i)}) g(T_\ell^{(i)})\bigr)\Biggr)^2 \ge \frac{2c_1 2^r}{n_r}\Biggr) \lesssim 2^{-r} e^{-(c_3 2^r \wedge \sqrt{n_r})}.\end{aligned} \tag{3.20}$$

Next, we deal with the second term on the right-hand side of (3.18). Let us introduce the shorthand notation $\tilde{Z}_{rk} = n_r^{-1} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \varepsilon_\ell^{(i)} \psi_{rk}(T_\ell^{(i)})$. Note that $\operatorname{cov}(\tilde{Z}_{rk}, \tilde{Z}_{rk'} | T) = 0$ for $|k - k'| \ge C$, for some large enough constant C, following from the disjoint support of the wavelet basis functions ψ_{rk} and $\psi_{rk'}$, and

$$\tilde{Z}_{rk} | T \sim N\Biggl(0, \frac{1}{n_r^2} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \psi_{rk}(T_\ell^{(i)})^2\Biggr).$$

Furthermore, let us denote by $\mathcal{B}_r$ the event that in each bin $I_{r,l} = [(l-1)2^{-r}, l2^{-r}]$, at most $2n_r/2^r$ observations $T_\ell^{(i)}$, $i \in A_{rk}$, $\ell = 1, \dots, m$, $k = 0, \dots, 2^r - 1$ fall. Since there are $2^{r - j_{B,n}} \le 2^r$ subgroups of machines at resolution level r, we note that in view of Lemma 5.2 we have that $\mathbb{P}_T(\mathcal{B}_r^c) \le 2^{2r+1} e^{-n_r 2^{-r-3}}$. Then, by recalling that for $r \le j_{B,n}$, $n_r \asymp n$, while for $r > j_{B,n}$, $n_r = nB/(2^r \log^2 n)$, we get that $n_r/2^r \gtrsim (nB/\log_2^2 n)^{\frac{2s_{\min}}{2+2s_{\min}}} \wedge n^{\frac{2s_{\min}}{1+2s_{\min}}}$; hence,

$$\mathbb{P}_T(\mathcal{B}_r^c) \lesssim e^{-n^\delta} \quad \text{for any } \delta < 2s_{\min}/(2 + 2s_{\min}), \tag{3.21}$$

and on $\mathcal{B}_r$ the inequality $n_r^{-2} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \psi_{rk}(T_\ell^{(i)})^2 \le C n_r^{-1}$ holds, for some sufficiently large $C > 0$. Let us denote the covariance matrix of the random vector $(\tilde{Z}_{r0}, \dots, \tilde{Z}_{r(2^r-1)}) | T$ by Σ_T. In view of the preceding argument, the in absolute value largest entry of Σ_T is

bounded from above by Cn_r^{-1} on the event $T \in \mathcal{B}_r$ and by noting that Σ_T has band size C, in view of Gershgorin circle Theorem [6] (see also Lemma C.3 in the Supplementary Material) the eigenvalues of Σ_T satisfy that $0 < \lambda_i \le Cn_r^{-1}$, $i = 1, .., 2^r$. Then, by the tail bounds of chi-square distributions (see, for instance, Theorem 4.1.9 of [7] or Lemma C.2 in the Supplementary Material),

$$\begin{aligned}&\mathbb{P}_{f_0}\left(\sum_{k=0}^{2^r-1}\tilde{Z}_{rk}^2 \ge \frac{C_1 2^r}{n_r}\Big|T=t\right)\\&\quad=\mathbb{P}\left(\sum_{i=1}^{2^r}\lambda_i\zeta_i^2 \ge \frac{C_1 2^r}{n_r}\right) \le \mathbb{P}\left(\sum_{i=1}^{2^r}\zeta_i^2 \ge C_2 2^r\right) \lesssim e^{-C_3 2^r},\end{aligned}$$

for some sufficiently large constants $C_1, C_2 > 0$ and small $C_3 > 0$, where $\zeta_i \overset{iid}{\sim} N(0,1)$. Hence, we can conclude that

$$\begin{aligned}&\mathbb{P}_{f_0,T}\left(\sum_{k=0}^{2^r-1}\left(\frac{1}{n_r}\sum_{i\in A_{rk}}\sum_{\ell=1}^{n/m}\varepsilon_\ell^{(i)}\psi_{rk}\big(T_\ell^{(i)}\big)\right)^2 \ge \frac{C_1 2^r}{n_r}\right)\\&\quad\le \int_{t\in\mathcal{B}_r}\mathbb{P}_{f_0}\left(\sum_{k=0}^{2^r-1}\tilde{Z}_{rk}^2 \ge \frac{C_1 2^r}{n_r}\Big|T=t\right)dt + \mathbb{P}_T\big(\mathcal{B}_r^c\big) \lesssim e^{-(C2^r\wedge n^\delta)},\end{aligned}$$

finishing the proof of the lemma. □

4. Proofs for the L_∞-norm.

4.1. *Proof of Theorem 2.8.* First of all, we note that in the nondistributed case, where all the information is available in the central machine, the minimax L_∞-risk is $L^{1/(1+2s)}(n/\log n)^{-s/(1+2s)}$. Since the class of distributed estimators is clearly a subset of the class of all estimators, this will be also a lower bound for the distributed case. The rest of the proof goes similarly to the proof of Theorem 3.1.

We consider the same subset of functions $\mathcal{F}_0$ as in the proof of Theorem 3.1, with functions given by (3.2). Note that each function $f_\beta \in \mathcal{F}_0$ belongs to the set $B^s_{\infty,\infty}(L)$, since

$$\|f_\beta\|_{B^s_{\infty,\infty}} = \sup_j 2^{(s+1/2)j}\sup_{k=0,\dots,2^j-1} f_{\beta,jk} = 2^{(s+1/2)j_n}L\delta_n^{1/2} \le L.$$

Furthermore, if $f_\beta \ne f_{\beta'}$, then there exists a $k \in K_{j_n}$ such that $\beta_k \ne \beta'_k$. Then, due to the disjoint support of the corresponding Daubechies wavelets $\psi_{j_n,k}$, $k \in K_{j_n}$, the L_∞-distance between the two functions is bounded from below by

$$\|f_\beta - f_{\beta'}\|_\infty \ge |f_{\beta,j_nk} - f_{\beta',j_nk}|\cdot\|\psi_{j_n,k}\|_\infty \gtrsim L2^{j_n/2}\delta_n^{1/2} \ge L\delta_n^{\frac{s}{1+2s}}.$$

