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

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

A major challenge and fundamental question for nonparametric distributed methods is whether or not it is possible to achieve a form of adaptive inference. In other words, whether we can design methods that do automatic, data-driven tuning in order to achieve the optimal bias-variance trade-off. We illustrated by example in [19] that naively using methods that are known to achieve optimal adaptation in non-distributed settings, can lead to sub-optimal performance in the distributed case. In the recent paper [25], which considers the same distributed signal-in-white-noise model and was written independently and at the same time as the present paper, it is in fact conjectured that adaptation in the considered particular distributed model is not possible.
In order to study convergence rates and adaptation for distributed methods in a meaningful way the class of methods should be restricted somehow. Indeed, if there is no limitation on communication or computation, then we could simply communicate all data from the various local machines to a central machine, aggregate it, and use some existing adaptive, rate-optimal procedure. In this paper we consider a setting in which the communication between the local and the global machines is restricted, much in the same way as the communication restrictions imposed in [23] in a parametric framework and recently in the simultaneously written paper [25] in the context of the distributed signal-in-white-noise model we introduced in [19].

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

These first results on rate-optimal distributed estimators are not adaptive, in the sense that the optimal procedures depend on the regularity of the unknown regression function. The same holds true for the procedure obtained in parallel in [25]. In this paper we go a step further and show that contrary perhaps to intuition, and contrary to the conjecture in [25], adaptation is in fact possible. Indeed, we exhibit in this paper an adaptive distributed method which involves a very specific grouping of the local machines, in combination with a Lepski-type method that is carried out in the central machine. We prove that the resulting distributed estimator adapts to a range of smoothness levels of the unknown regression function and that, up to logarithmic factors, it attains the minimax lower bound.

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

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

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

\[
\begin{align*}
X_{\ell}^{(i)} &= f_0(T_{\ell}^{(i)}) + \sigma \varepsilon_{\ell}^{(i)}, \\
T_{\ell}^{(i)} &\sim U(0, 1), \varepsilon_{\ell}^{(i)} \sim N(0, 1), \quad \ell = 1, \ldots, n/m, \quad i = 1, \ldots, m,
\end{align*}
\]

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

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

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

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

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

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

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

**Proof.** See Section 3.1

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

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

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

In regime (i) we have a large communication budget and by elementary computations we get that the minimum in (2.2) is attained in the second fraction and hence that $\delta_n = 1/(L^2 n)$. This means that in this case the derived lower bound corresponds to the usual non-distributed minimax rate $L^{2/(1+2s)} n^{-2s/(1+2s)}$. In the other extreme case, regime (iii), the minimum is taken at the first term in (2.2) and $\delta_n = m/(L^2 n \log_2 n)$, so the lower bound is of the order $L^{2/(1+2s)} (n \log_2(n)/m)^{-2s/(1+2s)}$. 


This rate is, up to the log₂ n factor, equal to the minimax rate corresponding to the sample size n/m. Consequently, in this case it does not make sense to consider distributed methods, since by just using a single machine the best rate can already be obtained (up to a logarithmic factor). In the intermediate case (ii) it is straightforward to see that δₙ = \( 2 \log_2 n \) already be obtained (up to a logarithmic factor). In the intermediate case (ii) it is.

Note that
\[ s \leq \frac{\log_2 n}{m} \] with at least

The sample size
\[ 2 \]
This rate is, up to the log₂ n factor, equal to the minimax rate corresponding to the sample size n/m. Consequently, in this case it does not make sense to consider distributed methods, since by just using a single machine the best rate can already be obtained (up to a logarithmic factor). In the intermediate case (ii) it.

\[ s \leq \frac{\log_2 n}{m} \]

\[ 2 \]

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

\[ \text{(i) if } B \geq (L^2 n)^{1/(1+2s)}/\log_2 n, \]

\[ \inf_{\hat{f} \in F_{\text{dist}}(B;...;B;2\log_2 n)} \sup_{f \in B_{2\log_2 n}} \mathbb{E}_{f_0,T} \| \hat{f} - f_0 \|_2^2 \gtrsim L^{1/2} n^{-\frac{s^2}{1+2s}}; \]

\[ \text{(ii) if } (L^2 n \log_2 n)/m^{2+2s} \gtrsim (1+2s)/(1+2s)/\log_2 n, \]

\[ \inf_{\hat{f} \in F_{\text{dist}}(B;...;B;2\log_2 n)} \sup_{f \in B_{2\log_2 n}} \mathbb{E}_{f_0,T} \| \hat{f} - f_0 \|_2^2 \gtrsim L^{1/2} \left( \frac{n^{1+2s}}{B \log_2 n} \right)^{1/2} n^{-\frac{s^2}{1+2s}}; \]

\[ \text{(iii) if } (L^2 n \log_2 n)/m^{2+2s} \gtrsim (1+2s)/(1+2s)/B, \]

\[ \inf_{\hat{f} \in F_{\text{dist}}(B;...;B;2\log_2 n)} \sup_{f \in B_{2\log_2 n}} \mathbb{E}_{f_0,T} \| \hat{f} - f_0 \|_2^2 \gtrsim L^{1/2} \left( \frac{n \log_2 n}{m} \right)^{-\frac{s^2}{1+2s}}. \]

2.2. Non-adaptive rate-optimal distributed procedures for L²-risk. Next we show that the derived lower bounds are (nearly) sharp by presenting distributed procedures that attain the bounds (up to logarithmic factors). We note that it is sufficient to consider only the case \( B \geq (L^2 n \log_2 n)/m^{2+2s} \), since otherwise distributed techniques do not perform better than standard techniques carried out on one of the local servers. In case (iii) therefore one would probably prefer to use a single local machine instead of a complicated distributed method with (possibly) worse performance.

As a first step let us consider Daubechies wavelets \( \psi_j \), \( j = 0, ..., k = 0, 1, ..., 2^j - 1 \) with at least \( s \) vanishing moments (for details, see Section B in the supplement).

Then let us estimate the wavelet coefficients of the underlying function \( f_0 \) in each local problem, i.e. for every \( j = 0, ..., \) and \( k = 0, 1, ..., 2^j - 1 \) let us construct

\[ f^{(i)}_{jk} = \frac{m}{n} \sum_{\ell=1}^{n/m} x^{(i)}_{\ell} \psi_{jk}(\ell) \]

and note that

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

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

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

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

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

**Proof.** See Section 3.4. \(\square\)

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

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

\[
\hat{f}_{jk} = \begin{cases} \frac{1}{m} \sum_{i=1}^{m} Y^{(i)}_{jk}, & \text{if } 2^j + k \leq (L^2 n)^{1/(1+2s)} \wedge (B / \log_2 n), \\ 0, & \text{else}. \end{cases}
\]

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

**Algorithm 2** Nonadaptive \(L_2\)-method, case (i)

1: In the local machines:
2: for \(i = 1\) to \(m\) do:
3: for \(2^j + k = 1\) to \((L^2 n)^{1/(1+2s)} \wedge (B / \log_2 n)\) do
4: \(Y^{(i)}_{jk} := \text{TransApprox}(\hat{y}^{(i)}_{jk})\)
5: In the central machine:
6: for \(2^j + k = 1\) to \((L^2 n)^{1/(1+2s)} \wedge (B / \log_2 n)\) do
7: \(\hat{f}_{jk} := \text{mean}(Y^{(i)}_{jk} : 1 \leq i \leq m)\).
8: Construct: \(\hat{f} = \sum \hat{f}_{jk} \psi_{jk}\).

We note again that the procedure outlined in Algorithm 2 is just a simple averaging, sometimes called “divide and conquer” or “embarrassingly parallel” in the learning literature (e.g. [24], [14]).
The following theorem asserts that the constructed estimator indeed attains the lower bound in case (i) (up to a logarithmic factor for $B$ close to the threshold). Note that in this upper bound result (and in the ones ahead), we do not have to assume the technical condition on the number of machines as in the lower bounds.

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

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

**Proof.** See Section 3.2.

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

We introduce the notation $\eta = [(L^2 n)^{1/2+2s}((\log_2 n)/B)^{1+2s}] \wedge m$. Then we group the local machines into $\eta$ groups and let the different groups work on different parts of wavelet domain as follows: the machines with numbers $1 \leq i \leq m/\eta$ each transmit the approximations $Y_{jk}^{(i)}$ of the estimated wavelet coefficients $f_{jk}$ for $1 \leq 2^j + k \leq \lceil B/\log_2 n \rceil$: the next machines, with numbers $m/\eta < i \leq 2m/\eta$, each transmit the approximations $Y_{jk}^{(i)}$ for $\lceil B/\log_2 n \rceil < 2^j + k \leq 2\lceil B/\log_2 n \rceil$, and so on. The last machines with numbers $(\eta - 1)m/\eta < i \leq m$ transmit the $Y_{jk}^{(i)}$ for $(\eta - 1)\lceil B/\log_2 n \rceil < 2^j + k \leq \eta\lceil B/\log_2 n \rceil$. Then in the central machine we average the corresponding transmitted noisy coefficients in the obvious way. Formally, using the notation $\mu_{jk} = \lceil (2^j + k)/B/\log_2 n \rceil^{-1} - 1$, the aggregated estimator $\hat{f}$ is the function with wavelet coefficients given by

$$
\hat{f}_{jk} = \begin{cases} 
\text{mean}\{Y_{jk}^{(i)} : \frac{\mu_{jk}m}{\eta} < i \leq \frac{(\mu_{jk}+1)m}{\eta}\}, & \text{if } 2^j + k \leq \eta \lceil B/\log_2 n \rceil, \\
0, & \text{else}.
\end{cases}
$$

The procedure is summarized as Algorithm 3.

**Algorithm 3** Nonadaptive $L_2$-method, case (ii)

1: In the local machines:
2: for $\ell = 1$ to $\eta$ do
3: \hspace{1cm} for $i = [(\ell - 1)m/\eta] + 1$ to $[(\ell m/\eta]$ do
4: \hspace{2cm} for $2^j + k = (\ell - 1)\lceil B/\log_2 n \rceil + 1$ to $\ell\lceil B/\log_2 n \rceil$ do
5: \hspace{3cm} $Y_{jk}^{(i)} := \text{TransApprox}(\hat{f}_{jk}^{(i)})$.
6: In the central machine:
7: for $2^j + k = 1$ to $\eta\lceil B/\log_2 n \rceil$ do
8: \hspace{1cm} $\hat{f}_{jk} := \text{mean}(Y_{jk}^{(i)} : \mu_{jk}m/\eta < i \leq (\mu_{jk}+1)m/\eta)$.
9: Construct: $\hat{f} = \sum \hat{f}_{jk} \varphi_{jk}$.

The following theorem asserts that this estimator attains the lower bound in case (ii) (up to a logarithmic factor). We also prove in Section 3.3 that Algorithm 2 is sub-optimal in this case.
Algorithm 3 belongs to the class of algorithms that are solutions to the equation
\( L \). Consider a large enough finite subset of this class. For the second term, we use the standard version of Fano’s inequality. We again remark that the right-hand side follows from the usual non-distributed minimax lower bound.

