Math. Stat. Learn. 5 (2022), 159–199
DOI 10.4171/MSL/33

Distributed function estimation: Adaptation using minimal communication

Botond Szabó and Harry van Zanten

Abstract. We investigate whether in a distributed setting, adaptive estimation of a smooth function at the optimal rate is possible under minimal communication. It turns out that the answer depends on the risk considered and on the number of servers over which the procedure is distributed. We show that for the L_∞-risk, adaptively obtaining optimal rates under minimal communication is not possible. For the L_2-risk, it is possible over a range of regularities that depends on the relation between the number of local servers and the total sample size.

1. Introduction

Distributed methods have attracted a lot of attention in the statistics and machine learning communities recently. There are several reasons for this, the most prominent ones being that they provide a way of dealing with large datasets and with privacy considerations. The theoretical literature on distributed methods is still rather minimal at the moment. A number of papers have recently investigated fundamental performance limits in distributed models, in particular pointing out issues that occur in high-dimensional or nonparametric problems, see for instance [1, 2, 4, 8, 17, 18, 21, 24, 26, 27, 30]. For example, optimal rates in distributed function estimation depend on the amount of communication that is allowed, and the relation of that amount with the regularity of the unknown function. The lower bounds obtained in [25, 28, 31] and the subsequent adaptation results in [25] show that in particular, automatically adapting to the smoothness of the unknown function is a complicated issue in communication restricted distributed settings. In the present paper we study this problem from a different, we think relevant and interesting perspective, not restricting communication a priori, but asking for rate-optimal procedures that require minimal communication.

To be specific, we will consider a distributed signal estimation problem in which we have m local machines and one central machine. At the ith machine we observe

2020 Mathematics Subject Classification. Primary 62G20; Secondary 62G05, 62G10, 94A15.
Keywords. Divide-and-conquer methods, minimax rates, adaptation, communication constraints, Besov spaces, nonparametric estimation.

the random function $X^{(i)}$ satisfying the stochastic differential equation

$$dX_t^{(i)} = f_0(t)\,dt + \sqrt{\frac{m}{n}}\,dW_t^{(i)}, \quad t \in [0,1], \quad i = 1,2,\ldots,m, \tag{1.1}$$

where $W^{(1)}, \ldots, W^{(m)}$ are independent standard Wiener processes. The goal is to estimate the function $f_0 \in L_2[0,1]$ which is assumed to have (Besov) regularity $s > 0$. Each local machine independently processes its own data and then communicate simultaneously in a single round an estimator, or statistic to a central machine. The central machine somehow aggregates all local estimators and produces a final estimator $\widehat{f}$ for the unknown signal f_0.

In the classical, non-distributed setting ($m = 1$) the minimax lower bound over Besov balls of regularity s is known to be of the order $n^{-s/(1+2s)}$ (e.g. [14]). Recently established minimax results for distributed nonparametric methods (see [25, 28, 31], and the appendix to this paper) show that this optimal rate can also be achieved in the distributed case ($m = m_n \to \infty$), but only if each local machine is allowed to communicate at least (up to a logarithmic factor) order $n^{1/(1+2s)}$ bits of information to the central machine (this is what the authors of [31] call the sufficient regime).

If the regularity s of the signal is known, a distributed strategy that achieves the rate $n^{-s/(1+2s)}$ under the restriction that the local machines communicate at most the minimal order $n^{1/(1+2s)}$ bits is easily constructed, see Theorem B.2 in the appendix. The situation changes in an interesting way if s is unknown however. We study in this paper to what degree it is in that case possible to estimate the signal at the optimal rate $n^{-s/(1+2s)}$, while at the same time only communicating order $n^{1/(1+2s)}$ bits of information between the local machines and the central one. The additional difficulty, on top of the fact that we ask for rate-adaptive estimation, is that the local machines must then ensure that they communicate at most of order $n^{1/(1+2s)}$ bits using only their local data, without knowing the regularity s. We stress that we *do not* put a priori communication restrictions on the considered estimation procedures, but that we study the question of estimation at the optimal rate, using minimal communication.

It turns out that whether or not this is possible for the L_2-risk depends on the relation between the number of machines m and the total sample size, or signal-to-noise ratio n. We prove that if $m = n^p$ for some $p \in (0, 1/2)$, then:

- There exists a distributed estimator that is adaptive over any range of regularities $[s_1, s_2]$ such that
$$0 < s_1 < s_2 < \frac{1}{4p} - \frac{1}{2},$$
achieving the optimal rate and transmitting the minimal amount of bits.
- If
$$s_2 > s_1 > \frac{1}{4p} - \frac{1}{2}$$

however, then there exists no single-round distributed procedure that achieves the optimal rate for every signal f with regularity in $\{s_1, s_2\}$, while transmitting the minimal amount of bits.

Stated differently, when considering L_2-risk, adaptively achieving the optimal rate using minimal communication over a range of regularities $[s_1, s_2]$ is possible if and only if

$$(2 + 4s_2) \log m < \log n.$$

This shows that it is problematic if either the number of machines is too large, or the range of regularities to which adaptation is required is too large.

The adaptive, minimal communication procedure that we propose in the first case implicitly exploits the fact that for the L_2-risk, there is a difference between lower bounds for estimation and testing, see for instance [14, 16]. Indeed, we employ the testing result of [10] to extract sufficient information about the regularity of the unknown signal in the local servers, which we then use in the subsequent estimation procedure. This approach depends crucially on the fact that we consider the L_2-risk. For the L_∞-risk there is no difference between testing and estimation rates and this approach breaks down. In fact, we prove that for the L_∞-norm, adaptive estimation at the optimal rate under minimal communication is never possible!

The impossibility results all derive from the fact that at the local servers, the sample size is too small to extract sufficient information about the regularity of a general signal. This suggests that if we restrict to a class of "nice" signals for which we do have access to such smoothness information from limited data, we should be able to obtain optimal rates and minimal communication adaptively. We prove that this is indeed the case if we consider the class of self-similar functions, cf. [13, 19], initially considered in the context of adaptive nonparametric confidence sets, where similar issues arise. See also for instance [6, 7, 13, 20, 22, 23]. While noticing the fact that very similar issues play a role, we are not able at the moment to obtain a more formal clarification of the connection between our impossibility results for adaptive distributed methods and the existing impossibility results for adaptive confidence sets.

The remainder of the paper is organised as follows. In the next section we present our main results. Theorems 2.1 and 2.2 and Corollary 2.3 assert that whether simultaneous adaptation over a range of regularities and minimal communication is possible for the L_2 risk, depends on the relation between the range of regularities and the number of local machines. Theorem 2.5 shows that simultaneous adaptation and minimal communication is not possible when L_∞ risk is considered. Finally, Theorem 2.6 asserts that it is possible under a self-similarity assumption. Proofs and auxiliary results are deferred to Sections 3–4 and the appendices.

1.1. Notations

For two positive sequences a_n, b_n we use the notation $a_n \lesssim b_n$ if there exists a 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. In the proofs we use the notation C and c for universal constants which value can differ from line to line and denote by $\#S$ or $|S|$ the cardinality of the finite set S. Furthermore, let $l(Y)$ denote the length of a binary string Y, and $\log x$ denote the logarithm with base 2, i.e. $\log_2 x$.

2. Main results

In our analysis we work with the distributed Gaussian white noise model also considered for instance in [24, 31], and which can be seen as an idealized version of the nonparametric regression model. Our results can in principle be derived in the regression context as well, similarly to we did in [25], or in a more general nonparametric setting, see [28]. However, since the additional technical issues would seriously lengthen the already long paper and would add no fundamental insight, we formulate everything in the signal in white noise setting in this paper.

As explained in the introduction, we assume that we have m local machines and that at the ith machine we observe the random function $X^{(i)}$ given by the stochastic differential equation (1.1). Parallel to each other, the local machines carry out a local statistical procedure and transmit the results to a central machine simultaneously, in one round, which constructs the final estimator $\widehat{f}$ for the functional parameter f_0 by somehow aggregating the local outcomes. The number of bits transmitted from machine i to the central machine is denoted by $\widehat{B}^{(i)}$. It is allowed to depend on the number of machines m and the local data $X^{(i)}$, i.e. $\widehat{B}^{(i)} = \widehat{B}^{(i)}(X^{(i)})$, and hence it is a random variable in general. We will impose a smoothness condition on f_0, assuming that either $f_0 \in B^s_{2,\infty}(L)$ or $f_0 \in B^s_{\infty,\infty}(L)$, see Appendix A for a rigorous introduction of these Besov classes. The first class is of Sobolev type, while the second one is of Hölder type.

2.1. Adaptation in L_2

We show that in case of the L_2-risk one can only adapt up to a limited range of smoothness levels (depending on the number of local machines). Outside that range one will achieve sub-optimal rates (where the rate is sub-optimal by a polynomial factor). The following theorem is our first main result.

Theorem 2.1. *Suppose that $m = n^p$ for some $p \in (0, 1/2)$. Then for any regularity parameters $s_2 > s_1 > 1/(4p) - 1/2$ there does not exist a single-round distributed*

estimator $\widehat{f}$ *with* $\widehat{B}^{(i)} \le (L^2 n)^{1/(1+2s_1)+\varepsilon_1} \log n$ *such that for* $l = 1, 2$, *we have*

$$\max_{i\in\{1,\dots,m\}} \sup_{f_0\in B^{s_l}_{2,\infty}(L)} E^{(i)}_{f_0} \widehat{B}^{(i)} \lesssim (L^2 n)^{\frac{1}{1+2s_l}+\varepsilon_1}, \tag{2.1}$$

$$\sup_{f_0\in B^{s_l}_{2,\infty}(L)} E_{f_0} \|\widehat{f} - f_0\|_2^2 \lesssim L^{\frac{1}{1+2s_l}+\varepsilon_2} n^{-\frac{2s_l}{1+2s_l}+\varepsilon_2}, \tag{2.2}$$

for some small enough constants $\varepsilon_1, \varepsilon_2 > 0$ *depending only on* s_1, s_2 *and* p.

Proof. See Section 3.1. ∎

The theorem tells us that considering even just two regularity classes (with regularities above some threshold level) there does not exist any single round distributed method that transmits the minimal amount of bits (multiplied by some (small) polynomial factor) and at the same time achieves the minimax rate in both smoothness classes (again up to a (small) polynomial factor). A formal definition of the minimal amount of transmitted bits is given in (B.1) in the Appendix. This negative result delivers a strong message. It shows that the issue of non-existence of an adaptive, rate-optimal distributed procedure using minimal communication cannot be resolved by allowing extra logarithmic factors, but is on the polynomial level.

The phenomenon behind the negative result is that in case of many local machines (large m) it is getting more difficult to test locally between the regularity classes (as the local "sample size" decreases in m) and also the "local regularity" of the function which one can judge at noise level m/n might be completely different than the "global regularity" of the truth which can be judged at a smaller noise level $1/n$.

Although full adaptation is not possible, it turns out that on a limited range of regularity levels it is possible to construct adaptive methods. Below we derive the complement of the results in Theorem 2.1 by describing a procedure which adapts to arbitrary two fixed regularities below the threshold $1/(4p) - 1/2$, i.e. $0 < s_1 < s_2 \le 1/(4p) - 1/2$, and transmits the minimal number of bits at the same time. The proposed procedure has two stages. First we "estimate" the smoothness of the underlying functional parameter of interest in every local machine parallel to each other and based on that transmit the right amount of information to the central machine. In the second stage we aggregate the locally transmitted information and provide a "global" adaptive estimator. We describe the procedure in more details below.

In our procedure we work with the equivalent sequence representation of the model using the Daubechies wavelets, i.e. in each machine $i = 1, \dots, m$ we observe the noise random variables

$$X^{(i)}_{jk} = \int_0^1 \psi_{jk}(t)\, dX^{(i)}(t) = f_{0,jk} + \sqrt{\frac{m}{n}} Z^{(i)}_{jk}, \quad j = 0, 1, 2, \dots;\ k = 1, \dots, 2^j, \tag{2.3}$$

where $\psi_{jk}(t)$, $t \in [0,1]$ are the Daubechies wavelet bases, $f_{0,jk} = \int_0^1 \psi_{jk}(t) f_0(t)\, dt$ are the wavelet coefficients of f_0 and $Z_{jk}^{(i)}$ are iid standard normal random variables. In Section A in the appendix we have collected a few properties of Daubechies wavelets which we will apply throughout the paper.

As a first step we split the data in all of the local models $i \in \{1, \ldots, m\}$ into two subsets $X_{jk}^{(i,1)}$, $X_{jk}^{(i,2)}$ for $j = 0, 1, 2, \ldots$, $k = 1, \ldots, 2^j$, such that they are pairwise independent and their variance is $2m/n$ (this can be done by adding and subtracting $\widetilde{Z}_{jk}^{(i)} \overset{\text{iid}}{\sim} N(0, m/n)$ from $X_{jk}^{(i)}$).

Next note that it was shown in [10] that there exists a consistent composite test between the classes $B_{2,\infty}^{s_2}(L)$ and $B_{2,\infty}^{s_1}(L)$ in the local problem using the first subset of observations $X^{(i,1)}$ if they are at least $L^{(1/2)/(1/2+2s_1)}(n/m)^{-s_1/(1/2+2s_1)}$ separated. The test proposed in Section 3 of [10] takes the form (in the local machines using the first subset of observations $X^{(i,1)}$)

$$\Psi_{n/m}^{(i)} = \Psi_{n/m}^{(i)}(\alpha, s_1, s_2) = 1 - \prod_{0 \le l \le \lfloor \frac{\log(L^2 n/(2m))}{2s_1+1/2} \rfloor} 1_{\{T_{n/m}^{(i)}(l) \le t_{n/m}(l, s_2, \alpha)\}}, \tag{2.4}$$

where

$$\begin{aligned}
t_{n/m}(l, s_2, \alpha) &= \frac{L^2}{2^{2ls_2}} + \frac{L}{2^{ls_2}}\tau_l + \frac{\tau_l^2}{4}, \\
\tau_l &= 24\sqrt{\frac{1}{\alpha}}\frac{2^{l+\lfloor \frac{\log(L^2 n/(2m))}{1/2+2s_2} \rfloor}}{\sqrt{n/(2m)}} \quad \text{for } l > 0, \\
\tau_0 &= 24\sqrt{\frac{1}{\alpha}}\frac{1}{\sqrt{n/(2m)}}, \\
T_{n/m}^{(i)}(l) &= \|\Pi_l \widehat{f}_{n/m}^{(i)}\|_2^2 - m2^{l+1}/n \quad \text{for } l > 0, \\
T_{n/m}^{(i)}(0) &= \|\Pi_0 \widehat{f}_{n/m}^{(i)}\|_2^2 - 2m/n, \\
\alpha &= n^{-\frac{2s_1(1/2-p(1+2s_1))}{(1+2s_1)(1/2+2s_1)}},
\end{aligned}$$

where $\Pi_l f$ denotes the projection of the function f to the resolution level l, i.e.

$$\Pi_l f = \sum_{k=1}^{2^l} f_{lk}\psi_{l,k},$$

see [10, (3.1), (3.2)], and $\widehat{f}_{n/m}^{(i)}$ is the wavelet estimate of f in the ith local machine using observations $X^{(i,1)}$, see [10, top of p. 6].

Using the test function above, we define the smoothness estimate at each local machine as

$$\hat{s}_{n/m}^{(i)} = \begin{cases} s_2, & \text{if } \Psi_{n/m}^{(i)} = 0, \\ s_1, & \text{if } \Psi_{n/m}^{(i)} = 1. \end{cases}$$

In each local model we take the first $(L^2 n)^{1/(1+2\hat{s}_{n/m}^{(i)})}$ coefficients in the second subset of observations in the sequence representation, i.e.

$$X_{jk}^{(i,2)} \quad \text{with } 2^j + k \le (L^2 n)^{\frac{1}{1+2\hat{s}_{n/m}^{(i)}}}.$$

Since these numbers might note have a finite binary representation we transmit their approximations $Y_{jk}^{(i)}$ following Algorithm 1, in Section 4.