Now, let F be a uniform random variable on the set $\mathcal{F}_0$. Then, in view of Fano's inequality (see, for instance, Theorem A.5 in the Supplementary Material with $\delta = \delta_n^{s/(1+2s)}$ and $p = 1$) we get that

$$\begin{aligned}&\inf_{\hat f\in\mathcal{F}_{\mathrm{dist}}(B^{(1)},\dots,B^{(m)};B^s_{\infty,\infty}(L))}\sup_{f_0\in B^s_{\infty,\infty}(L)}\mathbb{E}_{f_0,T}\big(\|\hat f - f_0\|_\infty\big)\\&\quad\gtrsim L\delta_n^{\frac{s}{1+2s}}\left(1-\frac{I(F;Y)+\log 2}{\log_2|\mathcal{F}_0|}\right).\end{aligned}$$

We conclude the proof by noting that the term in the bracket on the right-hand side of the preceding display is bounded from below by a constant; see the proof of Theorem 3.1.

4.2. *Proof of Theorem 2.10.* Similarly to the proof of Theorem 2.4, we get that $\mathbb{E}_{f_0,T}[l(Y_{jk}^{(i)})] \le (1/2+o(1))\log_2 n$, hence, we need at most $(1/2+o(1))B$ bits in expected value to transmit the $\lfloor B/\log_2 n\rfloor \wedge (L^2 n/\log_2 n)^{1/(1+2s)}$ approximated coefficients. Therefore, the total amount of transmitted bits per machine is bounded from above by B (for large enough n), hence $\hat{f} \in \mathcal{F}_{\mathrm{dist}}(B,\dots,B; B^s_{\infty,\infty}(L))$.

Similarly to the proof of Theorem 2.5, let $A_{jk} = \{\lfloor \mu_{jk} m/\eta\rfloor + 1, \dots, \lfloor(\mu_{jk}+1)m/\eta\rfloor\}$ be the collection of machines transmitting the (j,k)th approximated wavelet coefficient and note that the size of the set satisfies $|A_{jk}| \asymp m/\eta$. And recall that the aggregated estimator $\hat{f}$ satisfies for $2^j + k \le (\eta\lfloor B/\log_2 n\rfloor) \wedge (L^2 n/\log_2 n)^{1/(1+2s)}$ (i.e., the total number of different coefficients transmitted) that

$$\hat{f}_{jk} = \frac{1}{|A_{jk}|}\sum_{i\in A_{jk}} Y_{jk}^{(i)} = f_{0,jk} + Z_{jk} + W_{jk},$$

where $|W_{jk}| = |A_{jk}|^{-1}|\sum_{i\in A_{jk}} W_{jk}^{(i)}| \le n^{-1/2}$ and $Z_{jk} = |A_{jk}|^{-1}\sum_{i\in A_{jk}}(\hat{f}_{jk}^{(i)} - \mathbb{E}_{f_0,T}\hat{f}_{jk}^{(i)})$. We show below that for all $2^j \le n/\eta$,

$$\mathbb{E}_{f_0,T}\sup_k |Z_{jk}| \lesssim \sqrt{(\log_2 n)\eta/n}. \tag{4.1}$$

Next, note that by triangle inequality

$$\mathbb{E}_{f_0,T}\|f_0 - \hat{f}\|_\infty \le \|f_0 - \mathbb{E}_{f_0,T}\hat{f}\|_\infty + \mathbb{E}_{f_0,T}\|\hat{f} - \mathbb{E}_{f_0,T}\hat{f}\|_\infty.$$

We deal with the two terms on the right-hand side separately. Let us introduce the notation

$$j_n = \lfloor \log_2\big((\eta\lfloor B/\log_2 n\rfloor) \wedge (L^2 n/\log_2 n)^{1/(1+2s)}\big)\rfloor \le \log_2(n/\eta).$$

Then, by triangle inequality and noting that there exists a universal constant $C>0$, such that for each resolution level j the inequality $\|\sum_{k=0}^{2^j-1}|\psi_{jk}|\|_\infty \le C2^{j/2}$ holds,

$$\begin{aligned}
\|f_0 - \mathbb{E}_{f_0,T}\hat{f}\|_\infty &\le \Big\|\sum_{j=j_n}^{\infty}\sum_{k=0}^{2^j-1} f_{0,jk}\psi_{jk}\Big\|_\infty + \Big\|\sum_{j=0}^{j_n}\sum_{k=0}^{2^j-1}\mathbb{E}_{f_0,T}W_{jk}\psi_{jk}\Big\|_\infty \\
&\le \|f_0\|_{B^s_{\infty,\infty}}\sum_{j=j_n}^{\infty} 2^{-j(s+1/2)}\Big\|\sum_{k=0}^{2^j-1}|\psi_{jk}|\Big\|_\infty \\
&\quad + n^{-1/2}\sum_{j=0}^{j_n}\Big\|\sum_{k=0}^{2^j-1}|\psi_{jk}|\Big\|_\infty \\
&\lesssim L\sum_{j=j_n}^{\infty} 2^{-js} + \sqrt{2^{j_n}/n} \lesssim L2^{-j_n s} + \sqrt{2^{j_n}/n}.
\end{aligned} \tag{4.2}$$

Furthermore, in view of (4.1),

$$\begin{aligned}
&\mathbb{E}_{f_0,T}\|\hat{f} - \mathbb{E}_{f_0,T}\hat{f}\|_\infty \\
&\quad\le \sum_{j=0}^{j_n}\mathbb{E}_{f_0,T}\max_k\big(|Z_{jk}| + |W_{jk}|\big)\Big\|\sum_{k=0}^{2^j-1}|\psi_{jk}|\Big\|_\infty \\
&\quad\lesssim \sum_{j=0}^{j_n} 2^{j/2}\big(\sqrt{(\log_2 n)\eta/n} + \sqrt{1/n}\big) \lesssim \sqrt{2^{j_n}\eta(\log_2 n)/n},
\end{aligned} \tag{4.3}$$

providing the upper bound in the statement of the lemma.

