On Least Squares Estimation Under Heteroscedastic and Heavy-Tailed Errors

Arun K. Kuchibhotla and Rohit K. Patra

The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104.
e-mail: arunku@upenn.edu

Department of Statistics
University of Florida
221 Griffin-Floyd Hall
Gainesville, FL 32611
e-mail: rohitpatra@ufl.edu

Abstract: We consider least squares estimation in a general nonparametric regression model. The rate of convergence of the least squares estimator (LSE) for the unknown regression function is well studied when the errors are sub-Gaussian. We find upper bounds on the rates of convergence of the LSE when the errors have uniformly bounded conditional variance and have only finitely many moments. We show that the interplay between the moment assumptions on the error, the metric entropy of the class of functions involved, and the “local” structure of the function class around the truth drives the rate of convergence of the LSE. We find sufficient conditions on the errors under which the rate of the LSE matches the rate of the LSE under sub-Gaussian error. Our results are finite sample and allow for heteroscedastic and heavy-tailed errors.

Keywords and phrases: Dyadic peeling, finite sample tail probability bounds, local envelopes, maximal inequality, heavy tails.

1. Introduction

Suppose we have \( n \) i.i.d. observations \( \{(X_i, Y_i) \in \chi \times \mathbb{R}, 1 \leq i \leq n\} \) from the nonparametric regression model
\[
Y = f_0(X) + \epsilon,
\]
where \( \chi \) is a metric space, \( f_0 : \chi \to \mathbb{R} \) is an unknown measurable function, and \( \epsilon \) satisfies \( E(\epsilon|X) = 0 \) almost everywhere \( P_X \), the distribution of \( X \). In this paper, we consider the least squares estimator (LSE) for \( f_0 \) under the constraint that \( f_0 \in \mathcal{F} \), where \( \mathcal{F} \) denotes a class of real-valued functions on \( \chi \). Formally, the LSE is defined as
\[
\hat{f} := \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.
\]

The two most widely used metrics for assessing the error in estimation are the empirical loss (\( \|\hat{f} - f_0\|_n \)) and the prediction loss (\( \|\hat{f} - f_0\| \)), where for any function \( g : \chi \to \mathbb{R} \),
\[
\|g\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} g^2(X_i) \quad \text{and} \quad \|g\|^2 := \int_{\chi} g^2(x) dP_X(x).
\]

In this paper, we find upper bounds on the rate of convergence of \( \|\hat{f} - f_0\| \) and the corresponding risk, the mean integrated squared error,
\[
E(\|\hat{f} - f_0\|^2) := E \left[ \int_{\chi} \left( \hat{f}(x) - f_0(x) \right)^2 dP_X(x) \right].
\]
Why study the nonparametric LSE? Least squares is one of the most natural methods of estimation in regression. The study of the LSE has received considerable attention in statistics as well as machine learning; see [9, 36, 41, 54, 55] for important contributions. LSEs are particularly useful when \( \mathcal{F} \) is known to satisfy some shape constraints such as monotonicity, convexity, or unimodality. In such cases, the LSEs are tuning parameter-free, can be computed as the solution to convex optimization problems, and are adaptive, i.e., the rate of convergence of the LSE changes depending on the “structure” of \( f_0 \) [4, 23, 28, 43, 47]. For example, if \( \mathcal{F} \) is the class of monotone functions and \( f_0 \) is a strictly increasing function then \( \hat{f} \) converges at an \( n^{1/3} \) rate (i.e., \( \| \hat{f} - f_0 \| = O_p(n^{-1/3}) \)); however, if \( f_0 \equiv 0 \), then \( \hat{f} \) converges at an \( n^{1/2} \) rate [10, 25, 56].

Tsirelson [50] and van de Geer and Wegkamp [53] have established necessary and sufficient conditions on \( \mathcal{F} \) and \( \epsilon \) under which the LSE is “rate-optimal”; see discussions after Corollary 2.1 for more details, also see Example 3.1. Our results significantly expand the scenarios under which LSE can be proven to be rate-optimal or “be a safe choice” for the model at hand.

Theorem 3.2.5 of [55] implies that the LSE \( \hat{f} \) defined on \( \mathcal{F} \) satisfies \( \| \hat{f} - f_0 \| = O_p(r_n) \) for any \( r_n \) such that

\[
E \left[ \sup_{f \in \mathcal{F} : \| f - f_0 \| \leq r_n} \left| G_n \| 2\epsilon (f - f_0)(X) - (f - f_0)^2(X) \| \right| \right] \leq C \sqrt{n} r_n^2, \tag{3}
\]

where \( C \) denotes a constant\(^1\). In the rest of this paper, we make the convention that the constant \( C \) is not necessarily the same on each occurrence. van de Geer and Wainwright [52] show that the rate of convergence of the LSE is completely characterized by the empirical process above; hence, sharper bounds on the expectation in (3) leads to sharper rates for the LSE. Assuming that the functions in \( \mathcal{F} \) are uniformly bounded by \( \Phi < \infty \), the expectation in (3) can be bounded using symmetrization and contraction (Theorem 3.1.21 and Corollary 3.2.2 of [21], respectively) by

\[
E \left[ \sup_{f \in \mathcal{F} : \| f - f_0 \| \leq r_n} \left| G_n \left[ (|\epsilon| + \Phi)(f - f_0)(X) \right] \right| \right]. \tag{4}
\]

The path-breaking works by the authors of [7, 41, 54, 55] have provided sharp maximal inequalities to bound the expectation in (4). However, the assumptions are often too strong and might not be necessary: [6, 51, 54, 55] assume restrictive conditions (such as boundedness or sub-exponential tails) on the distribution of \( \epsilon \); [25, 26] assume that \( \epsilon \) is independent of \( X \); [37, 38] make minimal assumptions on \( \epsilon \) but make strong structural assumptions on \( \mathcal{F} \). Moreover study of the LSE in specific examples [2, 44, 56] has shown that such conditions are not necessary in general. In this work, we relax the assumptions needed on \( \epsilon \) and \( \mathcal{F} \) when providing sharp maximal inequalities to bound (4). This, in turn, helps us establish the rate of convergence of the LSE under weaker assumptions. We argue that there are three properties concerning \( \epsilon \) and \( \mathcal{F} \) that play a pivotal role when finding the rate of convergence of the LSE: (1) the tail behavior of \( \epsilon \); (2) the “complexity” of \( \mathcal{F} \); and (3) the “local” structure of \( \mathcal{F} \) in the neighborhood of \( f_0 \). In the following three subsections, we discuss these three aspects in detail and state our main assumptions.

1.1. Assumptions on \( \epsilon \)

In this work, we assume that there exists a \( \sigma > 0 \) such that

\[
E(\epsilon^2 | X) \leq \sigma^2 \text{ almost everywhere (a.e.) } P_X, \quad (\text{CVar})
\]

\(^1\)By “constant” we will always mean a quantity that does not depend on \( n \) but might depend on the various parameters introduced in our assumptions. In each occurrence of \( C \), we clarify what parameters of the model the constant depends on.
and there exists a finite \( q \geq 2 \) and \( K_q < \infty \) such that
\[
\mathbb{E}(|\epsilon|^q) \leq K_q^q. \tag{\mathcal{E}_q}
\]
Note that (CVar) allows for heteroscedastic errors and (\mathcal{E}_q) allows for heavy-tailed errors. Of course we are not the first to consider heavy-tailed errors (i.e., \( \epsilon \) with only finitely many moments). Both \cite{26} and \cite{38} allow for heavy-tailed errors, but require other strong assumptions on \( \epsilon \) and \( \mathcal{F} \), respectively. Shen and Wong \cite{44} and Chen and Shen \cite{13} also allow for heavy-tailed errors, but their results do not directly relate the rate of convergence of the LSE to the moment assumptions on \( \epsilon \).

