The Annals of Statistics
2019, Vol. 47, No. 4, 2286–2319
https://doi.org/10.1214/18-AOS1748

CONVERGENCE RATES OF LEAST SQUARES REGRESSION ESTIMATORS WITH HEAVY-TAILED ERRORS

BY QIYANG HAN AND JON A. WELLNER[1]

University of Washington

We study the performance of the least squares estimator (LSE) in a general nonparametric regression model, when the errors are independent of the covariates but may only have a pth moment ($p \geq 1$). In such a heavy-tailed regression setting, we show that if the model satisfies a standard "entropy condition" with exponent $\alpha \in (0, 2)$, then the L_2 loss of the LSE converges at a rate

$$\mathcal{O}_{\mathbf{P}}(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}).$$

Such a rate cannot be improved under the entropy condition alone.

This rate quantifies both some positive and negative aspects of the LSE in a heavy-tailed regression setting. On the positive side, as long as the errors have $p \geq 1 + 2/\alpha$ moments, the L_2 loss of the LSE converges at the same rate as if the errors are Gaussian. On the negative side, if $p < 1 + 2/\alpha$, there are (many) hard models at any entropy level α for which the L_2 loss of the LSE converges at a strictly slower rate than other robust estimators.

The validity of the above rate relies crucially on the independence of the covariates and the errors. In fact, the L_2 loss of the LSE can converge arbitrarily slowly when the independence fails.

The key technical ingredient is a new multiplier inequality that gives sharp bounds for the "multiplier empirical process" associated with the LSE. We further give an application to the sparse linear regression model with heavy-tailed covariates and errors to demonstrate the scope of this new inequality.

1. Introduction.

1.1. *Motivation and problems.* Consider the classical setting of nonparametric regression: suppose that

$$Y_i = f_0(X_i) + \xi_i \qquad \text{for } i = 1, \ldots, n, \tag{1.1}$$

where $f_0 \in \mathcal{F}$, a class of possible regression functions f where $f : \mathcal{X} \to \mathbb{R}$, $X_1, \ldots, X_n$ are i.i.d. P on $(\mathcal{X}, \mathcal{A})$, and $\xi_1, \ldots, \xi_n$ are i.i.d. "errors" independent of $X_1, \ldots, X_n$. We observe the pairs $\{(X_i, Y_i) : 1 \leq i \leq n\}$ and want to estimate f_0.

Received February 2018; revised May 2018.

[1]Supported in part by NSF Grant DMS-1566514.

MSC2010 subject classifications. Primary 60E15; secondary 62G05.

Key words and phrases. Multiplier empirical process, multiplier inequality, nonparametric regression, least squares estimation, sparse linear regression, heavy-tailed errors.

While there are many approaches to this problem, the most classical approach has been to study the least squares estimator (or LSE) $\hat{f}_n$ defined by

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2. \tag{1.2}$$

The LSE is well known to have nice properties (e.g., rate-optimality) when:

(E) the errors $\{\xi_i\}$ are sub-Gaussian or at least subexponential;

(F) the class $\mathcal{F}$ of regression functions satisfies a condition slightly stronger than a *Donsker* condition: namely, either a uniform entropy condition or a bracketing entropy condition with exponent $\alpha \in (0, 2)$:

$$\sup_{Q} \log \mathcal{N}\big(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)\big) \lesssim \varepsilon^{-\alpha},$$

where the supremum is over all finitely discrete measures Q on $(\mathcal{X}, \mathcal{A})$, or

$$\log \mathcal{N}_{[\,]}\big(\varepsilon, \mathcal{F}, L_2(P)\big) \lesssim \varepsilon^{-\alpha}.$$

See, for example, [10] and [64], Chapter 9, and Section 1.2 for notation. In spite of a very large literature, there remains a lack of clear understanding of the properties of $\hat{f}_n$ in terms of assumptions concerning the heaviness of the tails of the errors and the massiveness or "size" of the class $\mathcal{F}$.

Our interest here is in developing further tools and methods to study properties of $\hat{f}_n$, especially its convergence rate when the error condition (E) is replaced by:

(E′) the errors $\{\xi_i\}$ have only a p-moment for some $1 \le p < \infty$.

This leads to our first question.

QUESTION 1. *What determines the convergence rate b_n of $\hat{f}_n$ with respect to some risk or loss functions? When is this rate b_n determined by p (and hence the tail behavior of the ξ_i's), and when is it determined by α (and hence the size of $\mathcal{F}$)?*

There are a variety of measures of loss and risk in this setting. Two of the most common are:

(a) Empirical L_2 loss: $\|\hat{f}_n - f_0\|_{L_2(\mathbb{P}_n)}$,[2] and the corresponding risk $\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(\mathbb{P}_n)}$.

(b) Population (or prediction) L_2 loss $\|\hat{f}_n - f_0\|_{L_2(P)}$, and the corresponding risk $\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)}$.

[2]We write $\mathbb{P}_n$ for the empirical measure of the (X_i, Y_i) pairs: $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$.

Here, we will mainly focus on measuring loss or risk in the sense of the prediction loss (b) since it corresponds to the usual choice in the language of empirical risk minimization; see, for example, [8–10, 30, 31, 42, 61, 64, 66]. Thus we will (usually) measure loss or risk in $L_2(P)$, and hence study rates of convergence of

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \left[\int_{\mathcal{X}} |\hat{f}_n(x; (X_1, Y_1), \ldots, (X_n, Y_n)) - f_0(x)|^2 \, \mathrm{d}P(x)\right]^{1/2},$$

or, in somewhat more compact notation,

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)} = \mathbb{E}\left[\int_{\mathcal{X}} |\hat{f}_n(x) - f_0(x)|^2 \, \mathrm{d}P(x)\right]^{1/2}.$$

As we will see in Section 3, the rate of convergence of the LSE $\hat{f}_n$ under conditions (E′) and (F) is

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big). \tag{1.3}$$

So, the dividing line between p and α in determining the rate of convergence of the LSE is given by

$$p = 1 + 2/\alpha$$

in the following sense:

(R_α) If $p \geq 1 + 2/\alpha$, then for any function class with entropy exponent α, the rate of convergence of the LSE is $\mathcal{O}_{\mathbf{P}}(n^{-1/(2+\alpha)})$.

(R_p) If $p < 1 + 2/\alpha$, then there exist model classes $\mathcal{F}$ with entropy exponent α such that the rate of convergence of the LSE is $\mathcal{O}_{\mathbf{P}}(n^{-1/2+1/(2p)})$.

These rates in R_α and R_p indicate both some positive and negative aspects of the LSE in a heavy-tailed regression setting:

- If $p \geq 1 + 2/\alpha$, then the heaviness of the tails of the errors (E′) does not play a role in the rate of convergence of the LSE, since the rate in R_α coincides with the usual rate under the light-tailed error assumption (E) and the entropy condition (F).
- If $p < 1 + 2/\alpha$, there exist (many) hard models at any entropy level α for which the LSE converges only at a slower rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2+1/(2p)})$ compared with the faster (optimal) rate $\mathcal{O}_{\mathbf{P}}(n^{-1/(2+\alpha)})$—a rate that can be achieved by other robust estimation procedures. See Section 3 for examples and more details.

It should be noted that the assumption of independence of the errors ξ_i's and the X_i's in the regression model (1.1) is crucial for the above results to hold. In fact, when the errors ξ_i's can be dependent on the X_i's, there is no longer any universal moment condition on the ξ_i's alone that guarantees the rate-optimality of the LSE, as opposed to (R_α) (cf. Proposition 3).

To briefly introduce the main new tool we develop in Section 2 below, we first recall the classical methods used to prove consistency and rates of convergence of the LSE (and many other contrast-type estimators). These methods are based on a "basic inequality" which lead naturally to a multiplier empirical process. This is well known to experts in the area, but we will briefly review the basic facts here. Since $\hat{f}_n$ minimizes the functional $f \mapsto \mathbb{P}_n(Y - f(X))^2 = n^{-1}\sum_{i=1}^n (Y_i - f(X_i))^2$, it follows that

$$\mathbb{P}_n\big(Y - \hat{f}_n(X)\big)^2 \le \mathbb{P}_n\big(Y - f_0(X)\big)^2.$$

Adding and subtracting f_0 on the left-hand side, some algebra yields

$$\mathbb{P}_n\big(Y - f_0(X)\big)^2 + 2\mathbb{P}_n(Y - f_0)(f_0 - \hat{f}_n) + \mathbb{P}_n(f_0 - \hat{f}_n)^2 \le \mathbb{P}_n\big(Y - f_0(X)\big)^2.$$

Since $\xi_i = Y_i - f_0(X_i)$ under the model given by (1.1) we conclude that

$$\begin{aligned}\mathbb{P}_n\big(\hat{f}_n(X) - f_0(X)\big)^2 &\le 2\mathbb{P}_n\big(\xi\big(\hat{f}_n(X) - f_0(X)\big)\big)\\ &\le 2\sup_{f\in\mathcal{F}} \mathbb{P}_n\big(\xi\big(f(X) - f_0(X)\big)\big),\end{aligned} \tag{1.4}$$

where the process

$$f \mapsto n(\mathbb{P}_n - P)\big(\xi f(X)\big) = n\mathbb{P}_n\big(\xi f(X)\big) = \sum_{i=1}^n \xi_i f(X_i) \tag{1.5}$$

is a *multiplier empirical process*. This is exactly as in Section 4.3 of [64]. When the ξ_i's are Gaussian, the process in (1.5) is even a Gaussian process conditionally on the X_i's, and is relatively easy to analyze. If the $\{\xi_i\}$'s are integrable and $\mathcal{F}$ is a *Glivenko–Cantelli class* of functions, then the inequality (1.4) leads easily to consistency of the LSE in the sense of the loss and risk measures (a); see, for example, [64].

To obtain rates of convergence, we need to consider localized versions of the processes in (1.4), much as in Section 3.4.3 of [66]. As in Section 3.4.3 of [66], (but replacing their $\theta \in \Theta$ and ε by our $f \in \mathcal{F}$ and ξ) we consider

$$\mathbb{M}_n(f) = 2\mathbb{P}_n\xi(f - f_0) - \mathbb{P}_n(f - f_0)^2,$$

and note that $\hat{f}_n$ maximizes $\mathbb{M}_n(f)$ over $\mathcal{F}$. Since the errors have zero mean and are independent of the X_i's, this process has mean $M(f) \equiv -P(f - f_0)^2$. Since $\mathbb{M}_n(f_0) = 0 = M(f_0)$, centering then yields the process

$$\begin{aligned}f \mapsto \mathbb{Z}_n(f) &\equiv \mathbb{M}_n(f) - \mathbb{M}_n(f_0) - \big(M(f) - M(f_0)\big)\\ &= 2\mathbb{P}_n\xi(f - f_0) - (\mathbb{P}_n - P)(f - f_0)^2.\end{aligned}$$

Establishing rates of convergence for $\hat{f}_n$ then boils down to bounding

$$\mathbb{E} \sup_{f\in\mathcal{F}: P(f-f_0)^2 \le \delta^2} \mathbb{Z}_n(f)$$

as a function of n and δ; see, for example, [66] Theorem 3.4.1, pages 322–323. It is clear at least for $\mathcal{F} \subset L_\infty$ that this can be accomplished if we have good bounds for the multiplier empirical process (1.5) in terms of the empirical process itself

$$f \mapsto n(\mathbb{P}_n - P)\big(f(X)\big) = \sum_{i=1}^{n} \big(f(X_i) - Pf\big), \tag{1.6}$$

or, in view of standard symmetrization inequalities (as in Section 2.3 of [66]), its symmetrized equivalent,

$$f \mapsto \sum_{i=1}^{n} \varepsilon_i f(X_i), \tag{1.7}$$

where the ε_i are i.i.d. Rademacher random variables $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$ independent of the X_i's. This leads naturally to the following.

QUESTION 2. *Under what moment conditions on the ξ_i's can we assert that the multiplier empirical process* (1.5) *has* (*roughly*) *the same "size" as the empirical process* (1.6) (*or equivalently the symmetrized empirical process* (1.7) *for* (*nearly*) *all function classes $\mathcal{F}$ in a nonasymptotic manner?*

In Section 2 below, we provide simple moment conditions on the ξ_i's which yield a positive answer to Question 2, when the ξ_i's are independent from the X_i's. We then give some comparisons to the existing multiplier inequalities which illustrate the improvement possible via the new bounds in nonasymptotic settings, and show that our bounds also yield the asymptotic equivalence required for multiplier CLT's (cf. Section 2.9 of [66]). Further impossibility results are demonstrated, showing that there is no positive solution to Question 2 when the ξ_i's and the X_i's can be dependent.

In Section 3, we address Question 1 by applying the new multiplier inequality to derive the convergence rate of the LSE (1.3) in the context of the nonparametric regression model (1.1), and indicate in greater detail both the positive and negative aspects of the LSE due to this rate. We further show that no solution to Question 1 exists when the errors ξ_i's and the covariates X_i's can be dependent.