It remained to prove assertion (4.1). First, note that

$$Z_{jk}|T \sim N(\mu_{n,m,k,T}, \sigma^2_{n,m,k,T}) \quad \text{with}$$

$$\mu_{n,m,k,T} = \frac{\eta}{n} \sum_{i \in A_{jk}} \sum_{\ell=1}^{n/m} \psi_{jk}(T_\ell^{(i)}) f_0(T_\ell^{(i)}) - f_{0,jk} \lesssim 2^{j/2},$$

$$\sigma^2_{n,m,k,T} = \left(\frac{\eta}{n}\right)^2 \sum_{i \in A_{jk}} \sum_{\ell=1}^{n/m} \psi^2_{jk}(T_\ell^{(i)}) \lesssim 2^j \eta/n.$$

Using standard bounds on the maximum of Gaussian variables (see, for instance, Lemma 3.3.4 of [7]), we have that

$$\mathbb{E}_{f_0|T} \max_k |Z_{jk} - \mathbb{E}_{f_0|T} Z_{jk}| \leq \sqrt{2(j+1)} \max_k \sigma_{n,m,k,T}.$$

Furthermore, note that for $k \geq 2$

$$\mathbb{E}_T(\psi_{jk}(T_\ell^{(i)}) f_0(T_\ell^{(i)}))_+^k \leq \|f_0\|_\infty^k \|\psi_{jk}\|_\infty^{k-2} \mathbb{E}_T \psi_{jk}(T_\ell^{(i)})^2 \lesssim 2^{(k-2)j/2},$$

hence, in view of Bernstein's inequality (with $c = C2^{j/2}$ and $v = Cn/\eta$) (see Proposition 2.9 of [13] or Lemma C.5 in the Supplementary Material), we get that

$$\mathbb{P}_T\left(|\mu_{n,m,k,T}| \geq C\left(\sqrt{\frac{\gamma \eta \log_2 n}{n}} + \frac{2^{j/2}\eta}{n}\right)\right) \lesssim (n/\eta)^{-\gamma},$$

which implies for $2^j \leq n/\eta$ that

$$\mathbb{P}_T\left(\max_k |\mu_{n,m,k,T}| \geq C_\gamma \sqrt{(\log_2 n)\eta/n}\right) \lesssim (n/\eta)^{-\gamma+1}. \tag{4.4}$$

Therefore, one can deduce that

$$\begin{aligned}\mathbb{E}_T\left(\max_k |\mu_{n,m,k,T}|\right) &\leq C_\gamma \sqrt{(\log_2 n)\eta/n} + 2^{j/2}(n/\eta)^{-\gamma+1} \\ &\lesssim \sqrt{(\log_2 n)\eta/n},\end{aligned}$$

for large enough choice of $\gamma > 0$. Combining the above displays leads to

$$\begin{aligned}\mathbb{E}_{f_0,T} \max_k |Z_{jk}| &= \mathbb{E}_T\left(\mathbb{E}_{f_0|T}\left(\max_k |Z_{jk}|\right)\right) \\ &\leq \mathbb{E}_T\left(\max_k |\mu_{n,m,k,T}|\right) + \sqrt{2(j+1)}\mathbb{E}_T \max_k \sigma_{n,m,k,T} \\ &\leq c\left(\sqrt{(\log_2 n)\eta/n} + 2^{j/2}\sqrt{j}e^{-cn^\delta}\right) \leq C\sqrt{(\log_2 n)\eta/n},\end{aligned}$$

for some large enough constants $c, C > 0$ and $2^j \leq n/\eta$, where in the last line we have used that under the event $\mathcal{B}_j$ (i.e., the event that in each bin $I_{j,l} = [(l-1)2^{-j}, l2^{-j}]$, at most $2n/(\eta 2^j)$ observations $T_\ell^{(i)}$, $i \in A_{jk}$, $\ell = 1, \ldots, n/m$, $k = 0, \ldots, 2^j - 1$ fall) we have that $\max_k \sigma^2_{n,m,k,T} \leq C$, and $\mathbb{P}_T(\mathcal{B}_j^c) \leq Ce^{-cn^\delta}$; see (3.21).

4.3. *Proof of Theorem 2.16.* The proof of the theorem goes similarly to the proof of Theorem 2.13, here we only highlight the differences. First, recall that for every $s, L > 0$ and $f_0 \in B^s_{\infty,\infty}(L)$ we have $f_{0,jk} \leq L$, for all $j \geq 0, k \in \{0, 1, \ldots, 2^j - 1\}$, hence, following from the same argument as in Theorem 2.13, the estimator belongs to $\mathcal{F}_{\text{dist}}(B, \ldots, B; B^s_{\infty,\infty}(L))$.

Let us next introduce the notation $B(j, f_0) = 2^{-js}\|f_0\|_{B^s_{\infty,\infty}}$ and

$$j^* = \min\{j \in \mathcal{J} : B(j, f_0) \le \sqrt{j2^j/n_j}\}.$$

Then, by the definition of j^* we have

$$\sqrt{(j^*-1)2^{j^*-1}/n_{j^*-1}} < B(j^*-1, f_0) = 2^s B(j^*, f_0) \le 2^s\sqrt{j^*2^{j^*}/n_{j^*}}.$$

Distinguish again three cases according to the value of j^*, we get that

$$2^{j^*} \asymp \begin{cases} \left(\|f_0\|^2_{B^s_{\infty,\infty}} n/\log_2 n\right)^{1/(1+2s)} & \text{if } B \ge \overline{B}, \\ B/\log_2 n & \text{if } \underline{B} \le B < \overline{B}, \\ \left(nB\|f_0\|^2_{B^s_{\infty,\infty}}/\log_2^3 n\right)^{1/(2+2s)} & \text{if } B < \underline{B}, \end{cases} \tag{4.5}$$

and

$$n_{j^*} \gtrsim \begin{cases} n & \text{if } B \ge \overline{B}, \\ n/\log_2 n & \text{if } \underline{B} \le B < \overline{B}, \\ \|f_0\|^{-1/(1+s)}_{B^s_{\infty,\infty}} \left(nB/\log_2^{\frac{1+4s}{1+2s}} n\right)^{\frac{1+2s}{2+2s}} & \text{if } B < \underline{B}, \end{cases} \tag{4.6}$$

where $\underline{B} = (\|f_0\|^2_{B^s_{\infty,\infty}} n)^{\frac{1}{1+2s}} (\log_2 n)^{\frac{2s-1}{1+2s}}$ and $\overline{B} = 4(\|f_0\|^2_{B^s_{\infty,\infty}} n)^{\frac{1}{1+2s}} (\log_2 n)^{\frac{2s}{1+2s}}$ similarly to Section 3.6. Note that in all cases $j^* \le j_{\max}$ holds.