Let us denote by $\widetilde{N}$ the median of the values $(L^2 n)^{1/(1+2\hat{s}_{n/m}^{(i)})}$, $i = 1, \dots, m$ and $\hat{s}$ the corresponding regularity estimator at the central machine. The central machine then constructs the estimator $\widehat{f}$ as the average of the transmitted observations (for the first $\widetilde{N}$ coefficient), i.e.

$$\widehat{f}_{n,jk} = \begin{cases} \frac{1}{|M_{jk}|} \sum_{i \in M_{jk}} Y_{jk}^{(i)}, & \text{for } 2^j + k \le \widetilde{N}, \\ 0, & \text{for } 2^j + k > \widetilde{N}, \end{cases}$$

where M_{jk} is the collection of local machines satisfying $2^j + k \le (L^2 n)^{1/(1+2\hat{s}_{n/m}^{(i)})}$, i.e. the machines from which the local approximations $Y_{jk}^{(i)}$ are transmitted.

We state below that the above described procedure achieves the adaptive rate and transmits the minimum number of required bits $(L^2 n)^{1/(1+2s)}$ (up to a logarithmic factor).

Theorem 2.2. *For arbitrary $0 < s_1 < s_2 \le 1/(4p) - 1/2$ and $m \ge 5 \log n$ there exists a distributed estimator $\widehat{f}$ with number of transmitted bits $(\widehat{B}^{(1)}, \dots, \widehat{B}^{(m)})$, such that $\widehat{B}^{(i)} \le (L^2 n)^{1/(1+2s_1)} \log n$, $i = 1, \dots, m$, and for all $s \in \{s_1, s_2\}$, we have*

$$\max_{i \in \{1,\dots,m\}} \sup_{f_0 \in B_{2,\infty}^s(L)} E_{f_0}^{(i)} \widehat{B}^{(i)} \lesssim (L^2 n)^{\frac{1}{1+2s}} \log n,$$

$$\sup_{f_0 \in B_{2,\infty}^s(L)} E_{f_0} \|\widehat{f} - f_0\|_2^2 \lesssim L^{\frac{2}{1+2s}} n^{-\frac{2s}{1+2s}}.$$

Proof. See Section 3.2. ∎

The difficulty, as also discussed above, arises from the higher noise level in the local problems which results in less accurate tests between the smoothness classes. The existence of an estimator which can achieve adaptation (in a limited range of smoothness classes) is a consequence of the difference between the nonparametric

testing and estimation rates in the case of the L_2-norm, see for instance [14, 16]. Since one can test between smoothness classes with a faster rate than the corresponding estimation rate, it can compensate (up to some extent) for the higher local noise level m/n.

The preceding result can be extended to a scale of smoothness classes as well. Let us consider the collection of regularity classes $s_0 \in [s_1, s_2]$, where s_0 denotes the regularity of the truth we want to adapt to. The idea behind the approach is to introduce a fine enough grid of regularities in the interval $[s_1, s_2]$ and test between which two grid points the true regularity lies. Then one can apply the distributed method introduced above to these two grid points.

More concretely, similarly to the previous, simpler setting, we divide the data in each machine to two independent samples $X^{(i,1)}$ and $X^{(i,2)}$. Let $\mathcal{S}_n$ denote a $1/\log n$-grid of the interval $[s_1, s_2]$, i.e. $\mathcal{S}_n = \{s_1, s_1 + 1/\log n, \dots, s_2\}$. We will describe next a testing procedure for the regularity hyper-parameter s_0. Let us compute the test $\Psi^{(i)}_{n/m}(M_{n,t}^{-1}, t, s)$ for all $t < s, s, t \in \mathcal{S}_n$ and take $\widehat{s}^{(i)}_{n/m}$ to be the largest regularity s for which the null hypothesis was retained for every $t < s$, i.e.

$$\widehat{s}^{(i)}_{n/m} = \max\{s \in \mathcal{S}_n : \Psi^{(i)}_{n/m}(M_{n,t}^{-1}, t, s) = 0, \ \forall t < s\}. \tag{2.5}$$

The aggregated regularity estimator $\widehat{s}$ and the distributed estimator $\widehat{f}$ is then constructed the same way as above, but using the estimator $\widehat{s}^{(i)}_{n/m}$ given in (2.5). We state below that this procedure adapts both to the minimax risk and the minimum number of required communication.

Corollary 2.3. *Assume that $m = n^p$ for some $0 < p < 1/2$, then for arbitrary $0 < s_1 < s_2 < 1/(4p) - 1/2$ and $m \geq 5 \log n$ there exists a distributed estimator $\widehat{f}$ transmitting $\widehat{B}^{(i)}$ bits in the local machines $i = 1, \dots, m$ satisfying that $\widehat{B}^{(i)} \leq (L^2 n)^{1/(1+2s_1)} \log n$ and*

$$\max_{i=1,\dots,m} \sup_{s\in[s_1,s_2]} \sup_{f_0 \in B^s_{2,\infty}(L)} \frac{E^{(i)}_{f_0} \widehat{B}^{(i)}}{(L^2 n)^{1/(1+2s)} \log n} \lesssim 1,$$

$$\sup_{s\in[s_1,s_2]} \sup_{f_0 \in B^s_{2,\infty}(L)} \frac{E_{f_0} \|\widehat{f} - f_0\|_2^2}{L^{2/(1+2s)} n^{-2s/(1+2s)}} \lesssim 1.$$

Proof. See Section 3.3. ∎

Remark 2.4. We note that our procedure considers a fixed $L > 0$. In principle one could allow L to tend to infinity arbitrarily slowly and then our procedure would cover all regularity balls with constant radius for a price of an asymptotically negligible term. Alternatively one could modify our procedure to adapt to the radius L as well. The difficulty is that f_0 can belong to multiple Besov balls $B^s_{2,\infty}(L)$ corresponding

to different pairs of radius and regularity hyper-parameters and our goal is to adapt to the most appropriate one. We describe below our approach.

Let us assume that $L \in [1, n]$ and take the dyadic grid of this interval. For each grid point $L_j = 2^j$, $j = 0, \dots, \lfloor \log n \rfloor$, we construct the local regularity estimators $\hat{s}^{(i)}_{n/m}(L_j)$ following our approach described above and based on them we can construct the global estimators for the regularity and the functional parameter. All these pairs satisfy the guarantees given in Theorem 2.3 for radius L_j and corresponding regularity hyper-parameter s_j. After obtaining these $\log n$ pairs of local estimators $(L_j, \hat{s}^{(i)}_{n/m}(L_j))$ we simply transmit the local approximations of the first

$$\min_j (L_j^2 n)^{\frac{1}{1+2\hat{s}^{(i)}_{n/m}(L_j)}}$$

wavelet coefficients. Then by slightly adapting the aggregation technique described earlier we obtain an optimal procedure up to a logarithmic factor.[1]

2.2. Adaptation in L_∞

Next we deal with the L_∞-norm case. Here we show that in contrast to the L_2-case, adaptation is not possible even on a limited range of smoothness classes. The reason behind it is that in this case the minimax testing and estimation rates are the same and hence there is no room left to compensate for the higher local noise level.

Theorem 2.5. *Take any $0 < s_1 < s_2$ and assume that $m = n^p$, with $p \in (0, 1/2)$. Then there does not exist a distributed estimator $\widehat{f}$ with transmitted bits $\widehat{B}^{(i)} \le n^{1/(1+2s_1)+\varepsilon_1}$, $i = 1, \dots, m$, satisfying, for $\ell = 1, 2$, that*

$$\max_{i=1,\dots,m} \sup_{f_0 \in B^{s_\ell}_{\infty,\infty}(L)} E^{(i)}_{f_0} \widehat{B}^{(i)} \lesssim (L^2 n)^{\frac{1}{1+2s_\ell}+\varepsilon_1}, \tag{2.6}$$

$$\sup_{f_0 \in B^{s_\ell}_{\infty,\infty}(L)} E_{f_0} \|\widehat{f} - f_0\|_\infty \lesssim L^{\frac{1}{1+2s_l}+\varepsilon_2} (n/\log n)^{-\frac{s_\ell}{1+2s_\ell}+\varepsilon_2}, \tag{2.7}$$

for some sufficiently small $\varepsilon_1, \varepsilon_2 > 0$.

Proof. See Section 3.4. ∎

Next we introduce some additional restriction on the true function of interest under which adaptation is possible in the distributed setting. To do so, we consider the so-called self-similarity assumption, where loosely speaking we assume that the true

[1] For an alternative approach reaching adaptation in a somewhat modified and extended setting we refer to the recent manuscript [9] submitted after our paper.

function has similar smoothness at every resolution level. This will allow us to estimate the regularity s of the functional parameter of interest and therefore transmit the right amount of bits from the local machines to the central one.

We first introduce necessary notation. Let ψ_{jk} be the wavelet basis functions described in Appendix A. For $f \in L_2[0,1]$ and natural numbers $j_1 \leq j_2$ we define

$$f_{[j_1,j_2]} = \sum_{j=j_1}^{j_2} \sum_{k=1}^{2^j} f_{jk}\psi_{jk}.$$

Then following [5] we say that the function $f \in B^s_{\infty,\infty}(L)$ belongs to the self-similar class $S^s_\infty(L, \varepsilon, j_0, \rho)$ if

$$\|f_{[j,\rho j]}\|_{B^s_{\infty,\infty}} \geq \varepsilon L \quad \text{for } j \geq j_0 \text{ and } \rho > 1. \tag{2.8}$$

The self-similarity property was introduced (amongst other places) in the context of adaptive confidence bands. It was shown that under self-similarity one can construct adaptive L_∞ confidence bands whose size also adapts to the level of regularity, see for instance [5, 13, 19]. The underlying idea is the same as here. Under this assumption one can provide a consistent estimator for the smoothness and based on that construct the band corresponding the function class.

The following theorem shows that under the self-similarity assumption there exists a distributed method which adapts to regularity and at the same time transmits the minimal amount of bits (again up to logarithmic factors).

Theorem 2.6. *Consider the distributed Gaussian white noise model with $m \leq n^\delta$, for some $\delta \in (0,1)$ and assume that $f_0 \in B^s_{\infty,\infty}(L)$ for some $s \in [s_1, s_2]$ (where $0 < s_1 < s_2$ are arbitrary). Then there exists a distributed method such that the number of transmitted bits satisfies $\widehat{B}^{(i)} \leq (L^2 n/\log n)^{1/(1+2s_1)} \log n$ and*

$$\max_{i\in\{1,\dots,m\}} \sup_{s\in[s_1,s_2]} \sup_{f_0\in S^s_\infty(L,\varepsilon,j_0,\rho)} \frac{E^{(i)}_{f_0}\widehat{B}^{(i)}}{(L^2 n)^{\frac{1}{1+2s}}(\log n)^{\frac{2s}{1+2s}}} \lesssim 1,$$

$$\sup_{s\in[s_1,s_2]} \sup_{f_0\in S^s_\infty(L,\varepsilon,j_0,\rho)} \frac{E_{f_0}\|\widehat{f} - f_0\|_\infty}{L^{\frac{1}{1+2s}}(n/\log n)^{-\frac{s}{1+2s}}} \lesssim 1.$$

Proof. See Section 3.5. ∎

3. Proofs

3.1. Proof of Theorem 2.1

We argue by contradiction. We assume that the inequalities (2.1) and (2.2) hold. Then we construct a finite but large enough set $\mathcal{F}_0 \subset B^{s_1}_{2,\infty}(L)$ such that there does not exist

a consistent test between the elements of the set and the zero function, which clearly belongs to the smoother class $B^{s_2}_{2,\infty}(L)$. Using this non-existence result we arrive to contradiction with our assumptions.

As a first step we construct the set $\mathcal{F}_0$. Let us introduce the following notations

$$\widetilde{\delta}_n = \overline{\delta}_n \wedge (L^2 n/m)^{-\frac{1+2s_1}{1/2+2s_1}} (L^2 n)^{-\varepsilon_3}, \tag{3.1}$$

with

$$\begin{aligned}
\overline{\delta}_n &= L^{-2} \min\left\{ \frac{m}{n \log n}, \frac{1}{n[\overline{\delta}_n^{1/(1+2s_1)} b_n \log n \wedge 1]} \right\}, \\
b_n &= \left(\Gamma_n \vee (L^2 n)^{\varepsilon_1 - \frac{(s_1+1/4)\varepsilon_3}{1+2s_1}} (L^2 n)^{\frac{1}{1+2s_1}} \right) \log n, \\
\Gamma_n &= (L^2 n)^{\frac{1/2}{1+2s_1} + \frac{1/2}{1+2s_2} + \varepsilon_1},
\end{aligned}$$

and constants

$$\varepsilon_3 \in \left(0, \frac{p(1+2s_1) - 1/2}{1/2 + 2s_1}\right),$$

where $p(1+2s_1) - 1/2 > 0$ follows from the assumption $s_1 > 1/(4p) - 1/2$, and

$$\varepsilon_1 \in \left(0, \frac{s_2 - s_1}{(1+2s_1)(1+2s_2)} \wedge \frac{(s_1+1/4)\varepsilon_3}{1+2s_1}\right).$$

Note that $b_n \le (L^2 n)^{1/(1+2s_1)-\varepsilon_4} \log n$, with

$$\varepsilon_4 = \left(\frac{s_2 - s_1}{(1+2s_1)(1+2s_2)} - \varepsilon_1 \right) \wedge \left(\frac{(s_1+1/4)\varepsilon_3}{1+2s_1} - \varepsilon_1 \right) > 0.$$

In view of the definition of $\overline{\delta}_n$ this implies that

$$\overline{\delta}_n \ge \frac{(L^2 n)^{\varepsilon_4 \frac{1+2s_1}{2+2s_1}}}{L^2 n \log n} \gg (L^2 n)^{-1+\frac{\varepsilon_4}{2}}.$$

Furthermore,

$$\begin{aligned}
(L^2 n/m)^{-\frac{1+2s_1}{1/2+2s_1}} (L^2 n)^{-\varepsilon_3} &= (L^2 n)^{-(1-p)\frac{1+2s_1}{1/2+2s_1} - \varepsilon_3} \\
&= (L^2 n)^{-1} (L^2 n)^{\frac{p(1+2s_1)-1/2}{1/2+2s_1} - \varepsilon_3}.
\end{aligned}$$

Therefore, we can conclude that for large enough n, we obtain

$$\widetilde{\delta}_n \ge (L^2 n)^{-1+\varepsilon_5} \quad \text{with } \varepsilon_5 = (\varepsilon_4/2) \wedge \left(\frac{p(1+2s_1) - 1/2}{1/2 + 2s_1} - \varepsilon_3 \right) > 0. \tag{3.2}$$

The elements $f \in \mathcal{F}_0$ are then defined with the wavelet coefficients as

$$f_{jk} = \begin{cases} L\beta_k \widetilde{\delta}_n^{1/2}, & \text{if } j = j_n := \lfloor \frac{\log \widetilde{\delta}_n^{-1}}{1+2s_1} \rfloor, k = 1, \dots, 2^{j_n}, \\ 0, & \text{else,} \end{cases} \tag{3.3}$$

where $\beta_k \in \{-1, 1\}$. It is easy to check that $\mathcal{F}_0 \subset B^{s_1}_{2,\infty}(L)$ and besides, for every $f \in \mathcal{F}_0$, in view of the definition of $\widetilde{\delta}_n$,

$$\|0 - f\|_2^2 = \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} f_{jk}^2 = L^2 2^{j_n} \widetilde{\delta}_n \le L^2 \widetilde{\delta}_n^{\frac{2s_1}{1+2s_1}} = o\Big(L^{\frac{1}{1/2+2s_1}} (n/m)^{-\frac{2s_1}{1/2+2s_1}} \Big).$$

Next we take the average likelihood ratio over the class $\mathcal{F}_0$:

$$Z = \frac{1}{|\mathcal{F}_0|} \sum_{f \in \mathcal{F}_0} \frac{dP_f^{(i)}}{dP_0^{(i)}}, \quad \text{where } |\mathcal{F}_0| = 2^{j_n}.$$

In view of [14, (6.23)],

$$\inf_{\Psi^{(i)}} \left\{ E_0^{(i)} \Psi^{(i)} + \frac{1}{|\mathcal{F}_0|} \sum_{f \in \mathcal{F}_0} E_f^{(i)} (1 - \Psi^{(i)}) \right\} \ge (1 - \eta_n) \left(1 - \frac{\sqrt{E_0^{(i)} (Z-1)^2}}{\eta_n} \right), \tag{3.4}$$

for every $\eta_n \in (0, 1)$, where the infimum is taken over all local tests in the local problems. Furthermore, one can show by following the steps in the proof of Theorem 6.2.11 (c) on pages 493–494 of [14] (with $\gamma'_{L^2 n/m} = c_0^2 (n/m) \widetilde{\delta}_n$ and $\gamma_{n/m} = (L^2 n/m) \widetilde{\delta}_n^{(1/2+2s_1)/(1+2s_1)} \le (L^2 n)^{-((1/2+2s_1)/(1+2s_1))\varepsilon_3}$) that

$$E_0^{(i)} (Z-1)^2 \le \exp\{c' \gamma_{n/m}^2\} - 1 \lesssim \gamma_{n/m}^2 \lesssim (L^2 n)^{-\frac{1+4s_1}{1+2s_1}\varepsilon_3}.$$

By choosing $\eta_n = (L^2 n)^{-((1/4+s_1)\varepsilon_3)/(1+2s_1)}$, we get that

$$\inf_{\Psi^{(i)}} \left\{ E_0^{(i)} \Psi^{(i)} + \frac{1}{|\mathcal{F}_0|} \sum_{f \in \mathcal{F}_0} E_f^{(i)} (1 - \Psi^{(i)}) \right\} \ge (1 - C\eta_n)^2, \tag{3.5}$$

for some large enough constant $C > 0$, concluding the proof of the non-existence of consistent tests between $\mathcal{F}_0$ and the zero function.