1.2. Complexity of \( \mathcal{F} \)

The bound on \((4)\) depends on the “effective” number of elements in the supremum. The effective number is given by number of functions that are essentially “different”. This number is usually described in terms of metric entropy numbers. In the following sections, we use three of the most widely used entropy numbers. For any \( \zeta > 0 \), function class \( \mathcal{F} \), and metric \( d(\cdot, \cdot) \) on \( \mathcal{F} \times \mathcal{F} \), let \( N(\zeta, \mathcal{F}, d) \) be the minimum \( m \geq 1 \) for which there exist functions \( \{g_i\}_{i=1}^m \) such that for every \( f \in \mathcal{F} \) there exists a \( j \leq m \) such that \( d(f, g_j) \leq \zeta \). We only use metrics \( d(\cdot, \cdot) \) of the form \( d(f, g) = D(f - g) \) for some norm \( D(\cdot) \) and for these forms, we write \( N(\zeta, \mathcal{F}, d) \equiv N(\zeta, \mathcal{F}, D) \). \( N(\zeta, \mathcal{F}, d) \) and \( \log N(\zeta, \mathcal{F}, d) \) are called the \( \zeta \)-covering number and the \( \zeta \)-metric entropy of \( \mathcal{F} \) with respect to the metric \( d \), respectively. In Section 3, we study the LSE when \( \mathcal{F} \) satisfies
\[
\log N(\zeta, \mathcal{F}, \| \cdot \|_\infty) \leq A\zeta^{-\alpha}, \quad \text{for some } A > 0 \text{ and } \alpha \in [0, 2), \tag{L_\infty}
\]
where for any \( f : \chi \to \mathbb{R} \), \( f_\infty := \sup_{x \in \chi} |f(x)| \). When there is no scope for confusion, we will suppress the dependence on \( \zeta \) and call \( \log N(\zeta, \mathcal{F}, \| \cdot \|_\infty) \) the \( L_\infty \)-entropy. In Section 4, we assume that there exists a constant \( A > 0 \) such that \( \mathcal{F} \) satisfies
\[
\sup_{Q, F' \subseteq F - f_0} \log N(\zeta, F', \| \cdot \|_2, Q) \leq A\zeta^{-\alpha}, \quad \text{for } \alpha > 0, \text{ and } \beta \geq 0, \tag{VC}
\]
where \( F - f_0 := \{ f - f_0 : f \in \mathcal{F} \} \), \( F'(x) := \sup_{g \in \mathcal{F}'} |g(x)| \), the supremum in \( Q \) is taken over all finitely supported discrete measures on \( \chi \), and \( \| \cdot \|_2, Q \) denotes the \( L_2 \)-norm with respect to the measure \( Q \). If \( \mathcal{F} \) satisfies (VC), then \( \mathcal{F} - f_0 \) is said to be a uniform VC-type class.

The third entropy considered in the paper is the bracketing entropy. In contrast to covering numbers, the bracketing number \( N_{||}((\zeta, \mathcal{F}, d) \) is the smallest \( m \geq 1 \) such that there exist pairs of functions \( (g_{i}^U, g_{i}^L), \ldots, (g_{m}^U, g_{m}^L) \) that satisfy \( d(g_{i}^U, g_{i}^L) \leq \zeta \) for all \( j \leq m \) and for any \( f \in \mathcal{F} \) there exists a \( j \leq m \) such that \( g_{i}^U(x) \leq f(x) \leq g_{i}^L(x) \) for every \( x \in \chi \). In Section 2, we study the LSE when \( \mathcal{F} \) satisfies
\[
\log N_{||}((\zeta, \mathcal{F}, \| \cdot \|_2)) \leq A\zeta^{-\alpha}, \quad \text{for some } A > 0 \text{ and } \alpha \in [0, 2), \tag{L_2}
\]
where \( \| \cdot \|_2 \) is the \( L_2 \)-norm with respect to \( P_X \).

In (L_\infty), (L_2), and (VC), \( \alpha \) is known as the complexity parameter. For “simple” classes of functions, \( \alpha \) is small, while a larger \( \alpha \) corresponds to more “complex” \( \mathcal{F} \). For example, when \( \mathcal{F} \) is the class of real valued \( \gamma \)-Hölder functions on \( [0, 1]^d \) then \( \alpha = d/\gamma \); see \cite[Page 350]{21}. Note that for Hölder classes, larger \( \gamma \) (more smoothness) are “simpler” classes, thus \( \alpha \) is inversely proportional to \( \gamma \). See Table 3 for more examples.

1.3. Local structure of \( \mathcal{F} \)

The expression (4) can be rewritten as
\[
\mathbb{E} \left[ \sup_{g \in \mathcal{F}_n} \mathbb{E}_{n} \left[ |(\epsilon + \Phi)g(X)| \right] \right],
\]
where for any $\delta > 0$, $\mathcal{F}_\delta := \{ f - f_0 : f \in \mathcal{F} \text{ and } \| f - f_0 \| \leq \delta \}$. Because the supremum is over functions in $\mathcal{F}_\delta$, if the local ball in $\mathcal{F}$ (centered at $f_0$) is nicely behaved, then bounds on expectations that take into account the local structure will lead to sharper rate bounds. We account for the local structure via the following envelope function

$$F_\delta(x) := \sup_{f \in \mathcal{F}} \| (f - f_0)(x) \|.$$

We call $F_\delta$ the local envelope at $f_0$. Note that $F_\delta$ can depend on $f_0$, however, in many scenarios the standard upper bounds for $F_\delta$ do not depend on $f_0$. Because of this, and notational convenience, we have suppressed the dependence of $F_\delta$ on $f_0$ in our notation. The local envelope gives us an insight into the worst case behavior of functions in a $\delta$-neighborhood of $f_0$ in $\mathcal{F}$. If sup$_{f \in \mathcal{F}} \| f \|_\infty \leq \Phi$, it is clear that $\| F_\delta \|_\infty \leq 2\Phi$, however, if the functions in $\mathcal{F}$ are smooth (e.g., uniformly Lipschitz) then $2\Phi$ is a conservative bound. In fact, if $\chi$ is a bounded interval in $\mathbb{R}$ and the functions in $\mathcal{F}$ are uniformly Lipschitz with Lipschitz constant $L$ then $\| F_\delta \|_\infty \leq 2L^{1/3}\delta^{2/3}$; see Lemma 2 of [13] for a proof of this. In Figure 1 below, we plot $\mathcal{F}_\delta$ (left panel) and $F_\delta$ (right panel) when $\mathcal{F}$ is the class of 1-Lipschitz functions and $f_0(x) = x$.

![Illustration of $\mathcal{F}_\delta$ (left panel) and $F_\delta$ (right panel) when $\mathcal{F} := \{ f : [0,1] \to \mathbb{R} \mid \| f(x) - f(y) \| \leq |x - y| \}$ and $f_0(x) = x$ for $\delta = .2$ (solid gray) and $\delta = .05$ (solid black). Any 1-Lipschitz function $f$ that satisfies $\| f - f_0 \| \leq .2$ lies in the “band” created by the two solid gray lines. Here $s = 2/3$. The dashed line in the left panel is $f_0$.](image)

In general, if sup$_{f \in \mathcal{F}} \| f \|_\infty \leq \Phi$, then one can invoke the rich theory of interpolation inequalities [1, 30, 40] to show that

$$\| F_\delta \|_\infty \leq C\Phi^{1-s}\delta^s \quad \text{for some } 0 \leq s \leq 1,$$

where $C$ denotes a constant (see footnote 1 in page 2). Note that for a uniformly bounded class of functions, $s$ can only vary between 0 and 1. If $F_\delta$ does not shrink with $\delta$ (with respect to $\| \cdot \|_\infty$-norm) then $s \approx 0$. For a class of non-smooth functions $s$ will be small, and for the class of infinitely differentiable functions $s = 1$. Intuitively, smaller values of $s$ correspond to more “complex” models; see Table 1 for a few examples.