We split the risk into two parts

$$\begin{aligned} &\mathbb{E}_{f_0,T}\|f_0 - \hat{f}\|_\infty \\ &\quad = \mathbb{E}_{f_0,T}\|f_0 - \tilde{f}(\hat{j})\|_\infty 1_{\hat{j}>j^*} + \mathbb{E}_{f_0,T}\|f_0 - \tilde{f}(\hat{j})\|_\infty 1_{\hat{j}\le j^*} \end{aligned} \tag{4.7}$$

and deal with each term on the right-hand side separately. Note that in view of the definition of $\hat{j}$ and assertions (4.2) and (4.3)

$$\begin{aligned} &\mathbb{E}_{f_0,T}\|f_0 - \tilde{f}(\hat{j})\|_\infty 1_{\hat{j}\le j^*} \\ &\quad \le \mathbb{E}_{f_0,T}\|\tilde{f}(j^*) - \tilde{f}(\hat{j})\|_\infty 1_{\hat{j}\le j^*} + \mathbb{E}_{f_0,T}\|\tilde{f}(j^*) - f_0\|_\infty \\ &\quad \le \tau\sqrt{j^*2^{j^*}/n_{j^*}} + \|\mathbb{E}_{f_0,T}\tilde{f}(j^*) - f_0\|_\infty + \mathbb{E}_{f_0,T}\|\tilde{f}(j^*) - \mathbb{E}_{f_0,T}\tilde{f}(j^*)\|_\infty \\ &\quad \lesssim \sqrt{(\log_2 n)2^{j^*}/n_{j^*}} + \|f_0\|_{B^s_{\infty,\infty}} 2^{-j^*s}, \end{aligned}$$

which implies together with (4.5) and (4.6) that

$$\mathbb{E}_{f_0,T}\|f_0 - \hat{f}\|_\infty 1_{\hat{j}\le j^*} \lesssim \begin{cases} \|f_0\|^{1/(1+2s)}_{B^s_{\infty,\infty}} (n/\log_2 n)^{-s/(1+2s)} & \text{if } B \ge \overline{B}, \\ \sqrt{B(\log_2 n)/n} & \text{if } \underline{B} \le B \le \overline{B}, \\ \|f_0\|^{1/(1+s)}_{B^s_{\infty,\infty}} \left(nB/\log_2^3 n\right)^{-s/(2+2s)} & \text{if } B \le \underline{B}. \end{cases}$$

Noting that $\|f_0\|_{B^s_{\infty,\infty}} \le L$ leads to the claimed upper bounds.

Next, we deal with the first term on the right-hand side of (4.7). First, note that in view of (4.2),

$$\|f_0 - \mathbb{E}_{f_0,T}\tilde{f}(j)\|_\infty^2 \lesssim L^2 2^{-2js} + 2^j/n.$$

Furthermore, by using the upper bound $\psi_{lk}^2 \lesssim 2^l$

$$\begin{aligned}\mathbb{E}_{f_0,T}\|\tilde{f}(j) - \mathbb{E}_{f_0,T}\tilde{f}(j)\|_\infty^2 &\lesssim \mathbb{E}_{f_0,T}\left(\sup_{x\in[0,1]}\sum_{l=0}^{j}\sum_{k=0}^{2^l-1}|\psi_{lk}(x)|\big(|Z_{lk}| + |W_{lk}|\big)\right)^2 \\ &\lesssim 2^{2j}\mathbb{E}_{f_0,T}\sum_{l=0}^{j}\sum_{k=0}^{2^l-1}\big(Z_{lk}^2 + W_{lk}^2\big) \\ &\lesssim 2^{3j}\big(\mathbb{E}_{f_0,T}Z_{lk}^2 + n^{-1}\big) \lesssim 2^{3j}.\end{aligned}$$

Then, by Cauchy–Schwarz inequality and Lemma 4.1 we get that

$$\begin{aligned}&\mathbb{E}_{f_0,T}\|f_0 - \hat{f}\|_\infty 1_{\hat{j}>j^*} \\ &\le \sum_{j=j^*+1}^{j_{\max}} \mathbb{E}_{f_0,T}^{1/2}\|f_0 - \tilde{f}(j)\|_\infty^2 \mathbb{P}_{f_0,T}^{1/2}(\hat{j} = j) \\ &\lesssim \sum_{j=j^*+1}^{j_{\max}} 2^{(3/2)j}\mathbb{P}_{f_0,T}^{1/2}(\hat{j} = j) \lesssim 2^{j^*}e^{-c\tau^2 j^*} + 2^{(3/2)j_{\max}}n^{-2} \\ &= o\big(2^{-j^*s} + 1/\sqrt{n}\big),\end{aligned}$$

for sufficiently large choice of $\tau > 0$, resulting in the required upper bound and concluding the proof of our statement.

LEMMA 4.1. *Assume that $f_0 \in B_{\infty,\infty}^s(L)$, for some $s, L > 0$. Then, for every $C > 0$ there exist positive constants $c > 0$ such that for every $j > j^*$ and sufficiently large $\tau > 0$ we have*

$$\mathbb{P}_{f_0,T}(\hat{j} = j) \lesssim e^{-c\tau^2 j} + n^{-2}.$$

PROOF. Let us introduce the notation $j^- = j - 1$ and note that for every $j > j^*$ we have $j^- \ge j^*$. Then, by the definition of $\hat{j}$

$$\mathbb{P}_{f_0,T}(\hat{j} = j) \le \sum_{l=j}^{j_{\max}} \mathbb{P}_{f_0,T}\big(\|\tilde{f}(j^-) - \tilde{f}(l)\|_\infty > \tau\sqrt{l2^l/n_l}\big).$$