Next we show that (3.5) contradicts our assumptions. Let us define the test

$$\Psi^{(i)} = 1_{\widehat{B}^{(i)} \ge \Gamma_n}.$$

First note that following from Markov's inequality and assumption (2.1),

$$E_0^{(i)} \Psi^{(i)} = P_0^{(i)} (\widehat{B}^{(i)} \ge \Gamma_n) \le E_0^{(i)} (\widehat{B}^{(i)}) / \Gamma_n \le (L^2 n)^{\frac{1/2}{1+2s_2} - \frac{1/2}{1+2s_1}} = o(1).$$

Therefore in view of (3.5), we have that

$$\frac{1}{|\mathcal{F}_0|}\sum_{f\in\mathcal{F}_0} P_f^{(i)}(\widehat{B}^{(i)} < \Gamma_n) = \frac{1}{|\mathcal{F}_0|}\sum_{f\in\mathcal{F}_0} E_f^{(i)}(1-\Psi^{(i)}) \geq (1-C\eta_n)^2 - (L^2 n)^{\frac{1/2}{1+2s_2}-\frac{1/2}{1+2s_1}}.$$

As a consequence and in view of assumption $\widehat{B}^{(i)} \leq C(L^2n)^{1/(1+2s_1)+\varepsilon_1}\log n$, we have

$$\frac{1}{|\mathcal{F}_0|}\sum_{f\in\mathcal{F}_0} E_f^{(i)}\widehat{B}^{(i)} \lesssim \Gamma_n + (L^2n)^{\frac{1}{1+2s_1}+\varepsilon_1}(\log n)\Big(\eta_n + (L^2n)^{\frac{1/2}{1+2s_2}-\frac{1/2}{1+2s_1}}\Big) \lesssim b_n.$$

This means that the expected number (with respect to the joint distribution of the variables F and P_f, $f\in\mathcal{F}_0$) of transmitted bits on the class $\mathcal{F}_0$ is bounded from above by a multiple of b_n. So the distributed estimator satisfies assertion (B.7) in the proof of Theorem B.3 with $B^{(i)}$ replaced by Cb_n. Hence in view of the minimax lower bound derived in assertion (B.9) and the definition of $\widetilde{\delta}_n$ (with $B^{(i)}$ replaced by b_n in the definition of δ_n in the proof of Theorem B.3)

$$\sup_{f_0\in\mathcal{F}_0} E_{f_0}\|\widehat{f}-f_0\|_2^2 \gtrsim L^2\widetilde{\delta}_n^{\frac{2s_1}{1+2s_1}} \gg L^{\frac{2}{1+2s_1}+\varepsilon_2}n^{-\frac{2s_1}{1+2s_1}+\varepsilon_2},$$

with $\varepsilon_2 = 2\varepsilon_5 s_1/(1+2s_1)$, where the last inequality follows from (3.2). This contradicts assumption (2.2), finishing the proof of our statement.

3.2. Proof of Theorem 2.2

Let us then denote by $P_{X^{(i,1)}}$ and $P_{X^{(i,2)}}$ the distribution of the first and second subset of observations, respectively, and by $P_{X^{(i,2)}|X^{(i,1)}}$ the conditional distribution of the second subset given the first. The corresponding expected values are denoted by $E_{X^{(i,1)}}$, $E_{X^{(i,2)}}$, and $E_{X^{(i,2)}|X^{(i,1)}}$, respectively. Furthermore, let us introduce the notations $X_l = (X^{(1,l)},\dots,X^{(m,l)})$, $l=1,2$ and denote by P_{X_l} and E_{X_l} the corresponding probability distributions and expected values. Finally, we also note that we took $z_0 = 1$ in the test proposed in [10, Section 3] (since for notational convenience we take $J_0 = 0$, see Section A, hence we have $z_0 = 2^{J_0} = 1$).

Let us introduce the notation

$$R_\alpha^{s_1}(L) = \Big\{f\in B_{2,\infty}^{s_1}(L) : \|f - B_{2,\infty}^{s_2}(L)\|_2 \geq \widetilde{C}_\alpha L^{\frac{1/2}{1/2+2s_1}}(n/m)^{-\frac{s_1}{1/2+2s_1}}\Big\}.$$

In view of Lemma 4.4, we have for all $\alpha\in(0,1)$ and $0<m\leq n$ the test given in (2.4) satisfies that

$$\sup_{f\in B_{2,\infty}^{s_2}(L)} E_{X^{(i,1)}}\Psi_{n/m}^{(i)} + \sup_{f\in R_\alpha^{s_1}(L)} E_{X^{(i,1)}}(1-\Psi_{n/m}^{(i)}) \leq ce^{-0.5/\sqrt{\alpha}}, \tag{3.6}$$

with

$$\widetilde{C}_\alpha = 24\left(\frac{2^{s_1}}{\sqrt{1-2^{-2s_1}}} + 19\right)2^{\frac{s_1}{1+2s_1}}/\sqrt{\alpha}$$

and c not depending on α, n, m. Let $M_n = n^{\frac{2s_1(1/2-p(1+2s_1))}{(1+2s_1)(1/2+2s_1)}}$ tending to infinity (where the positivity of the exponent follows from the assumption $s_1 < 1/(4p) - 1/2$). Then the above test $\Psi^{(i)}_{n/m}$ (with $\alpha = M_n^{-1}$) is consistent in each local problem between the hypotheses

$$H_0 \colon f \in B^{s_2}_{2,\infty}(L) \quad \text{vs.} \quad H_1 \colon f \in R^{s_1}_{M_n^{-1}}(L).$$

Note that in view of Lemma 4.2 (with $\mu = f_{0,jk}$) we have that $l(Y^{(i)}_{jk}) \leq \log n$ with approximation error

$$|\varepsilon^{(i)}_{jk}| = |X^{(i,2)}_{jk} - Y^{(i)}_{jk}| \leq n^{-1/2}$$

on a set $\mathcal{E}^{(i)}_{jk}$ with $P_{X^{(i,2)}}((\mathcal{E}^{(i)}_{jk})^c) \leq e^{-c'n}$, for some $c' > 0$. Let us then introduce the notation

$$\mathcal{E} = \cap_{i=1}^m \cap_{j=0}^{\log n} \cap_{k=1}^{2^j} \mathcal{E}^{(i)}_{jk} \tag{3.7}$$

and note that $P_{X_2}(\mathcal{E}^c) \leq n^2 e^{-c'n} \lesssim e^{-cn}$, for any $0 < c < c'$. Hence, the number of transmitted bits conditioned on the first subsample $X^{(i,1)}$ is bounded from above by $l(Y^{(i)}) \leq (L^2 n)^{1/(1+2\widehat{s}^{(i)}_{n/m})} \log n$ almost surely.

We show that this procedure achieves the minimax convergence rate and transmits the optimal amount of bits (up to a logarithmic factor). First note that $\widehat{B}^{(i)} \lesssim (L^2 n)^{1/(1+2s_1)} \log n$ follows immediately by construction. Then recall that the test $\Psi^{(i)}_{n/m}$ is consistent, hence

$$\sup_{f \in B^{s_2}_{2,\infty}(L)} P_{X^{(i,1)}}(\widehat{s}^{(i)}_{n/m} = s_1) \leq C e^{-M_n^{1/2}/2}$$

and

$$\begin{aligned}
\sup_{f \in B^{s_2}_{2,\infty}(L)} E_{X^{(i,1)},X^{(i,2)}} \widehat{B}^{(i)} &\leq \sup_{f \in B^{s_2}_{2,\infty}(L)} E_{X^{(i,1)}} (L^2 n)^{\frac{1}{1+2\widehat{s}^{(i)}_{n/m}}} \log n \\
&\leq (L^2 n)^{\frac{1}{1+2s_2}} \log n + C e^{-M_n^{1/2}/2} (L^2 n)^{\frac{1}{1+2s_1}} \log n \\
&\leq (1 + o(1))(L^2 n)^{\frac{1}{1+2s_2}} \log n,
\end{aligned}$$

verifying that the number of transmitted bits is indeed optimal.

Next we provide optimal upper bounds for the risk. First let us consider the case $f \in B^{s_2}_{2,\infty}(L) \cup R^{s_1}_{M_n^{-1}}(L)$, where the estimator $\hat{s}^{(i)}_{n/m}$ is consistent, i.e.

$$\hat{s}^{(i)}_{n/m} = s_1 \quad \text{for } f \in R^{s_1}_{M_n^{-1}}(L),$$
$$\hat{s}^{(i)}_{n/m} = s_2 \quad \text{for } f \in B^{s_2}_{2,\infty}(L),$$

with $P_{X^{(i,1)}}$-probability at least $1 - ce^{-M_n^{1/2}/2}$. Let us introduce the notation M for the number of machines in $\{1, \dots, m\}$, where $\hat{s}^{(i)}_{n/m} \neq s_l$, $l = 2, 1$, for $f \in B^{s_2}_{2,\infty}(L)$ or $f \in R^{s_1}_{M_n^{-1}}(L)$, respectively. Note that M has a binomial distribution with parameters m and $p \le ce^{-M_n^{1/2}/2}$. Then by Hoeffding's inequality

$$\sup_{f \in R^{s_1}_{M_n^{-1}}(L)} P_{X_1}\big(\widetilde{N} \neq (L^2 n)^{\frac{1}{1+2s_1}}\big) + \sup_{f \in B^{s_2}_{2,\infty}(L)} P_{X_1}\big(\widetilde{N} \neq (L^2 n)^{\frac{1}{1+2s_2}}\big)$$
$$\le P\big(M \ge m/2\big) < e^{-m/5}. \tag{3.8}$$

Then in view of the almost sure inequality $\widetilde{N} \le (L^2 n)^{1/(1+2s_1)}$, we have that

$$\begin{aligned}
\sup_{f \in R^{s_1}_{M_n^{-1}}(L)} E_{X_1} \widetilde{N}^{-2s_1} &= (L^2 n)^{-\frac{2s_1}{1+2s_2}} P_{X_1}(M \ge m/2) \\
&\quad + (L^2 n)^{-\frac{2s_1}{1+2s_1}} P_{X_1}(M < m/2) \\
&\le (1 + o(1))(L^2 n)^{-\frac{2s_1}{1+2s_1}}, \qquad (3.9) \\
\sup_{f \in B^{s_2}_{2,\infty}(L)} E_{X_1} \widetilde{N} &= (L^2 n)^{\frac{1}{1+2s_2}} P_{X_1}(M < m/2) \\
&\quad + (L^2 n)^{\frac{1}{1+2s_1}} P_{X_1}(M \ge m/2) \\
&\le (L^2 n)^{1/(1+2s_2)} + (L^2 n)^{1/(1+2s_1)} e^{-m/5} \\
&\le (1 + o(1))(L^2 n)^{1/(1+2s_2)}, \qquad (3.10)
\end{aligned}$$

for $m \ge 5 \log(L^2 n) \ge \frac{10(s_2 - s_1)}{(2s_1+1)(2s_2+1)} \log(L^2 n)$.

Then similarly to the proof of Theorem B.2 (with m replaced by $|M_{jk}|$), we get on the set $\mathcal{E}$ (with $P_{X_2}(\mathcal{E}^c) \le e^{-cn}$), that

$$\hat{f}_{n,jk} = f_{0,jk} + \frac{1}{\sqrt{n}} Z_{jk} + \varepsilon_{jk},$$

with

$$Z_{jk} \overset{\text{iid}}{\sim} N\big(0, \sqrt{2m/|M_{jk}|}\big) \quad \text{and} \quad |\varepsilon_{jk}| \le n^{-1/2}.$$

Also note that $|\widehat{f}_{n,k}| \le \sqrt{n}$, since $|Y_{jk}^{(i)}| \le \sqrt{n}$ for all i, j, k. Using this reformulation of the estimator and the notation $\widetilde{j}_n = \lfloor \log \widetilde{N} \rfloor$, we get that

$$\begin{aligned}
\sup_{f \in B_{2,\infty}^{s_l}(L)} E_{X_2|X_1} \|\widehat{f} - f_0\|_2^2 1_{\mathcal{E}} &\le \sum_{j \ge \widetilde{j}_n} \sum_{k=1}^{2^j} f_{0,jk}^2 + \sum_{j=0}^{\widetilde{j}_n} \sum_{k=1}^{2^j} E\Big(\frac{1}{\sqrt{n}} Z_{jk} + \varepsilon_{jk}\Big)^2 1_{\mathcal{E}} \\
&\le \sum_{j \ge \widetilde{j}_n} 2^{-2js_l} \sup_{j \ge \widetilde{j}_n} 2^{2js_l} \sum_{k=1}^{2^j} f_{0,jk}^2 + \sum_{j=0}^{\widetilde{j}_n} \sum_{k=1}^{2^j} \frac{2E(Z_{jk}^2)}{n} + \frac{2}{n} \\
&\le L^2 2^{-2j_n s_l} + (2^{\widetilde{j}_n + 2} + 2)/n \asymp L^2 \widetilde{N}^{-2s_l} + \widetilde{N}/n, \\
\sup_{f \in B_{2,\infty}^{s_l}(L)} E_{X_2|X_1} \|\widehat{f} - f_0\|_2^2 1_{\mathcal{E}^c} &\le P_{X_2}(\mathcal{E}^c) 2^{\widetilde{j}_n + 1}(n + L^2) = o(n^{-1}), \qquad (3.11)
\end{aligned}$$

for $l = 1, 2$. Therefore, in view of assertion (3.9),

$$\begin{aligned}
\sup_{f \in B_{2,\infty}^{s_2}(L)} E_{X_1,X_2} \|\widehat{f} - f_0\|_2^2 &\lesssim \sup_{f \in B_{2,\infty}^{s_2}(L)} E_{X_1}\big(L^2 \widetilde{N}^{-2s_2} + \widetilde{N}/n\big) \\
&\lesssim L^{\frac{2}{1+2s_2}} n^{-2s_2/(1+2s_2)}, \\
\sup_{f \in R_{M_n^{-1}}^{s_1}(L)} E_{X_1,X_2} \|\widehat{f} - f_0\|_2^2 &\lesssim \sup_{f \in R_{M_n^{-1}}^{s_1}(L)} E_{X_1}\big(L^2 \widetilde{N}^{-2s_1} + \widetilde{N}/n\big) \\
&\lesssim L^{\frac{2}{1+2s_1}} n^{-2s_1/(1+2s_1)}.
\end{aligned}$$

It remains to deal with the intermediate set, i.e. $f_0 \in B_{2,\infty}^{s_1}(L) \backslash R_{M_n^{-1}}^{s_1}(L)$. Our local estimator $\widehat{s}_{n/m}^{(i)}$ will be either s_1 or s_2, hence for each machine the amount of transmitted bits is bounded from above by

$$(L^2 n)^{\frac{1}{1+2\widehat{s}_n^{(i)}}} \log n \le (L^2 n)^{\frac{1}{1+2s_1}} \log n$$