Note that the entropy conditions ($\mathcal{L}_2$, $\mathcal{L}_\infty$, or (VC)) and (5) complement each other in the sense that the entropy conditions give control over the “global” behavior of $\mathcal{F}$ and (5) provides control over the “local” behavior of $\mathcal{F}$, i.e., the behavior of $F_\delta$. It should be noted that [2, 13, 44] have implicitly used the property (5) when studying the LSE for certain specific examples. However, their results do not lead to a general relationship between $s$ and the rate of convergence of the LSE. Han and Wellner [25] use this notion of smoothness for finding the rate of convergence when $\epsilon$ is independent of $X$ and $\mathcal{F}$ satisfies (VC).

2This is shown in Lemma 2 of [13].

3See this [54, Page 19] for a definition.
1.4. Our contributions

When \( F \) satisfies \((L_2)\) or \((L_\infty)\) with complexity parameter \( \alpha \) and \( \epsilon \) is uniformly\(^4\) sub-Gaussian, then the LSE is known to be minimax rate optimal and it converges at an \( n^{1/(2+\alpha)} \) rate [54, Chapter 9]. In this paper, we show that for a wide variety of examples, the uniform sub-Gaussianity of \( \epsilon \) is not necessary for the \( n^{1/(2+\alpha)} \) rate of the LSE. In Sections 2 and 3, we relate the rate of convergence of the LSE to the behavior of \( F_\delta \) and the moment assumptions on \( \epsilon \) when \( F \) satisfies the global conditions \((L_2)\) and \((L_\infty)\), respectively. We will now briefly describe our results in Sections 2 and 3.

In Section 2, we consider classes of functions that satisfy \((L_2)\). We show that if \( F_\delta \) satisfies a \( L_q \) version of (5) (see (6)) with smoothness parameter \( s \), then the LSE converges at an \( n^{1/(2+\alpha)} \) rate if \( \epsilon \) has at least \( 2/s \) moments. In Section 2.1, we apply Theorem 2.1 to show that the convex LSE converges at the minimax rate of \( n^{2/5} \) if \( \mathbb{E}[|\epsilon|^3] \leq C < \infty \) a.e. \( P_X \) and \( f_0 \) is bounded. Previously, minimaxity of the convex LSE was known only under uniformly sub-Gaussian errors.

In Section 3, we show that if \( F \) satisfies \((L_\infty)\), then the LSE converges at an \( n^{1/(2+\alpha)} \) rate if \( \epsilon \) has at least \( 1+2/\alpha \) moments. However, only \( (2+\alpha(1-s))/(s+\alpha(1-s)) \) many moments for \( \epsilon \) are enough if \( F \) satisfies (5) with smoothness parameter \( s \) (Theorem 3.1). This is useful since classes with low \( \alpha \) (complexity) often\(^5\) have high \( s \) (local smoothness). In such scenarios \( (2+\alpha(1-s))/(s+\alpha(1-s)) \) will be significantly smaller than \( 1+2/\alpha \). In Section 3.1, we apply Theorem 3.1 to find moment conditions on \( \epsilon \) under which the LSE is minimax rate optimal for general \( d \)-dimensional H"older regression and its lower dimensional submodels.

In Section 4, we consider classes of functions that satisfy \((VC)\). We show that the LSE converges at a rate of \( n^{1/(2(2-s))} \) for any \( \alpha < 2 \) when \( \epsilon \) has just two moments. In Theorem 4.1, we also find the rate of convergence of the LSE when \( F \) is non-Donsker \((\alpha \geq 2)\) and \( \epsilon \) has only 2 moments. The results in this section are especially useful in proving adaptive properties for shape constrained LSEs; see Section 4.1 and Remark 4.3. Our main results in Sections 2–4 are summarized in Table 2.

\[ \text{Table 2} \]

| Entropy | Smoothness assumption | Moment assumption | Rate of convergence |
|---------|-----------------------|-------------------|--------------------|
| \((L_2)\) | \(|(\epsilon) + \Phi)|F_\delta(X)|_q \leq C\Phi^2\delta^s \) | \( q \geq 2/s \) | \( n^{1/(2+\alpha)} \) |
| \((L_\infty)\) | \(|F_\delta|_\infty \leq C\Phi^{1-s}\delta^s \) | \( q \geq \frac{2+\alpha(1-s)}{s+\alpha(1-s)} \) | \( n^{1/(2+\alpha)} \) |
| \((VC)\) | \(|F_\delta|_\infty \leq C\Phi^{1-s}\delta^s \) | \( q \geq 2 \) | \( n^{1/(2(2-s))} \) |

The first step in proving the results discussed above is to find a tight upper bound on (4). Such bounds are known as maximal inequalities. Then one uses the maximal inequality in conjunction

---

\(^4\) We say \( \epsilon \) is uniformly sub-Gaussian if the tail probability of \( \epsilon|X \) is uniformly (in \( X \in \chi \)) bounded by a constant multiple of the tail probability of a Gaussian random variable.

\(^5\) A simple counter example is the class of indicators of closed intervals on \([0,1]\), i.e., \( \{1_{[a,b]}(\cdot) : 0 \leq a \leq b \leq 1\} \).
with a peeling argument (see e.g., van der Vaart and Wellner [55, Theorem 3.2.5]) to bound the tail probability of the LSE. However, most existing maximal inequalities require \(|e(f - f_0)(x)|\) to be bounded or to have exponential tails. This is not true when \(\epsilon\) has heavy tails. The main innovation in the paper is a new peeling argument in Theorem C.1. The peeling argument uses a truncation device to split the bound on the tail probability into two parts. We use new (Proposition B.1) and existing ([55, Lemma 3.4.2] and Lemma F.1) maximal inequalities to bound the maximum of the truncated empirical process and the Markov inequality to control the unbounded remainder. Then we optimize over the truncation scale to find the rate of convergence; see the three steps in the proof of Theorem C.1. This new argument and some modifications of the existing and new maximal inequalities are at the core of our new results.

Our results in Sections 2–4 should be thought of as the “worst-case” rates. The results are trying to find the best rates for the LSE under only the entropy and smoothness conditions. There are, of course, specific examples of \(F\) (and \(f_0\)), when one can use the geometry/structure of the function class to show that the LSE converges at a rate faster than \(n^{1/(2 + \alpha)}\) or that the LSE converges at an \(n^{1/(2 + \alpha)}\) rate under weaker moment assumptions.

The paper is organized as follows. In Sections 2, 3, and 4, we find the rate of convergence of the LSE when \(F\) satisfies \((L_2), (L_\infty),\) and \((VC),\) respectively. Each section ends with an example, and in each of these examples we show (for the first time) that uniformly sub-Gaussian errors are not needed for the LSE to be minimax rate optimal. In Section 5, we briefly comment on the rate of the LSE under misspecification. In Section 6, we summarize the contributions of the paper and briefly discuss some future research directions. Appendix A proves three new interpolation inequalities used in our examples. Appendix B provides a new maximal inequality for maximums over finite sets and discusses an application that is of independent interest. Appendix C contains our main peeling result. The rest of the appendix provides proofs of the results in Sections 2–4.

2. Rates of convergence of the LSE using bracketing \(L_2(P_X)\)-entropy

Assumption \((L_2)\) is the most widely used entropy condition when establishing the rate of convergence of the LSE [21, 26, 55]. The following theorem (proved in Appendix D) finds an upper bound on the rate of convergence of the LSE when \(\epsilon\) is heavy-tailed and heteroscedastic.