By triangle inequality

$$\begin{aligned}\|\tilde{f}(j^-) - \tilde{f}(l)\|_\infty \le{}& \|\tilde{f}(j^-) - \mathbb{E}_{f_0,T}\tilde{f}(j^-)\|_\infty + \|\tilde{f}(l) - \mathbb{E}_{f_0,T}\tilde{f}(l)\|_\infty \\ &+ \|\mathbb{E}_{f_0,T}\tilde{f}(j^-) - \mathbb{E}_{f_0,T}\tilde{f}(l)\|_\infty.\end{aligned}$$

We deal with the terms on the right-hand side separately. First, note that

$$\begin{aligned}&\|\mathbb{E}_{f_0,T}\tilde{f}(j^-) - \mathbb{E}_{f_0,T}\tilde{f}(l)\|_\infty \\ &\le \left\|\sum_{r=j^-}^{l}\sum_{k=0}^{2^r-1} f_{0,rk}\psi_{rk}\right\|_\infty + \left\|\sum_{r=j^-}^{l}\sum_{k=0}^{2^r-1}\mathbb{E}_{f_0,T}W_{rk}\psi_{rk}\right\|_\infty\end{aligned}$$

$$\le c\big(\|f_0\|_{B^s_{\infty,\infty}} 2^{-j^- s} + \sqrt{2^l/n}\big) \le C\big(B(j^-, f_0) + \sqrt{2^l/n}\big)$$
$$\le C\big(B(j^*, f_0) + \sqrt{2^l/n}\big) \le C\big(\sqrt{j^* 2^{j^*}/n_{j^*}} + \sqrt{2^l/n}\big)$$
$$\le C\sqrt{l2^l/n_l}.$$

Furthermore,

$$\big\|\tilde f(l) - \mathbb{E}_{f_0,T}\tilde f(l)\big\|_\infty \le \sum_{j=0}^{l} \max_k\big(|Z_{jk}| + |W_{jk}|\big) \sup_{x\in[0,1]} \sum_{k=0}^{2^j-1} |\psi_{jk}(x)|$$
$$\le C\left(\sum_{j=0}^{l} 2^{j/2} \max_k |Z_{jk}| + \sqrt{2^l/n}\right).$$

We show below that for any $\gamma \ge 1$,

$$\mathbb{P}_{f_0,T}\big(\max_k\big(|Z_{lk}| \ge \tau\sqrt{\gamma l/n_l}\big) \lesssim n_l^{1-\gamma} + e^{-c\tau^2 l} \tag{4.8}$$

holds for some sufficiently large $\tau > 0$ and sufficiently small $c > 0$. By combining the above results we get that

$$\mathbb{P}_{f_0,T}\big(\big\|\tilde f(j^-) - \tilde f(l)\big\|_\infty \ge \tau\sqrt{l2^l/n_l}\big)$$
$$\lesssim \mathbb{P}_{f_0,T}\left(\big\|\tilde f(l) - \mathbb{E}_{f_0,T}\tilde f(l)\big\|_\infty \ge \frac{\tau - C}{2}\sqrt{l2^l/n_l}\right)$$
$$\lesssim \sum_{j=0}^{l} \mathbb{P}_{f_0,T}\left(\max_k |Z_{jk}| \ge \frac{\tau - 2C}{2C}\sqrt{l/n_l}\right)$$
$$\le l\mathbb{P}_{f_0,T}\left(\max_k |Z_{lk}| \ge \frac{\tau - 2C}{2C}\sqrt{l/n_l}\right)$$
$$\lesssim (\log_2 n) n_l^{1-\gamma} + e^{-(c/2)\tau^2 l}.$$

The above inequality together with the first display of the proof then implies that

$$\mathbb{P}_{f_0,T}(\hat j = j) \lesssim \sum_{l=j}^{j_{\max}} \big((\log_2 n) n_l^{1-\gamma} + e^{-(c/2)\tau^2 l}\big)$$
$$\lesssim (\log_2 n)^2 n_{j_{\max}}^{1-\gamma} + e^{-(c/2)\tau^2 j} \lesssim n^{-2} + e^{-(c/2)\tau^2 j},$$

for $\gamma \ge 5$, in view of $n_{j_{\max}} \gtrsim (nB/\log_2^2 n)^{\frac{1+2s_{\min}}{2+2s_{\min}}} \ge \sqrt{n}$, for any $s_{\min} > 0$, providing the statement of the lemma.

It remained to prove assertion (4.8). Note that by triangle inequality we get that

$$\max_k |Z_{lk}| \le \max_k |Z_{lk} - \mathbb{E}_{f_0|T} Z_{lk}| + \max_k |\mathbb{E}_{f_0|T} Z_{lk}|. \tag{4.9}$$

In view of assertion (4.4) with $\mathbb{P}_T$-probability at least $1 - Cn_l^{1-\gamma}$, the second term on the right-hand side is bounded from above by $C\sqrt{\gamma l/n_l}$. Furthermore, recall from the proof of Theorem 2.10 (i.e., Assertion (3.21)) that $n_l^{-2}\sum_{i\in A_{lk}}\sum_{\ell=1}^{n/m}\psi_{lk}^2(T_\ell^{(i)}) \lesssim n_l^{-1}$ holds with $\mathbb{P}_T$-probability at least $1 - Ce^{-cn^\delta}$, for some sufficiently small $\delta > 0$. Under the above event we have that there exists small enough constant $c > 0$ such that

$$\mathbb{P}_{f_0|T}\big(|Z_{l1} - \mathbb{E}_{f_0|T} Z_{l1}| \ge \tau\sqrt{l/n_l}\big) \le \exp\{-c\tau^2 l\}.$$

Therefore, the first term on the right-hand side of (4.9) is bounded from above by $\tau\sqrt{l/n_l}$ with $\mathbb{P}_{f_0|T}$-probability at least $1 - C2^l e^{-c\tau^2 l} \le 1 - Ce^{-(c/2)\tau^2 l}$ on $T \in \mathcal{B}_l$, for some sufficiently large constants $\tau, C > 0$ and sufficiently small positive constant c. □

5. Technical lemmas. In this section we provide the technical lemmas applied in the previous two sections.

5.1. *Proof of Lemma 3.2.* We are going to apply the general information bound given by Theorem A.13 in the Supplementary Material. To this end, we need a number of definitions.