Not surprisingly, the new bounds for the multiplier empirical process have applications to many settings in which the least squares criterion plays a role, for example, the Lasso in the sparse linear regression model. In Section 4, we give an application of the new bounds in a Lasso setting with both heavy-tailed errors and heavy-tailed covariates. Most detailed proofs are given in Section 5 and the Supplementary Material [26].

1.2. *Notation.* For a real-valued random variable ξ and $1 \le p < \infty$, let $\|\xi\|_p \equiv (\mathbb{E}|\xi|^p)^{1/p}$ denote the ordinary p-norm. The $L_{p,1}$ norm for a random variable ξ is defined by

$$\|\xi\|_{p,1} \equiv \int_0^\infty \mathbb{P}\big(|\xi| > t\big)^{1/p}\,\mathrm{d}t.$$

It is well known that $L_{p+\varepsilon} \subset L_{p,1} \subset L_p$ holds for any underlying probability measure, and hence a finite $L_{p,1}$ condition requires slightly more than a pth moment, but no more than any $p+\varepsilon$ moment; see Chapter 10 of [37].

For a real-valued measurable function f defined on $(\mathcal{X}, \mathcal{A}, P)$, $\|f\|_{L_p(P)} \equiv \|f\|_{P,p} \equiv (P|f|^p)^{1/p}$ denotes the usual L_p-norm under P, and $\|f\|_\infty \equiv \sup_{x\in\mathcal{X}}|f(x)|$. f is said to be P-centered if $Pf = 0$, and $\mathcal{F}$ is P-centered if all $f \in \mathcal{F}$ are P-centered. $L_p(g, B)$ denotes the $L_p(P)$-ball centered at g with radius B. For simplicity, we write $L_p(B) \equiv L_p(0, B)$. To avoid unnecessary measurability digressions, we will assume that $\mathcal{F}$ is countable throughout the article. As usual, for any $\phi : \mathcal{F} \to \mathbb{R}$, we write $\|\phi(f)\|_{\mathcal{F}}$ for $\sup_{f\in\mathcal{F}}|\phi(f)|$.

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of the normed space of real functions $f : \mathcal{X} \to \mathbb{R}$. For $\varepsilon > 0$ let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-covering number of $\mathcal{F}$, and let $\mathcal{N}_{[\,]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-bracketing number of $\mathcal{F}$; see page 83 of [66] for more details.

Throughout the article, $\varepsilon_1, \ldots, \varepsilon_n$ will be i.i.d. Rademacher random variables independent of all other random variables. C_x will denote a generic constant that depends only on x, whose numeric value may change from line to line unless otherwise specified. $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \le C_x b$ and $a \ge C_x b$, respectively, and $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$ ($a \lesssim b$ means $a \le Cb$ for some absolute constant C). For two real numbers a, b, $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. $\mathcal{O}_{\mathbf{P}}$ and $\mathfrak{o}_{\mathbf{P}}$ denote the usual big and small O notation in probability.

2. The multiplier inequality. Multiplier inequalities have a long history in the theory of empirical processes. Our new multiplier inequality in this section is closest in spirit to the classical multiplier inequality (cf. Section 2.9 of [66] or [23]), but strictly improves the classical one in a nonasymptotic setting (see Section 2.3).

Our work here is also related to [45], who derived bounds for the multiplier empirical process, assuming: (i) ξ_i's have a $2+\varepsilon$ moment, and (ii) $\{(\xi_i, X_i)\}$ are i.i.d. (i.e., ξ_i need not be independent from X_i). The bounds in [45] use techniques from generic chaining [58], and work particularly well for "sub-Gaussian classes" (defined in [45]). Our setting here will be different: we assume that: (i) ξ_i's have a $L_{p,1}$ ($p \ge 1$) moment and (ii) ξ_i's are independent from X_i's, but the ξ_i's need not be independent from each other.

We make this choice in view of a negative result of Alexander [1], stating that there is no universal moment condition on ξ_i's for a multiplier CLT to hold when ξ_i's need not be independent from X_i's, while a $L_{2,1}$ moment condition is known to be universal in the independent case [23, 37, 66]. The complication here makes

it more hopeful to work in the independent case for a precise understanding of the multiplier empirical process. In fact:

- In the independent case, we are able to quantify the exact *structural interplay* between the moment of the multipliers and the complexity of the indexing function class in the size of the multiplier empirical process (cf. Theorems 1–2), thereby giving a satisfactory answer to Question 2;
- Such an interplay fails when the X_i's may not be independent from the ξ_i's. Moreover, no simple moment condition on the ξ_i's alone can lead to a solution to Question 2 in the dependent case (cf. Proposition 1).

2.1. *Upper bound.* We first state the assumptions.

ASSUMPTION A. Suppose that $\xi_1, \dots, \xi_n$ are independent of the random variables $X_1, \dots, X_n$, and *either* of the following conditions holds:

(A1) $X_1, \dots, X_n$ are i.i.d. with law P on $(\mathcal{X}, \mathcal{A})$, and $\mathcal{F}$ is P-centered.

(A2) $X_1, \dots, X_n$ are permutation invariant, and $\xi_1, \dots, \xi_n$ are independent mean-zero random variables.

THEOREM 1. *Suppose Assumption A holds. Let $\{\mathcal{F}_k\}_{k=1}^n$ be a sequence of function classes such that $\mathcal{F}_k \supset \mathcal{F}_n$ for any $1 \le k \le n$. Assume further that there exists a nondecreasing concave function $\psi_n : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ with $\psi_n(0) = 0$ such that*

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \le \psi_n(k) \tag{2.1}$$

holds for all $1 \le k \le n$. Then

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} \le 4 \int_0^\infty \psi_n\left(\sum_{i=1}^{n} \mathbb{P}(|\xi_i| > t) \right) \mathrm{d}t. \tag{2.2}$$

The primary application of Theorem 1 to nonparametric regression problems in Section 3 involves a nonincreasing sequence of function classes $\mathcal{F}_1 \supset \cdots \supset \mathcal{F}_n$. It is also possible to use Theorem 1 for the case $\mathcal{F}_1 = \cdots = \mathcal{F}_n$; see Section 4 for an application to the sparse linear regression model.

The following corollary provides a canonical concrete application of Theorem 1.

COROLLARY 1. *Consider the same assumptions as in Theorem 1. Assume for simplicity that ξ_i's have the same marginal distributions. Suppose that for some $\gamma \ge 1$, and some constant $\kappa_0 > 0$,*

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \le \kappa_0 \cdot k^{1/\gamma} \tag{2.3}$$

holds for all $1 \le k \le n$. *Then for any* $p \ge 1$ *such that* $\|\xi_1\|_{p,1} < \infty$,

$$\mathbb{E}\left\|\sum_{i=1}^{n} \xi_i f(X_i)\right\|_{\mathcal{F}_n} \le 4\kappa_0 \cdot n^{\max\{1/\gamma, 1/p\}} \|\xi_1\|_{\min\{\gamma,p\},1}.$$

PROOF. First, consider $\gamma \le p$. In this case, letting $\psi_n(t) \equiv \kappa_0 t^{1/\gamma}$ in Theorem 1, we see that $\mathbb{E}\|\sum_{i=1}^n \xi_i f(X_i)\|_{\mathcal{F}_n} \le 4\kappa_0 \cdot n^{1/\gamma} \|\xi_1\|_{\gamma,1}$. On the other hand, if $\gamma > p$, we can take $\psi_n(t) \equiv \kappa_0 t^{1/p}$ to conclude that $\mathbb{E}\|\sum_{i=1}^n \xi_i f(X_i)\|_{\mathcal{F}_n} \le 4\kappa_0 \cdot n^{1/p} \|\xi_1\|_{p,1}$. Note that $\gamma \ge 1$ ensures the concavity of ψ_n. □

Corollary 1 says that the upper bound for the multiplier empirical process has two components: one part comes from the growth rate of the empirical process; another part comes from the moment barrier of the multipliers ξ_i's.

REMARK 1. One particular case for application of Theorem 1 and Corollary 1 is the following. Let $\delta_1 \ge \cdots \ge \delta_n \ge 0$ be a sequence of nonincreasing nonnegative real numbers, and $\mathcal{F}$ be an arbitrary function class. Let $\mathcal{F}_k \equiv \mathcal{F}(\delta_k) \equiv \{f \in \mathcal{F} : Pf^2 < \delta_k^2\}$ be the "local" set of $\mathcal{F}$ with L_2-radius at most δ_k. There exists a large literature on controlling such localized empirical processes; a classical device suited for applications in nonparametric problems is to use local maximal inequalities under either the uniform or bracketing entropy conditions (cf. Proposition 4).

An important choice in statistical applications for δ_k is given by

$$\mathbb{E}\left\|\sum_{i=1}^{k} \varepsilon_i f(X_i)\right\|_{\mathcal{F}(\delta_k)} \lesssim k\delta_k^2. \tag{2.4}$$

As will be seen in Section 3, the above choice $\{\delta_k\}$ corresponds to the rate of convergence of the LSE in the nonparametric regression model (1.1).

In this case, Theorem 1 and Corollary 1 yield that

$$\mathbb{E}\left\|\sum_{i=1}^{n} \xi_i f(X_i)\right\|_{\mathcal{F}(\delta_n)} \lesssim n\delta_n^2 \tag{2.5}$$

given sufficient moments of the ξ_i's.

REMARK 2. Choosing $\gamma \ge 2$ in Corollary 1 corresponds to the bounded Donsker regime[3] for the empirical process. In this case, we only need $\|\xi_1\|_{2,1} < \infty$ to ensure the multiplier empirical process to also be bounded Donsker. This moment condition is generally unimprovable in view of [36]. On the other hand, such a choice of γ can fail due to: (i) failure of integrability of the envelope functions of the classes $\{\mathcal{F}_k\}$, or (ii) failure of the classes $\{\mathcal{F}_k\}$ to be bounded Donsker. (i)

[3] $\mathcal{F}$ is said to be *bounded Donsker* if $\sup_{n\in\mathbb{N}} \mathbb{E}\sup_{f\in\mathcal{F}} |\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i)| < \infty$.

is related to the classical Marcinkiewicz–Zygmund strong laws of large numbers and the generalizations of those to empirical measures; see [2, 41, 43]. For (ii), some examples in this regard can be found in [57], Chapter 11 of [19]; see also Proposition 17.3.7 of [55].

Theorem 1 and Corollary 1 only concern the first moment of the suprema of the multiplier empirical process. For higher moments, we may use the following Hoffmann–Jørgensen/Talagrand-type inequality relating the qth moment estimate with the first moment estimate.

LEMMA 1 (Proposition 3.1 of [22]). *Let $q \geq 1$. Suppose $X_1, \ldots, X_n$ are i.i.d. with law P and $\xi_1, \ldots, \xi_n$ are i.i.d. with $\|\xi_1\|_{2\vee q} < \infty$. Let $\mathcal{F}$ be a class of functions with $\sup_{f\in\mathcal{F}} Pf^2 \leq \sigma^2$ such that either $\mathcal{F}$ is P-centered, or ξ_1 is centered. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i f(X_i)\right|^q \leq K^q\left[\left(\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \xi_i f(X_i)\right|\right)^q + q^{q/2}\big(\sqrt{n}\|\xi_1\|_2\sigma\big)^q + q^q\mathbb{E}\max_{1\leq i\leq n}|\xi_i|^q \sup_{f\in\mathcal{F}}|f(X_i)|^q\right].$$

Here, $K > 0$ is a universal constant.

2.2. *Lower bound.* Theorem 1 and Corollary 1 do not require any structural assumptions on the function class $\mathcal{F}$. [45] showed that for a "sub-Gaussian" class, a $2+\varepsilon$ moment on i.i.d. ξ_i's suffices to conclude that the multiplier empirical process behaves like the canonical Gaussian process. One may therefore wonder if the moment barrier for the multipliers in Corollary 1 is due to an artifact of the proof. Below in Theorem 2, we show that this barrier is intrinsic for general classes $\mathcal{F}$.

THEOREM 2. *Let $\mathcal{X} = [0,1]$ and P be a probability measure on $\mathcal{X}$ with Lebesgue density bounded away from 0 and ∞. Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables such that $\mathbb{E}\max_{1\leq i\leq n}|\xi_i| \geq \kappa_0 n^{1/p}$ for some $p > 1$ and some constant κ_0 independent of ξ_1. Then for any $\gamma > 2$, there exists a sequence of function classes $\{\mathcal{F}_k\}_{k=1}^n$ defined on $\mathcal{X}$ with $\mathcal{F}_k \supset \mathcal{F}_n$ for any $1 \leq k \leq n$ such that for n sufficiently large,*

$$\mathbb{E}\left\|\sum_{i=1}^k \varepsilon_i f(X_i)\right\|_{\mathcal{F}_k} \leq \kappa_1 \cdot k^{1/\gamma},$$

holds for all $1 \leq k \leq n$, and that

$$\mathbb{E}\left\|\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}_n} \geq \kappa_1^{-1} n^{\max\{1/\gamma, 1/p\}}.$$

Here, κ_1 is a constant depending on κ_0, γ and P.