**Theorem 2.1.** Suppose \(F\) satisfies \((L_2), \epsilon\) satisfies \((C\text{Var}),\) and \(f_0 \in F\). Let \(\Phi := \sup_{f \in F} \|f\|_\infty\). If there exists a constant \(C > 0\) such that

\[
\|(|\epsilon| + \Phi)F_\delta(X)\|_q \leq C\Phi^2\delta^s, \tag{6}
\]

for some \(s \in [0, 1]\). Then for any \(n \geq 1\) and \(D > 0\), we have

\[
P\left(r_n \|\hat{f} - f_0\| \geq D\right) \leq CD^{-q+1(s=1)/10}, \tag{7}
\]

where \(C\) depends only on \(\alpha, s, q, \) and \(\Phi\), and

\[
r_n := \min \left\{\left(\frac{n \Lambda^{-1/(2 + \alpha)}}{\sigma + \Phi^{2/(2 + \alpha)}}, \frac{n^{(q-1)/(q(2-s))}}{\Phi^{2/(2-s)}}, \frac{n^{1/(2 + \alpha + (2 - q)s)/(q - 1)}}{\Phi^{(2-q)+1(2 + \alpha)(q - 1)}}\right)\right\}. \tag{8}
\]

**Remark 2.1** (Assumptions in Theorem 2.1). We make a few observations on the assumptions of Theorem 2.1.

1. The covariate space \(X\) is not restricted to be Euclidean. The only assumption on \(X\) is that it be a metric space. This comment applies to all the results of the paper.
2. Observe that \((E_q)\) and \((5)\) together imply \((6)\), i.e., if \(\|F_\delta\|_\infty \leq C\Phi\delta^s\) and \(\mathbb{E}(|\epsilon|^q) \leq K_q^q\) then

\[
\|(|\epsilon| + \Phi)F_\delta(X)\|_q \leq (K_q + \Phi)C\Phi^{1-s}\delta^s \leq C\Phi^2\delta^s, \quad \text{where} \quad C := 1 + K_q/\Phi.
\]
3. There are cases when \( \|F_\epsilon\|_\infty \approx 1 \) but (CVar) and additional moment assumptions on \( \epsilon \) will imply that \( F_\epsilon \) satisfies (6); see Section 2.1 for an example of this.

4. The uniform boundedness assumption on \( F \) can be easily relaxed to only \( \|\hat{f}\|_\infty = O_p(1) \); see Section 2.1 for a detailed argument. Also see [26] and [34] for further examples.

5. The bound on the local envelopes in (6) can be relaxed to accommodate extra log factors. For example, if \( \| (\epsilon + \Phi) F_\epsilon(X) \|_q \leq C \Phi^q \delta^q \log^q(1/\delta) \) then \( r_n \) will increase by additional \( \log n \) factors; where the power of \( \log n \) will depend on \( \gamma, \alpha \), and \( q \). This dependence is computed explicitly in (40) in Appendix D.1.

Remark 2.2 (Conclusions of Theorem 2.1). The tail bound in (7) is a finite sample result and holds for all \( n \geq 1 \). Because \( q \geq 2 \), the tail probability bound in (7) implies that if \( s < 1 \), then there exists a constant \( C \) such that \( \mathbb{E}(r_n|\hat{f} - f_0|) \leq C \) for all \( n \geq 1 \). When \( s = 1 \), (7) implies that the tail probability decays at a polynomial rate with a coefficient of \( -q + 1/10 \). The \( 1/10 \) in the power is meant to represent a small constant. In fact when \( s = 1 \), we show that \( \mathbb{P}(r_n|\hat{f} - f_0| \geq D) \leq CD^{-n} \), for any \( n < q \); see (39) in Appendix D for a proof of this. Here the constant \( C \) depends on \( q, \alpha \), and \( \eta \) only.

Because all the assumptions and results are finite sample, \( F, A \), and \( \Phi \) are allowed to depend on \( n \). However, in applications, it is often the case that \( F, A \), and \( \Phi \) do not change with \( n \). In that case the dependence of \( r_n \) on \( A \) and \( \Phi \) in (8) can be ignored. Furthermore, if \( \Phi \ll \infty \) and \( \epsilon \) satisfies (E_q), then every uniformly bounded function class \( F \) satisfies (6) with \( s = 0 \). The following corollary finds the rate of the LSE if \( F \) does not satisfy the local smoothness/structural assumption of Section 1.3 (i.e., \( s = 0 \)) and \( A \) and \( \Phi \) are constants.

Corollary 2.1. Suppose \( f_0 \in F \) and \( \Phi := \sup_{f \in F} \| f \|_\infty \) is a constant. If there exist constants \( A \) and \( K_q \) such that \( \epsilon \) satisfies (CVar) and (E_q) and \( F \) satisfies (L_2), then for any \( n \geq 1 \) and \( D > 0 \), we have

\[
\mathbb{P}\left( n^{1/(\alpha + 2q/(q-1))}|\hat{f} - f_0| \geq D \right) \leq CD^{-q},
\]

where \( C \) is a constant depending only on \( A, \Phi, K_q, q, \sigma, \alpha \).

The above result is a direct application of Theorem 2.1 with \( s = 0 \). If \( \epsilon \) and \( X \) are further assumed to be independent then Theorem 3 of [26] shows that the LSE converges at a rate of \( n^{1/(2+\alpha)} \) when \( q \geq 1 + 2/\alpha \). Corollary 2.1 allows for heteroscedastic and heavy-tailed errors, but this relaxation comes at a cost. The rate of convergence obtained in (9) is strictly slower than the minimax rate for this setup\(^6\). A similar sub-\( n^{1/(2+\alpha)} \) rate was found Section 3.4.3.1 of [55] in the case of fixed design regression with heavy-tailed errors. There are two possible explanations for the rate bound in (9): (1) the LSE is not minimax rate optimal under the assumptions of Corollary 2.1 and the worst case\(^7\) rate is \( n^{1/(\alpha + 2q/(q-1))} \) \((< n^{1/(2+\alpha)})\); or (2) the LSE actually converges at an \( n^{1/(2+\alpha)} \) rate and the obtained rate is an artifact of the proof. The optimality of Corollary 2.1 is still an open problem.

Remark 2.3. Han and Wellner [26, Proposition 3 and Remark 10] argue that under (L_2), the rate of convergence of the LSE can be arbitrarily slow when \( \epsilon \) is heteroscedastic and has heavy tails. On surface, this might seem to be at odds with Corollary 2.1, but in their examples, both \( F \) and \( \mathbb{E}(\epsilon^2|X) \) are unbounded. This is important because the (essential) boundedness of \( \mathbb{E}(\epsilon^2|X) \) (CVar) is a crucial assumption in all our results. We use condition (CVar) to provide bracketing entropy bounds for \( \{\epsilon(f - f_0): f \in F\} \) based on the bracketing entropy bounds for \( \{f - f_0: f \in F\} \). Note that if \( [\ell, u] \) is the bracket for \( f - f_0 \), i.e., \( \ell \leq f - f_0 \leq u \) then

\[
\epsilon_+ - u - \epsilon \leq \epsilon(f - f_0) \leq \epsilon_+ u - \epsilon - \ell,
\]

\(^6\)There exist robust estimators such as the least absolute deviation estimator [42] or [55, Page 336] and median-of-means estimators [35] that converge at a \( n^{1/(2+\alpha)} \) rate under the assumptions of Corollary 2.1.

\(^7\)By the worst case rate we mean is there exists some dependence structure between \( \epsilon \) and \( X \) and choice of \( F \) such that rate in (9) is tight.
where \( \epsilon_+ \) and \( \epsilon_- \) are the positive and negative parts of \( \epsilon \), respectively. The width of this bracket is \(|\epsilon(u - \ell)|\). Under assumption (CVar), we have \( \|\epsilon(u - \ell)\| \leq \sigma\|u - \ell\| \). Therefore under (CVar), \( N_1(q, \{\epsilon(f - f_0) : f \in \mathcal{F}, \|\cdot\|\}) \leq N_1(q/\sigma, \{f - f_0 : f \in \mathcal{F}, \|\cdot\|\}) \). This crucial conclusion might not hold if (CVar) is not satisfied.