$P_{X^{(i,2)}}$-almost surely. Note that the median $\widetilde{N}$ also satisfies almost surely that

$$(L^2 n)^{\frac{1}{1+2s_1}} \ge \widetilde{N} \ge (L^2 n)^{\frac{1}{1+2s_2}}.$$

Then, using the notation $f_{0,j \le \widetilde{j}_n} = \sum_{j=0}^{\widetilde{j}_n} f_{0,jk} \psi_{jk}$, we get similarly to above, that

$$\begin{aligned}
E_{X_1,X_2} \|\widehat{f} - f_{0,j \le \widetilde{j}_n}\|_2^2 &\le E_{X_1} \sum_{j=0}^{\widetilde{j}_n} \sum_{k=1}^{2^j} E_{X_2|X_1}\Big(\frac{1}{\sqrt{n}} Z_{jk} + \varepsilon_{jk}\Big)^2 + o(n^{-1}) \\
&\lesssim E_{X_1} \widetilde{N}/n \le L^{\frac{2}{1+2s_1}} n^{-\frac{2s_1}{1+2s_1}}. \qquad (3.12)
\end{aligned}$$

To deal with the bias term let us denote by $\widetilde{f} \in B^{s_2}_{2,\infty}(L)$ a function satisfying

$$\|f_0 - \widetilde{f}\|_2^2 \lesssim \widetilde{C}^2_{M_n^{-1}} L^{\frac{1}{1/2+2s_1}} (n/m)^{-\frac{2s_1}{1/2+2s_1}},$$

then by recalling that

$$(n/m)^{\frac{1}{1/2+2s_1}} = n^{\frac{1-p}{1/2+2s_1}} = n^{\frac{1/2-p(1+2s_1)}{(1+2s_1)(1/2+2s_1)}} n^{\frac{1}{1+2s_1}},$$

we get that

$$\begin{aligned} E_{X_1}\|f_{0,j\le\widetilde{j}_n} - f_0\|_2^2 &\le E_{X_1} \sum_{j=\widetilde{j}_n}^{\infty} \sum_{k=1}^{2^j} f_{0,jk}^2 \\ &\le 2E_{X_1}\left(\sum_{j=\widetilde{j}_n}^{\infty} \sum_{k=1}^{2^j} (f_{0,jk} - \widetilde{f}_{jk})^2 + \sup_{j\ge\widetilde{j}_n} \left(2^{2js_2} \sum_{k=1}^{2^j} \widetilde{f}_{jk}^2 \right) \sum_{j=\widetilde{j}_n}^{\infty} 2^{-2js_2} \right) \\ &\lesssim \widetilde{C}^2_{M_n^{-1}} L^{\frac{1}{1/2+2s_1}} (n/m)^{-\frac{2s_1}{1/2+2s_1}} + E_{X_1} L^2 \widetilde{N}^{-2s_2} \lesssim L^{\frac{2}{1+2s_1}} n^{-\frac{2s_1}{1+2s_1}}, \quad (3.13) \end{aligned}$$

where the last inequality follows from $\widetilde{C}_{M_n^{-1}} \asymp n^{\frac{s_1(1/2-p(1+2s_1))}{(1+2s_1)(1/2+2s_1)}}$. Then by combining (3.12) and (3.13), we get that

$$E_{X_1,X_2}\|\widehat{f} - f_0\|_2^2 \lesssim L^{\frac{2}{1+2s_1}} n^{-\frac{2s_1}{1+2s_1}},$$

concluding the proof of the theorem. ∎

3.3. Proof of Corollary 2.3

Let us introduce the notation $\underline{s} = s_1 + \gamma_n/\log n$ for some $0 \le \gamma_n \le \lceil (s_2 - s_1)\log n \rceil$, $\gamma_n \in \mathbb{N}$, the lower bound of the $1/\log n$-bin containing s_0, i.e. $s_0 \in [\underline{s}, \underline{s} + 1/\log n]$. Then the probability of under smoothing is bounded from above by

$$(\gamma_n - 1)^2 \le (s_2 - s_1)^2 \log^2 n$$

times the probability of rejecting the correct null-hypothesis. Hence, in view of assertion (3.6) and the monotone decreasing property of the function $s \mapsto M_{n,s}$, we get that

$$P\left(\widehat{s}^{(i)}_{n/m} < \underline{s}\right) \lesssim (s_2 - s_1)^2 (\log n)^2 e^{-M_{n,s_2}^{1/2}/2} = o(1).$$

This implies, for all $i \in \{1, \dots, m\}$, that

$$\begin{aligned} E_{X^{(i,1)},X^{(i,2)}} \widehat{B}^{(i)} &= E_{X^{(i,1)}} \widehat{B}^{(i)} \le E_{X^{(i,1)}} (L^2 n)^{\frac{1}{1+2\widehat{s}^{(i)}_{n/m}}} \log n \\ &\lesssim (L^2 n)^{\frac{1}{1+2\underline{s}}} \log n + (L^2 n)^{\frac{1}{1+2s_1}} e^{-M_{n,s_2}^{1/2}/2} \log^2 n \\ &\lesssim (L^2 n)^{\frac{1}{1+2s_0}} \log n, \end{aligned}$$

and similarly to assertions (3.8) and (3.9) that

$$P_{X_1}(\widehat{s} < \underline{s}) = P_{X_1}\big(\widetilde{N} > (L^2 n)^{\frac{1}{1+2\underline{s}}}\big) \le e^{-m/5} \tag{3.14}$$
$$E_{X_1}\widetilde{N} < (L^2 n)^{\frac{1}{1+2\underline{s}}} + (L^2 n)^{\frac{1}{1+2s_1}} P_{X_1}\big(\widetilde{N} > (L^2 n)^{\frac{1}{1+2\underline{s}}}\big) \lesssim (L^2 n)^{\frac{1}{1+2s_0}},$$

for $m \ge 5 \log n$.

It remains to show that our procedure adapts to the minimax risk. First note that in view of assertion (3.12) and (3.14),

$$\sup_{f_0 \in B^{\underline{s}}_{2,\infty}} E_{X_1}\big(E_{X_2|X_1}\|\widehat{f} - f_{0,j\le \widetilde{j}_n}\|_2^2\big) \le E_{X_1}\widetilde{N}/n \lesssim L^{\frac{2}{1+2s_0}} n^{-\frac{2s_0}{1+2s_0}}.$$

Next, let $j_{n,s} = (1+2s)^{-1}\log(L^2 n)$, then for $\widetilde{j}_n = \lfloor \log \widetilde{N} \rfloor$, we have

$$\begin{aligned} &E_{X_1}\big(\|f_{0,j\le\widetilde{j}_n} - f_0\|_2^2\big) \\ &\quad = \Big(\sum_{s<\underline{s},\, s\in\mathcal{S}_n} + \sum_{s=\underline{s}} + \sum_{s>\underline{s},\, s\in\mathcal{S}_n}\Big) P_{X_1}(\widehat{s}=s) E_{X_1}\big(\|f_{0,j\le j_{n,s}} - f_0\|_2^2 \mid \widehat{s}=s\big) \\ &\quad = \Big(\sum_{s<\underline{s},\, s\in\mathcal{S}_n} + \sum_{s=\underline{s}} + \sum_{s>\underline{s},\, s\in\mathcal{S}_n}\Big) P_{X_1}(\widehat{s}=s) \sum_{j=j_{n,s}}^{\infty}\sum_{k=1}^{2^j} f_{0,jk}^2. \end{aligned} \tag{3.15}$$

We deal with the three terms on the right-hand side separately. In view of assertion (3.14) and $\|f_0\|_2^2 \le L^2$, we have that

$$\sum_{s<\underline{s}} P_{X_1}(\widehat{s}=s) \sum_{j=j_{n,s}}^{\infty}\sum_{k=1}^{2^j} f_{0,jk}^2 \le L^2 e^{-m/5} = o\big(n^{-\frac{2s_0}{1+2s_0}}\big).$$

Then it is also easy to see that

$$\begin{aligned} P_{X_1}(\widehat{s}=\underline{s}) \sum_{j=j_{n,\underline{s}}}^{\infty}\sum_{k=1}^{2^j} f_{0,jk}^2 &< \sum_{j=j_{n,\underline{s}}}^{\infty} 2^{-2j\underline{s}} \sup_{j\ge j_{n,\underline{s}}} 2^{2j\underline{s}} \sum_{k=1}^{2^j} f_{0,jk}^2 \\ &\le L^2 (L^2 n)^{-\frac{2\underline{s}}{1+2\underline{s}}} \lesssim (L^2 n)^{-\frac{2s_0}{1+2s_0}}. \end{aligned}$$

Then for arbitrary $s > \underline{s}$, $s \in \mathcal{S}_n$, using the notation

$$R^{\underline{s},s}_{M_{n,\underline{s}}^{-1}}(L) := \big\{f \in B^s_{2,\infty}(L) : \|f - B^s_{2,\infty}(L)\|_2 \ge \widetilde{C}_{M_{n,\underline{s}}^{-1}} L^{\frac{1/2}{1/2+2\underline{s}}} (n/m)^{-\frac{\underline{s}}{1/2+2\underline{s}}}\big\},$$

we have that

$$\begin{aligned} \sup_{f_0 \in R^{\underline{s},s}_{M_{n,\underline{s}}^{-1}}(L)} P_{X^{(i,1)}}\big(\widehat{s}^{(i)}_{n/m} \ge s\big) &\le \sup_{f_0 \in R^{\underline{s},s}_{M_{n,\underline{s}}^{-1}}(L)} E_{X^{(i,1)}}\big(1 - \Psi^{(i)}_{n/m}\big(M_{n,\underline{s}}^{-1}, \underline{s}, s\big)\big) \\ &\lesssim e^{-M_{n,\underline{s}}^{1/2}/2}. \end{aligned}$$

Therefore, by Hoeffding's inequality,

$$\sup_{f_0 \in R^{\underline{s},s}_{M^{-1}_{n,\underline{s}}}(L)} P_{X_1}(\widehat{s} \geq s) \leq e^{-m/5}, \tag{3.16}$$

hence by combining the preceding two displays, we get that

$$\sup_{f_0 \in R^{\underline{s},s}_{M^{-1}_{n,\underline{s}}}(L)} \sum_{j=j_{n,s}}^{\infty} \sum_{k=1}^{2^j} f_{0,jk}^2 P_{X_1}(\widehat{s} = s) \leq L^2 e^{-m/5} = o\big(n^{-\frac{2s_0}{1+2s_0}} / \log n\big).$$

For any $f_0 \in \mathcal{F}_s := B^{\underline{s}}_{2,\infty}(L) \backslash R^{\underline{s},s}_{M^{-1}_{n,\underline{s}}}(L)$, there exists an $\widetilde{f}_0 \in B^{\underline{s}}_{2,\infty}(L)$ such that

$$\|f_0 - \widetilde{f}\|_2 \leq \widetilde{C}_{M^{-1}_{n,\underline{s}}} L^{\frac{1/2}{1/2+2\underline{s}}} (n/m)^{-\frac{\underline{s}}{1/2+2\underline{s}}}.$$

Then similarly to assertion (3.13), we get that

$$\begin{aligned}
&\sup_{f_0 \in \mathcal{F}_s} \sum_{j=j_{n,s}}^{\infty} \sum_{k=1}^{2^j} f_{0,jk}^2 \\
&\quad \leq 2 \sup_{f_0 \in \mathcal{F}_s} \Bigg(\sum_{j=j_{n,s}}^{\infty} \sum_{k=1}^{2^j} (f_{0,jk} - \widetilde{f}_{0,jk})^2 + \sum_{j=j_{n,s}}^{\infty} 2^{-2js} \sup_{j \geq j_{n,s}} 2^{2js} \sum_{k=1}^{2^j} \widetilde{f}_{0,jk}^2 \Bigg) \\
&\quad \lesssim \widetilde{C}^2_{M^{-1}_{n,\underline{s}}} L^{\frac{1}{1/2+2\underline{s}}} (n/m)^{-\frac{2\underline{s}}{1/2+2\underline{s}}} + 2^{-2j_{n,s}s} \\
&\quad \lesssim L^{\frac{2}{1+2\underline{s}}} n^{-\frac{2\underline{s}}{1+2\underline{s}}} + L^{\frac{2}{1+2s}} n^{-\frac{2s}{1+2s}} \lesssim L^{\frac{2}{1+2s_0}} n^{-\frac{2s_0}{1+2s_0}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sup_{f_0 \in B^{\underline{s}}_{2,\infty}(L)} \sum_{s>\underline{s}}^{s_2} P_{X_1}(\widehat{s} = s) \sum_{j=j_{n,s}}^{\infty} \sum_{k=1}^{2^j} f_{0,jk}^2 \\
&\quad \lesssim \sum_{s>\underline{s}}^{s_2} \big(P_{X_1}(\widehat{s} = s) + o(1/\log n)\big) L^{\frac{2}{1+2s_0}} n^{-\frac{2s_0}{1+2s_0}} \lesssim L^{\frac{2}{1+2s_0}} n^{-\frac{2s_0}{1+2s_0}}.
\end{aligned}$$

Combining the upper bounds above, we get that

$$\begin{aligned}
&\sup_{f_0 \in B^{\underline{s}}_{2,\infty}(L)} E_{X_1,X_2} \|\widehat{f} - f_0\|_2^2 \\
&\quad \leq 2 \sup_{f_0 \in B^{\underline{s}}_{2,\infty}(L)} \big(E_{X_1} \|f_{0,j \leq \widetilde{j}_n} - f_0\|_2^2 + E_{X_1,X_2} \|\widehat{f} - f_{0,j \leq \widetilde{j}_n}\|_2^2\big) \\
&\quad \lesssim L^{\frac{2}{1+2s_0}} n^{-\frac{2s_0}{1+2s_0}},
\end{aligned}$$

concluding the proof of the corollary. ∎

3.4. Proof of Theorem 2.5

The proof follows the same lines of reasoning as the proof of Theorem 2.1, here we highlight only the differences.

First of all the set of functions $\mathcal{F}_0$ is defined slightly differently. Let us introduce the notations

$$\widetilde{\delta}_n = \overline{\delta}_n \wedge (L^{-2}m/n), \tag{3.17}$$

with

$$\overline{\delta}_n = L^{-2}\min\left\{\frac{m}{n\log m}, \frac{1}{n[\overline{\delta}_n^{1/(1+2s_1)}b_n \wedge 1]\log m}\right\},$$
$$b_n = \left(\Gamma_n \vee (L^2 n)^{\frac{1}{1+2s_1}-\varepsilon_1}\right)\log n \quad \text{and} \quad \Gamma_n = (L^2 n)^{\frac{1/2}{1+2s_1}+\frac{1/2}{1+2s_2}+\varepsilon_1},$$

with

$$\varepsilon_1 \in \left(0, \frac{s_2 - s_1}{(1+2s_1)(1+2s_2)} \wedge \frac{(1-p)/8}{1+2s_1}\right).$$

By elementary computations, one can deduce that $\overline{\delta}_n \geq (L^2 n)^{\varepsilon_1/2-1}$, and therefore

$$\widetilde{\delta}_n \geq (L^2 n)^{(\varepsilon_1/2 \wedge p)-1}. \tag{3.18}$$

Next, let us denote by K_j the largest set of Daubechies wavelets with disjoint supports at resolution level j. Note that $|K_j| \geq c_0 2^j$ (for large enough j and sufficiently small $c_0 > 0$). Then we consider the class of functions

$$\mathcal{F}_0 = \{f_k : k \in K_{j_n}\}, \quad \text{where } f_k = L\widetilde{\delta}_n^{1/2}\psi_{j_n,k}, \ \widetilde{j}_n = -\frac{\log(\widetilde{\delta}_n)}{1+2s_1}. \tag{3.19}$$

Since the functions in $\mathcal{F}_0$ have disjoint supports, we have

$$\sup_{f\in\mathcal{F}_0}\|0-f\|_\infty = \sup_{k\in K_{j_n}} L\widetilde{\delta}_n^{1/2}\|\psi_{j_n,k}\|_\infty \lesssim L2^{j_n/2}\widetilde{\delta}_n^{1/2}$$
$$\lesssim L\widetilde{\delta}_n^{s_1/(1+2s_1)} = o\big(L^{\frac{1}{1+2s_1}}(n/m)^{-\frac{s_1}{1+2s_1}}\big),$$

following from the definition of $\widetilde{\delta}_n$. Hence, it is not possible to test between the zero function and the set $\mathcal{F}_0$ in the local servers.