REMARK 3. The condition on the ξ_i's will be satisfied, for example, if the ξ_i's are i.i.d. with the tail condition $\mathbb{P}(|\xi_i| > t) \geq \kappa_0'/(1+t^p)$ for $t > 0$.

Combined with Corollary 1, it is seen that the growth rate $n^{\max\{1/\gamma, 1/p\}}$ of the multiplier empirical process cannot be improved in general. This suggests an interesting phase transition phenomenon from a worst-case perspective: if the complexity of the function class dominates the effect of the tail of the multipliers, then the multiplier empirical process essentially behaves as the empirical process counterpart; otherwise, the tail of the multipliers governs the growth of the multiplier empirical process.

REMARK 4. The function class we constructed that witnesses the moment barrier rate $n^{1/p}$ in Theorem 2 can be simply taken to be the class of indicators over closed intervals on $[0, 1]$. Although being the "simplest" function class in the theory of empirical processes, this class serves as an important running example that achieves the bad rate $n^{1/p}$.

2.3. *Comparison of Theorem 1 with the multiplier inequality in* [66]. In this section, we compare the classical multiplier inequality in Theorem 1 with the one in Section 2.9 of [66], which originates from [24, 25, 36]; see also [23]: for i.i.d. mean-zero ξ_i's and i.i.d. X_i's, and for any $1 \leq n_0 \leq n$,

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}} &\lesssim (n_0-1)\mathbb{E}\|f(X_1)\|_{\mathcal{F}}\frac{\mathbb{E}\max_{1\leq i\leq n}|\xi_i|}{\sqrt{n}} \\ &\quad + \|\xi_1\|_{2,1}\max_{n_0\leq k\leq n}\mathbb{E}\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^k \varepsilon_i f(X_i)\right\|_{\mathcal{F}}.\end{aligned} \tag{2.6}$$

2.3.1. *Nonasymptotic setting*. The major drawback of (2.6) is that it is not sharp in a nonasymptotic setting. For an illustration, let $\xi_1, \ldots, \xi_n$ be i.i.d. multipliers with $\|\xi_1\|_{p,1} < \infty$ ($p \geq 2$), X_i's be i.i.d. uniformly distributed on $[0, 1]$, and $\mathcal{F}$ be a uniformly bounded function class on $[0, 1]$ satisfying the entropy condition (F) with $\alpha \in (0, 2)$. We apply (2.6) with $\mathcal{F}(n^{-1/(2+\alpha)})$ (note that $n^{-1/(2+\alpha)}$ is the usual local radius for $1/\alpha$-smooth problems) and local maximal inequalities for the empirical process (Proposition 4 in Section 5 below) to see that

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}(n^{-1/(2+\alpha)})} &\lesssim \inf_{1\leq n_0\leq n} n_0 \cdot n^{-1/2+1/p} + n_0^{-\frac{(2-\alpha)}{2(2+\alpha)}} \\ &\asymp n^{-\frac{2-\alpha}{6+\alpha}(\frac{1}{2}-\frac{1}{p})} \equiv n^{-\delta_1(\alpha,p)}.\end{aligned} \tag{2.7}$$

On the other hand, Corollary 1 gives the rate

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}(n^{-1/(2+\alpha)})} \lesssim n^{-\min\{\frac{2-\alpha}{2(2+\alpha)}, 1/2-1/p\}} \equiv n^{-\delta_2(\alpha,p)}. \tag{2.8}$$

In the above inequalities, we used the following bound for the symmetrized empirical process (for illustration we only consider bracketing entropy):

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}(n^{-1/(2+\alpha)})}$$
$$\lesssim J_{[]}\big(n^{-1/(2+\alpha)},\mathcal{F},L_2(P)\big)\bigg(1+\frac{J_{[]}(n^{-1/(2+\alpha)},\mathcal{F},L_2(P))}{\sqrt{n}\cdot n^{-2/(2+\alpha)}}\bigg)\lesssim n^{\frac{2-\alpha}{2(2+\alpha)}},$$

where in the last line of the above display we used

$$J_{[]}\big(n^{-1/(2+\alpha)},\mathcal{F},L_2(P)\big)=\int_0^{n^{-1/(2+\alpha)}}\sqrt{1+\log\mathcal{N}_{[]}\big(\varepsilon,\mathcal{F},L_2(P)\big)}\,\mathrm{d}\varepsilon\lesssim n^{\frac{2-\alpha}{2(2+\alpha)}}.$$

It is easily seen that the bound (2.7) calculated from (2.6) is worse than (2.8) because $\delta_1(\alpha,p)<\delta_2(\alpha,p)$ for all $\alpha\in(0,2)$ and $p\geq 2$. Moreover, if $p\geq 1+2/\alpha$, the bound (2.8) becomes $n^{-\frac{2-\alpha}{2(2+\alpha)}}$, which matches the rate for the symmetrized empirical process $\mathbb{E}\|n^{-1/2}\sum_{i=1}^n\varepsilon_i f(X_i)\|_{\mathcal{F}(n^{-1/(2+\alpha)})}$.

2.3.2. *Asymptotic setting*. The primary application of (2.6) rests in studying asymptotic equicontinuity of the multiplier empirical process in the following sense. Suppose that $\mathcal{F}$ is Donsker. Then by the integrability of the empirical process (see Lemma 2.3.11 of [66]),[4] $\mathbb{E}\|n^{-1/2}\sum_{i=1}^n\varepsilon_i f(X_i)\|_{\mathcal{F}_\delta}\to 0$ as $n\to\infty$ followed by $\delta\to 0$. Now apply (2.6) via $n\to\infty$, $n_0\to\infty$ followed by $\delta\to 0$ we see that $\mathbb{E}\|n^{-1/2}\sum_{i=1}^n\xi_i f(X_i)\|_{\mathcal{F}_\delta}\to 0$ as $n\to\infty$ followed by $\delta\to 0$ if $\|\xi_1\|_{2,1}<\infty$. This shows that $(n^{-1/2}\sum_{i=1}^n\xi_i f(X_i))_{f\in\mathcal{F}}$ satisfies a CLT in $\ell^\infty(\mathcal{F})$ if $\mathcal{F}$ is Donsker and the ξ_i's are i.i.d. with $\|\xi_1\|_{2,1}<\infty$.

Our new multiplier inequality, Theorem 1, can also be used to study asymptotic equicontinuity of the multiplier empirical process with the help of the following lemma.

LEMMA 2. *Fix a concave function $\varphi:\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$, such that $\varphi(x)\to\infty$ as $x\to\infty$. Let $\{a_n\}\subset\mathbb{R}_{\geq 0}$ be such that $a_n\to 0$ as $n\to\infty$, and $\psi:\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ be the least concave majorant of $\{(n,a_n\varphi(n))\}_{n=0}^\infty$. Then $\psi(t)/\varphi(t)\to 0$ as $t\to\infty$.*

The proof of this lemma can be found in Appendix C in [26]. Take any sequence $\delta_n\to 0$ and let $a_n\equiv\mathbb{E}\|n^{-1/2}\sum_{i=1}^n\varepsilon_i f(X_i)\|_{\mathcal{F}_{\delta_n}}$. By Lemma 2, the least concave majorant function $\psi:\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ of the map $n\mapsto a_n n^{1/2}$ $(n\geq 0)$ satisfies $\psi(t)/t^{1/2}\to 0$ as $t\to\infty$. Now an application of Theorem 1 and the dominated convergence theorem shows that

$$\mathbb{E}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i f(X_i)\right\|_{\mathcal{F}_{\delta_n}}\leq 4\int_0^\infty\frac{\psi(n\mathbb{P}(|\xi_1|>t))}{\sqrt{n\mathbb{P}(|\xi_1|>t)}}\cdot\sqrt{\mathbb{P}\big(|\xi_1|>t\big)}\,\mathrm{d}t\to 0$$

as $n\to\infty$.

[4] Here, $\mathcal{F}_\delta\equiv\{f-g:f,g\in\mathcal{F},\|f-g\|_{L_2(P)}\leq\delta\}$.

We note that the moment conditions of Theorems 1 and 2 have a small gap: in essence we require an $L_{p,1}$ moment in Theorem 1, while an L_p moment is required in Theorem 2. In the context of multiplier CLTs discussed above, [36] showed that the $L_{2,1}$ moment condition is sharp—there exists a construction of a Banach space of X on which a multiplier CLT fails for ξX if $\|\xi_1\|_{2,1} = \infty$. It remains open in our setting if $L_{p,1}$ (or L_p) is the exact moment requirement.

2.4. *An impossibility result.* In this section, we formally prove an impossibility result, showing that the independence assumption between the X_i's and the ξ_i's is crucial for Theorem 1 and Corollary 1 to hold.

PROPOSITION 1. *Let $\mathcal{X} \equiv \mathbb{R}$. For every triple (δ, γ, p) such that $\delta \in (0, 1/2)$, $2 < \gamma < 1 + 1/(2\delta)$ and $2 \le p < \min\{4/\delta, 2\gamma/(1+\gamma\delta)\}$, there exist X_i's and ξ_i's satisfying*: (i) *$\{(X_i, \xi_i)\}$'s are i.i.d.*; (ii) *ξ_i is not independent from X_i but $\mathbb{E}[\xi_1|X_1] = 0$, $\|\xi_1\|_{p,1} < \infty$, and a sequence of function classes $\{\mathcal{F}_k\}_{k=1}^n$ defined on $\mathcal{X}$ with $\mathcal{F}_k \supset \mathcal{F}_n$ for any $1 \le k \le n$, such that*

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \lesssim k^{1/\gamma},$$

holds for all $1 \le k \le n$, and that

$$\mathbb{E}\left\| \sum_{i=1}^{n} \xi_i f(X_i) \right\|_{\mathcal{F}_n} \gtrsim_p \omega(n),$$

where $\omega(n) \ge n^{\beta} \cdot n^{\max\{1/\gamma, 1/p\}}$ for some $\beta = \beta(\delta, \gamma, p) > 0$. In other words, $\omega(n)$ grows faster than $n^{\max\{1/\gamma, 1/p\}}$ (= the upper bound in Theorem 1 and Corollary 1) by a positive power of n.

Proposition 1 is a negative result for the multiplier empirical processes in the similar vein as in [1], but more quantitatively: there is no universal moment condition for the multipliers that yield a positive solution to Question 2 when the X_i's and the ξ_i's are allowed to be dependent.

REMARK 5. The basic trouble for removing the independence assumption between the X_i's and the ξ_i's can be seen by the following example. Let X_i's be i.i.d. mean-zero random variables with a finite second moment. Then clearly $\sum_{i=1}^n X_i$ grows at a rate $\mathcal{O}_{\mathbf{P}}(n^{1/2})$ by the CLT. On the other hand, let $\xi_i = \varepsilon_i X_i$ where ε_i's are independent Rademacher random variables. Then the multiplier sum $\sum_{i=1}^n \xi_i X_i = \sum_{i=1}^n \varepsilon_i X_i^2$ may grow at a rate as fast as $\mathcal{O}_{\mathbf{P}}(n^{1-\delta})$, if $\varepsilon_1 X_1^2$ is in the domain of attraction of a symmetric stable law with index close to 1.

3. Nonparametric regression: Least squares estimation. In this section, we apply our new multiplier inequalities in Section 2 to study the least squares estimator (LSE) (1.2) in the nonparametric regression model (1.1) when the errors ξ_i's are heavy-tailed (E′), independent of the X_i's (but need not be independent of each other), and the model satisfies the entropy condition (F).

Our results here are related to the recent ground-breaking work of Mendelson and his coauthors [35, 44, 46, 47]. These papers proved rate-optimality of ERM procedures under a $2+\varepsilon$ moment condition on the errors, in a general structured learning framework that contains models satisfying sub-Gaussian/small-ball conditions. Their framework also allows arbitrary dependence between the errors ξ_i's and the X_i's. See [48] for some recent development. Here, the reasons for our focus on the different structure—models with entropy conditions, are twofold:

- Entropy is a standard and well-understood notion for the complexity of a large class of models; see examples in [23, 66].
- The moment condition on the errors needed to guarantee rate-optimality of the LSE in our setting is no longer a $2+\varepsilon$ moment. In fact, as we will show, $p \geq 1+2/\alpha$ (cf. Theorems 3–4) moments are needed for such a guarantee.

The reason that we work with independent errors is more fundamental: when the errors ξ_i's are allowed to be dependent on the X_i's, there is no universal moment condition on the ξ_i's alone that guarantees the rate-optimality of the LSE (cf. Proposition 3). In fact, even in the family of one-dimensional linear regression models with heteroscedastic errors of any finite pth moment, the convergence rate of the LSE can be as slow as specified (cf. Remark 10).

3.1. *Upper bound for the convergence rates of the LSE.*

THEOREM 3. *Suppose that $\xi_1, \ldots, \xi_n$ are mean-zero errors independent of $X_1, \ldots, X_n$ with the same marginal distributions, and $\|\xi_1\|_{p,1} < \infty$ for some $p \geq 1$. Further suppose that $\mathcal{F}$ is a P-centered function class (if the ξ_i's are i.i.d. $\mathcal{F}$ need not be P-centered) such that $\mathcal{F} - f_0 \subset L_\infty(1)$ satisfies the entropy condition* (F) *with some $\alpha \in (0,2)$. Then the LSE $\hat{f}_n$ in* (1.2) *satisfies*

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big). \tag{3.1}$$

Furthermore, if ξ_i's are i.i.d. and $p \geq 2$, then (3.1) *holds in expectation*:

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}\big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big). \tag{3.2}$$

One interesting consequence of Theorem 3 is a convergence rate of the LSE when the errors only have a $L_{p,1}$ moment ($1 < p \leq 2$).