The rates of convergence in Corollary 2.1 do not take into account any structure of \( \mathcal{F} \) other than the “size” of the function class. Theorem 2.1 improves upon Corollary 2.1 by using the local smoothness/structure of \( \mathcal{F} \) (around \( f_0 \)); see [2, 13, 25, 44] for results that use a similar smoothness property implicitly or explicitly. Theorem 2.1 shows that the LSE will converge at an \( n^{1/(2 + \alpha)} \) rate if \( \epsilon \) has enough moments. To better understand the rate in Theorem 2.1, let us assume that both \( A \) and \( \Phi \) are constants (do not change with \( n \)). In this case, \( r_n \) can be simplified to

\[
    r_n = \min\left\{ n^{1/(2 + \alpha)}, n^{(q-1)/(q(2-s))}, n^{(q-1)/(q(2-s)+\alpha(q-1))} \right\}.
\]

Furthermore, observe that

\[
    \frac{1}{2 + \alpha} - \frac{q-1}{q(2-s)+\alpha(q-1)} \leq \frac{q-1}{q(2-s)} \iff q \geq \frac{2}{s}.
\]

Thus

\[
    r_n = \min\left\{ n^{1/(2 + \alpha)}, n^{(q-1)/(q(2-s)+\alpha(q-1))} \right\}
\]

for all \( q \geq 2 \) and if \( q \geq 2/s \) then \( r_n = n^{1/(2 + \alpha)} \). The above calculations suggest an interesting interplay between \( \alpha, q, \) and \( s \). They show that if the \( \mathbb{E}(|\epsilon|^{2/s}|X) \leq C < \infty \) and \( \|F_\delta\|_{2/s} \leq C\delta^s \), then the rate of convergence of the LSE under the heavy-tailed heteroscedastic errors is \( n^{1/(2 + \alpha)} \) and this rate coincides with the rate under sub-Gaussian errors. This justifies the usage of least squares estimators under heavy-tailed errors in a wide variety of examples. However, if \( \epsilon \) has less than \( 2/s \) moments then Theorem 2.1 suggests that there are “hard” settings where the “noise” is too strong and the guaranteed rate of convergence for the LSE is slower than the minimax optimal rate of \( n^{1/(2 + \alpha)} \).

Table 3 shows some interesting applications of Theorem 2.1 and compares the results with Theorem 3 of [26]. Both of these theorems consider \( \mathcal{F} \) that satisfies \( (L_2) \). However, Theorem 2.1 uses the local structure/smoothness of the function when deriving the rates, while [26] does not assume any structure in \( \mathcal{F} \). Table 3 shows that when \( \mathcal{F} \) is class of Hölder or Sobolev functions, then the inherent smoothness of the functions involved can help significantly reduce the requirement on \( \epsilon \) for the optimal \( n^{1/(2 + \alpha)} \) rate of convergence when \( \alpha < 1 \). To see this, observe that for Hölder classes \( s = 2/(2 + \alpha) \). Thus when \( \alpha < 1 \), we have \( 2/s < 1 + 2/\alpha \), i.e., the moment requirements for Theorem 2.1 is smaller than that in [26, Theorem 3]. This is significant, as in contrast to the results of [26], Theorem 2.1 allows for heteroscedastic errors.

The proof of Theorem 2.1 (in Appendix D) is an application of our new refined peeling result, Theorem 2.1, in conjunction with a classical maximal inequality [55, Lemma 3.4.2] for bounded empirical processes. The classical inequality in [55, Lemma 3.4.2] applies only to bounded empirical process and cannot be used to control the unbounded empirical process in (4). In contrast to the standard peeling argument [55, Theorem 3.2.5], Theorem 2.1 incorporates a truncation step directly into the peeling argument (see Step 1 in the proof of Theorem 2.1) and thus allowing us to use the classical maximal inequality in this setting. To control the unbounded remainder, we observe that it has \( q \) moments and use a Markov inequality of \( q \)th order. The above two steps will show that \( \mathbb{P}(r_n \|f - f_0\| \geq D) = O(D^{-1}) \). To show that the probability of the tail is in fact of a much smaller order, we use Talagrand’s inequality [20, Proposition 3.1].

---

8These moment requirements are under stronger assumptions. Han and Wellner [26] assume independence between \( \epsilon \) and \( X \).

9Here \( s = 0 \), thus the upper bound on the rate of convergence of the LSE is \( n^{1/(2 + \alpha(q-1)/q)} \).
Different choices of $\mathcal{F}$, their corresponding values of $\alpha$ and $s$, and the number of moments of $\epsilon$ required for the LSE to converge at an $n^{1/(2+\alpha)}$ rate.

| $\chi$          | $\mathcal{F}$         | $\alpha$ | $s$       | Moments needed for an $n^{1/(2+\alpha)}$ rate |
|-----------------|------------------------|----------|-----------|---------------------------------------------|
| $[0, 1]^d$      | $\gamma$-Hölder class  | $d/\gamma$ | $2^{2\gamma}/(2\gamma+2)$ | $[26, \text{Theorem 3}]^{8}$ Theorem 2.1 |
| $[0, 1]^d$      | $\gamma$-Sobolev class | $d/\gamma$ | $2^{2\gamma-1}/(2\gamma+1)$ | $2 + 2^{d/(2\gamma-1)}$ |
| $[0, 1]$        | Uniformly Lipschitz    | $0$      | $2/3$     | $3$ |
| $[0, 1]$        | $\gamma$-Hölder class $\cup \{1_{[a, b]}: 0 \leq a \leq b \leq 1\}$ | $1/\gamma$ | $0$       | $1 + 2^{\gamma}$ | $\infty^9$ |

Remark 2.4. In each of the examples, the value of the complexity parameter is well known. For Hölder, Sobolev, and Lipschitz functions we use standard interpolation inequalities to find $s$; see e.g., [1, 13, 40, 44, 54, 55], also see Appendix A for newly derived interpolation inequalities. These works also contain other examples for which $\mathcal{F}$ satisfies the assumptions of Theorem 2.1.

2.1. Example 1: Univariate convex regression

We now find the rate of convergence of the convex LSE under heteroscedastic and heavy-tailed errors. Let $\mathcal{F}$ be the class of convex functions on $[0, 1]$ and $P_X$ be the uniform distribution on $[0, 1]$. Recall that $\hat{f} := \arg \min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$. Observe that $\hat{f}$ is only well-defined at the data points $\{X_i\}_{i=1}^{n}$. In this paper, we consider the canonical extension of $\hat{f}$, and define $\tilde{f}$ to be the unique left-continuous piecewise linear function on $[0, 1]$ with potential kinks at the data points. We are interested in finding the rate of convergence of $\tilde{f}$ when $f_0 \in \mathcal{F}$. The class of convex functions in $[0, 1]$ is unbounded. However, a simple modification of [25, Lemma 5] shows that $\|\hat{f}\|_{\infty} = O_p(1)$ if $\epsilon$ satisfies (CVar). Let $\mathcal{F}_n := \{f \in \mathcal{F} : \|f\|_{\infty} \leq C\sqrt{\log n}\}$, where $C$ is a constant. Because $\|\hat{f}\|_{\infty} = O_p(1)$, we have that $P(\hat{f} \notin \mathcal{F}_n) = o(1)$. Now define $\tilde{f} := \arg \min_{f \in \mathcal{F}_n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$, then $P(\tilde{f} = \hat{f}) = 1 - o(1)$. Thus the rate of convergence of $\tilde{f}$ coincides with the rate of convergence of $\hat{f}$, as for every $D > 0$

$$P(\|\tilde{f} - f_0\| > D) \leq P(\|\hat{f} - f_0\| > D) + o(1),$$

where the $o(1)$ term does not depend on $D$. If $\epsilon$ is uniformly sub-Gaussian or bounded then classical results (see [55, Section 3.4.3.2]) show that $\hat{f}$ converges at an $n^{2/5}$-rate up to a log $n$ factor. In this example, we will show that the light tail assumption is unnecessary and that Theorem 2.1 implies that $\tilde{f}$ converges at an $n^{2/5}$-rate (up to a polynomial in log $n$ factors) if $\epsilon$ satisfies (CVar) and $E(|\epsilon|^{3}|X)$ is uniformly bounded.