Using the notation Z for the likelihood ratio introduced in the proof of Theorem 2.1, we note that in view of the proof of Theorem 6.2.11 (b) on page 493 of [14], we have that

$$E(Z-1)^2 \leq (e^{\overline{\gamma}_n^2}-1)/|\mathcal{F}_0|, \quad \text{where } \overline{\gamma}_n = \sqrt{\widetilde{\delta}_n L^2 n/m}.$$

Then the infimum of the tests given in (3.4) is bounded from below by $(1 - C\eta_n)^2$ for $\eta_n = \widetilde{\delta}_n^{1/(4+8s_1)} \le (L^2 n)^{-(1-p)/(4+8s_1)} \le n^{-2\varepsilon_1}$. This leads to

$$\frac{1}{|\mathcal{F}_0|} \sum_{f \in \mathcal{F}_0} E_f^{(i)} \widehat{B}^{(i)} \lesssim \Gamma_n + (L^2 n)^{\frac{1}{1+2s_1}+\varepsilon_1} (\log n)\big(\eta_n + (L^2 n)^{\frac{1/2}{1+2s_2} - \frac{1/2}{1+2s_1}}\big) \lesssim b_n.$$

This means that the expected number (with respect to the joint distribution of the variables F and P_f, $f \in \mathcal{F}_0$) of transmitted bits on the class $\mathcal{F}_0$ is bounded from above by a multiple of b_n. So the distributed estimator satisfies assertion (B.7) in with $B^{(i)}$ replaced by Cb_n. Hence, in view of the minimax lower bound derived in assertion (B.13) (with $B^{(i)}$ replaced by b_n in the definition of δ_n in the proof of Theorem B.5) and the definition of $\widetilde{\delta}_n$, we have

$$\sup_{f_0 \in \mathcal{F}_0} E_{f_0} \|\widehat{f} - f_0\|_\infty \gtrsim L \widetilde{\delta}_n^{\frac{s_1}{1+2s_1}} \gg L^{\frac{1}{1+2s_1}+\varepsilon_2} n^{-\frac{s_1}{1+2s_1}+\varepsilon_2},$$

with $\varepsilon_2 = (\varepsilon_1/2 \wedge p)s_1/(1 + 2s_1)$, where the last inequality followed from (3.18). This contradicts assumption (2.7), finishing the proof of our statement. ∎

3.5. Proof of Theorem 2.6

First note that in Lemma 5.2 of [5], it was shown that the smoothness can be consistently estimated under the self-similarity condition, i.e. there exists an estimator $\widehat{s}_{n/m}^{(i)}$ such that for every $i \in \{1, \dots, m\}$ and $c > 0$ there exists $C > 0$ satisfying

$$\inf_{s \in [s_1, s_2]} \inf_{f_0 \in S_\infty^s(L, \varepsilon, j_0)} P_{f_0}\big(s - C/\log(n/m) \le \widehat{s}_{n/m}^{(i)} \le s\big) \lesssim (m/n)^c. \tag{3.20}$$

By choosing $c = 1/(1 - p)$, we have $(m/n)^c = 1/n$. Then we propose a similar estimation method as in Theorem 2.2. First we split the data into $X^{(i,1)}$ and $X^{(i,2)}$ and use the first sample $X^{(i,1)}$ to construct the estimator $\widehat{s}_{n/m}^{(i)}$ for the smoothness parameter s. Next transmit the approximation of the first $\widetilde{N}^{(i)} = (L^2 n/\log n)^{1/(1+2\widehat{s}_{n/m}^{(i)})}$ coefficients (instead of $(L^2 n)^{1/(1+2\widehat{s}_{n/m}^{(i)})}$ as in Theorem 2.2) of the second subset of observations $X^{(i,2)}$, following Algorithm 1. Then $\widehat{B}^{(i)} \le (L^2 n/\log n)^{1/(1+2s_1)} \log n$ and

$$\begin{aligned} E_{X^{(i,1)}, X^{(i,2)}} \widehat{B}^{(i)} &= E_{X^{(i,1)}} \widehat{B}^{(i)} = E_{X^{(i,1)}} \widetilde{N}^{(i)} \log n \\ &\le (L^2 n/\log n)^{\frac{1}{1+2s}} \log n + n^{-1} (L^2 n/\log n)^{\frac{1}{1+2s_1}} \log n \\ &\lesssim (L^2 n)^{\frac{1}{1+2s}} (\log n)^{\frac{2s}{1+2s}}. \end{aligned}$$

Besides, we also have that the median $\widetilde{N}$ of the values $\widetilde{N}^{(i)}$ satisfy that

$$P_{X_1}\big(L^{2/(1+2s)} n^{1/(1+2s)} \le \widetilde{N} \le C_1 L^{2/(1+2s)} n^{1/(1+2s)}\big) \ge 1 - C_2 e^{-m/5}, \tag{3.21}$$

for some large enough constants $C_1, C_2 > 0$.

Similarly to before, let $\widetilde{j}_n = \lfloor \log \widetilde{N} \rfloor$ and $f_{0,j\le\widetilde{j}_n} = \sum_{j\le\widetilde{j}_n}\sum_{k=1}^{2^j} f_{0,jk}\psi_{jk}$. Then using the notation $\mathcal{E}$ introduced in (3.7), we get that

$$\begin{aligned}
\|\widehat{f} - f_0\|_\infty 1_{\mathcal{E}} &\le \|\widehat{f} - f_{0,j\le\widetilde{j}_n}\|_\infty 1_{\mathcal{E}} + \|f_{0,j\le\widetilde{j}_n} - f_0\|_\infty \\
&\le \Big\| \sum_{j\le\widetilde{j}_n}\sum_{k=1}^{2^j} \frac{1}{|M_{jk}|}\sum_{i\in M_{jk}} \Big(\sqrt{\frac{m}{n}}Z_{jk}^{(i)} + \varepsilon_{jk}^{(i)}\Big)\psi_{jk}\Big\|_\infty 1_{\mathcal{E}} + \sum_{j=\widetilde{j}_n}^{\infty} 2^{j/2}\sup_{k\in K_j}|f_{0,jk}| \\
&\lesssim \sup_{j\le\widetilde{j}_n}\Big(\Big|\frac{1}{|M_{jk}|}\sum_{i\in M_{jk}}\sqrt{\frac{m}{n}}Z_{jk}^{(i)}\Big| + n^{-1/2}\Big)\sum_{j=0}^{\widetilde{j}_n} 2^{j/2} + \sum_{j=\widetilde{j}_n}^{\infty} 2^{j/2}\sup_{k\in K_j}|f_{0,jk}| \\
&\lesssim \sqrt{\frac{\widetilde{N}}{n}}\sup_{j\in\{1,\dots,\widetilde{j}_n\}}\sup_{k\in K_j}\big(|Z_{j,k}| + 1\big) + 2^{-\widetilde{j}_n s}\sum_{j=\widetilde{j}_n}^{\infty} 2^{j(s+1/2)}\sup_{k\in K_j}|f_{0,jk}|,
\end{aligned}$$

where

$$Z_{jk} := \frac{\sqrt{n}}{|M_{jk}|}\sum_{i\in M_{jk}}\sqrt{\frac{m}{n}}Z_{jk}^{(i)} \overset{\text{iid}}{\sim} N\Big(0, \frac{m}{|M_{jk}|}\Big), \quad 0 \le \varepsilon_{jk}^{(i)} \le 1/\sqrt{n} \text{ on } \mathcal{E}.$$

Therefore, in view of (3.21),

$$\begin{aligned}
E_{X_1,X_2}\|\widehat{f} - f_0\|_\infty &\lesssim E_{X_1}\sqrt{\frac{\widetilde{N}}{n}}\log\widetilde{N} + E_{X_1}L\widetilde{N}^{-s} + o(n^{-1}) \\
&\lesssim L^{\frac{1}{1+2s}}(n/\log n)^{-\frac{s}{1+2s}} + e^{-m/5} \lesssim L^{\frac{1}{1+2s}}(n/\log n)^{-\frac{s}{1+2s}}.
\end{aligned}$$

concluding the proof of our statement. ∎

4. Technical lemmas

The first lemma extends slightly the results of Shannon's source coding theorem by allowing also non-prefix codes, see [25, Lemma 5.1].

Lemma 4.1. *Let Y be a random finite binary string. Its expected length satisfies the inequality*

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

Let us take an arbitrary $x \in \mathbb{R}$ and write it in a scientific binary representation, i.e.

$$|x| = \sum_{k=-\infty}^{\log_2|x|} b_k 2^k,$$

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), i.e.

$$|y| = \sum_{k=-D\log_2 n}^{\log_2 |x|} b_k 2^k,$$

unless $|x| \geq \sqrt{n}$, in which case we set y to zero, see also Algorithm 1, a slightly modified version of Algorithm 1 from [25]. In the algorithm the function $x \mapsto \operatorname{sign}(x)$ is one if $x \geq 0$, and zero otherwise.

Algorithm 1 Transmitting a finite-bit approximation of a number

procedure TRANSAPPROX(x)
  if $|x| \geq n$ **then**
    Transmit: $\operatorname{sign}(x), b_{-\lfloor D\log n\rfloor+1}, \dots, b_{\lfloor \log |x|\rfloor}$.
    Construct: $y = (2\operatorname{sign}(x) - 1) \sum_{k=-D\log n+1}^{\log |x|} b_k 2^k$.
  else
    Transmit: 0.
    Construct: $y = 0$.

The next lemma gives an upper bound for the number of transmitted bits and the accuracy of the procedure described in Algorithm 1. It is a slightly reformulated version of Lemma 2.3 of [25] to accommodate almost sure upper bound on the code length.

Lemma 4.2. *For $X \sim N(\mu, \sigma^2)$, with $|\mu| \leq M$ and $\sigma \leq 1$ let the approximation Y of X given in Algorithm 1 and denote by $\mathcal{E}_X$ the event that $|X| \leq \sqrt{n}$. Then for large enough n, we have*

$$P_X(\mathcal{E}_X^c) = O(e^{-cn}), \quad |X - Y|1_{\mathcal{E}_X} < 2n^{-D}, \quad \text{and} \quad l(Y) \leq (D + 1/2)\log n,$$

for some $c > 0$.

Proof. It is straightforward to see that the last two inequalities of the statement hold. To prove the first one note that

$$P_X(\mathcal{E}_X^c) \leq P_X\big(|X| \geq \sqrt{n}\big) \leq P_X\big(|X - \mu| \geq \sqrt{n} - M\big) \lesssim e^{-cn}. \qquad \blacksquare$$

Next we provide an extended version of Lemma 4.2 of [10] with tighter upper bounds for small $\Delta > 0$. The main difference in the proof is that instead of Chebyshev's inequality we apply a more accurate concentration inequality, see [3, Lemma 8.1].

Lemma 4.3. *Let $\Delta > 0$. Then*

$$P\left\{\forall l : J_0 \le l \le j,\ |T_n(l) - \|\Pi_l f\|_2^2| \ge 4\sqrt{\frac{3z_0}{\Delta}\Big(\frac{2^{(j+l)/2}}{n^2} + 2^{l/4}\frac{\|\Pi_l f\|_2^2}{n}\Big)}\right\} \le 2e^{-c/\sqrt{\Delta}}$$

for $c = \sqrt{3/2}$ and $z_0 = 2^{J_0}$ the number of father wavelets (at resolution level J_0) and $\Pi_l f = \sum_{k=1}^{2^l} f_{lk}\psi_{lk}$ the projection of f into the wavelet resolution level l.

Proof. Note that for the wavelet estimator $\widehat{f}$ with signal-to-noise ration n, we get that

$$\|\Pi_l \widehat{f}\|_2^2 = \sum_k \widehat{f}_{lk}^2,$$

where $\widehat{f}_{lk} - f_{lk} \overset{\text{iid}}{\sim} N(0, 1/n)$.

Hence, in view of [3, Lemma 8.1] (with degree of freedom $D = 2^l$, non-centrality parameter $B = n\sum_{k=1}^{2^l} f_{lk}^2$ and $x = 1/(2\sqrt{\delta_l})$), we get for $\delta_l \le 1/4$ that

$$\begin{aligned}
&P\left\{\Big|\|\Pi_l \widehat{f}\|_2^2 - \frac{2^l}{n} - \|\Pi_l f\|_2^2\Big| \ge \sqrt{\frac{4}{\delta_l}\Big(\frac{2^l}{n^2} + \frac{\|\Pi_l f\|_2^2}{n}\Big)}\right\} \\
&\quad = P\left\{\Big|\sum_{k=1}^{2^l}\widehat{f}_{lk}^2 - \frac{2^l}{n} - \sum_{k=1}^{2^l} f_{lk}^2\Big| \ge \sqrt{\frac{4}{\delta_l}\Big(\frac{2^l}{n^2} + \frac{\sum_{k=1}^{2^l} f_{lk}^2}{n}\Big)}\right\} \\
&\quad \le P\left\{\Big|\sum_{k=1}^{2^l} n\widehat{f}_{lk}^2 - 2^l - n\sum_{k=1}^{2^l} f_{lk}^2\Big| \ge 2\sqrt{\Big(2^l + 2n\sum_{k=1}^{2^l} f_{lk}^2\frac{1}{2\sqrt{\delta_l}}\Big) + 2\frac{1}{2\sqrt{\delta_l}}}\right\} \\
&\quad \le 2e^{-0.5/\sqrt{\delta_l}}.
\end{aligned}$$

Similarly,

$$P\left\{\Big|\|\Pi_{J_0}\widehat{f}\|_2^2 - \frac{z_0}{n} - \|\Pi_{J_0} f\|_2^2\Big| \ge \sqrt{\frac{4}{\delta_{J_0}}\Big(\frac{z_0}{n^2} + \frac{\|\Pi_{J_0} f\|_2^2}{n}\Big)}\right\} \le 2e^{-0.5/\sqrt{\delta_{J_0}}}.$$

By the definition of $T_n(l)$ and union bound these results imply that

$$\begin{aligned}
&P\left\{\forall l : J_0 < l \le j, \big|T_n(l) - \|\Pi_l f\|_2^2\big| \ge \sqrt{\frac{4}{\delta_l}\Big(\frac{2^l}{n^2} + \frac{\|\Pi_l f\|_2^2}{n}\Big)},\right. \\
&\qquad \left.\big|T_n(J_0) - \|\Pi_{J_0} f\|_2^2\big| \ge \sqrt{\frac{4}{\delta_{J_0}}\Big(\frac{z_0}{n^2} + \frac{\|\Pi_{J_0} f\|_2^2}{n}\Big)}\right\} \le \sum_{J_0 \le l \le j} e^{-0.5/\sqrt{\delta_l}}.
\end{aligned}$$

Setting similarly to Lemma 4.2 of [10] the parameters $\delta_l = (2^{-(j-l)/2} + 2^{-l/4})\Delta/12$ and $\delta_{J_0} = \Delta/12$, we get in view of

$$\sum_{l=J_0}^{j} e^{-0.5/\sqrt{\delta_j}} \le \sum_{l=J_0}^{j} \left(e^{-\sqrt{3/2}\Delta^{-1/2}2^{(j-l)/4}} + e^{-\sqrt{3/2}\Delta^{-1/2}2^{l/8}}\right) \lesssim e^{-\sqrt{3/2}\Delta^{-1/2}},$$

which implies together with $z_0 \ge 1$ that

$$P\left\{\forall l : J_0 \le l \le j,\ |T_n(l) - \|\Pi_l f\|_2^2| \ge 4\sqrt{\frac{3z_0}{\Delta}\left(\frac{2^{(j+l)/2}}{n^2} + 2^{l/4}\frac{\|\Pi_l f\|_2^2}{n}\right)}\right\} \lesssim e^{-\sqrt{3/2}\Delta^{-1/2}},$$

concluding the proof of the lemma. ∎

The next lemma is a slightly rewritten version of Theorem 3.1 of [10] with tighter error bounds (for small $\alpha > 0$).