COROLLARY 2. *Suppose the assumptions in Theorem 3 hold with $p \in (1,2]$. Then*

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2}+\frac{1}{2p}}\big) = \mathfrak{o}_{\mathbf{P}}(1).$$

Consistency of the LSE has been a classical topic; see, for example, [60, 63] for sufficient and necessary conditions in this regard under a second moment assumption on the errors. Here, Theorem 3 provides a quantitative rate of convergence of the LSE when the errors may not even have a second moment (under stronger conditions on $\mathcal{F}$).

The connection between the proof of Theorem 3 and the new multiplier inequality in Section 2 is the following reduction scheme.

PROPOSITION 2. *Suppose that $\xi_1, \dots, \xi_n$ are mean-zero random variables independent of $X_1, \dots, X_n$, and $\mathcal{F} - f_0 \subset L_\infty(1)$. Further assume that*

$$\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f_0\|_{L_2(P)} \le \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (f - f_0)(X_i) \right| \lesssim \phi_n(\delta) \tag{3.3}$$

and

$$\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f_0\|_{L_2(P)} \le \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f - f_0)(X_i) \right| \lesssim \phi_n(\delta) \tag{3.4}$$

hold for some ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta$ is nonincreasing. Then $\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}(\delta_n)$ holds for any δ_n such that $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$. Furthermore, if $\xi_1, \dots, \xi_n$ are i.i.d. mean-zero with $\|\xi_1\|_p < \infty$ for some $p \ge 2$, then $\mathbb{E}\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}(\delta_n)$ for any $\delta_n \ge n^{-\frac{1}{2}+\frac{1}{2p}}$ such that $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$.

The remaining task in the proof of Theorem 3 is a calculation of the modulus of continuity of the (multiplier) empirical process involved in (3.3) and (3.4) using Theorem 1 and local maximal inequalities for the empirical process (see Proposition 4).

REMARK 6. Some remarks on the assumptions on $\mathcal{F}$.

1. The entropy condition (F) is standard in nonparametric statistics literature. The condition $\alpha \in (0, 2)$ additionally requires $\mathcal{F}$ to be a *Donsker* class. Although the proof applies to non-Donsker function classes with $\alpha \ge 2$, the first term in (3.1) becomes *suboptimal* in general; see [10].

2. $\mathcal{F}$ is assumed to be P-centered when the errors ξ_i's have an arbitrary dependence structure. It is known from [67] (see Theorem 1, p. 638) that for a centered function class, the minimax risk of estimating a regression function under arbitrary errors with second moments uniformly bounded, is no worse than that for i.i.d. Gaussian errors. If the errors are i.i.d., then $\mathcal{F}$ need not be P-centered (as stated in the theorem).

3. The uniform boundedness assumption on $\mathcal{F}$, including many classical examples (cf. Section 9.3 of [64]), should be primarily viewed as *a method of proof*: all that we need is $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$. In subsequent work of the authors [27], this

method is applied to shape-restricted regression problems in a heavy-tailed regression setting.

REMARK 7. Here, in Theorem 3 we focus on the regression model (1.1) with errors ξ_i's independent from X_i's. This is crucial: we show below in Proposition 3 that the independence assumption between the X_i's and ξ_i's cannot be relaxed for the rate in Theorem 3 to hold.

On the other hand, our Theorem 3 is useful in handling centered models with arbitrarily dependent errors in the regression model. This complements Mendelson's work [35, 44, 46, 47, 49] that allows arbitrary dependence between ξ_i and X_i's with independent observations in a learning framework.

REMARK 8. In Theorem 3, the results are "in probability" and "in expectation" statements. It is easy to see from the proof that a tail estimate can be obtained for $\|\hat{f}_n - f_0\|_{L_2(P)}$: if $\|\xi_1\|_{p,1} < \infty$ for some $p \geq 2$, then

$$\mathbb{P}\big(\delta_n^{-1}\|\hat{f}_n - f_0\|_{L_2(P)} > t\big) \leq Ct^{-p},$$

where $\delta_n \equiv n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}$. Constructing estimators other than the LSE that give rise to *exponential tail bound* under a heavy-tailed regression setting is also of significant interest. We refer the readers to, for example, [18, 38, 40] and references therein for this line of research.

3.2. *Lower bound for the convergence rates of the LSE.* At this point, (3.1) only serves as an *upper bound* for the convergence rates of the LSE. Since the rate $n^{-\frac{1}{2+\alpha}}$ corresponds to the optimal rate in the Gaussian regression case [68], it is natural to conjecture that this rate cannot be improved. On the other hand, the "noise" rate $n^{-\frac{1}{2}+\frac{1}{2p}}$ is due to the reduction scheme in Proposition 2, which relates the convergence rate of the LSE to the size of the multiplier empirical process involved. It is natural to wonder if this "noise rate" is a proof artifact due to some possible deficiency in Proposition 2.

THEOREM 4. *Let $\mathcal{X} = [0,1]$ and P be a probability measure on $\mathcal{X}$ with Lebesgue density bounded away from 0 and ∞, and ξ_i's are i.i.d. mean-zero errors independent of X_i's. Then for each $\alpha \in (0,2)$ and $2 \vee \sqrt{\log n} \leq p \leq (\log n)^{1-\delta}$ with some $\delta \in (0,1/2)$, there exists a function class $\mathcal{F} \equiv \mathcal{F}_n$, and some $f_0 \in \mathcal{F}$ with $\mathcal{F} - f_0$ satisfying the entropy condition* (F), *such that the following holds*: *there exists some law for the error ξ_1 with $\|\xi_1\|_{p,1} \lesssim \log n$, such that for n sufficiently large, there exists some least squares estimator f_n^* over $\mathcal{F}_n$ satisfying*

$$\mathbb{E}\|f_n^* - f_0\|_{L_2(P)} \geq \rho \cdot \big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big)(\log n)^{-2}.$$

Here, $\rho > 0$ is a (small) constant independent of n.

Theorem 4 has two claims. The first claim justifies the heuristic conjecture that the convergence rate for the LSE with heavy-tailed errors under entropy conditions, should be no better than the optimal rate in the Gaussian regression setting. Although here we give an existence statement, the proof is constructive: in fact, we use (essentially) a Hölder class. Other function classes are also possible if we can handle the Poisson (small-sample) domain of the empirical process indexed by these classes.

The second claim asserts that for any entropy level $\alpha \in (0, 2)$, there exist "hard models" for which the noise level dominates the risk for the least squares estimator. Here are some examples for these hard models.

EXAMPLE 1. A benchmark model witnessing the worst case rate $\mathcal{O}(n^{-\frac{1}{2}+\frac{1}{2p}})$ (up to logarithmic factors) is (almost) the one we used in Theorem 2, that is, the class of indicators[5] over closed intervals in $[0, 1]$.

EXAMPLE 2. Consider more general classes[5]

$$\mathcal{F}_k \equiv \Bigg\{ \sum_{i=1}^{k} c_i \mathbf{1}_{[x_{i-1},x_i]} : |c_i| \le 1, \\ 0 \le x_0 < x_1 < \cdots < x_{k-1} < x_k \le 1 \Bigg\}, \qquad k \ge 1.$$

The classes $\mathcal{F}_k$ also witness the worst case rate $\mathcal{O}(n^{-\frac{1}{2}+\frac{1}{2p}})$ (up to logarithmic factors) since they contain all indicators over closed intervals on $[0, 1]$, and are closely related to problems in the change-point estimation/detection literature. For instance, the case $k = 1$ is of particular importance in epidemic and signal processing applications; see [3, 69] from a testing perspective of the problem. From an estimation viewpoint, [11] proposed an ℓ_0-type penalized LSE for estimating regression functions in $\mathcal{F}_k$, where a (nearly) parametric rate is obtained under a sub-Gaussian condition on the errors. Our results here suggest that such least-squares type estimators may not work well for estimating step functions with multiple change-points if the errors are heavy-tailed.

EXAMPLE 3. Yet another class is given by the regression problem involving image restoration (or edge estimation); see, for example, [32, 33] or Example 9.3.7 of [64] (but we consider a random design). In particular, the class $\mathscr{C} \equiv \{\mathbf{1}_C : C \subset [0, 1]^d \text{ is convex}\}$[6] also witnesses the lower bound $\mathcal{O}(n^{-\frac{1}{2}+\frac{1}{2p}})$ (up to logarithmic factors) since it contains all indicators over hypercubes on $[0, 1]^d$.

[5]Excluding the indicators indexed by intervals that are too short.

[6]Excluding the indicators indexed by sets with too small volume.

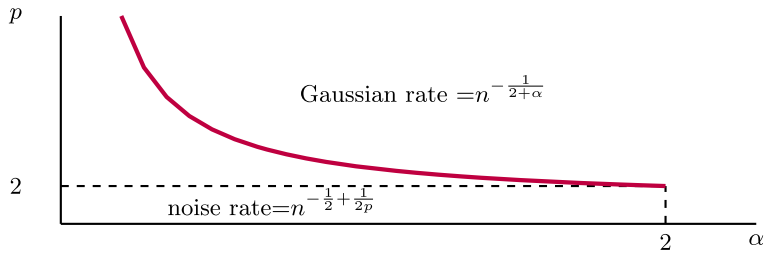

FIG. 1. *Tradeoff between the complexity of the function class and the noise level of the errors in the convergence rates for the LSE. The critical curve* (*purple*): $p = 1 + 2/\alpha$.

3.3. *Some positive and negative implications for the LSE.* Combining Theorems 3 and 4, we see that the tradeoff in the size of the multiplier empirical process between the complexity of the function class and the heaviness of the tail of the errors (multipliers) translates into the convergence rate of the LSE (cf. Figure 1). In particular, Theorems 3 and 4 indicate both some positive and negative aspects of the LSE in a heavy-tailed regression setting.

(*Positive implications for the LSE*): If $p \geq 1 + 2/\alpha$, then $\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}(n^{-\frac{1}{2+\alpha}})$. In this case, the noise level is "small" compared with the complexity of the function class so that the LSE achieves the optimal rate as in the case for i.i.d. Gaussian errors (see [68]).

(*Negative implications for the LSE*): If $p < 1 + 2/\alpha$, then $\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}(n^{-\frac{1}{2}+\frac{1}{2p}})$. In this case, the noise is so heavy-tailed that the *worst-case* rate of convergence of the LSE is governed by this noise rate (see above for examples). The negative aspect of the LSE is that this noise rate reflects a genuine deficiency of the LSE as an estimation procedure, rather than the difficulty due to the "hard model" in such a heavy-tailed regression setting. In fact, we can design simple robust procedures to outperform the LSE in terms of the rate of convergence.

To see this, consider the least-absolute-deviation (LAD) estimator $\tilde{f}_n$ (see, e.g., [20, 52, 53], or p. 336 of [66]) defined by $\tilde{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |Y_i - f(X_i)|$. It follows from a minor modification of the proof[7] in page 336 of [66] that as long as the errors $\xi_i \equiv M\eta_i$'s for some η_i admitting a smooth enough density, median zero and a first moment, and $M > 0$ not too small, then under the same conditions

[7]More specifically, we can proceed by replacing the empirical measure $\mathbb{P}_n$ by P, slightly restricting the suprema of the empirical process to $1/n \lesssim P(f - f_0)^2 < \delta^2$ in the third display on page 336 of [66], and noting that Theorem 3.4.1 of [66] can be strengthened to an expectation since the empirical processes involved are bounded.

as in Theorem 3, the LAD estimator $\tilde{f}_n$ satisfies

$$\sup_{f_0 \in \mathcal{F}} \mathbb{E}_{f_0} \|\tilde{f}_n - f_0\|_{L_2(P)} \leq \mathcal{O}\big(n^{-\frac{1}{2+\alpha}}\big),$$

where clearly the noise rate $\mathcal{O}(n^{-\frac{1}{2}+\frac{1}{2p}})$ induced by the moment of the errors does not occur. For statistically optimal procedures that do not even require a first moment on the errors, we refer the reader to [7].

It is worthwhile to note that the shortcomings of the LSE quantified here also rigorously justify the motivation of developing other robust procedures (cf. [4, 12, 13, 15, 16, 18, 28, 29, 38–40, 50]).

REMARK 9. Our Theorems 3 and 4 show that the moment condition

$$p \geq 1 + 2/\alpha$$

that guarantees the LSE to converge at the optimal rate (as in the case for Gaussian errors), is the best one can hope *under entropy conditions alone*. On the other hand, this condition may be further improved if additional structure is available. For instance, in the isotonic regression case ($\alpha = 1$), our theory requires $p \geq 3$ to guarantee an optimal $n^{-1/3}$ rate for the isotonic LSE, while it is known (cf. [70]) that a second moment assumption on the errors ($p = 2$) suffices. The benefits of this extra structure due to shape constraints are investigated in further work by the authors [27].