Theorem 3.1 of [16] shows that $\mathcal{F}_n$ satisfies $(L_2)$ with $A = C(\log n)^{1/4}$ and $\alpha = 1/2$. Further, if $\mathcal{F}_{n, \delta} := \{f - f_0 : f \in \mathcal{F}_n, \|f - f_0\| \leq \delta\}$ and $F_{n, \delta}(\cdot) := \sup_{g \in \mathcal{F}_{n, \delta}} |g(\cdot)|$, in Proposition A.1, we show that

$$\|F_{n, \delta}\|_{\infty} = C\sqrt{\log n} \quad \text{and} \quad \|F_{n, \delta}\|_{3} \leq 4\delta^{2/3} (\log(1/\delta))^{1/3} \sqrt{\log n};$$

Han and Wellner [25] assume that $\epsilon$ is independent of $X$. However, their proof ([25, Section 5.3.1]) goes through if we use the Etemadi’s maximal inequality [15, Proposition 1.1.2] and the fact that $\epsilon_i$’s satisfy (CVar) instead of Lévy’s inequality for sums of i.i.d random variables [15, Theorem 1.1.5].
see Fig. 2 for a plot of the local neighborhood and the local envelope. Suppose there exists a constant $C$ such that $\mathbb{E}(|\epsilon|^3|X) \leq C$ for a.e. $P_X$. Because $\Phi = C\sqrt{\log n}$, we have

$$\|(|\epsilon| + \Phi)F_{n,\delta}(X)\|_3 \leq C\delta^{2/3}\log(1/\delta)^{1/3}\log n.$$  

Thus $\hat{f}$, $\epsilon$, and $F_n$ satisfy the assumptions of Theorem 2.1 and item 5 of Remark 2.1 (also see Appendix D.1) with $\alpha = 1/2$, $s = 2/3$, $\nu = 1/3$, $\Phi = C\sqrt{\log n}$, and $A = (\log n)^{1/4}$. Hence by (40), a modification of (8), we have that $\hat{f}$ converges at an $n^{2/5}\log n$ rate when $\epsilon$ (essentially) has three moments, which in turn implies that $\|\hat{f} - f_0\|$ converges at the minimax optimal rate (up to a log $n$ factor) if $\mathbb{E}(|\epsilon|^3|X) \leq C$. This result seems to be new.

![Fig 2. Illustration of $F_{\beta}$ (left panel) and $F_{\gamma}$ (right panel) when $f_0(x) = x^2$ and $F := \{f : [0,1] \rightarrow \mathbb{R} \ | \|f\|_{\infty} \leq 2 \text{ and } f \text{ is convex}\}$ for $\beta = 0.5$ (solid black) and $\beta = 0.25$ (solid gray). Any convex function $f$ that is uniformly bounded by 2 and satisfies $\|f - f_0\| \leq 2$ lies in the band created by the solid gray lines. The dashed line in the left panel is $f_0$.](image)

### 3. Rates of convergence of the LSE using the $L_{\infty}$-entropy

Bracketing entropy condition ($L_2$) is the most widely used notion of complexity in the study of risk bounds for the LSE. However, often the function classes also satisfy the stronger entropy condition ($L_\infty$), especially when $\chi$ is bounded.\(^{11}\) Moreover, they often satisfy both ($L_2$) and ($L_\infty$) for the same value of the complexity parameter; e.g., the class of Hölder or Sobolev functions on $[0, 1]^d$ satisfy both ($L_2$) and ($L_\infty$) with the same complexity parameter. The following result (proved in Appendix E) shows that the rate of convergence of the LSE in Theorem 2.1 can be improved if $\mathcal{F}$ satisfies ($L_\infty$).

**Theorem 3.1.** Suppose $\mathcal{F}$ satisfies ($L_\infty$), $\epsilon$ satisfies (CVar) and ($\mathcal{E}_q$), and $f_0 \in \mathcal{F}$. Let $\Phi := \sup_{f \in \mathcal{F}} \|f\|_{\infty}$. If there exists a constant $C > 0$ such that

$$\|F_\delta\|_{\infty} \leq C\Phi^{1-s}\delta^s,$$  

for some $s$ in $[0, 1]$. Then for any $n \geq 1$ and $D > 0$, we have

$$\mathbb{P}\left(r_n\|\hat{f} - f_0\| \geq D\right) \leq CD^{-q+1(s=1)/10},$$  

where the constant $C > 0$ depends only on $q$, $s$, and $\alpha$, and

$$r_n := \min \left\{\left(\frac{nA^{-1}(2s+\alpha)}{\sigma + \Phi}(q-1)/(q(2-s))\right)\Phi^{2/(2-s)}, \left(\frac{nA^{-1}(2s+\alpha)}{\sigma + \Phi}(q-1)/(q(2-s))\right)\Phi^{(q(2-s)+\alpha s(q-1))/(q(2-s)+\alpha s(q-1))}\right\}.\tag{12}$$  

\(^{11}\)A counter example is the class univariate convex functions on $[0, 1]$. They satisfy ($L_2$) with $\alpha = 1/2$ but do not satisfy ($L_\infty$).
Assumption (10) on \( F_\delta \) is stronger than (6). Just as in Theorem 2.1, the tail bound in (11) is finite and holds for all \( n \geq 1 \) and the discussion in Remark 2.2 applies to (11) as well. The following corollary finds the rate of the LSE if \( \mathcal{F} \) does not satisfy any local smoothness/structural assumption of Section 1.3 (i.e., \( s = 0 \)) and \( A \) and \( \Phi \) are constants; cf. Corollary 2.1.

**Corollary 3.1.** Suppose \( f_0 \in \mathcal{F} \) and \( \Phi := \sup_{f \in \mathcal{F}} \|f\|_\infty \) is a constant. If there exist constants \( A \) and \( K_q \) such that \( \epsilon \) satisfies (CVar) and (\( \mathcal{E}_q \)) and \( \mathcal{F} \) satisfies (\( L_\infty \)), then for any \( n \geq 1 \) and \( D > 0 \), we have that \( \mathbb{P}(r_n \| \hat{f} - f_0 \| \geq D) \leq C D^{-q} \), where \( C \) is a constant depending only on \( q, \alpha, \sigma, A, \Phi, \) and \( K_q \) and \( r_n := \min \left\{ n^{1/(2+\alpha)}, n^{1/2-1/2q} \right\} \). \( \tag{13} \)

To prove Corollary 3.1, apply Theorem 3.1 with \( F_\delta \equiv 2 \Phi \) \((s = 0)\). If \( \mathcal{F} \) is such that \( s = 0 \) and satisfies (\( L_\infty \)) then Corollary 3.1 uses the stronger entropy condition to show that the LSE converges at an \( n^{1/(2+\alpha)} \) rate under heteroscedastic errors if \( q \geq 1 + 2/\alpha \); compare this to the rate of the LSE obtained in Corollary 2.1.\(^{12}\) It is well known, that the worst case rate for the LSE under only the entropy assumption (\( L_\infty \)) is \( n^{1/(2+\alpha)} \) when \( \epsilon \) is uniformly sub-Gaussian. Corollary 3.1 shows that the heavy-tailed (and heteroscedastic) nature of the \( \epsilon \) does not affect this rate as long as \( \epsilon \) has at least \( 1 + 2/\alpha \) moments and satisfies (CVar).

Theorem 3.1 shows that the upper bounds on the rate of convergence of the LSE in (13) can be reduced if \( \mathcal{F} \) satisfies the smoothness assumptions in (10). If \( A \) and \( \Phi \) are constants and \( \mathcal{F} \) satisfies (10), then Theorem 3.1 implies that the LSE converges at the rate \( \min \left\{ n^{1/(2+\alpha)}, n^{(q-1)/(q(2-s)+\alpha s(q-1))} \right\} \). \( \tag{14} \)

This implies if \( q \geq (2 + \alpha(1 - s))/(s + \alpha(1 - s)) \), then the LSE converges at an \( n^{1/(2+\alpha)} \) rate. Furthermore \( 2 + \alpha(1 - s)/s + \alpha(1 - s) \leq 1 + \frac{2}{\alpha} \) for all \( s \leq 1 \) and \( 0 \leq \alpha < 2 \). Thus if \( s > 0 \), then Theorem 3.1 shows that the LSE converges at an \( n^{1/(2+\alpha)} \) rate under weaker assumptions on \( \epsilon \) than in Corollary 3.1.