The proof of the theorem is very similar to the proof of Theorem 2.8. The first term on the right-hand side follows from the usual non-distributed minimax lower bound. For the second term we use the standard version of Fano’s inequality. We again consider a large enough finite subset of \( B_{2,\infty}^s(L) \). The effective resolution level for the \( L_1 \)-norm in the non-distributed case is \( (1 + 2s)^{-1} \log_2(n / \log_2 n) \). Similarly to the \( L_2 \) case the effective resolution level changes to \( (1 + 2s)^{-1} \log n \). The proof of the theorem is deferred to Section 3.5.

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

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

### 2.3. Distributed minimax results for \( L_\infty \)-risk

When we replace the \( L_2 \)-norm by the \( L_\infty \)-norm and correspondingly change the type of Besov balls we consider, we can derive a lower bound similar to Theorem 2.1 (see Section B in the supplement for the rigorous definition of Besov balls).

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

\[
\inf_{f \in \mathcal{F}_{dist}(B^{(1)}, \ldots, B^{(m)}; B_{\infty,\infty}^s(L))} \sup_{f_0 \in B_{\infty,\infty}^s(L)} E_{f_0,T}[\| \hat{f} - f_0 \|_\infty] \geq \frac{\eta}{\log_2 n} \left( \frac{n}{\log_2 n} \right)^{-\frac{s}{1+2s}} \lor L \delta_n^{1+2s}.
\]

**Proof.** See Section 4.1. \( \square \)

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

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

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

(iib) If $B \geq \left( L^2 n/(\log_2 n)^3+4s \right)^{1/(1+2s)}$, then

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

(iiib) If $(L^2 n \log_2(n)/m)^{1/(1+2s)} \leq B < \left( L^2 n/(\log_2 n)^3+4s \right)^{1/(1+2s)}$, then

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

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

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

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

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

$$\hat{f}_{jk} = \begin{cases} \text{mean}\{Y^{(i)}_{jk} : \frac{\mu_{jk} m}{\eta} < i \leq \frac{(\mu_{jk}+1)m}{\eta}\}, & \text{if } 2^j k \leq \eta \left\lfloor B/\log_2 n \right\rfloor \cap \left( n/\log_2 n \right)^{1/2+2s}, \\
0, & \text{else}, \end{cases}$$
where $\mu_{jk} = \lceil (2^j + k)B/\log_2 n \rceil^{-1} - 1$. The procedure is summarized as Algorithm 4 and the (up to a logarithmic factor) optimal behaviour is given in Theorem 2.10 below.

**Algorithm 4** Nonadaptive $L_\infty$-method, combined

1: In the local machines:
2: for $\ell = 1$ to $\eta$ do
3: for $i = \lfloor (\ell - 1)m/\eta \rfloor + 1$ to $\lfloor \ell m/\eta \rfloor$ do
4: for $2^j + k = (\ell - 1)[B/\log_2 n] + 1$ to $\ell[B/\log_2 n]$ do
5: $Y_{jk}^{(i)} := \text{TransApprox}(f_{jk}^{(i)})$.
6: In the central machine:
7: for $2^j + k = 1$ to $\eta[B/\log_2 n]$ do
8: $\hat{f}_{jk} := \text{mean}(Y_{jk}^{(i)}: \mu_{jk}m/\eta < i \leq (\mu_{jk} + 1)m}$.
9: Construct: $\hat{f} = \sum \hat{f}_{jk}\psi_{jk}$.

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

- for $B \geq n^{1/(1+2s)}(\log_2 n)^{2s/(1+2s)}$,
  $$\sup_{f_0 \in B_{s,L}^{\infty}(L)} \mathbb{E}_{f_0,T} \| \hat{f}_n - f_0 \|_\infty \lesssim L^{1/42} (n/\log_2 n)^{-\frac{s}{1+2s}};$$

- for $(n(\log_2 n)/m^{2+2s})^{1/(1+2s)} \sqrt{\log_2 n} \leq B < n^{1/(1+2s)}(\log_2 n)^{2s/(1+2s)}$,
  $$\sup_{f_0 \in B_{s,L}^{\infty}(L)} \mathbb{E}_{f_0,T} \| \hat{f}_n - f_0 \|_\infty \lesssim M_n L^{1/42} \left( \frac{n^{1/42}}{B(\log_2 n)^{3/42}} \right)^{\frac{s}{3+42}} (n/\log_2 n)^{-\frac{s}{1+2s}},$$
  with $M_n = (\log_2 n)^{s/3+42}$.

**Proof.** See Section 4.2.

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

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

2.4. **Adaptive distributed estimation.** The (almost) rate-optimal procedures considered so far have in common that they are non-adaptive, in the sense that they all use the knowledge of the regularity level $s$ of the unknown functional parameter of interest. In this section we exhibit a distributed algorithm attaining the lower
bounds (up to a logarithmic factor) across a whole range of regularities \( s \) simultaneously. In the non-distributed setting it is well known that this is possible, and many adaptation methods exist, including for instance the block Stein method, Lepski’s method, wavelet thresholding, and Bayesian adaptation methods, just to mention but a few (e.g. [7, 20]). In the distributed case the matter is more complicated. Using the usual adaptive tuning methods in the local machines will typically not work (see [19]) and in fact it was recently conjectured that adaptation, if at all possible, would require more communication than is allowed in our model (see [25]).

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

\[
s_{\min} = \arg \inf_{s > 0} \liminf_n \left\{ \left( \frac{L^2 n (\log_2 n)^2}{m^{2+2s}} \right)^{1/(1+2s)} \leq B \right\}
\]

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

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

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

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

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

\[
\begin{align*}
  j_{B,n} &:= \lfloor \log_2(B/\log_2 n) \rfloor \\
  j_{\text{max}} &:= \lfloor (2 + 2s_{\text{min}})^{-1} \log_2(nB) \rfloor \wedge \lfloor (1 + 2s_{\text{min}})^{-1} \log_2 n \rfloor.
\end{align*}
\]

The corresponding sets of indexes are denoted by \( I_0, I_1, \ldots, I_{\tilde{n}-1} \). Note that \( |I_t| \geq m/\log_2 n \), for \( t \in \{0, \ldots, \tilde{n} - 1\} \). Then the machines in the group \( I \) transmit the approximations \( Y_{jk}^{(i)} \) (with \( D = 1/2 \) in Algorithm 1) of the local estimators of the wavelet coefficients \( \hat{f}_{jk} \), for \( 0 \leq j \leq j_{B,n} - 1, k = 0, \ldots, 2^j - 1 \) to the central machine. The machines in group \( I_t, t \in \{0, \ldots, \tilde{n} - 1\} \), will be responsible for transmitting the coefficients at resolution level \( j = j_{B,n} + t \). First for every \( t \in \{0, \ldots, \tilde{n} - 1\} \), the machines in group \( I_t \) are split again into \( 2^t \) equal size groups (for simplicity each group has \( [2^{-t}((m/2)/\tilde{n})] \geq 1 \) machines and the leftovers are discarded again), denoted by \( I_{t,1}, I_{t,2}, \ldots, I_{t,2^t} \). A machine \( i \) in one of the groups \( I_{t,\ell} \) for \( \ell \in \{1, \ldots, 2^t\} \) transmits the approximations \( Y_{jk}^{(i)} \) (again with \( D = 1/2 \) in Algorithm 1) of the local estimators of the wavelet coefficients \( \hat{f}_{jk}^{(i)} \), for \( j = j_{B,n} + t \) and \( (\ell - 1)2^{j_{B,n}} \leq k < \ell 2^{j_{B,n}} \) to the central machine.

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

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

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

\[
(2.5) \quad \hat{f}(j) = \sum_{l=j-1}^{2^t-1} \sum_{k=0}^{2^t-1} \hat{f}_{lk} \psi_{lk}.
\]

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

\[
(2.6) \quad \hat{j} = \min \{ j \in J : \| \hat{f}(j) - \hat{f}(l) \|_{\lambda}^2 \leq \tau \lambda \|n/m, \forall l > j, l \in J \},
\]

for some sufficiently large parameter \( \tau > 1 \) (defined later) and \( n_j = |I_{j-j_{B,n}+1}|n/m \asymp \frac{nB}{2^t(\log_2 n)^2} \), for \( j \geq j_{B,n} \) and \( n_j = \|I\|n/m \asymp n \) for \( j < j_{B,n} \). Then we construct our final estimator \( \hat{f} \) simply by taking \( \hat{f} = \hat{f}(\hat{j}) \).
We summarize the above procedure (without discarding servers for achieving exactly equal size subgroups) in Algorithm 5, below.

**Algorithm 5** Adaptive $L_2$-method

1: In the local machines:
2: for $i = 1$ to $m/2$ do
3:  for $j = 0$ to $j_{B,n} - 1$ do
4:    for $k = 0$ to $2^j - 1$ do
5:      $Y_{jk}^{(i)} := \text{TransApprox}(\hat{f}_{jk}^{(i)})$.
6:  for $t = 0$ to $\tilde{n} - 1$ do
7:    Let $j := j_{B,n} + t$.
8:   for $\ell = 1$ to $2^t$ do
9:     for $i = m/2 + t \left\lfloor \frac{m/2}{\eta} \right\rfloor + (\ell - 1) \left(2^t \left\lfloor \frac{m/2}{\eta} \right\rfloor + 1\right)$ to $m/2 + t \left\lfloor \frac{m/2}{\eta} \right\rfloor + \tilde{\eta} - 1$ do
10:    for $k = (\ell - 1)2^j + 1$ to $\ell 2^j - 1$ do
11:      $Y_{jk}^{(i)} := \text{TransApprox}(\hat{f}_{jk}^{(i)})$.
12: In the central machine:
13: (1) Averaging the local observations:
14: for $j = 0$ to $j_{\text{max}}$ do
15:  for $k = 0$ to $2^j - 1$ do
16:    $f_{jk} := \text{mean}\{Y_{jk}^{(i)}: i \leq m/2\}$.
17: for $t = 0$ to $\tilde{n} - 1$ do
18:  Let $j := j_{B,n} + t$.
19:   for $\ell = 1$ to $2^t$ do
20:     for $k = (\ell - 1)2^j + 1$ to $\ell 2^j - 1$ do
21:       $f_{jk} := \text{mean}\{Y_{jk}^{(i)}: m/2 + t \left\lfloor \frac{m/2}{\eta} \right\rfloor + (\ell - 1) \left(2^t \left\lfloor \frac{m/2}{\eta} \right\rfloor + 1\right) < i \leq m/2 + t \left\lfloor \frac{m/2}{\eta} \right\rfloor + \tilde{\eta} - 1\}$.
22:     $\tilde{f}_j := \sum_{j=0}^{j_{\text{max}}} \sum_{k=0}^{2^j - 1} f_{jk} \psi_{jk}$.
23: (2) Lepski’s method:
24: for $j = 0$ to $j_{\text{max}}$ do
25:  $\hat{f}(j) := \sum_{j'=j-1}^{j_{\text{max}}} \sum_{k=0}^{2^j - 1} \hat{f}_{jk} \psi_{jk}$.
26:  Let $j := j_{\text{max}}$, $\text{stop} := \text{FALSE}$.
27: while $\text{stop} = \text{FALSE}$ and $j \geq 0$ do
29:   while $\text{stop} = \text{FALSE}$ and $l \leq j_{\text{max}}$ do
30:     if $\|\hat{f}(j) - \hat{f}(l)\|_2 \leq \tau 2^l/n_l$ then
32:     else $\text{stop} := \text{TRUE}$.
33:   if $\text{stop} = \text{FALSE}$ then
34:     $j := j - 1$.
35: Construct: $\hat{f} = \hat{f}(j)$.
Theorem 2.13. For every \( s, L > 0 \) the distributed method \( \hat{f} \) described above belongs to \( \mathcal{F}_{dist}(B, \ldots, B; B_{2,\infty}^{s}(L)) \) and for all \( s \in \left[s_{\min}, s_{\max}\right) \)

\[
\sup_{f_{0} \in B_{2,\infty}^{s}(L)} \mathbb{E}_{f_{0}, T} \left\| \hat{f} - f_{0} \right\|_{2} \lesssim \begin{cases} 
L^{1/(1+2s)} n^{-s/(1+2s)}, & \text{if } B \geq 4(L^{2}n)^{1/(1+2s)} \log_{2} n, \\
M_{n} L^{1/(1+2s)} \left( n^{1/(1+2s)} / B \log_{2} n \right)^{2/s}, & \text{if } B < 4(L^{2}n)^{1/(1+2s)} \log_{2} n,
\end{cases}
\]

with \( M_{n} = (\log_{2} n)^{s/(1+2s)} \).