3.4. *An impossibility result.* In this section, dual to the impossibility result in Proposition 1 for the multiplier empirical process, we formally prove that the independence assumption between the X_i's and the ξ_i's is necessary for the rate in Theorems 3 and 4 to hold.

PROPOSITION 3. *Consider the regression model* (1.1) *without assuming independence between the X_i's and the ξ_i's. Let $\mathcal{X} \equiv \mathbb{R}$. For every triple (δ, α, p) such that $\delta \in (0, 1/2)$, $4\delta < \alpha < 2$ and $2 \leq p < \min\{4/\delta, (2+4/\alpha)/(1+(1+2/\alpha)\delta)\}$, there exist*:

- *X_i's and ξ_i's satisfying*: (i) *$\{(X_i, \xi_i)\}$'s are i.i.d.*; (ii) *ξ_i is not independent from X_i but $\mathbb{E}[\xi_1|X_1] = 0$, $\|\xi_1\|_{p,1} < \infty$;*
- *a function class $\mathcal{F} \equiv \mathcal{F}_n$, and some $f_0 \in \mathcal{F}$ with $\mathcal{F} - f_0$ satisfying the entropy condition* (F),

such that the following holds: *for n sufficiently large, there exists some least squares estimator f_n^* over $\mathcal{F}_n$ satisfying*

$$\mathbb{E}\|f_n^* - f_0\|_{L_2(P)} \geq \delta_n,$$

where $\delta_n \geq n^{\beta} \cdot (n^{-1/(2+\alpha)} \vee n^{-1/2+1/(2p)})$ for some $\beta = \beta(\delta, \alpha, p) > 0$. In other words, δ_n shrinks to 0 slower than $n^{-1/(2+\alpha)} \vee n^{-1/2+1/(2p)}$ (= the rate of the LSE in Theorems 3 and 4) by a positive power of n.

Proposition 3 is a negative result on the LSE: there is no universal moment condition on ξ_i's that guarantees the rate-optimality of the LSE when the errors ξ_i's can be dependent on the X_i's.

REMARK 10. One basic model underlying the construction of Proposition 3 is the following: consider the (one-dimensional) linear regression model with heteroscedastic errors

$$Y_i = \alpha_0 X_i + \xi_i, \qquad i = 1, \dots, n,$$

where $\xi_i = \varepsilon_i X_i$ for some independent Rademacher random variables ε_i's. Clearly, $\mathbb{E}[\xi_i | X_i] = 0$, but ξ_i is (highly) dependent on X_i. The least squares estimator $\hat{\alpha}_n \equiv \arg\min_{\alpha \in \mathbb{R}} n^{-1} \sum_{i=1}^n (Y_i - \alpha X_i)^2$ has a closed form:

$$\hat{\alpha}_n \equiv \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \alpha_0 + \frac{\sum_{i=1}^n \varepsilon_i X_i^2}{\sum_{i=1}^n X_i^2}.$$

Suppose X_i's have a finite second moment, then by the SLLN, $\hat{\alpha}_n \to \alpha_0$ a.s., but the convergence rate of $\|\hat{f}_n - f_0\|_{L_2(P)} = |\hat{\alpha}_n - \alpha_0| \|X_1\|_2$ can be as slow as any $n^{-\delta}$: note that $\sum_{i=1}^n X_i^2 = \mathcal{O}(n)$ under the assumed second moment condition on X_i's, while the sum of the centered random variables $\sum_{i=1}^n \varepsilon_i X_i^2$ may have a growth rate $\mathcal{O}(n^{1-\delta})$ if $\varepsilon_1 X_1^2$ is in the domain of attraction of a symmetric stable law with index close to 1 (recall Remark 5).

A simple modification of the construction along the lines of the proof of Proposition 1 allows the situation where ξ_i's have a finite pth moment ($p \geq 2$), while the convergence rate of the LSE can be as slow as $n^{-\delta}$.

So in order to derive the rate-optimality of the LSE under any universal moment condition on the errors ξ_i's, in a framework that allows arbitrary dependence between the ξ_i's and the X_i's, it is necessary to impose conditions on the model $\mathcal{F}$ to exclude the counterexamples (as in [35, 44, 46, 47, 49]).

4. Sparse linear regression: Lasso revisited. In this section, we consider the sparse linear regression model:

$$Y = X\theta_0 + \xi, \tag{4.1}$$

where $X \in \mathbb{R}^{n \times d}$ is a (random) design matrix and $\xi = (\xi_1, \dots, \xi_n)$ is a mean-zero noise vector independent of X. When the true signal $\theta_0 \in \mathbb{R}^d$ is sparse, one popular estimator is the Lasso [59]:

$$\hat{\theta}(\lambda) \equiv \arg\min_{\theta \in \mathbb{R}^d} \left(\frac{1}{n} \|Y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right). \tag{4.2}$$

The lasso estimator has been thoroughly studied in an already vast literature; we refer readers to the monograph [14] for a comprehensive overview.

Our main interest here concerns the following question: *under what moment conditions on the distributions of X and ξ can the lasso estimator enjoy the optimal rate of convergence?* In particular, neither X nor ξ need be light tailed a priori (i.e., not sub-Gaussian), and the components $\xi_1, \dots, \xi_n$ of the vector ξ need not be independent.

Previous work guaranteeing rate-optimality of the Lasso estimator typically assumes that both X and ξ are sub-Gaussian; see [14, 51, 62]. Relaxing the sub-Gaussian conditions in the Lasso problem is challenging: [35] showed how to remove the sub-Gaussian assumption on ξ in the case X is sub-Gaussian. The problem is even more challenging if we relax the sub-Gaussian assumption on the design matrix X. Our goal in this section is to demonstrate how the new multiplier inequality in Theorem 1, combined with (essentially) existing techniques, can be used to give a systematic treatment to the above question, in a rather straightforward fashion.

Before stating the result, we need some notion of the *compatibility condition*: For any $L > 0$ and $S \subset \{1, \dots, d\}$, define

$$\phi(L, S) = \sqrt{|S|} \min\left\{ \frac{1}{\sqrt{n}} \|X\theta_S - X\theta_{S^c}\|_2 : \|\theta_S\|_1 = 1, \|\theta_{S^c}\|_1 \le L \right\}.$$

Here, for any $\theta = (\theta_i) \in \mathbb{R}^d$, $\theta_S \equiv (\theta_i \mathbf{1}_{i \in S})$ and $\theta_{S^c} \equiv (\theta_i \mathbf{1}_{i \notin S})$. Let $B_0(s)$ be the set of s-sparse vectors in $\mathbb{R}^d$, that is, $\theta \in B_0(s)$ if and only if $|\{i : \theta_i \neq 0\}| \le s$. Further let $\Sigma = \mathbb{E}\hat{\Sigma}$ where $\hat{\Sigma} = X^\top X/n$ is the sample covariance matrix, and $\underline{\sigma}_d = \sigma_{\min}(\Sigma)$ and $\bar{\sigma}_d = \sigma_{\max}(\Sigma)$ be the smallest and largest singular value of the population covariance matrix, respectively. Here, $d = d_n$ and $s = s_n$ can either stay bounded or blow up to infinity in asymptotic statements.

THEOREM 5. *Let X be a design matrix with i.i.d. mean-zero rows, and $0 < \liminf \underline{\sigma}_d \le \limsup \bar{\sigma}_d < \infty$. Suppose that*

$$\min_{|S| \le s} \phi(3, S) \ge c_0 \tag{4.3}$$

holds for some $c_0 > 0$ with probability tending to 1 as $n \to \infty$, and that for some $1/4 \le \alpha \le 1/2$,

$$\limsup_{n \to \infty} \frac{\log d \cdot (M_4(X) \vee \log^2 d)}{n^{2-4\alpha}} < \infty, \tag{4.4}$$

where $M_4(X) \equiv \mathbb{E} \max_{1 \le j \le d} |X_{1j}|^4$. Then for $\hat{\theta}^L \equiv \hat{\theta}(2L\|\boldsymbol{\xi}_n\|_{1/\alpha,1}\sqrt{\log d/n})$,

$$\lim_{L \to \infty} \limsup_{n \to \infty} \sup_{\theta_0 \in B_0(s)} \mathbb{P}_{\theta_0}\left(\frac{1}{n} \|X(\hat{\theta}^L - \theta_0)\|_2^2 > \frac{16L^2 \|\boldsymbol{\xi}_n\|_{1/\alpha,1}^2}{c_0^2} \cdot \frac{s \log d}{n} \right) = 0. \tag{4.5}$$

Here, $\|\boldsymbol{\xi}_n\|_{1/\alpha,1} \equiv \int_0^\infty (\frac{1}{n} \sum_{i=1}^n \mathbb{P}(|\xi_i| > t))^\alpha \, \mathrm{d}t$.

The rate $\sqrt{s \log d/n}$ in the above theorem is well known to be (nearly) minimax optimal for prediction in the sparse linear regression model (e.g., [54]). The quantity $\|\boldsymbol{\xi}_n\|_{1/\alpha,1}$ should be thought as the "noise level" of the regression problem. For instance, if the ξ_i's are i.i.d., and $\alpha = 1/2$, then $\|\boldsymbol{\xi}_n\|_{1/\alpha,1} = \|\xi_1\|_{2,1}$.

Although in Theorem 5 we only consider prediction error, the estimation error $\|\hat{\theta}^L - \theta_0\|_1$ can be obtained using completely similar arguments by noting that Lemma 3 below also holds for estimation error.

REMARK 11. Two technical remarks.

1. As in Theorem 3, we assume in Theorem 5 that the rows of X have zero-mean as vectors in $\mathbb{R}^d$ so that arbitrary dependence structure among ξ_i's can be allowed. For i.i.d. errors, the zero-mean assumption is not needed.

2. (4.5) is of an asymptotic nature mainly due to the weak asymptotic assumptions made in (4.3) and (4.4). It is clear from the proof that concrete probability estimates can be obtained if a probability estimate for (4.3) is available.

As an illustration of the scope of Theorem 5, we consider several different scaling regimes for the parameter space (d, n, s). For simplicity of discussion, we assume that the errors $\xi_1, \ldots, \xi_n$ have the same marginal distributions and the design matrix X has i.i.d. entries such that X_{11} has a Lebesgue density bounded away from ∞ and $\mathbb{E}X_{11}^2 = 1$.

EXAMPLE 4. Consider the scaling regime $d/n \to \lambda \in (0, 1)$. We claim that $\mathbb{E}|X_{11}|^{4+\varepsilon} \vee \|\xi\|_{4,1} < \infty$ for some $\varepsilon > 0$ guarantees the validity of (4.5). First, (4.3) holds under the finite fourth moment condition; see [6]. Second, (4.4) holds under the assumed moment conditions. Note that a fourth moment condition on X_{11} is necessary: if $\mathbb{E}X_{11}^4 = \infty$, then $\limsup \bar{\sigma}_d = \infty$ a.s.; see [5]. This corollary of Theorem 5 appears to be a new result; [38] considered a different "tournament" Lasso estimator with best tradeoff between confidence statement and convergence rate under heavy-tailed designs and errors.

EXAMPLE 5. If $\|X_{11}\|_p \lesssim p^\beta$ for some $\beta \geq 1/2$ and all $p \lesssim \log n$, then Theorem E of [34] showed that the compatibility condition (4.3) holds under $n \gtrsim s \log d \vee (\log d)^{(4\beta-1)}$. Condition (4.4) is satisfied if $\|\xi\|_{2+\varepsilon} < \infty$ and $\log d \lesssim \log n$.

The condition $\log d \lesssim \log n$ requires polynomial growth of d with n; this can be improved if X_{11} is light tailed. In particular, if $\mathbb{E}\exp(\mu|X_{11}|^\gamma) < \infty$ for some $\mu, \gamma > 0$, then we can take $\beta = 1/\gamma$ so that (4.3) holds under $n \gtrsim s \log d \vee (\log d)^{(4/\gamma)-1}$, while (4.4) is satisfied if $\|\xi\|_{2+\varepsilon} < \infty$ and $d \leq \exp(n^{c_{\varepsilon,\gamma}})$ for some constant $c_{\varepsilon,\gamma} > 0$. Different choices of γ lead to:

- If the entries of X have subexponential tails, then we may take $\gamma = 1$. In this case, (4.5) is valid under $\|\xi\|_{2+\varepsilon} < \infty$ subject to $n \gtrsim s \log d \vee \log^3 d$ and

$d \le \exp(n^{c_{\varepsilon,1}})$ for some constant $c_{\varepsilon,1} > 0$. This seems to be a new result; the recent result of [56] considered the similar tail condition on X along with a subexponential tail for the errors ξ_i's, while their rates come with additional logarithmic factors.