Without loss of generality we can assume that $T_1^{(i)} \le T_2^{(i)} \le \cdots \le T_{n/m}^{(i)}$, $i = 1, \ldots, m$, and let $\ell_k = \ell_k^{(i)} = \max\{j \in \{1, \ldots, n/m\} : T_j^{(i)} \in I_k\}$ denote the index of the largest element $T_j^{(i)}$ in the interval $I_k = [(k-1)C_2 2^{-j_n}, kC_2 2^{-j_n}]$, $k = 1, \ldots, |K_{j_n}| = 2^{j_n}/C_2$. Note that $T_{\ell_{k-1}+1}^{(i)}, \ldots, T_{\ell_k}^{(i)} \in I_k$. For convenience let us introduce the following notation:

$$
\begin{aligned}
X_{[j_1:j_2]}^{(i)} &= \big(X_{j_1}^{(i)}, X_{j_1+1}^{(i)}, \ldots, X_{j_2}^{(i)}\big),\\
d &= |K_{j_n}|,\\
F_{-k} &= (F_1, .., F_{k-1}, F_k, \ldots, F_d),\\
\delta &= L\delta_n^{1/2} 2^{j_n/2} \|\psi\|_\infty,\\
a^2 &= \frac{2^5 n\delta^2}{dm/\log(dm)},\\
\mu_k(t) &= \big(L\delta_n^{1/2}\psi_{j_n,k}(t_j)\big)_{j=(\ell_{k-1}+1),\ldots,\ell_k},\\
B_k(t) &= \big\{x \in \mathbb{R}^{\ell_k - \ell_{k-1}} : \big|\mu_k(t)^T x\big| \le a\big\},\\
\mathcal{B} &= \left\{t \in [0,1]^{n/m} : \frac{n}{2dm} \le \ell_k - \ell_{k-1} \le \frac{2n}{dm}, k = 1, \ldots, d\right\}.
\end{aligned}
$$

Note that $X_{[(\ell_{k-1}+1):\ell_k]}^{(i)} | (T^{(i)}, F_k)$ is independent of F_{-k} and

$$
X_{[(\ell_{k-1}+1):\ell_k]}^{(i)} \big| \big(T^{(i)} = t, F_k = \beta_k\big) \sim \mathbb{P}_{\beta_k|T^{(i)}=t}^{(i)} = N_{\ell_k - \ell_{k-1}}\big(\beta_k \mu_k(t), I\big).
$$

Furthermore, note that the inequalities $\delta^2 \le \frac{0.4^2 md}{2^7 n \log(dm)}$ (in view of $C_1 \ge 0.4^{-2} 2^8 L^2 \|\psi\|_\infty^2 C_2$) and $n/m \ge 2^6 d \log(n/m)$ (in view of $m = O(n^{\frac{2s}{1+2s}}/\log^2 n)$) hold.

Then, by the definition of $B_k(t)$ we have for all $t \in [0,1]^{n/m}$ and $k = 1, \ldots, d$ that

$$
\begin{aligned}
\sup_{x \in B_k(t)} \frac{\varphi_{\mu_k(t)}(x)}{\varphi_{-\mu_k(t)}(x)} &= \sup_{x \in B_k(t)} \exp\left\{\frac{\big|\|x - \mu_k(t)\|_2^2 - \|x + \mu_k(t)\|_2^2\big|}{2}\right\}\\
&= \sup_{x \in B_k(t)} \exp\{2|x^T \mu_k(t)|\} = \exp\{2a\},
\end{aligned}
$$

where φ_μ denotes the density function of a normal distribution with mean vector μ and identity covariance matrix. Then, by Theorem A.13 in the Supplementary Material (with

$\mathcal{F}_0 = \{\beta = (\beta_k)_{k=1..d} : \beta_k \in \{-1, 1\}, k = 1, \dots, d\})$ we have that

$$
\begin{aligned}
I\big(F; Y^{(i)}\big) &= \int_{[0,1]^{n/m}} I\big(F; Y^{(i)} | T^{(i)} = t\big)\, dt \\
&\le \sum_{k=1}^{d} (\log 2) \int_{[0,1]^{n/m}} \sqrt{\mathbb{P}^{(i)}_{\beta_k | T^{(i)}=t}\big(X^{(i)}_{[(\ell_{k-1}+1):\ell_k]} \notin B_k(t)\big)}\, dt \\
&\quad + \sum_{k=1}^{d} \int_{[0,1]^{n/m}} \mathbb{P}^{(i)}_{\beta_k | T^{(i)}=t}\big(X^{(i)}_{[(\ell_{k-1}+1):\ell_k]} \notin B_k(t)\big)\, dt \\
&\quad + 2C^2(C-1)^2 I\big(X^{(i)}; Y^{(i)} | T^{(i)}\big),
\end{aligned} \tag{5.1}
$$

with $C = \exp\{2^{7/2}\delta\sqrt{n\log(dm)}/\sqrt{dm}\}$.

Note that $I(X^{(i)}; Y^{(i)}|T^{(i)}) \le H(Y^{(i)}|T^{(i)}) \le H(Y^{(i)})$. In view of Lemma 5.2, we have that $\mathbb{P}_T(T^{(i)} \in \mathcal{B}) \ge 1 - 2de^{-n/(8md)} \ge 1 - 2(md)^{-4}$ following from the inequality $n/m \ge 2^6 d \log(n/m)$. Besides, for arbitrary $t \in \mathcal{B}$ we have in view of

$$
\big\|\mu_k(t)\big\|_2^2 \le \sum_{j=\ell_{k-1}+1}^{\ell_k} \delta_n \psi_{j_n,k}(t_j)^2 \le \|\psi\|_\infty^2 \delta_n 2^{j_n}(\ell_k - \ell_{k-1}) \le 2n\delta^2/(md)
$$