Proof. See Section 3.6. \( \square \)

Remark 2.14. Compared to the lower bound in Corollary 2.2 one can observe that in case \( B \geq n^{1/(1+2s)} \log_{2} n \) the upper bound is sharp. For \( B < 4(L^{2}n)^{1/(1+2s)} \log_{2} n \) we might get an extra slowly varying term of order at most \( O((\log n)^{s/(1+2s)}) \). Also not that our method is sharp in the radius of the Besov ball \( L \).

Remark 2.15. The computational complexity of the adaptive algorithm in each local machine is \( O(Bn/(m \log_{2} n)) \), since \( B/\log_{2} n \) empirical wavelet coefficients \( \hat{f}_{j,k}^{(i)} \) are computed and each of them requires \( O(n/m) \) operations. In the central machine the computational complexity of the estimators \( \hat{f}_{j,k} \) for \( j < j_{B,n} \) is \( O(m) \), while for \( j_{B,n} \leq j \leq j_{\max} \) is \( O(m2^{j_{B,n} - j}/\log_{2} n) \), hence the total computational complexity of the estimators \( \hat{f}_{j,k} \), \( j \leq j_{\max} \), \( 0 \leq k \leq 2^{j_{1}} - 1 \) is \( O(mB/\log_{2} n) \). Then to compute \( \| \hat{f}(j) - \hat{f}(l) \|_{2}^{2}, j < l < j_{\max} \), requires \( O(2^{l}) = O(n^{1/(1+2s_{\min})}) \) operations, hence a conservative upper bound for the computational complexity of \( \hat{f} \) is \( O(n^{1/(1+2s_{\min})}\log_{2} n) \), but this could be further reduced by saving the values \( \| \hat{f}(j) - \hat{f}(j+1) \|_{2}^{2} \) and reusing them multiple times.

A slight modification of the above algorithm also leads to a (up to a logarithmic factor) minimax adaptive estimation rate in the \( L_{\infty} \)-norm. We construct the truncation estimator \( \hat{f}(j) \) as in Algorithm 5, see (2.5). The only difference to the \( L_{2} \)-case is that we introduce an extra \( l \) term in the definition of \( \hat{j} \), i.e.

\[
\hat{j} = \min \left\{ j \in \mathcal{J} : \| \hat{f}(j) - \hat{f}(l) \|_{\infty} \leq \tau \sqrt{l2^l/n_{j}}, \forall l > j, l \in \mathcal{J} \right\}.
\]

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

Theorem 2.16. For every \( L, s > 0 \) the distributed method \( \hat{f} \) described above belongs to \( \mathcal{F}_{dist}(B, \ldots, B; B_{2,\infty}^{s}(L)) \). Furthermore for all \( s \in \left[s_{\min}, s_{\max}\right) \)

\[
\sup_{f_{0} \in B_{2,\infty}^{s}(L)} \mathbb{E}_{f_{0}, T} \left\| \hat{f} - f_{0} \right\|_{\infty} \lesssim \begin{cases} 
L^{1/(1+2s)} (n/\log_{2} n)^{-s/(1+2s)}, & \text{if } B \geq \overline{B}, \\
M_{n} L^{1/(1+s)} \left( n^{1/(1+2s)} / B(\log_{2} n) \right)^{2/s}, & \text{if } B < \overline{B},
\end{cases}
\]

with \( \overline{B} = 4(L^{2}n(\log_{2} n)^{2s})^{1/(1+2s)} \) and \( M_{n} = (\log_{2} n)^{1+2s} \).

Proof. See Section 4.3. \( \square \)

3. Proofs for the \( L_{2} \)-norm.
3.1. Proof of Theorem 2.1. Note that without loss of generality we can multiply \( \delta_n \) with an arbitrary constant. In the proof we define \( \delta_n \) as the solution to

\[
\delta_n = C_1^{-1} L^{-2} \min \left\{ \frac{m}{n \log^2 n}, \frac{m}{n \sum_{i=1}^{m} (|\log \delta_n| / \log (n)) B(i)} \right\},
\]

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

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

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

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

\[
\|f_{\beta}\|_{B_{2, \infty}^2}^2 = \sup_j 2^{2j} \sum_{k=0}^{2^j-1} f_{\beta, jk}^2 \leq L^2 2^{2j_j} |J_{j_n}| \delta_n \leq L^2.
\]

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

\[
N^{\max}_t = \max_{f_{\beta} \in \mathcal{F}_0} \left| \{f_{\beta'} \in \mathcal{F}_0 : \|f_{\beta} - f_{\beta'}\|_2 \leq t\} \right|,
\]

\[
N^{\min}_t = \min_{f_{\beta} \in \mathcal{F}_0} \left| \{f_{\beta'} \in \mathcal{F}_0 : \|f_{\beta} - f_{\beta'}\|_2 \leq t\} \right|,
\]

satisfy \( N^{\max}_t = N^{\min}_t = \sum_{i=0}^{\left\lfloor (K_{j_n})^2 \right\rfloor} \frac{|J_{j_n}|}{2} \) for \( \tilde{t} = t^2/(4 \delta_n L^2) < |J_{j_n}|/2 \) (and therefore \( N^{\max}_t < |\mathcal{F}_0| - N^{\min}_t \)).

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

\[
F \rightarrow (T, X) \rightarrow Y.
\]

In this setting, the following general theorem applies. It is a slight extension of Corollary 1 of [5], see also Theorem A.6 with the corresponding proof in the supplement.
THEOREM 3.1. If the semimetric space \( (\mathcal{F}, d) \) contains a finite set \( \mathcal{F}_0 \) and \( |\mathcal{F}_0| - N_\ell^\text{min} > N_\ell^\text{max} \), then for all \( p, t > 0 \),

\[
\inf_{\hat{f} \in \mathcal{E}(Y)} \sup_{f \in \mathcal{F}} \mathbb{E}_f d^p(\hat{f}, f) \geq t^p \left( 1 - \frac{I(F; Y) + \log 2}{\log(|\mathcal{F}_0|/N_\ell^\text{max})} \right),
\]

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

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

\[
(3.4) \quad \inf_{\hat{f} \in \mathcal{F}_{\text{dist}}(B(1), \ldots, B(m); B^2_{\infty}(L))} \sup_{f_0 \in B^2_{\infty}(L)} \mathbb{E}_{f_0, T} \|\hat{f} - f_0\|_2^2 \geq L^2 \delta_n |K_{jn}| \left( 1 - \frac{I(F; Y) + \log 2}{\log(|\mathcal{F}_0|/N_\ell^\text{max})} \right),
\]

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 \( N_\ell^\text{max} = \sum_{i=0}^{\tilde{t}} (|K_{jn}|) < 2^{(|K_{jn}|)} \leq 2(e|K_{jn}|/\tilde{t})^\tilde{t} \) and therefore, for \( \tilde{t} = |K_{jn}|/6 \) (i.e. \( t^2 = 2L^2 \delta_n |K_{jn}|/3 \)),

\[
\log(|\mathcal{F}_0|/N_\ell^\text{max}) \geq |K_{jn}| \log(2(6e)^{-1/6}2^{-1/|K_{jn}|}) \geq |K_{jn}|/6.
\]

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

\[
(3.5) \quad I(F; Y) \leq |K_{jn}|/8 + O(1).
\]

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

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

\[
(3.6) \quad I(F; Y) \leq 4L^2 \mathbb{E}_f \|\psi\|_2^2 \delta_n |K_{jn}| \sum_{i=1}^m \left( 2^{12 \log_2(n)} |K_{jn}|^{-1} B^{(i)} \right) + O(1).
\]

Since in view of the definition of \( \delta_n \) we have that

\[
\delta_n \leq \frac{2^{12C_1-L-2m}}{n \sum_{i=1}^m \left( 2^{12 \log_2(n)} |K_{jn}|^{-1} B^{(i)} \right) + 1},
\]

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

3.2. Proof of Theorem 2.4. First note that by using Cauchy-Schwartz inequality we get that

\[
\mathbb{E}_{f_0, T} (\log_2 |\hat{f}^{(i)}_{jk}| \vee 1) \leq 1 + \mathbb{E}_{f_0, T} |\hat{f}^{(i)}_{jk}| = 1 + \mathbb{E}_{f_0, T} X^{(i)}_1 \psi_{jk}(T^{(i)}_1)
\]

\[
\leq 1 + \|f_0\|_2 \|\psi_{jk}\|_2 + \|\psi_{jk}\|_2 \mathbb{E}_{f_0} |\epsilon^{(i)}_1| = O(1).
\]
Hence in view of Lemma 2.3 (with $D = 1/2$) the approximation satisfies

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

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

Next for convenience we introduce the notation for the approximation error $W_{jk}^{(i)} = Y_{jk}^{(i)} - f_{jk}^{(i)}$, satisfying $|W_{jk}^{(i)}| \leq n^{-1/2}$. The estimator $\hat{f}$ is given by its wavelet coefficients $\hat{f}_{jk}$, $j \in \mathbb{N}, k \in \{0, 1, \ldots, 2^j - 1\}$. For $2^j + k > (L^2 n)^{1/(1+2s)} \wedge [B/\log_2 n]$ we have $\hat{f}_{jk} = 0$, while for $2^j + k \leq (L^2 n)^{1/(1+2s)} \wedge [B/\log_2 n]$, we need at most $(1/\sqrt{n})B$ bits in expected value to transmit $\{\hat{f}_{jk} \in \{0, 1, \ldots, 2^j\} \}$.