- If the entries of X have sub-Gaussian tails, then we may take $\gamma = 2$. In this case, (4.5) is valid under $\|\xi\|_{2+\varepsilon} < \infty$ subject to $n \gtrsim s \log d$ and $d \le \exp(n^{c_{\varepsilon,2}})$ for some constant $c_{\varepsilon,2} > 0$. This recovers a recent result of [35] in the case where X and ξ are independent (up to the mild dimension constraint on d).

Now we prove Theorem 5. The following reduction (basic inequality) is well known; cf. Theorem 6.1 of [14].

LEMMA 3. *On the event* $\mathcal{E}_L \equiv \{\max_{1\le j\le d} |\frac{2}{n}\sum_{i=1}^n \xi_i X_{ij}| \le L\sqrt{\log d/n}\}$, *with tuning parameter* $\lambda \equiv 2L\sqrt{\log d/n}$, *it holds that* $n^{-1}\|X(\hat\theta^L - \theta_0)\|_2^2 \le 16L^2\phi^{-2}(3, S_0)\cdot s_0 \log d/n$ *where* $S_0 = \{i : (\theta_0)_i \ne 0\}$ *and* $s_0 = |S_0|$.

The difficulty involved here is that *both* X and ξ can be heavy tailed. By Theorem 1, to account for the effect of the ξ_i's, we only need to track the size of $\mathbb{E}\max_{1\le j\le d}|\sum_{i=1}^k \varepsilon_i X_{ij}|$ at each scale $k \le n$. This is the content of the following Gaussian approximation lemma.

LEMMA 4. *Let* $X_1, \dots, X_n$ *be i.i.d. random vectors in* $\mathbb{R}^d$ *with covariance matrix* Σ. *If* $\sup_d \sigma_{\max}(\Sigma) < \infty$, *then for all* $k, d \in \mathbb{N}$,

$$\mathbb{E}\max_{1\le j\le d}\left|\sum_{i=1}^k \varepsilon_i X_{ij}\right| \lesssim \big(k\log^3 d\cdot\big(M_4(X)\vee \log^2 d\big)\big)^{1/4} + (k\log d)^{1/2}.$$

The proof of the lemma is inspired by the recent work [17] who considered Gaussian approximation of the maxima of high-dimensional random vectors by exploiting *second* moment information for the X_i's. We modify their method by taking into account the *third* moment information of X_i's induced by the symmetric Rademacher ε_i's; such a modification proves useful in identifying certain sharp moment conditions considered in the examples (in particular Example 4). See Appendix C.2 in [26] for a detailed proof.

PROOF OF THEOREM 5. By Lemma 3 and the assumption on the compatibility condition (4.3), we see that with the choice for tuning parameter $\lambda \equiv 2L\|\boldsymbol{\xi}_n\|_{1/\alpha,1}\sqrt{\log d/n}$, the left-hand side of (4.5) can be bounded by

$$\begin{aligned}&\mathbb{P}_{\theta_0}\left(\frac{1}{n}\|X(\hat\theta^L - \theta_0)\|_2^2 > \frac{16L^2\|\boldsymbol{\xi}_n\|_{1/\alpha,1}^2}{\phi^2(3, S_0)}\cdot\frac{s\log d}{n}\right) + o(1)\\ &\le \mathbb{P}\left(\max_{1\le j\le d}\left|\frac{2}{n}\sum_{i=1}^n \xi_i X_{ij}\right| > L\|\boldsymbol{\xi}_n\|_{1/\alpha,1}\sqrt{\frac{\log d}{n}}\right) + o(1).\end{aligned} \tag{4.6}$$

By Lemma 4, we can apply Theorem 1 with $\mathcal{F}_1 = \cdots = \mathcal{F}_n \equiv \{\pi_j : \mathbb{R}^d \to \mathbb{R},\ j = 1, \ldots, d\}$ where $\pi_j(x) = x_j$ for any $x = (x_l)_{l=1}^d \in \mathbb{R}^d$, and

$$\psi_n(k) \equiv C\big(k^\alpha \big(\log^3 d \cdot \big(M_4 \vee \log^2 d\big)\big)^{1/4} + k^{1/2}\sqrt{\log d}\big)$$

for any $1/4 \le \alpha \le 1/2$ such that (4.4) holds and $\|\boldsymbol{\xi}_n\|_{1/\alpha,1} < \infty$, to conclude that

$$\begin{aligned}\mathbb{E}\max_{1\le j\le d}\left|\sum_{i=1}^n \xi_i X_{ij}\right| &\lesssim n^\alpha \big(\log^3 d \cdot \big(M_4 \vee \log^2 d\big)\big)^{1/4}\|\boldsymbol{\xi}_n\|_{1/\alpha,1} + n^{1/2}\sqrt{\log d}\|\boldsymbol{\xi}_n\|_{2,1}\\ &\lesssim \big(n^\alpha \big(\log^3 d \cdot \big(M_4 \vee \log^2 d\big)\big)^{1/4} + n^{1/2}\sqrt{\log d}\big)\|\boldsymbol{\xi}_n\|_{1/\alpha,1}.\end{aligned}$$

By Markov's inequality, (4.6) can be further bounded (up to constants) by

$$\frac{1}{L}\left(\frac{\log d \cdot (M_4 \vee \log^2 d)}{n^{2-4\alpha}} \vee 1\right)^{1/4} + \mathfrak{o}(1).$$

The claim of Theorem 5 therefore follows from the assumption (4.4). □

5. Proofs for the main results: Main steps. In this section, we outline the main steps for the proofs of our main theorems. Proofs for many technical lemmas will be deferred to later sections.

5.1. *Preliminaries.* Let

$$J(\delta, \mathcal{F}, L_2) \equiv \int_0^\delta \sup_Q \sqrt{1 + \log \mathcal{N}\big(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\big)}\,\mathrm{d}\varepsilon \tag{5.1}$$

denote the *uniform* entropy integral, where the supremum is taken over all discrete probability measures, and

$$J_{[]}\big(\delta, \mathcal{F}, \|\cdot\|\big) \equiv \int_0^\delta \sqrt{1 + \log \mathcal{N}_{[]}\big(\varepsilon, \mathcal{F}, \|\cdot\|\big)}\,\mathrm{d}\varepsilon \tag{5.2}$$

denote the *bracketing* entropy integral. The following local maximal inequalities for the empirical process play a key role throughout the proof.

PROPOSITION 4. *Suppose that $\mathcal{F} \subset L_\infty(1)$, and $X_1, \ldots, X_n$'s are i.i.d. random variables with law P. Then with $\mathcal{F}(\delta) \equiv \{f \in \mathcal{F} : Pf^2 < \delta^2\}$,*

1. *If the uniform entropy integral* (5.1) *converges, then*

$$\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}(\delta)} \lesssim \sqrt{n} J(\delta, \mathcal{F}, L_2)\left(1 + \frac{J(\delta, \mathcal{F}, L_2)}{\sqrt{n}\delta^2\|F\|_{P,2}}\right)\|F\|_{P,2}. \tag{5.3}$$

2. *If the bracketing entropy integral* (5.2) *converges, then*

$$\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}(\delta)} \lesssim \sqrt{n} J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big)\left(1 + \frac{J_{[]}(\delta, \mathcal{F}, L_2(P))}{\sqrt{n}\delta^2}\right). \tag{5.4}$$

PROOF. (5.3) follows from [65]; see also Section 3 of [21], or Theorem 3.5.4 of [23]. (5.4) follows from Lemma 3.4.2 of [66]. □

We will primarily work with $F \equiv 1$ in the above inequalities. A two-sided estimate for the empirical process will be important for proving lower bounds in Theorems 2 and 4. The following definition is from [21], page 1167.

DEFINITION 1. A function class $\mathcal{F}$ is *α-full* ($0 < \alpha < 2$) if and only if there exists some constant $K_1, K_2 > 1$ such that both

$$\log \mathcal{N}\big(\varepsilon \|F\|_{L_2(\mathbb{P}_n)}, \mathcal{F}, L_2(\mathbb{P}_n)\big) \le K_1 \varepsilon^{-\alpha}, \qquad \text{a.s.}$$

for all $\varepsilon > 0, n \in \mathbb{N}$, and

$$\log \mathcal{N}\big(\sigma \|F\|_{L_2(P)}/K_2, \mathcal{F}, L_2(P)\big) \ge K_2^{-1} \sigma^{-\alpha}$$

hold. Here, $\sigma^2 \equiv \sup_{f \in \mathcal{F}} Pf^2$, F denotes the envelope function for $\mathcal{F}$, and $\mathbb{P}_n$ is the empirical measure for i.i.d. samples $X_1, \dots, X_n$ with law P.

The following lemma, giving a sharp two-sized control for the empirical process under the α-full assumption, is proved in Theorem 3.4 of [21].

LEMMA 5. *Suppose that $\mathcal{F} \subset L_\infty(1)$ is α-full with $\sigma^2 \equiv \sup_{f \in \mathcal{F}} Pf^2$. If $n\sigma^2 \gtrsim_\alpha 1$ and $\sqrt{n}\sigma(\frac{\|F\|_{L_2(P)}}{\sigma})^{\alpha/2} \gtrsim_\alpha 1$, then there exists some constant $K > 0$ depending only on α, K_1, K_2 such that*

$$K^{-1}\sqrt{n}\sigma\left(\frac{\|F\|_{L_2(P)}}{\sigma}\right)^{\alpha/2} \le \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} \le K\sqrt{n}\sigma\left(\frac{\|F\|_{L_2(P)}}{\sigma}\right)^{\alpha/2}.$$

Note that the right-hand side of the inequality can also be derived from (5.3) [taking supremum over all finitely discrete probability measures only serves to get rid of the random entropy induced by $L_2(\mathbb{P}_n)$ norm therein].

The following lemma guarantees the existence of a particular type of α-full class that serves as the basis of the construction in the proof of Theorems 2 and 4. The proof can be found in Appendix D in [26].

LEMMA 6. *Let $\mathcal{X}, P$ be as in Theorem 2. Then for each $\alpha > 0$, there exists some function class $\mathcal{F}$ defined on $\mathcal{X}$ which is α-full and contains $\mathcal{G} \equiv \{\mathbf{1}_{[a,b]} : 0 \le a \le b \le 1\}$.*

5.2. *Proof of Theorem 1.* The key ingredient in the proof of Theorem 1 is the following, which may be of independent interest.

PROPOSITION 5. *Suppose Assumption* A *holds. For any function class* $\mathcal{F}$,

$$\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\|_{\mathcal{F}} \le \mathbb{E}\left[\sum_{k=1}^{n}\bigl(|\eta_{(k)}|-|\eta_{(k+1)}|\bigr)\mathbb{E}\left\|\sum_{i=1}^{k}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right], \tag{5.5}$$

where $|\eta_{(1)}| \ge \cdots \ge |\eta_{(n)}| \ge |\eta_{(n+1)}| \equiv 0$ *are the reversed order statistics for*: (i) [*under* (A1)] $\{2|\xi_i|\}_{i=1}^n$, (ii) [*under* (A2)] $\{|\xi_i - \xi_i'|\}_{i=1}^n$ *with* $\{\xi_i'\}$ *being an independent copy of* $\{\xi_i\}$.

PROOF. We drop $\mathcal{F}$ from the notation for supremum norm over $\mathcal{F}$ and write $\|\cdot\|$ for $\|\cdot\|_{\mathcal{F}}$. We first consider the condition (A1). Note that for $(X_1', \ldots, X_n')$ being an independent copy of $(X_1, \ldots, X_n)$, we have

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\| &= \mathbb{E}_{\boldsymbol{\xi},\boldsymbol{X}}\left\|\sum_{i=1}^{n}\xi_i\bigl(f(X_i)-\mathbb{E}_{\boldsymbol{X}'}f(X_i')\bigr)\right\| \\ &\le \mathbb{E}\left\|\sum_{i=1}^{n}\xi_i\bigl(f(X_i)-f(X_i')\bigr)\right\|.\end{aligned}$$

Here in the first equality we used the centeredness assumption on the function class $\mathcal{F}$ in (A1). Now conditional on $\boldsymbol{\xi}$, for fixed $\varepsilon_1, \ldots, \varepsilon_n$, the map $(X_1, \ldots, X_n, X_1', \ldots, X_n') \mapsto \|\sum_{i=1}^n \xi_i\varepsilon_i(f(X_i)-f(X_i'))\|$ is a permutation of the original map (without ε_i's). Since $(X_1, \ldots, X_n, X_1', \ldots, X_n')$ is the coordinate projection of a product measure, it follows by taking expectation over $\varepsilon_1, \ldots, \varepsilon_n$ that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}\left\|\sum_{i=1}^{n}\xi_i\bigl(f(X_i)-f(X_i')\bigr)\right\| \\ &\quad= \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{X},\boldsymbol{X}'}\left\|\sum_{i=1}^{n}\xi_i\varepsilon_i\bigl(f(X_i)-f(X_i')\bigr)\right\|.\end{aligned} \tag{5.6}$$