Lemma 4.4. *Let $\alpha > 0$. The test $\Psi_n(\alpha)$ satisfies, for all $\alpha > 0$ and $n > 0$, that*

$$\sup_{f\in H_0} E_f \Psi_n + \sup_{f\in H_1} E_f(1-\Psi_n) \le 2e^{-1/\sqrt{\alpha}},$$

where

$$H_0: f \in B^{s_2}_{2,\infty}(L) \quad \text{and} \quad H_1: f \in \left\{B^{s_1}_{2,\infty}(L) : \|f - B^{s_2}_{2,\infty}(L)\|_2 \ge \rho_n\right\},$$

with

$$\rho_n = \widetilde{C}_\alpha n^{-s_1/(1/2+2s_1)} \quad \text{and} \quad \widetilde{C}_\alpha = 24\left(\frac{2^{s_1}L}{\sqrt{1-2^{-2s_1}}+19}\sqrt{1/\alpha}\right).$$

Proof. The proof goes the same way as of Theorem 3.1 of [10], with the only difference that we apply Lemma 4.3 instead of Lemma 4.2 of [10] and replacing (4.3) and (4.4) in [10] with the (slightly) sharper bounds

$$\sum_{l=J_0}^{j} c\frac{2^{(j+l)/8}}{\sqrt{n}} \le \frac{c}{\sqrt{n}}2^{j/4}\left(1 + \frac{1}{1-2^{-1/8}}\right) \le 14cL^{\frac{1/2}{1/2+2t}}n^{-\frac{t}{1/2+2t}},$$

$$c\frac{L}{\sqrt{1-2^{-2t}}}2^{-jt} \le \frac{c}{\sqrt{1-2^{-2t}}}L^{\frac{1/2}{1/2+2t}}n^{-\frac{t}{1/2+2t}},$$

for $j = (1/2+2t)^{-1}\log(L^2 n)$, respectively. ∎

A. Definitions and notations for wavelets

In this section, we collect some notations and definitions about wavelets, a more detailed description can be found for instance in [14, 15].

We consider the Cohen, Daubechies and Vial construction of compactly supported, orthonormal, N-regular wavelet basis of $L_2[0,1]$, see for instance [11] and let the us use the notation $\{\psi_{jk} : j = 0, 1, \ldots,\ k = 1, \ldots, 2^j\}$. For an arbitrary function $f \in L_2[0,1]$, we can consider the wavelet representation

$$f = \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} f_{jk} \psi_{jk},$$

with $f_{jk} = \langle f, \psi_{jk} \rangle$. Following from the orthonormality of the wavelet basis, we have that

$$\|f\|_2^2 = \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} f_{jk}^2.$$

In our analysis we work with the Besov spaces $B^s_{2,\infty}$ and $B^s_{\infty,\infty}$. The corresponding Besov norms for $s \in (0, N)$ are defined as

$$\|f\|^2_{B^s_{2,\infty}} = \sup_{j \geq j_0} 2^{2js} \sum_{k=0}^{2^j-1} f_{jk}^2 \quad \text{and} \quad \|f\|_{B^s_{\infty,\infty}} = \sup_{j \geq 0, k} \{2^{j(s+1/2)} |f_{jk}|\}.$$

Then the Besov spaces $B^s_{2,\infty}$, $B^s_{\infty,\infty}$ and the corresponding Besov balls $B^s_{2,\infty}(L)$, $B^s_{\infty,\infty}(L)$ of radius $L > 0$ are defined as

$$\begin{aligned}
B^s_{2,\infty} &= \{f \in L_2[0,1] : \|f\|_{B^s_{2,\infty}} < \infty\},\\
B^s_{2,\infty}(L) &= \{f \in L_2[0,1] : \|f\|_{B^s_{2,\infty}} < L\},\\
B^s_{\infty,\infty} &= \{f \in L_2[0,1] : \|f\|_{B^s_{\infty,\infty}} < \infty\},\\
B^s_{\infty,\infty}(L) &= \{f \in L_2[0,1] : \|f\|_{B^s_{\infty,\infty}} < L\},
\end{aligned}$$

respectively. We note that the Besov space $B^s_{2,\infty}$ is larger than the standard Sobolev space where instead of the supremum one would take the sum over the resolution levels j. For $s \neq N$, $B^s_{\infty,\infty}$ is equivalent to the classical Hölder space with regularity s, while for integer s they are equivalent to the so-called Zygmond spaces, see [11].

B. Minimax bounds for the distributed white noise model

B.1. Distributed minimax rates

As explained in the introduction, the results in this paper are motivated by minimax lower bounds that we have for estimation in the distributed white noise model under communication constraints. The analogous results for the distributed nonparametric regression model were derived in the paper [25], cf. [31] for the white noise case. Since the minimax bounds put the results of the present paper into context, we give the formulations and proofs for the setting of the white noise model in this appendix for the sake of completeness.

The setting is as before that we have m local machines and at the ith machine we observe the random function $X^{(i)}$ given by the stochastic differential equation (1.1). The local machines carry out a local statistical procedure and transmit the results to a central machine, which constructs the final estimator. Now we add the restriction that local machine i is allowed to send at most $B^{(i)}$ bits (on average) to the central machine. The central machine will then collect the transmitted bits from the local computers and combine them to a global, aggregated answer. More formally, for a target function class $\mathcal{F}$, we write $\widehat{f} \in \mathcal{F}_{\text{dist}}(B^{(1)}, \dots, B^{(m)}; \mathcal{F})$ if $\widehat{f}$ is a measurable function of messages of length $\widehat{B}^{(i)}$ sent from the local machines and for every $f_0 \in \mathcal{F}$ it holds that $E_{f_0}\widehat{B}^{(i)} \le B^{(i)}$ for every i. For simplicity, we will focus on the case $B^{(1)} = \cdots = B^{(m)}$ that the communication restriction is the same for every local machine.

Theorem B.1. *Let $s, L > 0$.*

- *If $B \ge n^{1/(1+2s)}/\log m$, then*

$$\inf_{\widehat{f} \in \mathcal{F}_{\text{dist}}(B,\dots,B;B^s_{2,\infty}(L))} \sup_{f_0 \in B^s_{2,\infty}(L)} E_{f_0}\|\widehat{f} - f_0\|_2^2 \ge cL^{\frac{2}{1+2s}} n^{-\frac{2s}{1+2s}};$$

- *If $(n\log(n)/m^{2+2s})^{1/(1+2s)} \le B \le n^{1/(1+2s)}/\log m$, then*

$$\inf_{\widehat{f} \in \mathcal{F}_{\text{dist}}(B,\dots,B;B^s_{2,\infty}(L))} \sup_{f_0 \in B^s_{2,\infty}(L)} E_{f_0}\|\widehat{f} - f_0\|_2^2 \ge cL^{\frac{1}{1+s}} \left(\frac{B\log n}{n^{1/(1+2s)}}\right)^{-\frac{s}{1+s}} n^{-\frac{2s}{1+2s}};$$

- *If $B \le (n\log(n)/m^{2+2s})^{1/(1+2s)}$, then*

$$\inf_{\widehat{f} \in \mathcal{F}_{\text{dist}}(B,\dots,B)} \sup_{f_0 \in B^s_{2,\infty}(L)} E_{f_0}\|\widehat{f} - f_0\|_2^2 \ge cL^{\frac{2}{1+2s}} \left(\frac{n}{m\log n}\right)^{-\frac{2s}{1+2s}},$$

for some $c > 0$ small enough not depending on L.

Proof. See Section B.2. ∎

The result shows that it is indeed only possible to obtain the optimal rate $n^{-s/(1+2s)}$ over Besov balls of regularity s if, up to a logarithmic factor, every machine is allowed to transmit order $n^{1/(1+2s)}$ bits to the central machine. Also for completeness, and to prepare for our new adaptation results, we recall the following theorem that shows that this result is indeed sharp (up to log-factors), i.e. if order $n^{1/(1+2s)}$ bits are allowed, then the optimal rate can indeed be achieved with some procedure.

In fact, the theorem considers the first two cases of the preceding one, i.e.

$$\Big(\frac{n\log(n)}{m^{2+2s}}\Big)^{\frac{1}{1+2s}} \leq B.$$

The third case is not interesting since in that case distributed methods do not perform better than any standard technique applied on a single, local server.

Theorem B.2. *Let $s, L > 0$, $m \leq n$. Then there exists a distributed estimator $\widehat{f} \in \mathcal{F}_{\text{dist}}(B, \ldots, B; B^s_{2,\infty}(L))$ satisfying:*

- *for $B \geq (L^2 n)^{1/(1+2s)}/\log n$:*

$$\sup_{f_0 \in B^s_{2,\infty}(L)} \mathbb{E}_{f_0} \|\widehat{f} - f_0\|_2^2 \leq c\big(L^{\frac{2}{1+2s}} n^{-\frac{2s}{1+2s}}\big) \vee \big(L^2 (B/\log n)^{-2s}\big),$$

- *for $(L^2 n \log(n)/m^{2+2s})^{1/(1+2s)} \vee \log n \leq B \leq (L^2 n)^{1/(1+2s)}/\log n$:*

$$\sup_{f_0 \in B^s_{2,\infty}(L)} \mathbb{E}_{f_0} \|\widehat{f} - f_0\|_2^2 \leq c M_n L^{\frac{2}{1+s}} \left(\frac{n^{1/(1+2s)}}{B\log n}\right)^{\frac{2s}{2+2s}} n^{-\frac{2s}{1+2s}},$$

with $M_n = (\log n)^{2s}$ and $c > 0$ not depending on L.

Proof. See Section B.3. ∎

Based on the above distributed minimax lower and upper bounds we define the minimum communication required to reach the minimax squared-L_2 estimation rate $L^{2/(1+2s)} n^{-2s/(1+2s)}$ over the Besov class $B^s_{2\infty}(L)$ as

$$\arg\inf_{B>0} \inf_{\widehat{f} \in \mathcal{F}_{\text{dist}}(B,\ldots,B;B^s_{2\infty}(L))} \sup_{f_0 \in B^s_{2,\infty}(L)} E_{f_0} \|\widehat{f} - f_0\|_2^2 \leq M L^{\frac{2}{1+2s}} n^{-\frac{2s}{1+2s}}, \quad \text{(B.1)}$$

for some given, large enough constant $M > 0$. We note that Theorems B.1 and B.2 show that the optimal communication is (up to a logarithmic factor) order $(L^2 n)^{1/(1+2s)}$. Furthermore, the choice of M only influences the constant factor, not the rate of B, hence we omit it from our notation.

One can also derive similar matching lower and upper bounds for the L_∞-norm for $f_0 \in B^s_{\infty,\infty}(L)$ in case of the Gaussian white noise model, as in [25] where the nonparametric regression model was considered. Since our focus in this paper is not on deriving minimax rates, we have deferred this result to Section B.4 in the appendix.

B.2. Proof of Theorem B.1

The proof of the theorem follows from the following, more general theorem with taking $B^{(1)} = \cdots = B^{(m)} = B$. The proof is a slight extension for a larger set of estimators and adaptation to the Gaussian white noise setting of the proof of Theorem 2.1 in [25].

Theorem B.3. *Let the sequence $\delta_n = o(1)$ be defined as*

$$\delta_n = 2^{-15} L^{-2} \min\left\{ \frac{m}{n \log n}, \frac{m}{n \sum_{i=1}^m [\delta_n^{\frac{1}{1+2s}} B^{(i)} \log n \wedge 1]} \right\}. \tag{B.2}$$

Then in the distributed Gaussian white noise model (1.1), we have for any $s > 0$ that

$$\inf_{\hat{f} \in \mathcal{F}_{\text{dist}}(B^{(1)}, \ldots, B^{(m)})} \sup_{f_0 \in B^s_{2,\infty}(L)} E_{f_0} \| \hat{f} - f_0 \|_2^2 \geq c L^2 \delta_n^{\frac{2s}{1+2s}},$$

for some $c > 0$ not depending on L.

Proof of Theorem B.3. We prove the desired lower bound for the minimax risk using a modified version of Fano's inequality, given in Theorem B.7. 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 Appendix A and define $j_n = \lfloor (\log \delta_n^{-1})/(1+2s) \rfloor$. For $\beta \in \{-1, 1\}^{2^{j_n}}$, let $f_\beta \in L_2[0,1]$ be the function with wavelet coefficients

$$f_{\beta,jk} = \begin{cases} L \beta_k \delta_n^{1/2}, & \text{if } j = j_n, k = 1, \ldots, 2^{j_n}, \\ 0, & \text{else.} \end{cases} \tag{B.3}$$

Now define $\mathcal{F}_0 = \{ f_\beta : \beta \in \{-1, 1\}^{2^{j_n}} \}$. Note that $\mathcal{F}_0 \subset B^s_{2,\infty}(L)$, since

$$\| f_\beta \|^2_{B^s_{2,\infty}} = \sup_j 2^{2sj} \sum_{k=1}^{2^j} f^2_{\beta,jk} = L^2 2^{(2s+1) j_n} \delta_n \leq L^2.$$

Therefore, for an arbitrary set of estimators $\widehat{\mathcal{F}}$ we have that

$$\inf_{\hat{f} \in \widehat{\mathcal{F}}} \sup_{f_0 \in B^s_{2,\infty}(L)} \mathbb{E}_{f_0} \| \hat{f} - f_0 \|_2^2 \geq \inf_{\hat{f} \in \widehat{\mathcal{F}}} \sup_{f_0 \in \mathcal{F}_0} \mathbb{E}_{f_0} \| \hat{f} - f_0 \|_2^2.$$

To prove the statement of the theorem we take the set of distributed estimators $\widehat{\mathcal{F}} = \mathcal{F}_{\text{dist}}(B^{(1)}, \ldots, B^{(m)}; B^s_{2,\infty}(L))$, but the inequality holds more generally.

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

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

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

satisfy

$$N_t^{\max} = N_t^{\min} \quad \text{and} \quad N_t^{\max} = \sum_{i=o}^{\widetilde{t}} \binom{2^{j_n}}{i} < \frac{|\mathcal{F}_0|}{2}$$

for $\widetilde{t} := t^2/4\delta_n L^2 < 2^{j_n-1}$ (and therefore $N_t^{\max} < |\mathcal{F}_0| - N_t^{\min}$).

Recall the notations $X = (X^{(1)}, \ldots, X^{(m)})$ for the data available at the local machines and $Y = (Y^{(1)}, \ldots, Y^{m)})$ for the binary messages transmitted to the central machine satisfying the distribution protocol, and consider the Markov chain $F \to X \to Y$, where F is a uniform random element in $\mathcal{F}_0$. It then follows from Theorem B.7 (with $t^2 = L^2\delta_n 2^{j_n+1}/3$ and $d(f,g) = \|f-g\|_2$) that

$$\inf_{\widehat{f}\in\widehat{\mathcal{F}}} \sup_{f_0\in\mathcal{F}_0} \mathbb{E}_{f_0}\|\widehat{f} - f_0\|_2^2 \gtrsim L^2\delta_n 2^{j_n}\left(1 - \frac{I(F;Y) + \log 2}{\log(|\mathcal{F}_0|/N_t^{\max})}\right), \tag{B.4}$$

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

$$N_t^{\max} = \sum_{i=1}^{\widetilde{t}} \binom{2^{j_n}}{i} < 2\binom{2^{j_n}}{\widetilde{t}} \le 2\Big(\frac{e2^{j_n}}{\widetilde{t}}\Big)^{\widetilde{t}}$$

and therefore, for $\widetilde{t} = 2^{j_n-1}/3$ (i.e. $t^2 = L^2\delta_n 2^{j_n+1}/3$),

$$\log\Big(\frac{|\mathcal{F}_0|}{N_t^{\max}}\Big) \ge 2^{j_n}\log\big(2(6e)^{-1/6}2^{-2^{-j_n}}\big) \ge 2^{j_n-1}/3.$$

Hence, recalling that $2^{j_n} = \delta_n^{-1/(1+2s)}$, we see that to prove

$$\inf_{\widehat{f}\in\widehat{\mathcal{F}}} \sup_{f_0\in\mathcal{F}_0} \mathbb{E}_{f_0}\|\widehat{f} - f_0\|_2^2 \gtrsim L^2\delta_n^{2s/(1+2s)} \tag{B.5}$$

and as a consequence to derive the statement of the theorem it is sufficient to show that

$$I(F;Y) \le \delta_n^{-1/(1+2s)}/8 + O(1). \tag{B.6}$$

Observe that for the class of distributed estimators $\widehat{\mathcal{F}} = \mathcal{F}_{\text{dist}}(B^{(1)}, \ldots, B^{(m)}; B^s_{2,\infty}(L))$, by definition the following inequality holds

$$E^{(i)}l\big(Y^{(i)}\big) = \frac{1}{|\mathcal{F}_0|}\sum_{f\in\mathcal{F}_0} E_f^{(i)}l\big(Y^{(i)}\big) \le B^{(i)}, \tag{B.7}$$

where the expectation is taken over the joint distribution of the random variable F and $P_f^{(i)}$, $f \in \mathcal{F}_0$. Next note that for $\delta_n \le m/(2^{11}L^2 n\log n)$ the conditions of Lemma B.8 are satisfied hence by applying the lemma (with $\delta^2 = L^2\delta_n$ and $d = \delta_n^{-1/(1+2s)}$),

we get

$$I(F;Y) \le 2L^2 n\delta_n m^{-1} \sum_{i=1}^{m} \min\left\{2^{10}\log\left(m\delta_n^{-\frac{1}{1+2s}}\right)H(Y^{(i)}), \delta_n^{-\frac{1}{1+2s}}\right\} + 4\log 2$$
$$\le 2L^2 n\delta_n m^{-1}\delta_n^{-\frac{1}{1+2s}} \sum_{i=1}^{m}\left(2^{11}\log(n)\delta_n^{\frac{1}{1+2s}} B^{(i)} \wedge 1\right) + O(1), \qquad \text{(B.8)}$$

where the last inequality follows from Lemma 4.1 and assertion (B.7). Since from the definition of δ_n it follows that

$$\delta_n \le \frac{2^{-4}L^{-2}mn^{-1}}{\sum_{i=1}^{m}\left[2^{11}\log(n)\delta_n^{\frac{1}{1+2s}} B^{(i)} \wedge 1\right]},$$

the right-hand side of (B.8) is further bounded by $2^{-3}\delta_n^{-1/(1+2s)} + O(1)$, finishing the proof of assertion (B.6) and concluding the proof of the theorem.