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

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

$$\mathbb{E}_{f_0,T} Z_{jk}^2 \leq 2 \mathbb{E}_{f_0,T} \left( \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{m} f_0(T^{(i)}_\ell) \psi_j(T^{(i)}_\ell) - \mathbb{E}_{f_0,T} f_0(T^{(i)}_1) \psi_j(T^{(i)}_1) \right)^2$$

$$+ 2 \mathbb{E}_{f_0,T} \left( \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{m} \varepsilon^{(i)}_\ell \psi_j(T^{(i)}_\ell) \right)^2$$

$$\leq 2n^{-1} \mathbb{E}_T \left( f_0(T^{(i)}_1) \psi_j(T^{(i)}_1) - \mathbb{E}_T f_0(T^{(i)}_1) \psi_j(T^{(i)}_1) \right)^2$$

$$+ 2n^{-1} \mathbb{E}_{f_0} (\varepsilon^{(i)}_1)^2 \mathbb{E}_T \psi_j^2(T^{(i)}_1)$$

$$\leq 2n^{-1} \int_0^1 f_0^2(t) \psi_j^2(t) dt + 2n^{-1} \leq 2(M + 1)/n.$$

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

$$\mathbb{E}_{f_0,T} \| \hat{f} - f_0 \|^2 \leq \sum_{j \geq j_n} \sum_{k=0}^{2^j - 1} \mathbb{E}_{f_0,T} Z_{jk}^2 + \sum_{j=0}^{2^j - 1} \mathbb{E}_{f_0,T} W_{jk}^2$$

$$\leq \sum_{j \geq j_n} 2^{-2js} \sup_{j \geq j_n} \sum_{k=0}^{2^j - 1} f_0^2_{jk} + \sum_{j=0}^{2^j - 1} \sum_{k=0}^{j_n} n^{-1}$$

$$\leq L^2 2^{-2js} + 2n^2/n \leq (L^2 n^{-2s/(1+2s)}) \vee (L^2 (B/\log_2 n)^{-2s}).$$

### 3.3. Proof of Theorem 2.5

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

Next let $A_{jk} = \{\lfloor jk/m \rfloor + 1, \ldots, \lfloor (jk + 1)m/n \rfloor\}$ be the collection of machines transmitting the $(j, k)$th approximated wavelet coefficient $Y_{jk}^{(i)}$ and note that the size of the set satisfies $|A_{jk}| \asymp m/n$. Then our aggregated estimator $\hat{f}$ satisfies for $i = 1, \ldots, \hat{f}_n \in \mathcal{F}_{dist}(B, \ldots, B; B^2_{2,\infty}(L))$, hence $\hat{f}_n \in \mathcal{F}_{dist}(B, \ldots, B; B^2_{2,\infty}(L))$. 

$$\hat{f}_n \in \mathcal{F}_{dist}(B, \ldots, B; B^2_{2,\infty}(L)).$$
\[ 2^j + k \leq \eta |B/\log_2 n| \] (i.e. the total number of different coefficients transmitted) that
\[
\hat{f}_{jk} = \frac{1}{|A_{jk}|} \sum_{i \in A_{jk}} Y_{jk}^{(i)} = f_{0,jk} + Z_{jk} + W_{jk},
\]
where \(|W_{jk}| = \frac{1}{|A_{jk}|} \sum_{i \in A_{jk}} W_{jk}^{(i)}| \leq n^{-1/2}\) and \(Z_{jk} = \frac{1}{|A_{jk}|} \sum_{i \in A_{jk}} (\hat{f}_{jk}^{(i)} - E_{f_{0,T}}\hat{f}_{jk}^{(i)})\).

Note that similarly to above
\[
E_{f_0,T}Z_{jk}^2 \leq 2E_{f_0,T}\left( \frac{m}{n|A_{jk}|} \sum_{i \in A_{jk}} \sum_{\ell = 1}^{n/m} f_0(T_{\ell}^{(i)}\psi_{jk}(T_{\ell}^{(i)}) - E_{f_0,T}f_0(T_{\ell}^{(i)}\psi_{jk}(T_{\ell}^{(i)})) \right)^2
\]
\[
+ 2E_{f_0,T}\left( \frac{m}{n|A_{jk}|} \sum_{i \in A_{jk}} \sum_{\ell = 1}^{n/m} \epsilon_{\ell}^{(i)}\psi_{jk}(T_{\ell}^{(i)}) \right)^2
\]
\[
\leq \frac{2m}{n|A_{jk}|} E_{f_0,T}\left( f_0(T_{1}^{(1)}\psi_{jk}(T_{1}^{(1)}) - E_{f_0,T}f_0(T_{1}^{(1)}\psi_{jk}(T_{1}^{(1)})) \right)^2
\]
\[
+ \frac{2m}{n|A_{jk}|} E_{f_0,T}\left( \epsilon_{1}^{(1)} \right)^2 E_{f_0,T}\psi_{jk}^2(T_{1}^{(1)}) \leq \frac{2(M^2 + 1)m}{n|A_{jk}|}.
\]

Let \(j_n = |\log_2(\eta |B/\log_2 n|)|\). Then similarly to (3.7) the risk of the aggregated estimator is bounded as
\[
E_{f_0,T}\|\hat{f} - f_0\|^2 \leq \sum_{j \geq j_n} \sum_{k = 0}^{2^j - 1} f_{0,jk}^2 + 2 \sum_{j = 0}^{j_n} \sum_{k = 0}^{2^j - 1} E_{f_0,T}(Z_{jk}^2 + W_{jk}^2)
\]
\[
\lesssim \sum_{j \geq j_n} 2^{-2js} \sup_{j \geq j_n} 2^{2js} \sum_{k = 0}^{2^j - 1} f_{0,jk}^2 + \sum_{j = 0}^{j_n} \sum_{k = 0}^{2^j - 1} \eta/n
\]
\[
\lesssim L^2 \left( B\eta \log_2 n \right)^{-2s} + \frac{Bn^2}{n \log_2 n}
\]
\[
\times \left( L^{\frac{2}{1+s}} (nB/\log_2 n)^{-\frac{s}{1+s}} \right) \vee \left( L^2 \left( \frac{Bm}{\log_2 n} \right)^{-2s} \right)
\]
\[
= \left( L^{\frac{2}{1+s}} (\log_2 n)^{\frac{2s}{1+s}} \left( \frac{n^{1/1+2s}}{B \log_2 n} \right)^{\frac{s}{1+s}} \right) \vee \left( L^2 \left( \frac{Bm}{\log_2 n} \right)^{-2s} \right),
\]
concluding the proof of the theorem.

Finally we show that Algorithm 2 is in general suboptimal in this case. Consider the function \(f_0 \in B^2_{2,\infty}(1)\) with wavelet coefficients \(f_{0,jk} = 2^{-j(s+1)/2}\), \(j \in \mathbb{N}, k = 0, ..., 2^j - 1\), and take \(j_n = |\log_2(B/\log_2 n)|\), then
\[
E_{f_0,T}\|\hat{f} - f_0\|^2 \geq \sum_{j \geq j_n} \sum_{k = 0}^{2^j - 1} f_{0,jk}^2 \geq \sum_{k = 0}^{2^{j_n}(2s+1)} 2^{-j_n(2s+1)}
\]
\[
\gtrsim \left( \frac{B}{\log_2 n} \right)^{-2s} = \tilde{M}_n \left( \frac{n^{1/1+2s}}{B \log_2 n} \right)^{\frac{s}{1+s}} n^{-\frac{2s}{1+s}},
\]
where the multiplication factor \(\tilde{M}_n = \left( \frac{n\log_2 n}{B^{1+2s}} \right)^{\frac{s}{1+s}}\) tends to infinity and can be of polynomial order, yielding a highly sub-optimal rate.
3.4. **Proof of Lemma 2.3.** One can easily see by construction that

\[
0 \leq |X - Y| \leq n^{-D}.
\]

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

\[
\mathbb{E}(1 + (1 \lor \log_2 |X|) + D \log_2 n) = 1 + D \log_2(n) + \mathbb{E}(1 \lor \log_2 |X|)
\]

\[
= (D + o(1)) \log_2 n.
\]

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

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

\[
P_{\theta_0}(E_{jk}^c) \leq m \sum_{i=1}^m P_{\theta_0}(|\tilde{f}_{jk}^{(i)}| \geq n) \leq m P_{\theta_0}(\frac{\sum_{i=1}^m X_{\ell}^{(i)} \psi_{jk}(T_{\ell}^{(i)})}{n} \geq n)
\]

\[
\leq m P_{\theta_0}(2^{j/2} \|\psi\|_{\infty} \max_{\ell} |Z_{\ell}^{(i)}| + M) \geq n)
\]

\[
\leq n P_{\theta_0}(|Z_{\ell}^{(i)}| \geq c_1 \sqrt{n}) \leq ne^{-c_2 n} = o(n^{-2}),
\]

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

\[
\mathbb{E}_{f_0, T} \|\tilde{f} - f_0\|_2^2 = \sum_{j \geq j_n} \sum_{k=0}^{2^j-1} f_{0,jk}^2 + \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} \mathbb{E}_{f_0, T}(f_0,jk - \tilde{f}_{jk})^2 1_{E_{jk}}
\]

\[
+ \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} \mathbb{E}_{f_0, T}(f_0,jk - \tilde{f}_{jk})^2 1_{E_{jk}^c}
\]

\[
\leq \sum_{j \geq j_n} 2^{-2j} \sup_{l \geq j_n} 2^{2l} \sum_{k=0}^{2^l-1} f_{0,jk}^2 + \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} \mathbb{E}_{f_0, T}(f_0,jk - \tilde{f}_{jk})^2
\]

\[
+ \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} f_{0,jk}^2 P_{\theta_0}(E_{jk}^c)
\]

\[
\leq L^2 2^{-2j_n} + \sum_{j=0}^{j_n} \sum_{k=0}^{2^j-1} \mathbb{E}_{f_0, T}(Z_{jk}^2 + W_{jk}^2) + o(n^{-1}).
\]

We conclude the proof by noting that the first two terms on the right hand side have the required upper bounds, see the proofs of Theorems 2.4 and 2.5, while the third term is of smaller order than the previous one.
3.6. Proof of Theorem 2.13. First recall that for every \( s, L > 0 \) and \( f_0 \in B_{2,\infty}^s(L) \) we have \( f_{0,jk}^2 \leq L^2 \), \( j \geq 0, k \in \{0, \ldots, 2^j - 1\} \). Therefore, in view of Lemma 2.3 (with \( D = 1/2 \)) we have \( \mathbb{E}_{f_0}^{T} [l(Y_{jk}^{(i)})] \leq (1/2 + o(1)) \log_2 n \). Since the machines in group \( I \) and the machines in \( I_{t,\ell}, t \in \{0, \ldots, \tilde{\eta} - 1\}, \ell \in \{1, \ldots, 2^j\} \) transmit at most \( \lceil B/\log_2 n \rceil \) coefficients we have that in expected value at most

\[
\lceil B/\log_2 n \rceil \left( \frac{1}{2} + o(1) \right) \log_2 n \leq B
\]