This entails that

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\| &\le 2\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\varepsilon},\boldsymbol{X}}\left\|\sum_{i=1}^{n}|\xi_i|\operatorname{sgn}(\xi_i)\varepsilon_i f(X_i)\right\| \\ &= 2\mathbb{E}\left\|\sum_{i=1}^{n}|\xi_i|\varepsilon_i f(X_i)\right\|,\end{aligned} \tag{5.7}$$

where the equality follows since the random vector $(\operatorname{sgn}(\xi_1)\varepsilon_1, \ldots, \operatorname{sgn}(\xi_n)\varepsilon_n)$ has the same distribution as that of $(\varepsilon_1, \ldots, \varepsilon_n)$ and is independent of $\xi_1, \ldots, \xi_n$. We will simply write $|\xi_i|$ without the absolute value in the sequel for notational convenience. Let π be a permutation over $\{1, \ldots, n\}$ such that $\xi_i = \xi_{(\pi(i))}$. Then the

right-hand side of (5.7) equals

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{i=1}^{n}\xi_{(\pi(i))}\varepsilon_i f(X_i)\right\| \\
&\quad= \mathbb{E}\left\|\sum_{i=1}^{n}\xi_{(i)}\varepsilon_{\pi^{-1}(i)} f(X_{\pi^{-1}(i)})\right\| \qquad \text{(by relabeling)} \\
&\quad= \mathbb{E}\left\|\sum_{i=1}^{n}\xi_{(i)}\varepsilon_i f(X_i)\right\| \qquad \big(\text{by invariance of } (P_X\otimes P_\varepsilon)^n\big).
\end{aligned} \tag{5.8}$$

Now write $\xi_{(i)} = \sum_{k\geq i}(\xi_{(k)} - \xi_{(k+1)})$ where $\xi_{(n+1)} \equiv 0$. The above display can be rewritten as

$$\mathbb{E}\left\|\sum_{i=1}^{n}\sum_{k=i}^{n}(\xi_{(k)} - \xi_{(k+1)})\varepsilon_i f(X_i)\right\| = \mathbb{E}\left\|\sum_{k=1}^{n}(\xi_{(k)} - \xi_{(k+1)})\sum_{i=1}^{k}\varepsilon_i f(X_i)\right\|. \tag{5.9}$$

The claim under (A1) follows by combining (5.7)–(5.9). For (A2), let ξ_i''s be an independent copy of ξ_i's. Then the analogy of (5.7) becomes

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\| &= \mathbb{E}\left\|\sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i')f(X_i)\right\| \leq \mathbb{E}\left\|\sum_{i=1}^{n}(\xi_i - \xi_i')f(X_i)\right\| \\
&= \mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i|\xi_i - \xi_i'|f(X_i)\right\| = \mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i|\eta_i|f(X_i)\right\|,
\end{aligned}$$

where $\eta_i \equiv \xi_i - \xi_i'$. The claim for (A2) follows by repeating the arguments in (5.8) and (5.9). □

PROOF OF THEOREM 1. First, consider (A1). Using Proposition 5, we see that

$$\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\|_{\mathcal{F}_n} \leq 2\mathbb{E}\left[\sum_{k=1}^{n}\big(|\xi_{(k)}| - |\xi_{(k+1)}|\big)\mathbb{E}\left\|\sum_{i=1}^{k}\varepsilon_i f(X_i)\right\|_{\mathcal{F}_n}\right]. \tag{5.10}$$

By the assumption that $\mathcal{F}_k \supset \mathcal{F}_n$ for any $1 \leq k \leq n$,

$$\mathbb{E}\left\|\sum_{i=1}^{k}\varepsilon_i f(X_i)\right\|_{\mathcal{F}_n} \leq \mathbb{E}\left\|\sum_{i=1}^{k}\varepsilon_i f(X_i)\right\|_{\mathcal{F}_k} \leq \psi_n(k). \tag{5.11}$$

Collecting (5.10)–(5.11), we see that

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\|_{\mathcal{F}_n} &\leq 2\mathbb{E}\left[\sum_{k=1}^{n}\big(|\xi_{(k)}| - |\xi_{(k+1)}|\big)\psi_n(k)\right] \\
&= 2\mathbb{E}\sum_{k=1}^{n}\int_{|\xi_{(k+1)}|}^{|\xi_{(k)}|}\psi_n(k)\,\mathrm{d}t
\end{aligned}$$

$$\le 2\mathbb{E}\int_0^\infty \psi_n\big(\big|\{i : |\xi_i| \ge t\}\big|\big)\,\mathrm{d}t$$
$$\le 2\int_0^\infty \psi_n\left(\sum_{i=1}^n \mathbb{P}\big(|\xi_i| > t\big)\right)\mathrm{d}t,$$

where the last inequality follows from Fubini's theorem and Jensen's inequality, completing the proof for the upper bound for (A1). For (A2), mimicking the above proof, we have

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}_n} &\le \int_0^\infty \psi_n\left(\sum_{i=1}^n \mathbb{P}\big(|\xi_i - \xi_i'| \ge t\big)\right)\mathrm{d}t \\
&\le \int_0^\infty \psi_n\left(\sum_{i=1}^n \mathbb{P}\big(|\xi_i| \ge t/2\big) + \mathbb{P}\big(|\xi_i'| \ge t/2\big)\right)\mathrm{d}t \\
&= \int_0^\infty \psi_n\left(2\sum_{i=1}^n \mathbb{P}\big(|\xi_i| \ge t/2\big)\right)\mathrm{d}t \\
&= 2\int_0^\infty \psi_n\left(2\sum_{i=1}^n \mathbb{P}\big(|\xi_i| > t\big)\right)\mathrm{d}t.
\end{aligned}$$

The proof of the claim for (A2) is completed by noting that $\psi_n(2x) \le 2\psi_n(x)$ due to the concavity of ψ_n and $\psi_n(0) = 0$. □

5.3. *Proof of Theorem 2.* We need the following lemma.

LEMMA 7. *Suppose that $\xi_1, \dots, \xi_n$ are i.i.d. mean-zero random variables independent of i.i.d. $X_1, \dots, X_n$. Then*

$$\|\xi_1\|_1 \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} \le 2\mathbb{E}\left\|\sum_{i=1}^n \xi_i f(X_i)\right\|_{\mathcal{F}}.$$

PROOF. The proof follows that of the left-hand side inequality in Lemma 2.9.1 of [66], so we omit the details. □

LEMMA 8. *Let $X_1, \dots, X_n$ be i.i.d. random variables distributed on $[0,1]$ with a probability law P admitting a Lebesgue density bounded away from ∞. Let $\{I_i\}_{i=1}^n$ be a partition of $[0,1]$ such that $I_i \cap I_j = \varnothing$ for $i \ne j$ and $\bigcup_{i=1}^n I_i = [0,1]$, and $L^{-1}n^{-1} \le |I_i| \le Ln^{-1}$ for some absolute value $L > 0$. Then there exists some $\tau \equiv \tau_{L,P} \in (0,1)$ such that for n sufficiently large,*

$$\mathbb{P}\big(X_1, \dots, X_n \text{ lie in at most } \tau n \text{ intervals among } \{I_i\}_{i=1}^n\big) \le 0.5^{n-1}.$$

The proof of Lemma 8 can be found in Appendix D in [26]. Now we are in position to prove Theorem 2.

PROOF OF THEOREM 2. The proof will proceed in two steps. The first step aims at establishing a lower bound for the multiplier empirical process on the order of $n^{1/\gamma}$.

Let $\alpha = 2/(\gamma - 1)$, and $\tilde{\mathcal{F}}$ be an α-full class on $\mathcal{X}$ in Lemma 6. Further let $\delta_k = k^{-1/(2+\alpha)}$ and $\tilde{\mathcal{F}}_k \equiv \tilde{\mathcal{F}}(\delta_k) = \{f \in \tilde{\mathcal{F}} : Pf^2 < \delta_k^2\}$. Then it follows from Lemma 5 that there exists some constant $K > 0$,

$$K^{-1}k^{\alpha/(2+\alpha)} \le \mathbb{E}\left\|\sum_{i=1}^{k} \varepsilon_i f(X_i)\right\|_{\tilde{\mathcal{F}}_k} \le K k^{\alpha/(2+\alpha)}.$$

Lemma 7 now guarantees that $\mathbb{E}\|\sum_{i=1}^{n} \xi_i f(X_i)\|_{\tilde{\mathcal{F}}_n}$ can be bounded from below by a constant multiple of $n^{\alpha/(2+\alpha)} = n^{1/\gamma}$ where the constant depends on $\|\xi_1\|_1$. This completes the first step of the proof.

In the second step, we aim at establishing a lower bound of order $n^{1/p}$. To this end, let $\{I_j\}_{j=1}^n$ be a partition of $\mathcal{X}$ such that $L^{-1}n^{-1} \le |I_j| \le Ln^{-1}$. On the other hand, let $f_j \equiv \mathbf{1}_{I_j} \in \tilde{\mathcal{F}}_n$ for $1 \le j \le n$ (increase δ_n by constant factors if necessary), and $\mathcal{E}_n$ denote the event that $X_1, \ldots, X_n$ lie in $N \ge \tau n$ sets among $\{I_j\}_{j=1}^n$. Then Lemma 8 entails that $\mathbb{P}(\mathcal{E}_n) \ge 1 - 0.5^n \ge 1/2$ for n sufficiently large. Furthermore, let $\mathcal{I}_j \equiv \{i : X_i \in I_j\}$ and pick any $X_{\iota(j)} \in I_j$. Note that $\mathcal{I}_j$'s are disjoint, and hence conditionally on $\boldsymbol{X}$ we have

$$\begin{aligned}
\mathbb{E}\max_{1\le j\le \tau n}|\xi_j| &\le \mathbb{E}\max_{1\le j\le N}|\xi_{\iota(j)}| && \text{(by i.i.d. assumption on } \xi_i\text{'s)}\\
&\le \mathbb{E}\max_{1\le j\le N}\left|\xi_{\iota(j)} + \mathbb{E}\sum_{i\in\mathcal{I}_j\setminus\iota(j)}\xi_i\right| && (\mathcal{I}_j\text{'s are disjoint and } \mathbb{E}\xi_i = 0)\\
&\le \mathbb{E}\max_{1\le j\le N}\left|\sum_{i\in\mathcal{I}_j}\xi_i\right| && \text{(by Jensen's inequality)}.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\|_{\tilde{\mathcal{F}}_n} &\ge \mathbb{E}\left[\max_{1\le j\le n}\left|\sum_{i=1}^{n}\xi_i f_j(X_i)\right|\right] \ge \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{\xi}}\max_{1\le j\le N}\left|\sum_{i\in\mathcal{I}_j}\xi_i\right|\mathbf{1}_{\mathcal{E}_n}\right]\\
&\ge \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{\xi}}\max_{1\le j\le \tau n}|\xi_j|\mathbf{1}_{\mathcal{E}_n}\right] \ge \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\max_{1\le j\le \tau n}|\xi_j|
\end{aligned}$$

for n sufficiently large. Now the second step follows from the assumption, and hence completing the proof. □

5.4. *Proof of Theorem 3.* We first prove Proposition 2.

PROOF OF PROPOSITION 2. Let $\mathbb{M}_n f \equiv \frac{2}{n}\sum_{i=1}^n (f-f_0)(X_i)\xi_i - \frac{1}{n}\sum_{i=1}^n (f-f_0)^2(X_i)$, and $Mf \equiv \mathbb{E}[\mathbb{M}_n(f)] = -P(f-f_0)^2$. Here, we used the fact that $\mathbb{E}\xi_i = 0$ and the independence assumption between $\{\xi_i\}$ and $\{X_i\}$. Then it is easy to see that

$$\big|\mathbb{M}_n f - \mathbb{M}_n f_0 - (Mf - Mf_0)\big| \le \left|\frac{2}{n}\sum_{i=1}^n (f-f_0)(X_i)\xi_i\right| + \big|(\mathbb{P}_n - P)(f-f_0)^2\big|.$$

The first claim (i.e., convergence rate in probability) follows by standard symmetrization and contraction principle for the empirical process indexed by a uniformly bounded function class, followed by an application of Theorem 3.2.5 of [66].