that

$$
\begin{aligned}
\mathbb{P}^{(i)}_{f_k}\big(X^{(i)}_{[(\ell_{k-1}+1):\ell_k]} \notin B_k(t) | T^{(i)} = t\big) &= \mathbb{P}^{(i)}_{f}\big(|\mu_k(t)^T X^{(i)}_{[(\ell_{k-1}+1):\ell_k]}| > a | T^{(i)} = t\big) \\
&\le 2\exp\bigg\{-\frac{(a - \|\mu_k(t)\|_2^2)^2}{2\|\mu_k(t)\|_2^2}\bigg\} \\
&\le 2\exp\bigg\{-\frac{a^2}{4\|\mu_k(t)\|_2^2}\bigg\} \le 2(md)^{-4}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\int_{[0,1]^{n/m}} \sqrt{\mathbb{P}^{(i)}_{\beta_k | T^{(i)}=t}\big(X^{(i)}_{[(\ell_{k-1}+1):\ell_k]} \notin B_k(t)\big)}\, dt \\
&\quad \le \int_{\mathcal{B}} \sqrt{\mathbb{P}^{(i)}_{\beta_k | T^{(i)}=t}\big(X^{(i)}_{[(\ell_{k-1}+1):\ell_k]} \notin B_k(t)\big)}\, dt + \mathbb{P}_T\big(T^{(i)} \notin \mathcal{B}\big) \\
&\quad \le \sqrt{2}(md)^{-2} + 2(md)^{-4} \le 2(md)^{-2},
\end{aligned}
$$

and similarly $\int_{[0,1]^{n/m}} \mathbb{P}^{(i)}_{f|T^{(i)}=t}(X^{(i)}_{[(\ell_{k-1}+1):\ell_k]} \notin B_k(t))\, dt \le 4(md)^{-4}$. Then, by plugging in the above inequalities into (5.1) and using the inequalities $e^x \le 1 + 2x$ for $x \le 0.4$ and $C^2 \le 2$ we get that

$$
I\big(F; Y^{(i)}\big) \le \frac{4\log 2}{m^2 d} + \frac{2^{12}\delta^2 n \log(dm)}{md} H\big(Y^{(i)}\big).
$$

Furthermore, from the data-processing inequality and the convexity of the KL divergence

$$
\begin{aligned}
I(F;Y^{(i)}) &\le I(F;(T^{(i)},X^{(i)})) \le I(F;X^{(i)}|T^{(i)}) + I(F;T^{(i)}) \\
&= \int_{t\in[0,1]^{n/m}} I(F;X^{(i)}|T^{(i)}=t)\,dt \\
&\le \int_{t\in[0,1]^{n/m}} \frac{1}{|\mathcal{F}_0|^2} \sum_{\beta,\beta'\in\mathcal{F}_0} K\big(\mathbb{P}^{(i)}_{\beta|T^{(i)}=t} \| \mathbb{P}^{(i)}_{\beta'|T^{(i)}=t}\big)\,dt \qquad (5.2) \\
&\le \frac{1}{2|\mathcal{F}_0|^2} \sum_{\beta,\beta'\in\mathcal{F}_0} \sum_{\ell=1}^{n/m} \sum_{k\in K_{j_n}} (\beta'_k-\beta_k)^2 L^2\delta_n \int_0^1 \psi^2_{j_n,k}(t_\ell)\,dt_\ell \\
&\le 2\delta^2 n/m.
\end{aligned}
$$

Then, by combining the previous upper bounds and using the data processing inequality $I(F;Y) \le \sum_{i=1}^m I(F;Y^{(i)})$ we get that

$$
\begin{aligned}
I(F;Y) &\le \frac{4\delta^2 n}{m} \sum_{i=1}^m \min\big\{2^{10}\log(md)d^{-1}H\big(Y^{(i)}\big),1\big\} + 4\log 2 \\
&\le \frac{4L^2\delta_n 2^{j_n}\|\psi\|_\infty^2 n}{m} \sum_{i=1}^m \min\big\{2^{10}\log(md)d^{-1}H\big(Y^{(i)}\big),1\big\} + 4\log 2.
\end{aligned}
$$

Finally, we arrive to our statement by using Lemma 5.3 and $2^{j_n} = C_2 d$.

REMARK 5.1. We note that in [22] it is sufficient to provide the upper bound (5.2) for the mutual information as there is no limitation in the amount of transmitted bits. In our setting one has to take into account the code length as well, hence, sharper upper bounds are required which is actually the core and most challenging part of the proof of Lemma 3.2.

LEMMA 5.2. *Let $X_1, X_2, \ldots, X_n$ be independent and uniformly distributed over $\{1, 2, \ldots, r\}$, and denote by $\chi_k = \{\ell \in \{1,\ldots,n\} : X_\ell = k\}$ the index set of the observations belonging to the kth bin, $k = 1,\ldots,r$. Then,*

$$
P\big(2^{-1}n/r \le |\chi_k| \le 2n/r, k = 1,\ldots,r\big) \ge 1 - 2re^{-n/(8r)}.
$$

PROOF. We start with the proof of the upper bound. Note that by Chernoff's bound

$$
P\Big(\sup_{k=1,\ldots,r} |\chi_k| \ge 2n/r\Big) \le \sum_{k=1}^r P\big(|\chi_k| \ge 2n/r\big) \le re^{-n/(3r)},
$$

and similarly for the lower bound

$$
P\Big(\inf_{k=1,\ldots,r} |\chi_k| \le 2^{-1}n/r\Big) \le re^{-n/(8r)}. \qquad \square
$$

5.2. *Entropy of a finite binary string.* In the proof of Theorem 2.1 we need to bound the entropy of transmitted finite binary string $Y^{(i)}$. Since we do not want to restrict ourself only to prefix codes, we can not use a standard version of Shannon's source coding theorem for this purpose. Instead, we use the following result:

LEMMA 5.3. *Let Y be a random finite binary string. Its entropy and expected length satisfy the inequality*

$$
H(Y) \le 2\mathbb{E}l(Y) + 1.
$$

PROOF. We construct and auxiliary random string U such that $l(U)$ and $l(Y)$ have the same distribution and such that, given its length, U has a uniform distribution on the set of strings with that length. Specifically, let $N = l(Y)$ and consider a random binary string U with distribution $U|N = n \sim \text{Unif}(\{0,1\}^n)$. Let S denote the set of all finitely long binary strings. Then, the KL-divergence between Y and U is given by

$$\begin{aligned} K(Y,U) &= \sum_{s\in S} \mathbb{P}(Y=s)\log\frac{\mathbb{P}(Y=s)}{\mathbb{P}(U=s)} \\ &= \sum_{s\in S} \mathbb{P}(Y=s)\log\frac{1}{\mathbb{P}(U=s)} - H(Y) \\ &= \sum_{n}\sum_{s\in\{0,1\}^n} \mathbb{P}(Y=s)\log\frac{1}{\mathbb{P}(U=s)} - H(Y). \end{aligned}$$

Now, for every n and $s \in \{0,1\}^n$, we have $\mathbb{P}(U=s) = \mathbb{P}(U=s \mid N=n)\mathbb{P}(N=n) = 2^{-n}\mathbb{P}(N=n)$. It follows that

$$\sum_{s\in\{0,1\}^n} \mathbb{P}(Y=s)\log\frac{1}{\mathbb{P}(U=s)} = \mathbb{P}(N=n)\log\frac{2^n}{\mathbb{P}(N=n)}.$$

Hence,

$$K(Y,U) \le (\log 2)\mathbb{E}N + H(N) - H(Y).$$

The nonnegativity of the KL-divergence thus implies that $H(Y) \le \mathbb{E}N + H(N)$.