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

Next we show that the estimator \( \hat{f} \) achieves the minimax rate. First let us introduce the notation \( |W_{jk}^{(i)}| = |Y_{jk}^{(i)} - f_{jk}^{(i)}| \leq n^{-1/2} \). Then note that for \( j \leq j_{\text{max}} \) and \( k \in \{0, \ldots, 2^j - 1\} \) the aggregated quantities \( \hat{f}_{jk} \) defined in (2.4) are equal to

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

where

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

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

Note that the squared bias satisfies

\[
\| \mathbb{E}_{f_0,T} \hat{f}(j) - f_0 \|^2_2 \lesssim \|K(f_0,j) - f_0\|^2_2 + 2^j/n \lesssim 2^{-j_s}\|f_0\|^2_{B_{2,\infty}^s} + 2^j/n,
\]

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

\[
\mathbb{E}_{f_0,T} \| \hat{f}(j) - \mathbb{E}_{f_0,T} \hat{f}(j) \|_2^2 \lesssim \sum_{\ell \leq j-1} \sum_{k=0}^{2^\ell-1} \left( \mathbb{E}_{f_0,T} Z_\ell^2 + \mathbb{E}_{f_0,T} W_\ell^2 \right)
\]

\[
\lesssim \sum_{\ell \leq j-1} \sum_{k=0}^{2^\ell-1} n_\ell^{-1} \leq 2^j/n_j.
\]

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

\[
\begin{align*}
\hat{j}^* = \min \{ & j \in J : B(j, f_0) \leq 2^j/n_j \},
\end{align*}
\]

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

\[
B(j, f_0) \leq 2^j/n_j, \text{ for } j \geq \hat{j}^* \quad \text{and} \quad B(j, f_0) > 2^j/n_j, \text{ for } j < \hat{j}^*.
\]

Therefore

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

We summarize these findings in the following displays

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

and

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

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

Let us split the risk into two parts

\begin{equation}
E_{f_0,T}\|f_0 - \hat{f}\|_2 = E_{f_0,T}\|f_0 - \hat{f}(\hat{j})\|_{21_{j < j^*}} + E_{f_0,T}\|f_0 - \hat{f}(\hat{j})\|_{21_{j \leq j^*}},
\end{equation}

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

\begin{equation}
E_{f_0,T}\|f_0 - \hat{f}(\hat{j})\|_{21_{j < j^*}} \leq 2E_{f_0,T}\|\hat{f}(\hat{j}) - \hat{f}(\hat{j}^*)\|_{21_{j < j^*}} + 2E_{f_0,T}\|\hat{f}(\hat{j}^*) - f_0\|_2^2 \\
\lesssim 2^{j^*/n_{j^*}} + \|E_{f_0,T}\hat{f}(\hat{j}^*) - f_0\|_2^2 + E_{f_0,T}\|\hat{f}(\hat{j}^*) - E_{f_0,T}\hat{f}(\hat{j}^*)\|_2^2 \\
\lesssim 2^{j^*/n_{j^*}} + \|f_0\|_{B^2_{2,\infty}}^2 2^{-2j^*},
\end{equation}

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

\begin{equation}
E_{f_0,T}\|f_0 - \hat{f}\|_{21_{j \leq j^*}} \lesssim \begin{cases}
\|f_0\|_{B^2_{2,\infty}}^2 n^{-2s/(1+2s)}, & \text{if } B \geq \overline{B}, \\
B/n, & \text{if } B \leq B \leq \overline{B}, \\
\|f_0\|_{B^2_{2,\infty}}^2 \left(nB/\log_2 n\right)^{-2s/(2+2s)}, & \text{if } B < B.
\end{cases}
\end{equation}

Since $\|f_0\|_{B^2_{2,\infty}} \leq L$ the preceding upper bounds are bounded from above by the ones stated in the theorem.
Next we deal with the first term on the right hand side of (3.16). By Cauchy-Schwarz inequality and Lemma 3.3 we get that
\[
\mathbb{E}_{f_0,T} \| f_0 - \hat{f} \|_2^2 1_{j > j^*} \leq \sum_{j = j^*+1}^{j_{\text{max}}} \mathbb{E}_{f_0,T}^{1/2} \| f_0 - \hat{f}(j) \|^2_2 \mathbb{P}_{f_0,T}^{1/2}(\hat{j} = j) \\
\lesssim \sum_{j = j^*+1}^{j_{\text{max}}} \mathbb{P}_{f_0,T}(\hat{j} = j) \lesssim j_{\text{max}} e^{-(cn^s \wedge \sqrt{m})} + \sum_{k=1}^{\infty} e^{-(c/2)2^s k} \\
= o(n^{-1}) + o(2^{-j^*}),
\]
resulting in the required upper bound in view of (3.17), concluding the proof of our statement.

**Lemma 3.3.** Assume that \( f_0 \in B^s_{2,\infty}(L) \), for some \( s, L > 0 \). Then there exists a universal constants \( c, \delta > 0 \) such that for every \( j > j^* \) we have
\[
\mathbb{P}_{f_0,T}(\hat{j} = j) \lesssim e^{-(c^2/n^4 \wedge \sqrt{m})}.
\]

**Proof.** Let us introduce the notation \( j^- = j - 1 \) and note that for every \( j > j^* \) we have \( j^- \geq j^* \). Then by the definition of \( \hat{j} \)
\[
\mathbb{P}_{f_0,T}(\hat{j} = j) \leq \sum_{l = j}^{j_{\text{max}}} \mathbb{P}_{f_0,T}(\| \hat{f}(j^-) - \hat{f}(l) \|^2_2 > 2^l/n_l).
\]
Note that the left hand side term in the probability in view of Parseval’s inequality can be given in the form
\[
\| \hat{f}(j^-) - \hat{f}(l) \|^2_2 = \sum_{r = j^-}^{l-1} \sum_{k=0}^{2^r-1} \left( f_{0,r,k} + Z_{r,k} + W_{r,k} \right)^2 \\
\leq 3 \sum_{r = j^-}^{l-1} \sum_{k=0}^{2^r-1} \left( f_{0,r,k}^2 + Z_{r,k}^2 + W_{r,k}^2 \right).
\]
We deal with the three terms on the right hand side separately. Note that the functions \( j \mapsto B(j, f_0) \) and \( j \mapsto n_j \) are monotone decreasing, hence by the definition of \( j^* \) we get for \( l \geq j^- \geq j^* \)
\[
\sum_{r = j^-}^{l-1} \sum_{k=0}^{2^r-1} f_{0,r,k}^2 \leq B(j^+, f_0) \leq B(j^*, f_0) \leq 2^{j^*}/n_{j^*} \leq 2^l/n_l.
\]
Furthermore \( \sum_{r = j^-}^{l-1} \sum_{k=0}^{2^r-1} W_{r,k}^2 \leq 2^l/n \leq 2^l/n_l. \)

Let \( S(r) = \{ \sum_{k=0}^{2^l-1} b_{lk} \psi_{lk} : \sum_{k=0}^{2^l-1} b_{lk}^2 = 1 \} \) denote the unite sphere in the linear subspace spanned by the basis functions \( \psi_{lk} \), \( l \leq r, 0 \leq k \leq 2^l - 1 \). Then in view of Lemma 5.3 of [3], see also Lemma C.4 in the supplement, and the
\[ (\alpha + \beta)^2 \leq 2\alpha^2 + 2\beta^2 \]

we get that
\[
\sum_{k=0}^{2^r-1} Z_{rk}^2 = \sum_{k=0}^{2^r-1} \left( \frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} (Y_T(i) \psi_{jk}(T_{\ell}^{(i)}) - \mathbb{E} f_0 T Y_T(i) \psi_{jk}(T_{\ell}^{(i)})) \right)^2 \\
\leq 2 \sup_{g \in S(r)} \left( \frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} (f_0(T_{\ell}^{(i)}) g(T_{\ell}^{(i)}) - \mathbb{E}_T f_0(T_{\ell}^{(i)}) g(T_{\ell}^{(i)}))^2 \right) \\
+ 2 \sum_{k=0}^{2^r-1} \left( \frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} (\varepsilon_{\ell}^{(i)} \psi_{jk}(T_{\ell}^{(i)}))^2 \right)^2.
\]

We deal with the two terms on the right hand side separately, starting with the first one. Note that for every \( g \in S(r) \) the inequality \( \|g\|_\infty \leq C 2^{r/2} \) holds, for some universal constant \( C > 0 \) and
\[
\sup_{g \in S(r)} \mathbb{V}_T (f_0(T_{1}^{(i)}) g(T_{1}^{(i)})) \leq \|f_0\|_\infty^2.
\]

Next for convenience let us introduce the notation
\[
\nu(g) = \frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \left( f_0(T_{\ell}^{(i)}) g(T_{\ell}^{(i)}) - \mathbb{E}_T f_0(T_{\ell}^{(i)}) g(T_{\ell}^{(i)}) \right).
\]

Then by the definition of \( S(r) \) and Cauchy-Schwarz inequality
\[
\mathbb{E}_T \sup_{g \in S(r)} |\nu(g)| \leq \sum_{k=0}^{2^r-1} \mathbb{E}_T (\nu(\psi_{rk,k})^2) = \sum_{k=0}^{2^r-1} \frac{1}{n_r} \mathbb{V}_T (f_0(T_1^{(i)}) \psi_{rk}(T_1^{(i)})) \leq \frac{\|f_0\|_\infty^2 2^r}{n_r}.
\]

Therefore in view of Lemma 5 of [11], see also Lemma C.1 in the supplement, there exist constants \( c_1, c_2, c_3 > 0 \) such that
\[
\mathbb{E}_T \sup_{g \in S(r)} \left[ \left( \frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} (f_0(T_{\ell}^{(i)}) g(T_{\ell}^{(i)}) - \mathbb{E}_T f_0(T_{\ell}^{(i)}) g(T_{\ell}^{(i)}))^2 \right) - c_1 2^r / n_r \right] \\
\leq c_2 \frac{1}{n_r} e^{-c_3 2^r} + c_4 \frac{2^r}{n_r^2} e^{-\sqrt{n_r}} \leq \frac{1}{n_r} e^{-(c_3 2^r \wedge \sqrt{n_r})}.
\]

Therefore by Markov’s inequality we get that
\[
\mathbb{P}(T \left( \sup_{g \in S(r)} \left( \frac{1}{n_r} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} (f_0(T_{\ell}^{(i)}) g(T_{\ell}^{(i)}) - \mathbb{E}_T f_0(T_{\ell}^{(i)}) g(T_{\ell}^{(i)}))^2 \right) \right) \geq \frac{2c_1 2^r}{n_r} \right) \leq 2^{-r} e^{-(c_3 2^r \wedge \sqrt{n_r})}.
\]