Now assume that $\xi_1, \dots, \xi_n$ are i.i.d. mean-zero errors with $\|\xi_1\|_p < \infty$ for some $p \ge 2$. Fix $t \ge 1$. For $j \in \mathbb{N}$, let $\mathcal{F}_j \equiv \{f \in \mathcal{F} : 2^{j-1} t\delta_n \le \|f - f_0\|_{L_2(P)} < 2^j t\delta_n\}$. Then by a standard peeling argument, we have

$$\mathbb{P}\big(\|\hat{f}_n - f_0\|_{L_2(P)} \ge t\delta_n\big) \le \sum_{j\ge 1} \mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \big(\mathbb{M}_n(f) - \mathbb{M}_n(f_0)\big) \ge 0\Big).$$

Each probability term in the above display can be further bounded by

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{f\in\mathcal{F}_j} \big(\mathbb{M}_n(f) - \mathbb{M}_n(f_0) - (Mf - Mf_0)\big) \ge 2^{2j-2}t^2\delta_n^2\Big) \\
&\quad\le \mathbb{P}\left(\sup_{f\in\mathcal{F}-f_0:\|f\|_{L_2(P)}\le 2^j t\delta_n} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right| \ge 2^{2j-4}t^2\sqrt{n}\delta_n^2\right) \\
&\qquad + \mathbb{P}\left(\sup_{f\in\mathcal{F}-f_0:\|f\|_{L_2(P)}\le 2^j t\delta_n} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f^2(X_i) - Pf^2\big)\right| \ge 2^{2j-3}t^2\sqrt{n}\delta_n^2\right).
\end{aligned}$$

By the contraction principle and moment inequality for the empirical process (Lemma 1), we have

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\substack{f\in\mathcal{F}-f_0:\\ \|f\|_{L_2(P)}\le 2^j t\delta_n}} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i f(X_i)\right|^2\right) \vee \mathbb{E}\left(\sup_{\substack{f\in\mathcal{F}-f_0:\\ \|f\|_{L_2(P)}\le 2^j t\delta_n}} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f^2(X_i)\right|^2\right) \\
&\qquad \lesssim \big[\phi_n(2^j t\delta_n)\big]^2 + (1\vee\|\xi_1\|_2)^2 2^{2j}t^2\delta_n^2 + (1\vee\|\xi_1\|_p)^2 n^{-1+2/p}.
\end{aligned}$$

In the above calculation, we used the fact that $\mathbb{E}\max_{1\le i\le n}|\xi_i|^2 \le \|\xi_1\|_p^2 n^{2/p}$ under $\|\xi_1\|_p < \infty$. By Chebyshev's inequality,

$$\begin{aligned}
&\mathbb{P}\big(\|\hat{f}_n - f_0\|_{L_2(P)} \ge t\delta_n\big) \\
&\qquad \le C_\xi \sum_{j\ge 1}\left[\left(\frac{\phi_n(2^j t\delta_n)}{2^{2j}t^2\sqrt{n}\delta_n^2}\right)^2 \vee \frac{1}{2^{2j}t^2 n\delta_n^2} \vee \frac{1}{2^{4j}t^4 n^{2-2/p}\delta_n^4}\right].
\end{aligned}$$

Under the assumption that $\delta_n \geq n^{-\frac{1}{2}+\frac{1}{2p}}$, and noting that $\phi_n(2^j t\delta_n) \leq 2^j t\phi_n(\delta_n)$ by the assumption that $\delta \mapsto \phi_n(\delta)/\delta$ is nonincreasing, the right-hand side of the above display can be further bounded up to a constant by $\sum_{j\geq 1}(\frac{\phi_n(\delta_n)}{2^j t\sqrt{n}\delta_n^2})^2 + \frac{1}{t^2} \lesssim \frac{1}{t^2}$ for $t \geq 1$. The expectation bound follows by integrating the tail estimate. □

The following lemma calculates an upper bound for the multiplier empirical process at the target rate in Theorem 3. The proof can be found in Appendix D in [26].

LEMMA 9. *Suppose that Assumption A holds with i.i.d.* $X_1, \ldots, X_n$*'s with law* P*, and* $\mathcal{F} \subset L_\infty(1)$ *satisfies the entropy condition* (F) *with* $\alpha \in (0, 2)$*. Further assume for simplicity that* ξ_i*'s have the same marginal distributions with* $\|\xi_1\|_{p,1} < \infty$*. Then with* $\delta_n \equiv n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}$ *we have*

$$\mathbb{E}\sup_{Pf^2\leq\rho^2\delta_n^2}\left|\sum_{i=1}^n \xi_i f(X_i)\right| \vee \mathbb{E}\sup_{Pf^2\leq\rho^2\delta_n^2}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \leq \bar{K}_\alpha\big(\rho^{1-\alpha/2} \vee \rho^{-\alpha}\big)\begin{cases} n^{\frac{\alpha}{2+\alpha}}\big(1 \vee \|\xi_1\|_{1+2/\alpha,1}\big), & p \geq 1+2/\alpha, \\ n^{\frac{1}{p}}\big(1 \vee \|\xi_1\|_{p,1}\big), & 1 \leq p < 1+2/\alpha. \end{cases}$$

PROOF OF THEOREM 3. The claim follows immediately from Lemma 9 by noting that the rate δ_n chosen therein corresponds to the condition (3.3) in Proposition 2, along with Proposition 4 handling (3.4). □

Acknowledgments. The authors would like to thank Vladimir Koltchinskii, Richard Samworth, two referees and an Associate Editor for helpful comments and suggestions on an earlier version of the paper. We also thank Shahar Mendelson for sending us a copy of his paper [48].

SUPPLEMENTARY MATERIAL

Supplement: Additional proofs (DOI: 10.1214/18-AOS1748SUPP; .pdf). In the supplement [26], we provide detailed proofs for (i) Theorem 4, (ii) the impossibility results Propositions 1 and 3 and (iii) all remaining lemmas.

REFERENCES

[1] ALEXANDER, K. S. (1985). The nonexistence of a universal multiplier moment for the central limit theorem. In *Probability in Banach Spaces*, *V* (*Medford*, *Mass*., 1984). *Lecture Notes in Math*. **1153** 15–16. Springer, Berlin. MR0821974

[2] ANDERSEN, N. T., GINÉ, E. and ZINN, J. (1988). The central limit theorem for empirical processes under local conditions: The case of Radon infinitely divisible limits without Gaussian component. *Trans. Amer. Math. Soc*. **308** 603–635. MR0930076

[3] ARIAS-CASTRO, E., DONOHO, D. L. and HUO, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* **51** 2402–2425. MR2246369

[4] AUDIBERT, J.-Y. and CATONI, O. (2011). Robust linear least squares regression. *Ann. Statist.* **39** 2766–2794. MR2906886

[5] BAI, Z. D., SILVERSTEIN, J. W. and YIN, Y. Q. (1988). A note on the largest eigenvalue of a large-dimensional sample covariance matrix. *J. Multivariate Anal.* **26** 166–168. MR0963829

[6] BAI, Z. D. and YIN, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.* **21** 1275–1294. MR1235416

[7] BARAUD, Y., BIRGÉ, L. and SART, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.* **207** 425–517. MR3595933

[8] BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554

[9] BARTLETT, P. L. and MENDELSON, S. (2006). Empirical minimization. *Probab. Theory Related Fields* **135** 311–334. MR2240689

[10] BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719

[11] BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348

[12] BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann. Statist.* **43** 2507–2536. MR3405602

[13] BUBECK, S., CESA-BIANCHI, N. and LUGOSI, G. (2013). Bandits with heavy tail. *IEEE Trans. Inform. Theory* **59** 7711–7717. MR3124669

[14] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

[15] CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407

[16] CATONI, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. Preprint. Available at arXiv:1603.05229.

[17] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

[18] DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558

[19] DUDLEY, R. M. (2014). *Uniform Central Limit Theorems*, 2nd ed. *Cambridge Studies in Advanced Mathematics* **142**. Cambridge Univ. Press, New York. MR3445285

[20] GAO, X. and HUANG, J. (2010). Asymptotic analysis of high-dimensional LAD regression with Lasso. *Statist. Sinica* **20** 1485–1506. MR2777333

[21] GINÉ, E. and KOLTCHINSKII, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143–1216. MR2243881

[22] GINÉ, E., LATAŁA, R. and ZINN, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability, II* (*Seattle, WA*, 1999). *Progress in Probability* **47** 13–38. Birkhäuser, Boston, MA. MR1857312

[23] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285

[24] GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12** 929–998. MR0757767

[25] GINÉ, E. and ZINN, J. (1986). Lectures on the central limit theorem for empirical processes. In *Probability and Banach Spaces* (*Zaragoza*, 1985). *Lecture Notes in Math*. **1221** 50–113. Springer, Berlin. MR0875007

[26] HAN, Q. and WELLNER, J. A. (2019). Supplement to "Convergence rates of least squares regression estimators with heavy-tailed errors." DOI:10.1214/18-AOS1748SUPP.

[27] HAN, Q. and WELLNER, J. A. (2018). Robustness of shape-restricted regression estimators: An envelope perspective. Preprint. Available at arXiv:1805.02542.

[28] HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40. MR3491112

[29] JOLY, E., LUGOSI, G. and OLIVEIRA, R. I. (2017). On the estimation of the mean of a random vector. *Electron. J. Stat.* **11** 440–451. MR3619312

[30] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

[31] KOLTCHINSKII, V. and PANCHENKO, D. (2000). Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability, II* (*Seattle, WA*, 1999). *Progress in Probability* **47** 443–457. Birkhäuser, Boston, MA. MR1857339

[32] KOROSTELËV, A. P. and TSYBAKOV, A. B. (1992). Asymptotically minimax image reconstruction problems. In *Topics in Nonparametric Estimation*. *Adv. Soviet Math.* **12** 45–86. Amer. Math. Soc., Providence, RI. MR1191691

[33] KOROSTELËV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. *Lecture Notes in Statistics* **82**. Springer, New York. MR1226450

[34] LECUÉ, G. and MENDELSON, S. (2017). Sparse recovery under weak moment assumptions. *J. Eur. Math. Soc.* (*JEMS*) **19** 881–904. MR3612870

[35] LECUÉ, G. and MENDELSON, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.* **46** 611–641. MR3782379

[36] LEDOUX, M. and TALAGRAND, M. (1986). Conditions d'intégrabilité pour les multiplicateurs dans le TLC banachique. *Ann. Probab.* **14** 916–921. MR0841593

[37] LEDOUX, M. and TALAGRAND, M. (2011). *Probability in Banach Spaces*: *Isoperimetry and Processes*. Springer, Berlin. MR2814399

[38] LUGOSI, G. and MENDELSON, S. (2018). Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*. To appear.

[39] LUGOSI, G. and MENDELSON, S. (2018). Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc. (JEMS)*. To appear.

[40] LUGOSI, G. and MENDELSON, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794. MR3909950

[41] MASON, D. M. (1983). The asymptotic distribution of weighted empirical distribution functions. *Stochastic Process. Appl.* **15** 99–109. MR0694539

[42] MASSART, P. and NÉDÉLEC, É. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366. MR2291502

[43] MASSART, P. and RIO, E. (1998). A uniform Marcinkiewicz–Zygmund strong law of large numbers for empirical processes. In *Asymptotic Methods in Probability and Statistics* (*Ottawa, ON*, 1997) 199–211. North-Holland, Amsterdam. MR1661481

[44] MENDELSON, S. (2015). Learning without concentration. *J. ACM* **62** Art. 21, 25. MR3367000

[45] MENDELSON, S. (2016). Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl.* **126** 3652–3680. MR3565471

[46] MENDELSON, S. (2017). "Local" vs. "global" parameters—breaking the Gaussian complexity barrier. *Ann. Statist.* **45** 1835–1862. MR3718154

[47] MENDELSON, S. (2017). On aggregation for heavy-tailed classes. *Probab. Theory Related Fields* **168** 641–674. MR3663627

[48] MENDELSON, S. (2017). Extending the small-ball method. Preprint. Available at arXiv:1709.00843.

[49] MENDELSON, S. (2017). On multiplier processes under weak moment assumptions. In *Geometric Aspects of Functional Analysis*. *Lecture Notes in Math*. **2169** 301–318. Springer, Cham. MR3645129

[50] MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

[51] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

[52] POLLARD, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199. MR1128411

[53] PORTNOY, S. and KOENKER, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statist. Sci.* **12** 279–300. MR1619189

[54] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

[55] SHORACK, G. R. and WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics*. *Classics in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR3396731

[56] SIVAKUMAR, V., BANERJEE, A. and RAVIKUMAR, P. K. (2015). Beyond sub-Gaussian measurements: High-dimensional structured estimation with sub-exponential designs. In *Advances in Neural Information Processing Systems* 2206–2214.

[57] STRASSEN, V. and DUDLEY, R. M. (1969). The central limit theorem and ε-entropy. In *Probability and Information Theory* (*Proc. Internat. Sympos., McMaster Univ., Hamilton, Ont.*, 1968) 224–231. Springer, Berlin. MR0279872

[58] TALAGRAND, M. (2014). *Upper and Lower Bounds for Stochastic Processes*: *Modern Methods and Classical Problems*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. 3. *Folge. A Series of Modern Surveys in Mathematics* **60**. Springer, Heidelberg. MR3184689

[59] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[60] VAN DE GEER, S. (1987). A new approach to least-squares estimation, with applications. *Ann. Statist.* **15** 587–602. MR0888427

[61] VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343

[62] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

[63] VAN DE GEER, S. and WEGKAMP, M. (1996). Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.* **24** 2513–2523. MR1425964

[64] VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge Univ. Press, Cambridge.

[65] VAN DER VAART, A. and WELLNER, J. A. (2011). A local maximal inequality under uniform entropy. *Electron. J. Stat.* **5** 192–203. MR2792551

[66] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

[67] YANG, Y. (2001). Nonparametric regression with dependent errors. *Bernoulli* **7** 633–655. MR1849372

[68] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

[69] YAO, Q. W. (1993). Tests for change-points with epidemic alternatives. *Biometrika* **80** 179–191. MR1225223

[70] ZHANG, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195-4322
USA
E-MAIL: royhan@uw.edu
jaw@stat.washington.edu
URL: http://www.stat.washington.edu/jaw/