The proofs of Theorems 2.1 and 3.1 are similar. But in the case of Theorem 2.1, we could apply the readily available maximal inequality \([55, \text{Lemma 3.4.2}]\). Existing maximal inequalities, however, cannot take the \( L_\infty \)-covering number into account. For this purpose, we use generic chaining \([48]\) in conjunction with a new maximal inequality for the maximum over a finite set; see Proposition B.1. Proposition B.1 is also of independent interest and compares favorably to Lemma 8 of \([14]\). Our result shows that the maximum of \( N \) centered averages converges at the rate of \( \sqrt{n^{-1} \log N} \), if \( \log N = o(n) \) and the envelope has finite \( q \geq 2 \) moments; see (\( 24 \)). On the other hand, \([14, \text{Lemma 8}]\) requires \( \log N = O(n^{1-2/q}) \) for a \( \sqrt{n^{-1} \log N} \) rate of convergence; see Remark B.1 in Appendix B for more details.

### 3.1. Example 2: Multivariate and multiple index smooth regression models

In this section, we consider the example of multivariate regression when the unknown function is known to be smooth. Let \( \chi := [0, 1]^d \) and \( P_X \) be the uniform distribution on \( \chi \). For any vector \( k = (k_1, \ldots, k_d) \) of \( d \) positive integers, define the differential operator \( D^k := \frac{\partial^{\mid k\mid} f}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}} \), where

\[^{12}\] Note that \( \log N_\delta(\zeta, \mathcal{F}, \| \cdot \|) \leq \log N(\zeta, \mathcal{F}, \| \cdot \|_\infty) \). Thus, the assumptions of Corollary 2.1 are weaker than the assumptions of Corollary 3.1.

\[^{13}\] We can ignore the middle term in (12) because \( q(2 - s) \leq q(2 - s) + s(q - 1) \) for all \( q \geq 2 \).

\[^{14}\] This can be easily relaxed to assume that \( \chi \) is a bounded and convex subset of \( \mathbb{R}^d \).
implies that Lemma 2, we have that in (2) with then $M$.

Thus by Theorem 3.2, we have that the LSE is minimax rate optimal under heteroscedastic and heavy-tailed errors. Note that $C$.

Multivariate smooth regression Suppose $f_0 \in \mathcal{F}_{\gamma,d}$ for some $\gamma > d/2$ and $\hat{f}$ is defined as in (2) with $\mathcal{F} = \mathcal{F}_{\gamma,d}$. We will apply Theorem 3.1 to show that $\hat{f}$ is minimax rate optimal under heteroscedastic and heavy-tailed errors. Note that $\mathcal{F}_{\gamma,d}$ also satisfies the assumptions of Theorem 2.1 but its application would have led to the same rate of convergence but under a stronger moment condition $q \geq 2 + d/\gamma$; see Table 3.

Multiple index smooth regression In the above setup the rate of convergence of the LSE is strongly affected by the dimension (curse of dimensionality). When $d$ is large, a widely used semiparametric alternative that ameliorates the curse of dimensionality is the multiple index model [29, 34]. In multiple index model the true regression function is assumed to belong to $\mathcal{M}_{\gamma,d,d} := \{ x \mapsto f(Bx) : f \in \mathcal{F}_{\gamma,d} \text{ and } B \in \mathbb{R}^{d_1 \times d} \text{ satisfying } \|B\|_2 \leq 1 \}$, for some $d_1 \leq d$. Note that minimax optimal rate in the multiple index model is $n^{\gamma/(2\gamma+d_1)}$; Stone [46, Page 129], a much faster rate than that in (15). We now show that the LSE achieves the minimax rate.

Because $\|B\|_2 \leq 1$, it can be easily shown that there exists a constant $C$ (depending on $d$) such that

$$\log N(\nu, \mathcal{M}_{\gamma,d,d_1}, \| \cdot \|_\infty) \leq C \nu^{-d_1/\gamma} \text{ for all } \nu > 0.$$ 

Thus $\mathcal{M}_{\gamma,d,d_1}$ is much less “complex” than $\mathcal{F}_{\gamma,d}$, when $d_1$ is smaller than $d$. By Proposition A.3, we have that $\mathcal{M}_{\gamma,d,d_1}$ satisfies (10) with $s = 2\gamma/(2\gamma + d_1)$. Thus Theorem 3.1 shows that if

$$\hat{f} := \arg\min_{f \in \mathcal{M}_{\gamma,d,d_1}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

then

$$n^{\gamma/(2\gamma+d_1)} \| \hat{f} - f_0 \| = O_p(1), \quad \text{when } q \geq 2 + \frac{2d_1\gamma - d_1^2}{2\gamma^2 + d_1^2}.$$ 

Additive model regression An even simpler function class than $\mathcal{M}_{\gamma,d,d_1}$ is given by functions that are separable in their coordinates [8, 17]. Formally, define

$$\mathcal{A}_\gamma := \{ x \in \mathbb{R}^d \mapsto f(x) = \sum_{j=1}^{d} f_j(x_j) : f_j \in \mathcal{F}_{\gamma,1} \}.$$ 

In this case it can be shown that there exists a constant $C$ such that

$$\log N(\nu, \mathcal{A}_\gamma, \| \cdot \|_\infty) \leq C \nu^{-1/\gamma} \text{ for all } \nu > 0.$$
By Proposition A.2 it follows that $A_\gamma$ satisfies (10) with $s = 2\gamma/(2\gamma + 1)$. Thus Theorem 3.1 shows that if

$$\hat{f} := \arg\min_{f \in A_\gamma} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

then

$$n^{\gamma/(2\gamma+1)} \| \hat{f} - f_0 \| = O_p(1) \quad \text{when} \quad q \geq 2 + \frac{2\gamma - 1}{2\gamma^2 + 1}.$$

The function classes above can also be replaced by other smoothness classes such as Sobolev or Besov spaces [39]. Furthermore, using the proofs of Propositions A.2 and A.3, one can also consider combination of function spaces $M_{\gamma,d,d_t}$ and $A_\gamma$, wherein some coordinates are modeled through linear combinations and the remaining coordinates are modeled through additive model.

4. Rate of convergence for VC-type classes

In Sections 2 and 3, we saw that the local envelope $F_\delta$ affects the rate of convergence of the LSE. We saw that if $F$ satisfies $(L_2)$ or $(L_\infty)$ with $\alpha$ and $F_\delta$ is “small”, then the LSE converges at an $n^{1/(2+\alpha)}$ even when $\epsilon$ has only few moments. If $F$ is the class of smooth functions (e.g., Sobolev, Hölder, or Besov spaces), then $s$ depends only on the smoothness of the functions in the class and not on the choice of $f_0$ (recall that $F_\delta$ is the local neighborhood of $f_0$ in $F$). However, it turns out that for certain choices of $F$ (e.g., class of monotone or convex functions) the size of $F_\delta$ can depend on $f_0$. For example, in Proposition A.1, we showed that if $F$ is the class uniformly bounded convex functions on $[0,1]$, then $F_\delta(x) \leq C \Phi^{1/3} \delta^{2/3} \max\{x^{-1/3}, (1-x)^{-1/3}\}$ for any $f_0 \in F$. But if $f_0$ is a linear function (or piecewise linear) then [22, Lemma A.3] shows that $F_\delta(x) \leq C \delta \max\{x^{-1/2}, (1-x)^{-1/2}\}$ (note that $F_\delta$ is smaller when $f_0$ is linear). This change in local behavior of $F_\delta$ when $f_0$ belongs to a particular subclass of functions (e.g., piecewise constant functions when $F$ is the set of monotone functions or piecewise linear functions when $F$ is the set of convex functions) drives the adaptive behavior of the LSE in shape-constrained regression; see e.g., [5, 10, 12, 23] and references therein. Furthermore, in these examples it turns out that $F$ satisfies (VC), when $f_0$ belongs to these special subclasses of $F$. In the following theorem (proved in Appendix F) we find the worst-case rate of convergence of the LSE when $F$ and $f_0$ satisfy (VC) and $\epsilon$ satisfies (CVar). In this section, we do not make any assumptions on the higher order moments of $\epsilon$. This is done with the goal of keeping the result simple. Furthermore, it turns out that LSE is rate optimal in certain scenarios with just two finite moments.