Next we deal with the second term on the right hand side of (3.18). Let us introduce the shorthand notation \( \hat{Z}_{rk} = n_r^{-1} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \varepsilon_{\ell}^{(i)} \psi_{rk}(T_{\ell}^{(i)}) \). Note that \( \text{cov}(\hat{Z}_{rk}, \hat{Z}_{rk'}|T) = 0 \) for \( |k - k'| \geq C \), for some large enough constant \( C \), following from the disjoint support of the wavelet basis functions \( \psi_{rk} \) and \( \psi_{rk'} \), and
\[
\hat{Z}_{rk}|T \sim N(0, \frac{1}{n_r^2} \sum_{i \in A_{rk}} \sum_{\ell=1}^{n/m} \psi_{rk}(T_{\ell}^{(i)})^2).
\]
Furthermore, let us denote by $B_r$ the event that in each bin $I_{r,t} = [(l-1)2^{-r}, l2^{-r}]$, at most $2n_r/2^r$ observations $T^{(i)}_{\xi_k^r}$, $i \in A_k$, $\xi = 1, \ldots, m$, $k = 0, \ldots, 2^r - 1$ fall. Since there are $2^r - 1$ $n_r$ subgroups of machines at resolution level $r$ we note that in view of Lemma 5.2 we have that $\mathbb{P}(B_r) \leq 2^{2r+1} e^{-n_r^2 - r^3}$. Then by recalling that for $r \leq jB_n$, $n_r \geq n$ while for $r > jB_n$, $n_r = nB/(2^r \log^2 n)$, we get that $n_r/2^r \geq (nB/\log^2 n)^{1/2\delta_{\min}} \wedge n^{1/2\delta_{\min}}$, hence

\begin{equation}
\mathbb{P}(B_r) \lesssim e^{-n^\delta}, \quad \text{for any } \delta < 2\delta_{\min}/(2 + 2\delta_{\min}),
\end{equation}

and on $B_r$ the inequality $n_r^{-2} \sum_{i \in A_k} \sum_{\ell=1}^{n/m} \psi_r(T^{(i)}_\ell)^2 \leq Cn_r^{-1}$ holds, for some sufficiently large $C > 0$. Let us denote the covariance matrix of the random vector $(\hat{Z}_{r0}, \ldots, \hat{Z}_{r(2^r - 1)})^T$ by $\Sigma_r$. In view of the preceding argument the in absolute value largest entry of $\Sigma_r$ is bounded from above by $Cn_r^{-1}$ on the event $T \in B_r$, and by noting that $\Sigma_r$ has band size $C$, in view of Gershgorin circle theorem [6], see also Lemma C.3 in the supplement, the eigenvalues of $\Sigma_r$ satisfy that $0 < \lambda_i \leq Cn_r^{-1}$, $i = 1, \ldots, 2^r$. Then by the tail bounds of chi-square distributions, see for instance Theorem 4.1.9 of [7] (or Lemma C.2 in the supplement),

\[ \mathbb{P}(f_0 \left( \sum_{k=0}^{2^r-1} \hat{Z}_{rk}^2 | T = t \right) = \mathbb{P}(\sum_{i=1}^{2^r} \lambda_i \zeta_i \geq C_1 2^r \mathbb{P}(\sum_{i=1}^{2^r} \zeta_i^2 \geq C_2 2^r) \lesssim e^{-C_3 2^r},
\]

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

\[ \mathbb{P}(f_0, T \left( \sum_{k=0}^{2^r-1} \left( \frac{1}{n_r} \sum_{i \in A_k} \sum_{\ell=1}^{n/m} \zeta^{(i)}_\ell \psi_r(T^{(i)}_\ell) \right)^2 \geq C_1 2^r \right) \]

\[ \leq \int_{t \in B_r} \mathbb{P}(f_0 \left( \sum_{k=0}^{2^r-1} \hat{Z}_{rk}^2 | T = t \right) dt + \mathbb{P}(B_r^c) \lesssim e^{-(C2^r \wedge n^\delta)},
\]

finishing the proof of the lemma. \hfill $\square$

4. Proofs for the $L_\infty$-norm.

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

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

\[ \|f_\beta\|_{B_{r, \infty}} = \sup_j 2^{(s+1/2)j} \sup_{k=0, \ldots, 2^r-1} f_{\beta,jk} = 2^{(s+1/2)j} \Delta_0^{1/2} \leq L.
\]

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

\[ \|f_\beta - f_\beta'\|_{\infty} \geq |f_{\beta,jn_k} - f_{\beta',jn_k}| \cdot \|\psi_{jn_k}\|_{\infty} \gtrsim L2^{jn} \Delta_0^{1/2} \geq \Delta_0^{\frac{s}{1+2s}}.
\]
Now let $F$ be a uniform random variable on the set $\mathcal{F}_0$. Then in view of Fano’s
inequality (see for instance Theorem A.5 in the supplement with $\delta = \delta_n^{s/(1+2s)}$ and
$p = 1$) we get that

$$\inf_{\hat{f} \in \mathcal{F}_{\text{dist}}(B^{(1)}, \ldots, B^{(m)}; B_{\infty, \infty}(L))} \sup_{f_0 \in B_{\infty, \infty}(L)} \mathbb{E}_{f_0, T} \left( \| \hat{f} - f_0 \|_\infty \right) \geq L \delta_n^{\frac{1}{2s}} \left( 1 - \frac{I(F; Y) + \log 2}{\log_2 |\mathcal{F}_0|} \right).$$

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

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

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

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

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

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

Next note that by triangle inequality

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

We deal with the two terms on the right hand side separately. Let us introduce the
notation $j_n = \lfloor \log_2 \left( (\eta B/ \log_2 n) \land (L^2/n/ \log_2 n)^{1/(1+2s)} \right) \rfloor \leq \log_2 (n/\eta)$. Then by triangle inequality and noting that there exists a universal constant $C > 0$
such that for each resolution level $j$ the inequality $\| \sum_{k = 0}^{2^j - 1} |\psi_{jk}| \|_\infty \leq C 2^{j/2}$ holds,

$$\| f_0 - \mathbb{E}_{f_0, T}\hat{f} \|_\infty \leq \| \sum_{j = j_n}^{\infty} \sum_{k = 0}^{2^j - 1} f_{0,jk} \psi_{jk} \|_\infty + \| \sum_{j = j_n}^{\infty} \sum_{k = 0}^{2^j - 1} \mathbb{E}_{f_0, T} W_{jk} \psi_{jk} \|_\infty$$

$$\leq \| f_0 \|_{B_{\infty, \infty}} \sum_{j = j_n}^{\infty} 2^{-j(s+1/2)} \| \sum_{k = 0}^{2^j - 1} |\psi_{jk}| \|_\infty + n^{-1/2} \sum_{j = j_n}^{\infty} \sum_{k = 0}^{2^j - 1} \| \psi_{jk} \|_\infty$$

$$\lesssim L \sum_{j = j_n}^{\infty} 2^{-js} + \sqrt{2^{jn}/n} \lesssim L 2^{-jn_s} + \sqrt{2^{jn}/n}. \tag{4.2}$$
Furthermore, in view of (4.1),
\[
\mathbb{E}_{f_0,T} \| \hat{f} - \mathbb{E}_{f_0,T} \hat{f} \|_\infty \leq \sum_{j=0}^{j_n} \mathbb{E}_{f_0,T} \max_k (|Z_{jk}| + |W_{jk}|) \sum_{k=0}^{2^j-1} |\psi_{jk}| \| \psi_{jk} \|_\infty
\]
(4.3)
\[
\lesssim \sum_{j=0}^{j_n} 2^{j/2} \left( \sqrt{\log_2 n} \eta/n + \sqrt{1/n} \right) \lesssim \sqrt{2^{j_n} \eta (\log_2 n)/n},
\]
providing the upper bound in the statement of the lemma.

It remained to prove assertion (4.1). First note that
\[
Z_{jk}|T \sim N(\mu_{n,m,k,T}, \sigma^2_{n,m,k,T}), \quad \text{with}
\]
\[
\mu_{n,m,k,T} = \frac{\eta}{n} \sum_{i \in A_{jk}} \psi_{jk}(T^{(i)}_\ell) f_0(T^{(i)}_\ell) - f_0,jk \lesssim 2^{j/2},
\]
\[
\sigma^2_{n,m,k,T} = (\frac{\eta}{n})^2 \sum_{i \in A_{jk}} \psi_{jk}(T^{(i)}_\ell) \lesssim 2^{j} \eta/n.
\]
Using standard bounds on the maximum of Gaussian variables (see for instance Lemma 3.3.4 of [7]) we have that
\[
\mathbb{E}_{f_0,T} \max_k |Z_{jk} - \mathbb{E}_{f_0,T} Z_{jk}| \leq \sqrt{2(j + 1)} \max_k \sigma_{n,m,k,T}.
\]
Furthermore, note that for \( k \geq 2 \)
\[
\mathbb{E}_T (\psi_{jk}(T^{(i)}_\ell) f_0(T^{(i)}_\ell))^k \leq \|I_0\|^k \|\psi_{jk}\|_\infty^k \mathbb{E}_T (\psi_{jk}(T^{(i)}_\ell))^2 \lesssim 2^{(k-2)j/2},
\]
hence in view of Bernstein’s inequality (with \( c = C2^{j/2} \) and \( v = Cn/\eta \), see Proposition 2.9 of [13] (or Lemma C.5 in the supplement), we get that
\[
\mathbb{P}_T \left( |\mu_{n,m,k,T}| \geq C \sqrt{\gamma (\log_2 n) \eta/n} \right) \lesssim (n/\eta)^{-\gamma},
\]
which implies for \( 2^j \leq n/\eta \) that
(4.4)
\[
\mathbb{P}_T \left( \max_k |\mu_{n,m,k,T}| \geq C \sqrt{\log_2 n} \eta/n \right) \lesssim (n/\eta)^{-\gamma+1}.
\]
Therefore one can deduce that
\[
\mathbb{E}_T \left( \max_k |\mu_{n,m,k,T}| \right) \leq C \sqrt{\log_2 n} \eta/n + 2^{j/2} (n/\eta)^{-\gamma+1}
\]
\[
\lesssim \sqrt{\log_2 n} \eta/n,
\]
for large enough choice of \( \gamma > 0 \). Combining the above displays leads to
\[
\mathbb{E}_{f_0,T} \max_k |Z_{jk}| = \mathbb{E}_T (\mathbb{E}_{f_0,T} (\max_k |Z_{jk}|))
\]
\[
\leq \mathbb{E}_T \left( \max_k |\mu_{n,m,k,T}| \right) + \sqrt{2(j + 1)} \mathbb{E}_T \max_k \sigma_{n,m,k,T}
\]
\[
\leq c \sqrt{\log_2 n} \eta/n + 2^{j/2} \sqrt{\gamma e^{-cn^2}} \leq C \sqrt{\log_2 n} \eta/n,
\]
for some large enough constants \( c, C > 0 \) and \( 2^j \leq n/\eta \), where in the last line we have used that under the event \( B_j \) (i.e. the event that in each bin \( I_{j,k} = [(l - 1)2^{-j}, l2^{-j}], \) at most \( 2n/(\eta 2^j) \) observations \( T^{(i)}_\ell, i \in A_{jk}, \ell = 1, \ldots, n/m, k = 0, \ldots, 2^j - 1 \) fall) we have that \( \max_k \sigma^2_{n,m,k,T} \leq C \), and \( \mathbb{P}_T (B_j) \leq Ce^{-cn^2} \), see (3.21).
4.3. **Proof of Theorem 2.16.** The proof of the theorem goes similarly to the proof of Theorem 2.13, here we only highlight the differences. First recall that for every \( s, L > 0 \) and \( f_0 \in \mathcal{B}_{\infty,\infty}^s (L) \) we have \( f_{0,jk} \leq L \), for all \( j \geq 0, k \in \{0, 1, ..., 2^j - 1\} \), hence following from the same argument as in Theorem 2.13, the estimator belongs to \( \mathcal{F}_{dist}(B, \ldots, B; \mathcal{B}_{\infty,\infty}^s (L)) \).