Note that we have used the properties of the distributed estimation class $\widehat{\mathcal{F}}$ only in assertion (B.7), hence for any distributed method satisfying this inequality we have that

$$\inf_{\widehat{f}\in\widehat{\mathcal{F}}} \sup_{f_0\in B^s_{2,\infty}(L)} E_{f_0}\|\widehat{f} - f_0\|_2^2 \gtrsim L^2\delta_n^{\frac{2s}{1+2s}}. \qquad \text{(B.9)}$$

We note that all the computations above hold for arbitrary $\delta_n' \le \delta_n$ as well. ∎

B.3. Proof of Theorem B.2

First we give the algorithm achieving the upper bound. Let us introduce the notation

$$\eta = \left(\left\lfloor\left((L^2 n)^{\frac{1}{1+2s}}\log(n)/B\right)^{\frac{1+2s}{2+2s}}\right\rfloor \vee 1\right) \wedge m.$$

Then we group the local machines into η groups and let the different groups work on different parts of the signal as follows: the machines with indexes $1 \le i \le m/\eta$ each transmit the approximations $Y_{jk}^{(i)}$ of the observations $X_{jk}^{(i)}$ for $1 \le 2^j + k \le (B/\log n) \wedge n^{1/(1+2s)}$ using Algorithm 1. If $\eta > 1$ then the next machines, with indexes $m/\eta < i \le 2m/\eta$, each transmit the approximations $Y_{jk}^{(i)}$ for $B/\log n < 2^j + k \le 2B/\log n$, and so on. The last machines with numbers $(\eta-1)m/\eta < i \le m$ transmit $Y_{jk}^{(i)}$ for $(\eta-1)B/\log n < 2^j + k \le \eta B/\log n$. Then in the central machine we average the corresponding transmitted approximated noisy coefficients in the obvious way. Formally, using the notation

$$\mu_{jk} = \left\lceil (2^j + k)\log(n)/B\right\rceil - 1,$$

the aggregated estimator $\widehat{f}$ is the function with wavelet coefficients given by

$$\widehat{f}_{jk} = \begin{cases} \text{mean}\{Y_{jk}^{(i)} : \frac{\mu_{jk}m}{\eta} < i \leq \frac{(\mu_{jk}+1)m}{\eta}\}, & \text{if } 2^j + k \leq \eta B/\log n, \\ 0, & \text{else.} \end{cases}$$

The procedure is summarized as Algorithm 2.

Algorithm 2 Algorithm for the L_2-norm

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 $2^j + k = \lfloor(\ell-1)B/\log n\rfloor + 1$ to $\lfloor \ell B/\log n\rfloor$ **do**
      $Y_{jk}^{(i)}$:=TransApprox($X_{jk}^{(i)}$)
In the central machine:
for $2^j + k = 1$ to $\lfloor(\eta B/\log n) \wedge n^{1/(1+2s)}\rfloor$ **do**
  $\widehat{f}_{jk} := \text{mean}\{Y_{jk}^{(i)} : \mu_{jk}m/\eta < i \leq (\mu_{jk}+1)m/\eta\}$
Construct: $\widehat{f} = \sum \widehat{f}_{jk}\psi_{jk}$.

In the algorithm described above each machine transmits the approximations of at most $n^{1/(1+2s)} \wedge (B/\log n)$ noisy coefficients. Note that for any $f \in B^s_{2,\infty}(L)$, we have that

$$f_{jk}^2 \leq \sup_j 2^{js} \sum_k f_{jk}^2 \leq L^2,$$

hence in view of Lemma 4.2 (with $|\mu| = |f_{0,jk}| \leq L$) the approximation satisfies

$$0 \leq |X_{jk}^{(i)} - Y_{jk}^{(i)}|1_{\mathcal{E}} \leq 1/\sqrt{n}, \quad |Y_{jk}^{(i)}| \leq \sqrt{n}, \quad \text{and} \quad l(Y_{jk}^{(i)}) \leq \log n,$$

where the set $\mathcal{E}$ was defined in (3.7) and satisfies that $P_X(\mathcal{E}) \leq e^{-cn}$, for some $c > 0$. Therefore, we need at most B bits to transmit $n^{1/(1+2s)} \wedge (B/\log n)$ coefficients, hence $\widehat{f} \in \mathcal{F}_{\text{dist}}(B, \ldots, B; B^s_{2,\infty}(L))$.

Next for convenience we introduce the notation

$$A_{jk} = \{\lfloor \mu_{jk}m/\eta\rfloor + 1, \ldots, \lfloor(\mu_{jk}+1)m/\eta\rfloor\}$$

for the collection of machines transmitting the (j,k)th coefficient and note that $\#(A_{jk}) \asymp m/\eta$. Then our aggregated estimator $\widehat{f}$ on the set $\mathcal{E}$ satisfies for $2^j + k \leq \eta B/\log n$ (i.e. the total number of different coefficients transmitted) that

$$\widehat{f}_{jk} = \frac{1}{\#(A_{jk})} \sum_{i \in A_{jk}} Y_{jk}^{(i)} = f_{0,jk} + \sqrt{\frac{m}{n\#(A_{jk})}} Z_{jk} - \varepsilon_{jk},$$

where

$$\varepsilon_{jk} = \frac{1}{\#(A_{jk})} \sum_{i \in A_{jk}} \varepsilon_{jk}^{(i)} \in [0, n^{-1/2}] \quad \text{and} \quad Z_{jk} \overset{\text{iid}}{\sim} N(0,1).$$

Let $j_n = \lfloor \log\big((L^2 n)^{1/(1+2s)} \wedge (\eta B / \log n)\big) \rfloor$. Then the risk of the aggregated estimator is bounded as

$$\begin{aligned}
\mathbb{E}_{f_0} \|\widehat{f} - f_0\|_2^2 1_{\mathcal{E}} &\le \sum_{j=j_n}^{\infty} \sum_{k=1}^{2^j} f_{0,jk}^2 + \sum_{j=0}^{j_n} \sum_{k=1}^{2^j} \mathbb{E}_{f_0}\Big(\frac{m}{n\#(A_{jk})} Z_{jk}^2 + \varepsilon_{jk}^2\Big) 1_{\mathcal{E}} \\
&\lesssim \sum_{j=j_n}^{\infty} 2^{-2js} \sup_{j \ge j_n} 2^{2js} \sum_{k=1}^{2^j} f_{0,jk}^2 + \sum_{j=0}^{j_n} \sum_{k=1}^{2^j} \eta / n \\
&\lesssim L^2 \Big(\frac{\eta B}{\log_2 n} \wedge (L^2 n)^{\frac{1}{1+2s}}\Big)^{-2s} + \frac{\eta}{n}\Big(\frac{\eta B}{\log_2 n} \wedge (L^2 n)^{\frac{1}{1+2s}}\Big) \\
&\asymp \Big\{(\log n)^{\frac{2s}{1+s}} L^{\frac{2}{1+s}} \Big(\frac{n^{1/(1+2s)}}{B \log n}\Big)^{\frac{s}{1+s}} \vee L^{\frac{2}{1+2s}}\Big\} n^{-\frac{2s}{1+2s}} \vee \Big(\frac{mB}{\log n}\Big)^{-2s} \\
&\lesssim \Big\{(\log n)^{2s} L^{\frac{2}{1+s}} \Big(\frac{n^{1/(1+2s)}}{B \log n}\Big)^{\frac{s}{1+s}} \vee L^{\frac{2}{1+2s}}\Big\} n^{-\frac{2s}{1+2s}}, \qquad \text{(B.10)}
\end{aligned}$$

where we have used that for $f_0 \in B_{2,\infty}^s(L)$, we have $|f_{0,jk}| \le L$ for any $j \ge 0$, $k = 1, \dots, 2^j$. The above inequality together with

$$\mathbb{E}_{f_0} \|\widehat{f} - f_0\|_2^2 1_{\mathcal{E}^c} \lesssim n \mathbb{P}_{f_0}(\mathcal{E}^c) \lesssim L^2 n e^{-cn} = o(n^{-1})$$

concludes the proof of the theorem. ∎

B.4. Minimax bounds for distributed methods in L_∞-norm

Similarly to the L_2-case we consider the situation where all communication budgets are the same, i.e. $B^{(1)} = \cdots = B^{(m)} = B$.

Theorem B.4. *Consider $s, L > 0$, communication constraint $B^{(1)} = \cdots = B^{(m)} = B > 0$, then*

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

$$\inf_{\widehat{f} \in \mathcal{F}_{\text{dist}}(B,\dots,B; B_{\infty,\infty}^s(L))} \sup_{f_0 \in B_{\infty,\infty}^s(L)} \mathbb{E}_{f_0} \|\widehat{f} - f_0\|_\infty \ge c L^{\frac{1}{1+2s}} (n/\log n)^{-\frac{s}{1+2s}};$$

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

$$\inf_{\widehat{f}\in\mathcal{F}_{\text{dist}}(B,\dots,B;B^s_{\infty,\infty}(L))}\ \sup_{f_0\in B^s_{\infty,\infty}(L)} \mathbb{E}_{f_0}\|\widehat{f}-f_0\|_\infty \ge cL^{\frac{1}{1+s}}\left(\frac{n^{1/(1+2s)}}{B(\log n)^{\frac{3+4s}{1+2s}}}\right)^{\frac{s}{2+2s}}\left(\frac{n}{\log n}\right)^{-\frac{s}{1+2s}};$$

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

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

for some $c>0$ *not depending on* L.

This theorem is actually a direct consequence of the following more general theorem where the communication thresholds can vary between the machines.

Theorem B.5. *Consider* $s, L>0$, *communication constraints* $B^{(1)},\dots,B^{(m)}>0$ *and let the sequence* $\delta_n = o(1)$ *be defined as the solution to the equation* (B.2). *Then in the distributed Gaussian white noise model* (1.1) *there exists* $c>0$ *not depending on* L *such that*

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

Proof. First of all we note that in the non-distributed case where all the information is available in the global machine the minimax L_∞-risk is $(n/\log n)^{-\frac{s}{1+2s}}$. 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 B.2.

First we construct a finite subset $\mathcal{F}_0 \subset B^s_{\infty,\infty}(L)$ and then give a lower bound for the minimax risk over it. Let us denote by K_j the largest set of Daubechies wavelets at resolution level j with disjoint supports. Note that $|K_j| \ge c_0 2^j$ (for large enough j and sufficiently small $c_0>0$). Let us again multiply δ_n with a sufficiently small constant and work with this δ_n in the rest of the proof

$$\delta_n := c_0 2^{-13} L^{-2}\min\left\{\frac{m}{n\log n}, \frac{m}{n\sum_{i=1}^m[\delta_n^{1/(1+2s)}\log(n)B^{(i)}\wedge 1]}\right\}. \tag{B.11}$$

Let $j_n = \lfloor(\log\delta_n^{-1})/(1+2s)\rfloor$ and for $\beta\in\{-1,1\}^{|K_{j_n}|}$ let $f_\beta \in L_\infty[0,1]$ be the function with wavelet coefficients

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

Now let $\mathcal{F}_0 = \{f_\beta : \beta_k \in \{-1, 1\}, k \in K_{j_n}\}$.

Note that each function $f_\beta \in \mathcal{F}_0$ belongs to the set $B^s_{\infty,\infty}(L)$, since

$$\begin{aligned}\|f_\beta\|_{B^s_{\infty,\infty}} &= \sup_{j,k} 2^{(s+1/2)j} f^2_{\beta,jk} \\ &= 2^{(s+1/2)j_n} \sup_{k\in K_{j_n}} L\delta_n^{1/2} = L2^{(s+1/2)j_n}\delta_n^{1/2} \le L.\end{aligned}$$

Furthermore, if $f_\beta \neq f_{\beta'}$, then there exists a $k' \in K_{j_n}$ such that $\beta_{k'} \neq \beta'_{k'}$. 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

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

Next observe that for an arbitrary set of estimators $\widehat{\mathcal{F}}$, we have

$$\inf_{\widehat{f}\in\widehat{\mathcal{F}}} \sup_{f_0\in B^s_{\infty,\infty}(L)} \mathbb{E}_{f_0}\|\widehat{f} - f_0\|_\infty \ge \inf_{\widehat{f}\in\widehat{\mathcal{F}}} \sup_{f_0\in\mathcal{F}_0} \mathbb{E}_{f_0}\|\widehat{f} - f_0\|_\infty.$$

Now let F be a uniform random variable on the set $\mathcal{F}_0$. Then in view of Fano's inequality (see Theorem B.7 with $t = L\delta_n^{s/(1+2s)}$ and $p = 1$), we get that

$$\inf_{\widehat{f}\in\widehat{\mathcal{F}}} \sup_{f_0\in\mathcal{F}_0} \mathbb{E}_{f_0}\|\widehat{f} - f_0\|_\infty \gtrsim L\delta_n^{\frac{s}{1+2s}} \left(1 - \frac{I(F;Y) + \log 2}{\log|\mathcal{F}_0|}\right).$$

Hence, since $\log|\mathcal{F}_0| \ge |K_{j_n}| \ge c_0 2^{j_n} = c_0\delta_n^{-1/(1+2s)}$, it remains to show that

$$I(F;Y) \le (c_0/2)\delta_n^{-\frac{1}{1+2s}} + O(1).$$

In view of Lemma B.8 (applied with $\delta = \delta_n^{1/2}$, $d = |K_{j_n}| = c_0\delta_n^{-1/(1+2s)}$, $X = X^{(i)}$, $Y = Y^{(i)}$, $i = 1, \dots, m$, and noting that $\delta_n \le m/(2^{11}L^2 n \log n)$, hence the conditions are fulfilled),

$$\begin{aligned}I(F;Y) &\le 2L^2 n\delta_n m^{-1}\delta_n^{-\frac{1}{1+2s}} \sum_{i=1}^m \big(2^{10}\log(n)\delta_n^{\frac{1}{1+2s}} H(Y^{(i)}) \wedge c_0\big) + 4\log 2, \\ &\le 2^{12}L^2 n\delta_n m^{-1}\delta_n^{-\frac{1}{1+2s}} \sum_{i=1}^m \big(\log(n)\delta_n^{\frac{1}{1+2s}} B^{(i)} \wedge 1\big) + O(1) \\ &\le (c_0/2)\delta_n^{-\frac{1}{1+2s}} + O(1),\end{aligned}$$

where the second inequality follows from Theorem 4.1 and assertion (B.7) for $\widehat{\mathcal{F}} = \mathcal{F}_{\mathrm{dist}}(B^{(1)}, \dots, B^{(m)}; B^s_{\infty,\infty}(L))$ and the third by the definition of δ_n, see (B.11). Hence, we can conclude that

$$\inf_{\widehat{f}\in\mathcal{F}_{\mathrm{dist}}(B^{(1)},\dots,B^{(m)};\mathcal{F}_0)} \sup_{f_0\in\mathcal{F}_0} \mathbb{E}_{f_0}\|\widehat{f} - f_0\|_\infty \gtrsim L\delta_n^{\frac{s}{1+2s}}. \tag{B.12}$$