To complete the proof we show that $H(N) \le \mathbb{E}N + 1$. To do so, consider the index set $I = \{i : \mathbb{P}(N=i) \ge e^{-i}\}$ and note that the function $x \mapsto x\log(1/x)$ is monotone increasing for $x \le e^{-1}$. Then,

$$\begin{aligned} H(N) &= \sum_{i\in I} \mathbb{P}(N=i)\log\frac{1}{\mathbb{P}(N=i)} + \sum_{i\in I^c} \mathbb{P}(N=i)\log\frac{1}{\mathbb{P}(N=i)} \\ &\le \sum_{i\in I} \mathbb{P}(N=i)i + \sum_{i\in I^c} e^{-i}i \le \mathbb{E}N + 1. \end{aligned}$$

This completes the proof. □

Acknowledgements. We would like to thank the Associate Editor and the referees for their careful review of the various versions of this paper and for their valuable comments and suggestions.

Both authors were supported by the Netherlands Organization of Scientific Research NWO.

The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

SUPPLEMENTARY MATERIAL

Supplement to "Adaptive distributed methods under communication constraints" (DOI: 10.1214/19-AOS1890SUPP; .pdf). Supplementary information

REFERENCES

[1] BATTEY, H., FAN, J., LIU, H., LU, J. and ZHU, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *Ann. Statist.* **46** 1352–1382. MR3798006 https://doi.org/10.1214/17-AOS1587

[2] CARPENTIER, A. (2013). Honest and adaptive confidence sets in L_p. *Electron. J. Stat.* **7** 2875–2923. MR3148371 https://doi.org/10.1214/13-EJS867

[3] CHAGNY, G. (2013). Penalization versus Goldenshluger–Lepski strategies in warped bases regression. *ESAIM Probab. Stat.* **17** 328–358. MR3066383 https://doi.org/10.1051/ps/2011165

[4] DEISENROTH, M. P. and NG, J. W. (2015). Distributed Gaussian processes. ArXiv E-prints.

[5] DUCHI, J. C. and WAINWRIGHT, M. J. (2013). Distance-based and continuum Fano inequalities with applications to statistical estimation. ArXiv E-prints.

[6] GERSHGORIN, S. A. (1931). Uber die abgrenzung der eigenwerte einer matrix. *Bulletin de L'Académie des Sciences de L'URSS. Classe des Sciences Mathématiques et na* **6** 749–754.

[7] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[8] GUHANIYOGI, R., LI, C., SAVITSKY, T. D. and SRIVASTAVA, S. (2017). A divide-and-conquer Bayesian approach to large-scale kriging. ArXiv E-prints.

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

[10] KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. MR3248677 https://doi.org/10.1111/rssb.12050

[11] LACOUR, C. (2008). Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.* **99** 787–814. MR2405092 https://doi.org/10.1016/j.jmva.2007.04.006

[12] LEE, J. D., LIU, Q., SUN, Y. and TAYLOR, J. E. (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res.* **18** Paper No. 5, 30. MR3625709

[13] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR2319879

[14] ROSENBLATT, J. D. and NADLER, B. (2016). On the optimality of averaging in distributed statistical learning. *Inf. Inference* **5** 379–404. MR3609865 https://doi.org/10.1093/imaiai/iaw013

[15] SCOTT, S. L., BLOCKER, A. W., BONASSI, F. V., CHIPMAN, H., GEORGE, E. and MCCULLOCH, R. (2013). Bayes and big data: The consensus Monte Carlo algorithm. In *EFaBBayes* 250 *Conference* **16**.

[16] SHANG, Z. and CHENG, G. (2015). A Bayesian splitotic theory for nonparametric models. ArXiv E-prints.

[17] SRIVASTAVA, S., CEVHER, V., TRAN-DINH, Q., DUNSON and WASP, D. B. (2015). Scalable Bayes via barycenters of subset posteriors. In *AISTATS*.

[18] SZABÓ, B. and VAN ZANTEN, H. (2020). Supplement to "Adaptive distributed methods under communication constraints." https://doi.org/10.1214/19-AOS1890SUPP.

[19] SZABÓ, B. and VAN ZANTEN, H. (2019). An asymptotic analysis of distributed nonparametric methods. *J. Mach. Learn. Res.* **20** Paper No. 87, 30. MR3960941

[20] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

[21] WANG, J., CHEN, J. and WU, X. (2010). On the sum rate of Gaussian multiterminal source coding: New proofs and results. *IEEE Trans. Inform. Theory* **56** 3946–3960. MR2752477 https://doi.org/10.1109/TIT.2010.2050960

[22] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500 https://doi.org/10.1214/aos/1017939142

[23] ZHANG, Y., DUCHI, J., JORDAN, M. I. and WAINWRIGHT, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems* 2328–2336.

[24] ZHANG, Y., WAINWRIGHT, M. J. and DUCHI, J. C. (2012). Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems* 25 1502–1510 (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.). Curran Associates, Inc.

[25] ZHU, Y. and LAFFERTY, J. (2018). Quantized minimax estimation over Sobolev ellipsoids. *Inf. Inference* **7** 31–82. MR3801518 https://doi.org/10.1093/imaiai/iax007

[26] ZHU, Y. and LAFFERTY, J. (2018). Distributed nonparametric regression under communication constraints. In *Proceedings of the* 35*th International Conference on Machine Learning*, *ICML* 2018, *Stockholmsmässan*, *Stockholm*, *Sweden*, *July* 10–15, 2018 6004–6012.