Let us next introduce the notations \( B(j, f_0) = 2^{-js} \| f_0 \|_{\mathcal{B}_{\infty,\infty}^s} \) and

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

Then by the definition of \( j^* \) we have

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

Distinguish again three cases according to the value of \( j^* \) we get that

\[
2^j \asymp \begin{cases} (\| f_0 \|_{\mathcal{B}_{\infty,\infty}^s} n / \log_2 n)^{1/(1+2s)}, & \text{if } B \geq \overline{B}, \\ B / \log_2 n, & \text{if } B \leq B < \overline{B}, \\ (nB\| f_0 \|_{\mathcal{B}_{\infty,\infty}^s}^2 / \log_2 n)^{1/(2+2s)}, & \text{if } B < \underline{B}, \end{cases}
\]

and

\[
n_{j^*} \asymp \begin{cases} n, & \text{if } B \geq \overline{B}, \\ n / \log_2 n, & \text{if } B \leq B < \overline{B}, \\ \| f_0 \|_{\mathcal{B}_{\infty,\infty}^s}^{1/(1+s)} (nB / \log_2 n)^{\frac{1+4s}{1+2s}}, & \text{if } B < \underline{B}. \end{cases}
\]

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

We split the risk into two parts

\[
\mathbb{E}_{f_0,T} \| f_0 - \hat{f} \|_\infty = \mathbb{E}_{f_0,T} \| f_0 - \tilde{f}(\hat{j}) \|_\infty 1_{j > j^*} + \mathbb{E}_{f_0,T} \| f_0 - \tilde{f}(\hat{j}) \|_\infty 1_{j \leq j^*}
\]

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

\[
\mathbb{E}_{f_0,T} \| f_0 - \tilde{f}(\hat{j}) \|_\infty 1_{j \leq j^*} \leq \mathbb{E}_{f_0,T} \| \tilde{f}(j^*) - \tilde{f}(\hat{j}) \|_\infty 1_{j \leq j^*} + \mathbb{E}_{f_0,T} \| \tilde{f}(j^*) - f_0 \|_\infty + \mathbb{E}_{f_0,T} \| \tilde{f}(j^*) - \mathbb{E}_{f_0,T} \tilde{f}(j^*) \|_\infty
\]

\[
\lesssim \sqrt{(\log_2 n)2^{j^*}/n_{j^*}} + \| f_0 \|_{\mathcal{B}_{\infty,\infty}^s} 2^{-j^*s},
\]

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

\[
\mathbb{E}_{f_0,T} \| f_0 - \hat{f} \|_\infty 1_{j \leq j^*} \lesssim \begin{cases} \| f_0 \|_{\mathcal{B}_{\infty,\infty}^s}^{1/(1+2s)} (n / \log_2 n)^{-s/(1+2s)}, & \text{if } B \geq \overline{B}, \\ \sqrt{B(\log_2 n)/n}, & \text{if } B \leq B < \overline{B}, \\ \| f_0 \|_{\mathcal{B}_{\infty,\infty}^s}^{1/(1+s)} (nB / \log_2 n)^{-s/(2+2s)}, & \text{if } B < \underline{B}. \end{cases}
\]

Noting that \( \| f_0 \|_{\mathcal{B}_{\infty,\infty}} \leq L \) leads to the claimed upper bounds.

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

\[
\| f_0 - \mathbb{E}_{f_0,T} \tilde{f}(j^*) \|^2 \lesssim \| f_0 \|^2 - \mathbb{E}_{f_0,T} \tilde{f}(j^*) \|^2 \lesssim L^2 2^{-2js} + 2^j/n.
\]
Furthermore, by using the upper bound $\psi_{lk}^2 \lesssim 2^j$

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

$$\lesssim 2^{2j}\mathbb{E}_{f_0,T}\sum_{l=0}^{j} \sum_{k=0}^{2^l-1} (Z_{lk}^2 + W_{lk}^2)$$

$$\lesssim 2^{3j}(\mathbb{E}_{f_0,T}Z_{lk}^2 + n^{-1}) \lesssim 2^{3j}. $$

Then by Cauchy-Schwarz inequality and Lemma 4.1 we get that

$$\mathbb{E}_{f_0,T}\|f_0 - \hat{f}\|_\infty 1_{j > j^*} \lesssim \sum_{j=j^*+1}^{j_{\max}} \mathbb{E}_{f_0,T}\|f_0 - \hat{f}(j)\|_\infty^2 \mathbb{P}_{f_0,T}(\hat{j} = j) $$

$$\lesssim \sum_{j=j^*+1}^{j_{\max}} 2^{(3/2)j}\mathbb{P}_{f_0,T}(\hat{j} = j) \lesssim 2^{j^*e^{-c\tau j^*}} + 2^{(3/2)j_{\max}H^{-2}} $$

$$= o(2^{-j^*s} + 1/\sqrt{n}),$$

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

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

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

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

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

By triangle inequality

$$\|\tilde{f}(j^-) - \hat{f}(l)\|_\infty \leq \|\tilde{f}(j^-) - \mathbb{E}_{f_0,T}\tilde{f}(j^-)\|_\infty + \|\mathbb{E}_{f_0,T}\tilde{f}(j^-) - \mathbb{E}_{f_0,T}\tilde{f}(l)\|_\infty + \|\mathbb{E}_{f_0,T}\tilde{f}(l)\|_\infty.$$

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

$$\|\mathbb{E}_{f_0,T}\tilde{f}(j^-) - \mathbb{E}_{f_0,T}\tilde{f}(l)\|_\infty \leq \|\sum_{r=j^-}^{l} \sum_{k=0}^{2^r-1} f_0,rkW_{rk}\|_\infty + \|\sum_{r=j^-}^{l} \sum_{k=0}^{2^r-1} \mathbb{E}_{f_0,T}W_{rk}\psi_{lk}\|_\infty$$

$$\leq c\|f_0\|_{B_{s,\infty}} 2^{-j^-s} + \sqrt{2^l/n_l}) \leq C(B(j^-, f_0) + \sqrt{2^l/n_l})$$

$$\leq C(B(j^*, f_0) + \sqrt{2^l/n_l}) \leq C(\sqrt{j^*2^{j^*}}/n_{j^*} + \sqrt{2^l/n_l})$$

$$\leq C\sqrt{l2^{l}/n_l}.$$
Furthermore, 
\[
\|\tilde{f}(l) - \mathbb{E}_{f_0,T} \tilde{f}(l)\|_\infty \leq \sum_{j=0}^{l} \max_k (|Z_{jk}| + |W_{jk}|) \sup_{x \in [0,1]} \sum_{k=0}^{2^{j-1}} |\psi_{jk}(x)| \\
\leq C \left( \sum_{j=0}^{l} 2^{j/2} \max_k |Z_{jk}| + \sqrt{2^j/n} \right).
\]

We show below that for any \( \gamma \geq 1 \), 
\begin{equation}
\mathbb{P}_{f_0,T} \left( \max_k (|Z_{lk}| \geq \tau \sqrt{l/n}) \right) \lesssim n_1^{1-\gamma} + e^{-cr^2l}
\end{equation}
holds for some sufficiently large \( \tau > 0 \) and sufficiently small \( c > 0 \). By combining the above results we get that 
\[
\mathbb{P}_{f_0,T} \left( \|\tilde{f}(j^-) - \tilde{f}(l)\|_\infty \geq \tau \sqrt{l^2/n} \right) \lesssim \mathbb{P}_{f_0,T} \left( \|\tilde{f}(l) - \mathbb{E}_{f_0,T} \tilde{f}(l)\|_\infty \geq \frac{\tau - C}{2} \sqrt{l^2/n} \right) \\
\lesssim \sum_{j=0}^{l} \mathbb{P}_{f_0,T} \left( \max_k |Z_{jk}| \geq \frac{\tau - 2C}{2C} \sqrt{l/n} \right) \\
\leq l \mathbb{P}_{f_0,T} \left( \max_k |Z_{lk}| \geq \frac{\tau - 2C}{2C} \sqrt{l/n} \right) \\
\lesssim \left( \log_2 n \right) n_1^{1-\gamma} + e^{-c(\gamma/2) \tau^2 l}.
\]

The above inequality together with the first display of the proof then implies that 
\[
\mathbb{P}_{f_0,T}(\hat{j} = j) \lesssim \left( \log_2 n \right) n_1^{1-\gamma} + e^{-c(\gamma/2) \tau^2 j} \\
\lesssim (\log_2 n)^2 n_{j_{\max}}^{1-\gamma} + e^{-c(\gamma/2) \tau^2 j} \lesssim n^{-2} + e^{-c(\gamma/2) \tau^2 j},
\]
for \( \gamma \geq 5 \), in view of \( n_{j_{\max}} \gtrsim (nB/\log_2 n)^{1+2s_{\min}} \gtrsim \sqrt{n} \), for any \( s_{\min} > 0 \), providing the statement of the lemma.

It remained to prove assertion (4.8). Note that by triangle inequality we get that 
\begin{equation}
\max_k |Z_{lk}| \leq \max_k |Z_{lk} - \mathbb{E}_{f_0,T} Z_{lk}| + \max_k |\mathbb{E}_{f_0,T} Z_{lk}|.
\end{equation}

In view of assertion (4.4) with \( \mathbb{P}_T \)-probability at least \( 1 - Cn_1^{1-\gamma} \) the second term on the right hand side is bounded from above by \( C \sqrt{\gamma l/n} \). Furthermore recall from the proof of Theorem 2.10 (i.e. assertion (3.21)) that \( n_i^{-2} \sum_{i \in A_{jk}} \sum_{\ell=1}^{n/m} \psi_{jk}^{(i)}(T_{\ell}^{(i)}) \lesssim n_i^{-1} \) holds with \( \mathbb{P}_T \)- probability at least \( 1 - Ce^{-c\delta} \), for some sufficiently small \( \delta > 0 \). Under the above event we have that there exists small enough constant \( c > 0 \) such that 
\[
\mathbb{P}_{f_0,T}(|Z_{1k} - \mathbb{E}_{f_0,T} Z_{1k}| \geq \tau \sqrt{l/n}) \leq \exp\{ -cr^2 l \}.
\]

Therefore the first term on the right hand side of (4.9) is bounded from above by \( \tau \sqrt{l/n} \) with \( \mathbb{P}_{f_0,T} \)-probability at least \( 1 - C2^l e^{-c\tau^2 l} \leq 1 - Ce^{-(c/2) \tau^2 l} \) on \( T \in B_l \), for some sufficiently large constants \( \tau, C > 0 \) and sufficiently small positive constant \( c \).
5. Technical lemmas. In this section we provide the technical lemmas applied in the previous two sections.