Note that we have used the properties of the distributed estimation class $\widehat{\mathcal{F}}$ only in assertion (B.7), hence for any class of distributed estimator $\widehat{\mathcal{F}}$ satisfying this inequality, we have that

$$\inf_{\widehat{f}\in\widehat{\mathcal{F}}} \sup_{f_0\in B^s_{\infty,\infty}(L)} E_{f_0}\|\widehat{f}-f_0\|_\infty \gtrsim L\delta_n^{\frac{s}{1+2s}}, \tag{B.13}$$

which concludes the proof. ∎

Next we give an algorithm providing matching upper bounds in the first two cases. Note that the last case, similarly to the L_2-norm is less relevant as using the data available only on a single machine would provide at least as good an estimator as any distributed algorithm. The algorithm is very similar to the L_2-case, i.e. Algorithm 2, and is basically the rewrite of Algorithm 4 of [25] tailored to the Gaussian white noise model. Here we just highlight the differences compared to Algorithm 2. We divide the machines into $\eta = (\lfloor (L^2 n(\log_2 n)^{2s}/B^{1+2s})^{1/(2+2s)}\rfloor \wedge m)\vee 1$ equal sized groups ($\eta = 1$ corresponds to case (ib), while $\eta > 1$ corresponds to case (iib)). Similarly to before machines with indexes $1 \le i \le m/\eta$ transmit the approximations $Y_{jk}^{(i)}$ 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 $Y_{jk}^{(i)}$ for

$$\big((\eta-1)\lfloor B/\log_2 n\rfloor\big) \wedge (n/\log_2 n)^{\frac{1}{1+2s}} < 2^j + k \le \big(\eta\lfloor B/\log_2 n\rfloor\big) \wedge (n/\log_2 n)^{\frac{1}{1+2s}}.$$

Then in the central machine we average the corresponding transmitted coefficients in the obvious way, similarly to the L_2-norm case. The procedure is summarized as Algorithm 3 and the (up to a logarithmic factor) optimal behaviour is given in Theorem B.6 below.

Theorem B.6. *Let $s, L > 0$, then the distributed estimator $\widehat{f}$ described in Algorithm 3 belongs to $\mathcal{F}_{\text{dist}}(B,\dots,B; B^s_{\infty,\infty}(L))$ and satisfies*

- *for $B \ge (L^2 n)^{1/(1+2s)}(\log_2 n)^{2s/(1+2s)}$, we have*

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

- *for $(L^2 n(\log_2 n)/m^{2+2s})^{\frac{1}{1+2s}} \vee \log_2 n \le B < (L^2 n)^{\frac{1}{1+2s}}(\log_2 n)^{\frac{2s}{1+2s}}$, we have*

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

with $M_n = (\log_2 n)^{s\vee\frac{3s}{2+2s}}$ and $c > 0$ not depending on L.

Algorithm 3 Non-adaptive 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**
      $Y_{jk}^{(i)}$:=TransApprox($X_{jk}^{(i)}$).
In the central machine:
for $2^j + k = 1$ to $\eta \lfloor B/\log_2 n \rfloor$ **do**
  $\hat{f}_{jk} := mean\{Y_{jk}^{(i)} : \mu_{jk} m/\eta < i \le (\mu_{jk}+1)m/\eta\}$.
Construct: $\hat{f} = \sum \hat{f}_{jk}\psi_{jk}$.

The proof of the theorem follows the same reasoning as the proof of Theorem B.3 but for the L_∞-norm and it basically follows from the proof of Theorem 2.8 of [25] tailored to the Gaussian white noise model.

B.5. Technical lemmas

First we recall a slight modification of Fano's inequality, see [12, Corollary 1] or [25, Theorem A.6]. Given a finite set $\mathcal{F}_0 \subset \mathcal{F}$, we use the notations

$$N_t^{\max} = \max_{f \in \mathcal{F}_0} \{\#\{\tilde{f} \in \mathcal{F}_0 : d(f,\tilde{f}) \le t\}\},$$
$$N_t^{\min} = \min_{f \in \mathcal{F}_0} \{\#\{\tilde{f} \in \mathcal{F}_0 : d(f,\tilde{f}) \le t\}\}.$$

Theorem B.7. *If $\mathcal{F}$ 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 all estimators depending only on Y and the function class $\mathcal{F}$, and F is a uniformly distributed random variable on $\mathcal{F}_0$.

The next lemma gives an upper bound for the mutual information between the uniform random variable F on $\mathcal{F}_0 \subset \mathbb{R}^d$ and the set of observations on all local machines $Y = (Y^{(1)}, \dots, Y^{(m)})$ in the d-dimensional many normal means model.

Lemma B.8. *Let $F = \delta\beta$, with $\delta^2 \le 2^{-10} m/(n \log(md))$ and β a uniformly distributed random variable over $\{-1,1\}^d$. Furthermore, suppose that $X = (X^{(1)}, \dots, X^{(m)})$, where $X^{(i)}$s are d-dimensional random variables satisfying that $X_j^{(i)} \mid F_j$ and F_j are*

independent of F_{-j}, and $X_j^{(i)} \mid (F = f) \sim \mathbb{P}_{f_j}^{(i)} = N(f_j, m/n)$. Then

$$I(F;Y) \le \sum_{i=1}^{m} \frac{2\delta^2}{m/n} \min\bigl\{2^{10}\log(md)H(Y^{(i)}), d\bigr\} + 4\log 2,$$

where $I(F;Y)$ is the mutual information between F and Y in the Markov chain $F \to X \to Y$.

Proof. Let us introduce the notation $a^2 = 2^4 \log(md)m/n$ and note that

$$\sup_{|x|\le a} \frac{\varphi_{\delta,m/n}(x)}{\varphi_{-\delta,m/n}(x)} \le \sup_{|x|\le a} e^{\frac{n|(x-\delta)^2-(x+\delta)^2|}{2m}} \le \sup_{|x|\le a} e^{\frac{2n\delta|x|}{m}} \le e^{\frac{2an\delta}{m}},$$

where φ_{μ,σ^2} denotes the density function of a normal distribution with mean μ and variance σ^2. Furthermore, let us introduce the notation $B_j = \{|x_j| \le a\}$, $j = 1, \dots, d$. Then by Theorem B.9 (with $\mathcal{F}_0 = \{f = \delta\beta : \beta \in \{-1,1\}^d\}$), we have that

$$I(F;Y^{(i)}) \le d(\log 2)\sqrt{P_{X_j^{(i)}}(X_j^{(i)} \notin B_j)} + d^2 P_{X_j^{(i)}}(X_j^{(i)} \notin B_j) + 2C^2(C-1)^2 I(X^{(i)};Y^{(i)}), \quad \text{(B.14)}$$

with $C = e^{2^3|\delta|\sqrt{\log(md)n/m}}$. Next note that for $Z \sim N(0, m/n)$, we have

$$P_{X_j^{(i)}}(X_j^{(i)} \notin B_j) \le P\bigl(|Z| \ge a - \delta\bigr) \le 2e^{-\frac{(a-\delta)^2 n}{2m}} \le 2e^{-\frac{a^2 n}{4m}} \le 2(md)^{-4},$$

and the inequality $I(X^{(i)};Y^{(i)}) \le H(Y^{(i)})$ holds. Then by plugging in the above inequalities into (B.14) and using the inequalities $e^x \le 1 + 2x$ for $x \le 0.4$ and $C^2 \le 2$, we get that

$$I(F;Y^{(i)}) \le \sqrt{2}(\log 2)m^{-2}d^{-1} + 2(\log 2)m^{-4}d^{-2} + 2^{11}\delta^2 \frac{\log(md)n}{m} H(Y^{(i)}).$$

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

$$\begin{aligned} I(F;Y^{(i)}) \le I(F;X^{(i)}) &\le \frac{1}{|\mathcal{F}_0|^2} \sum_{f,f' \in \mathcal{F}_0} K\bigl(\mathbb{P}_f^{(i)} \| \mathbb{P}_f^{(i)}\bigr) \\ &= \frac{\delta^2}{2m/n} \frac{1}{|\mathcal{F}_0|^2} \sum_{f,f' \in \mathcal{F}_0} \|\beta - \beta'\|_2^2 \le 2(n/m)d\delta^2. \end{aligned}$$

We conclude our statement by noting that

$$I(F;Y) \le \sum_{i=1}^{m} I(F;Y^{(i)})$$

■

The next theorem provide an upper bound for the mutual information, see [25, Theorem A.9] or [29, Lemma 3].

Theorem B.9. *Let us consider the Markov chain $F \to X^{(i)} \to Y^{(i)}$, where F is the uniform distribution on $\mathcal{F}_0 \subset \mathbb{R}^d$ and $X^{(i)} \mid (F = f) \sim P_{X^{(i)}|F=f}$ is a d-dimensional random variable. Assume that $X_j^{(i)} \mid F_j$ and F_j are independent of F_{-j}. For $C \geq 1$, define*

$$B_j = \left\{ x_j : \max_{f \neq f'} \frac{p(x_j \mid f_j)}{p(x_j \mid f_j')} \leq C \right\}$$

for a constant $C \geq 1$ and density $p(x_j | f_j)$. Then

$$I(F; Y^{(i)}) \leq \sum_{j=0}^{d} \Big((\log 2) \sqrt{P_{X_j^{(i)}}(X_j^{(i)} \notin B_j)} + \log |\mathcal{F}_0| P_{X_j^{(i)}}(X_j^{(i)} \notin B_j) \Big) + 2C^2 (C-1)^2 I(X^{(i)}; Y^{(i)}),$$

where $I(X^{(i)}; Y^{(i)})$ is the mutual information between $X^{(i)}$ and $Y^{(i)}$.

Funding. Research supported by the Netherlands Organization of Scientific Research NWO. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101041064 and 320637).

References

[1] L. P. Barnes, Y. Han, and A. Ozgur, Learning distributions from their samples under communication constraints. 2019, arXiv:1902.02890v1

[2] H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu, Distributed testing and estimation under sparse high dimensional models. *Ann. Statist.* **46** (2018), no. 3, 1352–1382 Zbl 1392.62060 MR 3798006

[3] L. Birgé, An alternative point of view on Lepski's method. In *State of the art in probability and statistics (Leiden, 1999)*, pp. 113–133, IMS Lecture Notes Monogr. Ser. 36, Inst. Math. Statist., Beachwood, OH, 2001 Zbl 1373.62142 MR 1836557

[4] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *STOC'16 – Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1011–1020, ACM, New York, 2016 Zbl 1373.68235 MR 3536632

[5] A. D. Bull, Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** (2012), 1490–1516 Zbl 1295.62049 MR 2988456

[6] A. D. Bull and R. Nickl, Adaptive confidence sets in L^2. *Probab. Theory Related Fields* **156** (2013), no. 3-4, 889–919 Zbl 1273.62105 MR 3078289

[7] T. T. Cai and M. G. Low, An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** (2004), no. 5, 1805–1840 Zbl 1056.62060 MR 2102494

[8] T. T. Cai and H. Wei, Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. 2020, arXiv:2001.08877

[9] T. T. Cai and H. Wei, Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation. *Ann. Statist.* **50** (2022), no. 2, 698–725 Zbl 1486.62099 MR 4404917

[10] A. Carpentier, Testing the regularity of a smooth signal. *Bernoulli* **21** (2015), no. 1, 465–488 Zbl 1320.94021 MR 3322327

[11] A. Cohen, I. Daubechies, and P. Vial, Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** (1993), no. 1, 54–81 Zbl 0795.42018 MR 1256527

[12] J. C. Duchi and M. J. Wainwright, Distance-based and continuum Fano inequalities with applications to statistical estimation. 2013, arXiv:1311.2669

[13] E. Giné and R. Nickl, Confidence bands in density estimation. *Ann. Statist.* **38** (2010), no. 2, 1122–1170 Zbl 1183.62062 MR 2604707

[14] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*. Camb. Ser. Stat. Probab. Math., 40, Cambridge Univ. Press, New York, 2016 Zbl 1358.62014 MR 3588285

[15] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov, *Wavelets, approximation, and statistical applications*. Lect. Notes Stat. 129, Springer, New York, 1998 Zbl 0899.62002 MR 1618204

[16] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*. Lect. Notes Stat. 169, Springer, New York, 2003 Zbl 1013.62049 MR 1991446

[17] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan, A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** (2014), no. 4, 795–816 Zbl 07555464 MR 3248677

[18] J. D. Lee, Q. Liu, Y. Sun, and J. E. Taylor, Communication-efficient sparse regression. *J. Mach. Learn. Res.* **18** (2017), Paper No. 5 Zbl 1434.62157 MR 3625709

[19] D. Picard and K. Tribouley, Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** (2000), no. 1, 298–335 Zbl 1106.62331 MR 1762913

[20] J. Robins and A. van der Vaart, Adaptive nonparametric confidence sets. *Ann. Statist.* **34** (2006), no. 1, 229–253 Zbl 1091.62039 MR 2275241

[21] J. D. Rosenblatt and B. Nadler, On the optimality of averaging in distributed statistical learning. *Inf. Inference* **5** (2016), no. 4, 379–404 Zbl 1426.68241 MR 3609865

[22] J. Rousseau and B. Szabo, Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *Ann. Statist.* **48** (2020), no. 4, 2155–2179 Zbl 1471.62350 MR 4134790

[23] B. Szabó, A. W. van der Vaart, and J. H. van Zanten, Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** (2015), no. 4, 1391–1428 Zbl 1317.62040 MR 3357861

[24] B. Szabó and H. van Zanten, An asymptotic analysis of distributed nonparametric methods. *J. Mach. Learn. Res.* **20** (2019), Paper No. 87 Zbl 1434.68457 MR 3960941

[25] B. Szabó and H. van Zanten, Adaptive distributed methods under communication constraints. *Ann. Statist.* **48** (2020), no. 4, 2347–2380 Zbl 1455.62097 MR 4134798

[26] B. Szabó, L. Vuursteen, and H. van Zanten, Optimal distributed composite testing in high-dimensional Gaussian models with 1-bit communication. *IEEE Trans. Inf. Theory* **68** (2022), no. 6, 4070–4084 Zbl 07555916 MR 4433269

[27] B. Szabó, L. Vuursteen, and H. van Zanten, Optimal high-dimensional and nonparametric distributed testing under communication constraints. 2022, arXiv:2202.00968

[28] A. Zaman and B. Szabó, Distributed nonparametric estimation under communication constraints. 2022, arXiv:2204.10373

[29] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright, Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *NeurIPS 2013 – Advances in Neural Information Processing Systems 26*, pp. 2328–2336, Curran Associates, Inc., 2013

[30] Y. Zhang, J. C. Duchi, and M. J. Wainwright, Communication-efficient algorithms for statistical optimization. In *NeurIPS 2012 – Advances in Neural Information Processing Systems 25*, pp. 1502–1510, Curran Associates, Inc., 2012

[31] Y. Zhu and J. Lafferty, Distributed nonparametric regression under communication constraints. In *ICML 2018 – Proceedings of the 35th International Conference on Machine Learning*, pp. 6004–6012, Proceedings of Machine Learning Research 80, PMLR, 2018

Received 3 October 2020; revised 12 June 2022.

Botond Szabó
Department of Decision Sciences, Bocconi University, via Rontgen 1, 20136 Milano, Italy;
and Bocconi Institute for Data Science and Analytics (BIDSA); botond.szabo@unibocconi.it

Harry van Zanten
Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1111,
1081 HV Amsterdam, Netherlands; j.h.van.zanten@vu.nl