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

Without loss of generality we can assume that \( T^{(i)}_{1} \leq T^{(i)}_{2} \leq \ldots \leq T^{(i)}_{n/m}, i = 1, \ldots, m \), and let \( \ell_{k} = \ell_{k}^{(i)} = \max \{ j \in \{ 1, \ldots, n/m \} : T^{(i)}_{j} \in I_{k} \} \) denote the index of the largest element \( T^{(i)}_{j} \) in the interval \( I_{k} = [(k-1)C_{2}^{-2^{-j_{n}}, kC_{2}^{-2^{-j_{n}}}], k = 1, \ldots, |K_{j_{n}}| = 2^{j_{n}}/C_{2} \). Note that \( T^{(i)}_{\ell_{k-1}+1}, \ldots, T^{(i)}_{\ell_{k}} \in I_{k} \). For convenience let us introduce the following notations

\[
X^{(i)}_{[j_{1}, j_{2}]}, X^{(i)}_{j_{1}+1}, \ldots, X^{(i)}_{j_{2}}
\]

d = |K_{j_{n}}|,

\[
F_{-k} = (F_{1}, \ldots, F_{k-1}, F_{k}, \ldots, F_{d}),
\]

\[
\delta = L_{n}^{1/2}2^{j_{n}/2}/\|\psi\|_{\infty},
\]

\[
a^{2} = \frac{2^{5}n\delta^{2}}{dm/\log(dm)},
\]

\[
\mu_{k}(t) = (L_{n}^{1/2}\psi_{\ell_{k}}(t_{j}))_{j=(\ell_{k-1}+1), \ldots, \ell_{k}},
\]

\[
B_{k}(t) = \{ x \in \mathbb{R}^{\ell_{k}-\ell_{k-1}} : \| x - \mu(t) \|^{2} \leq a \},
\]

\[
\mathcal{B} = \{ t \in [0, 1]^{n/m} : \frac{n}{2dm} \leq \ell_{k} - \ell_{k-1} \leq \frac{2n}{dm}, k = 1, \ldots, d \}.
\]

Note that \( X_{[(\ell_{k-1}+1):\ell_{k}]}(T(t), F_{k}) \) is independent of \( F_{-k} \) and

\[
X_{[(\ell_{k-1}+1):\ell_{k}]}(T(t) = t, F_{k} = \beta_{k}) \sim \mathbb{P}_{\beta_{k}|T(t)=t} = N_{\ell_{k}-\ell_{k-1}}(\beta_{k}\mu(t), I).
\]

Furthermore, note that the inequalities \( \delta^{2} \leq \frac{0.42\log(n)}{2^{2/\log dm}} \) (in view of \( C_{1} \geq 0.4^{-2}2^{8}L_{n}^{2}\|\psi\|_{\infty}^{2}C_{2} \)) and \( n/m \geq 2^{6}d \log(n/m) \) (in view of \( m = O(n^{2^{-n}}/\log^{2} n) \)) hold.

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

\[
\sup_{x \in B_{k}(t)} \frac{\varphi_{\mu(t)}(x)}{\varphi_{-\mu(t)}(x)} = \sup_{x \in B_{k}(t)} \exp \left\{ \frac{\| x - \mu(t) \|^{2} - \| x + \mu(t) \|^{2}}{2} \right\}
\]

\[
= \sup_{x \in B_{k}(t)} \exp \{ 2|x^{T}\mu_{k}(t)| \} = \exp \{ 2a \},
\]

where \( \varphi_{\mu} \) denotes the density function of a normal distribution with mean vector \( \mu \) and identity covariance matrix. Then by Theorem A.13 in the supplement (with \( F_{0} = \{ \beta = (\beta_{k})_{k=1, \ldots, d} : \beta_{k} \in \{-1, 1\}, k = 1, \ldots, d \} \) we have that

\[
(5.1) \quad I(F; Y^{(i)}) = \int_{[0,1]^{n/m}} I(F; Y^{(i)}|T^{(i)} = t) dt
\]

\[
\leq \sum_{k=1}^{d} (\log 2) \int_{[0,1]^{n/m}} \sqrt{\mathbb{P}_{\beta_{k}|T(t)=t}(X_{[(\ell_{k-1}+1):\ell_{k}]}, t \notin B_{k}(t))} dt
\]

\[
+ \sum_{k=1}^{d} \int_{[0,1]^{n/m}} \mathbb{P}_{\beta_{k}|T(t)=t}(X_{[(\ell_{k-1}+1):\ell_{k}]} \notin B_{k}(t)) dt
\]

\[
+ 2C^{2}(C-1)^{2} I(X^{(i)}; Y^{(i)}|T^{(i)}),
\]
with \( C = \exp\{2^{7/2}\delta\sqrt{n\log(dm)}/\sqrt{dm}\} \).

Note that \( I(X^{(i)}; Y^{(i)}|T^{(i)}) \leq H(Y^{(i)}|T^{(i)}) \leq H(Y^{(i)}) \). In view of Lemma 5.2 we have that \( \Pr(T^{(i)} \in B) \geq 1 - 2de^{-n/(8md)} \geq 1 - 2(\ell d)^{-4} \) following from the inequality \( n/m \geq 2\ell d \log(n/m) \). Besides for arbitrary \( t \in B \) we have in view of

\[
\|\mu_k(t)\|^2_2 \leq \sum_{j=k-1+1}^{t_k} \delta_n \psi_{j,n,k}(t_j)^2 \leq \|\psi\|^2_{\infty} \delta_n 2\delta_n (t_k - t_{k-1}) \leq 2n\delta^2/(md)
\]

that

\[
P_{f_k}^{(i)}(X^{(i)} \not\in B_k(t)|T^{(i)} = t) = \mathbb{P}_{f}^{(i)}(\|\mu_k(t)\|^2_2 X^{(i)}_{[t_{k-1}+1]:t_k} > a|T^{(i)} = t) \leq 2 \exp\{-\frac{(a - \|\mu_k(t)\|^2_2)^2}{2\|\mu_k(t)\|^2_2}\}
\]

\[
\leq 2 \exp\{-\frac{a^2}{4\|\mu_k(t)\|^2_2}\} \leq 2(\ell d)^{-4}.
\]

Therefore

\[
\int_{[0,1]^{n/m}} \sqrt{\Pr_{f_k}^{(i)}|T^{(i)} = t} (X^{(i)}_{[t_{k-1}+1]:t_k} \not\in B_k(t)) \, dt 
\]

\[
\leq \int_B \sqrt{\Pr_{f_k}^{(i)}|T^{(i)} = t} (X^{(i)}_{[t_{k-1}+1]:t_k} \not\in B_k(t)) \, dt + \Pr(T^{(i)} \not\in B) 
\]

\[
\leq \sqrt{2(\ell d)^{-2}} + 2(\ell d)^{-4} \leq 2(\ell d)^{-4},
\]

and similarly \( \int_{[0,1]^{n/m}} \mathbb{P}_{f_T^{(i)}}(X^{(i)}_{[t_{k-1}+1]:t_k} \not\in B_k(t)) \, dt \leq 4(\ell d)^{-4} \). Then by plugging in the above inequalities into (5.1) and using the inequalities \( e^x \leq 1 + 2x \) for \( x \leq 0.4 \) and \( C^2 \leq 2 \) we get that

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

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

\[
I(F; Y^{(i)}) \leq I(F; (T^{(i)}, X^{(i)})) \leq I(F; X^{(i)}|T^{(i)}) + I(F; T^{(i)})
\]

\[
= \int_{t \in [0,1]^{n/m}} I(F; X^{(i)}|T^{(i)} = t) \, dt 
\]

\[
\leq \int_{t \in [0,1]^{n/m}} \frac{1}{|F_0|^2} \sum_{\beta, \beta^\prime \in F_0} K(\mathbb{P}_{\beta'|T^{(i)} = t}^{(i)} \| \mathbb{P}_{\beta|T^{(i)} = t}^{(i)}) \, dt 
\]

\[
\leq \frac{1}{2|F_0|^2} \sum_{\beta, \beta^\prime \in F_0} \sum_{t \in [0,1]^{n/m}} \sum_{k \in K_j} (\beta_k' - \beta_k)^2 L^2 \delta_n \int_0^1 \psi_{j,n,k}(t) \, dt 
\]

\[
\leq 2\delta^2 n/m.
\]

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

\[
I(F; Y) \leq \frac{4\delta^2 n}{m} \sum_{i=1}^m \min \left\{ 2^{10} \log(md) d^{-1} H(Y^{(i)}), 1 \right\} + 4 \log 2 
\]

\[
\leq \frac{4L^2 \delta_n 2^{10n} \|\psi\|_{\infty}^2}{m} \sum_{i=1}^m \min \left\{ 2^{10} \log(md) d^{-1} H(Y^{(i)}), 1 \right\} + 4 \log 2.
\]
Finally we arrive to our statement by using Lemma 5.3 and $2^{h_n} = C_2 d$.

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

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

$$P(2^{-1} n/r \leq |\chi_k| \leq 2n/r, k = 1, ..., r) \geq 1 - 2r e^{-n/(8r)}.$$  

**Proof.** We start with the proof of the upper bound. Note that by Chernoff’s bound

$$P\left( \sup_{k=1,...,r} |\chi_k| \geq 2n/r \right) \leq \sum_{k=1}^{r} P(|\chi_k| \geq 2n/r) \leq re^{-n/(3r)},$$

and similarly for the lower bound

$$P\left( \inf_{k=1,...,r} |\chi_k| \leq 2^{-1} n/r \right) \leq re^{-n/(8r)}.$$  

\[ \square \]

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

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

$$H(Y) \leq 2E l(Y) + 1.$$  

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

$$K(Y, U) = \sum_{s \in S} \mathbb{P}(Y = s) \log \frac{\mathbb{P}(Y = s)}{\mathbb{P}(U = s)}$$

$$= \sum_{s \in S} \mathbb{P}(Y = s) \log \frac{1}{\mathbb{P}(U = s)} - H(Y)$$

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

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

Hence,

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

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

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

$$H(N) = \sum_{i \in I} \mathbb{P}(N = i) \log \frac{1}{\mathbb{P}(N = i)} + \sum_{i \notin I} \mathbb{P}(N = i) \log \frac{1}{\mathbb{P}(N = i)} \leq \sum_{i \in I} \mathbb{P}(N = i)i + \sum_{i \notin I} e^{-i}i \leq \mathbb{E}N + 1.$$ 

This completes the proof. \( \square \)

**Acknowledgement.** We would like to thank the AE and the referees for their careful review of the various versions of this paper and for their valuable comments and suggestions.

**References.**


Mathematical Institute
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands
E-mail: b.t.szabo@math.leidenuniv.nl

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 107
1098 XG Amsterdam
The Netherlands
E-mail: hvzanten@uva.nl