**Theorem 4.1.** Suppose $F$ satisfies (VC), $\epsilon$ satisfies (CVar), and $f_0 \in F$. Let $\Phi := \sup_{f \in F} \| f \|_\infty$. Assume that $\sigma, \Phi$, and $A$ (in (VC)) are constants. If there exists a constant $C > 0$ such that

$$\| F_\delta \| \leq C \Phi^{1-s} \delta^s,$$

for some $s$ in $[0,1]$. Then for any $n \geq 1$ and $D > 0$, we have

$$P\left( r_n \| \hat{f} - f_0 \| \geq D \right) \leq CD^{-4(2-s)/3},$$

where the constant $C > 0$ depends only on $\sigma, \Phi, A$, and $\alpha$, and

$$r_n := \begin{cases} n^{1/(2(2-s))} & \text{if } \alpha \in [0,2) \text{ and } \beta \geq 0, \\
^{1/2} / \log n)^{1/(2-s)} & \text{if } \alpha = 2 \text{ and } \beta = 0, \\
n^{1/(\alpha(2-s))} & \text{if } \alpha > 2 \text{ and } \beta = 0. \end{cases}$$

**Remark 4.1.** We make some observations about the assumptions and conclusions of Theorem 4.1.
1. The assumption (16) has a different structure than those in Theorems 2.1 and 3.1. In contrast to assumption (10), we only require a bound on the $L_2$-norm of $F_0$. No control is required for higher moments of $\epsilon$.

2. The above theorem provides the rates of convergence of the LSE even when $\mathcal{F}$ is non-Donsker.

3. The assumption that $A$ and $\Phi$ do not change with $n$ is made to keep the presentation simple. In the proof of the result (Appendix F), we provide explicit finite sample tail bounds that allow $A$ and $\Phi$ to depend on $n$. See (54), (56), and (57) to find the exact relationship between $r_n$, $\Phi$, and $A$ for the three situations considered in (17).

4. If $s = 1$, then it is clear that the obtained rate of convergence of the LSE cannot be improved when $\alpha \in [0, 2)$ and $\beta \geq 0$.

5. If $s < 1$ and $\epsilon$ satisfies higher moment assumptions, then the rates obtained in (17) can be improved using the tools developed in this paper.

6. Just as in Theorems 2.1 and 3.1, we can relax the bound on $\|F_0\|$ in (16) to be of the form $C\Phi\delta^s(\log(1/\delta))^{\nu}$. This relaxation will increase $r_n$ in (17) by a polynomial in $\log n$ factor; the order of the polynomial will depend only on $s, \nu$, and $\alpha$. In particular, if $\alpha \in [0, 2)$ and $\beta \geq 0$, then $r_n = n^{1/(2s-1)}(\log n)^{\nu}$ if $\|F_0\| \leq C\Phi\delta^s(\log(1/\delta))^{\nu}$.

7. To prove Theorem 4.1, we use the refined Dudley’s chaining inequality (Lemma F.1) in conjunction with our refined peeling result in Appendix C.

Remark 4.2 (Comparison with Theorem 1 of [25]). Theorem 4.1 is an improvement over Theorem 2 of [25]. Our result allows $\epsilon$ to depend on $X$ and they are required to have only 2-moments, whereas [25, Theorem 1] requires the error to have $L_{2,1}$ moments and be independent of $X$. Our proof of Theorem 4.1 uses our new refined peeling result (Theorem C.1) and is very different from the proof of [25, Theorem 1]. Furthermore, Theorem 4.1 and discussion in Section 2.1 (see footnote 10) can be used to improve upon [25, Theorem 3] and establish the rate adaptive behavior of the LSE [5, 10, 25] when $\mathbb{E}(|\epsilon|^p X) \leq \sigma^p$ and $\epsilon$ is allowed to depend on $X$.

4.1. Example 3: Univariate isotonic regression

Let $\mathcal{F}$ be the set of nondecreasing functions on $[0, 1]$, $P_X$ be any nonatomic probability measure on $[0, 1]$, and $f_0 \equiv 0$ (or any other constant). Under the fixed design version of (1), [56] shows that $\|f - f_0\|_1 = O_p(\sqrt{\log n/n})$ when $\epsilon$ has finite variance. Han and Wellner [25] consider the model (1) (random design) and show that $\|f - f_0\| = O_p(\sqrt{\log n/n})$ if $\epsilon$ has finite $L_{2,1}$ moment (see footnote 15 for a definition) and is independent of $X$. We will show that both the independence and the finite $L_{2,1}$ moment assumption in [25] can be removed if $\epsilon$ satisfies (CVar).

Note that $\mathcal{F}$ is unbounded, but the discussion in Section 2.1 (see footnote 10 in page 9) and [25, Lemma 5] show that if $\epsilon$ satisfies (CVar), then $\|f\|_\infty = O_p(1)$. Thus, following the arguments of Section 2.1, it is easy to see that the rate of convergence of the isotonic LSE matches (up to a polynomial in $\log n$ factor) the rate of convergence of LSE when $\mathcal{F}$ is the set of nondecreasing functions uniformly bounded by 1. Giné and Koltchinskii [19, Example 3.8] show that if $f_0 \equiv 0$ and $\mathcal{F} := \{f : [0, 1] \to [-1, 1] | f \text{ is nondecreasing} \}$, then

$$F_\delta(x) = \min \left\{ 1, \delta \max \left( P_X[0, x], P_X[x, 1] \right)^{-1/2} \right\} \text{ for all } x \in [0, 1],$$

(18)

where for every $0 \leq a \leq b \leq 1$, $P_X[a, b] := \mathbb{P}(X \in [a, b])$. Furthermore, Giné and Koltchinskii [19, Example 3.8] show that $\|F_\delta\| \leq C\delta \sqrt{\log(1/\delta)}$. Therefore, Theorem 4.1 in conjunction with the arguments in item 6 of Remark 4.1 (also see item 5 in Remark 2.1) implies that $\hat{f}$ converges at an $n^{1/2}$ rate (up to a polynomial in $\log n$ factor) when $\epsilon$ satisfies (CVar). Similar results exist

\footnote{The $L_{2,1}$ moment for $\epsilon$ is $\int_0^\infty \sqrt{\mathbb{P}(|\epsilon| > t)} \, dt$. Finite $L_{2,1}$ moments imply finite second moments. But the converse is not true.}
when $\epsilon$ is independent of $X$ [25] or under the fixed design setting ($X_1, \ldots, X_n$ are fixed and non-random) [10, 18, 56]. However, to the best of our knowledge our result here is new and reduces the assumptions on $\epsilon$ for optimal convergence of the LSE.

**Remark 4.3** (Extension to piecewise constant functions and adaptive rates). In the above example, we showed that the isotonic LSE converges at a parametric rate (up to $\log n$ factors) when $f_0$ is a constant function. The above result can be generalized to case when $f_0$ is piecewise constant functions with $K$-pieces to show that that the $\|\hat{f} - f_0\|^2 \leq CKn^{-1}\log^2 n$ with high probability. Furthermore, if $f_0$ can be “approximated well” by a piecewise constant function then Theorem 4.1 can be used to find sharp rate upper bounds on $\|\hat{f} - f_0\|$. See [25, Section 3 and Theorem 3] for an excellent and elaborate discussion on this.

Finally as discussed in the beginning of Section 4, Guntuboyina and Sen [22, Lemma A3] show that $F_5$ satisfies (18) when $f_0$ is a linear function and $F$ is the class of uniformly bounded convex function. Thus a similar almost parametric rate can be proved if $f_0$ is a linear (piecewise linear or well approximated by piecewise linear function) function on $[0, 1]$ and $F$ is the set of convex functions on $[0, 1]$.

5. Misspecification

The results in the previous sections find the rate of convergence of the LSE $\hat{f}$ when $f_0 \in F$ and errors have finite number of moments. A crucial step in finding upper bounds for $\|\hat{f} - f_0\|$ is proving

$$\mathbb{E}[(Y - f(X))^2] - \mathbb{E}[(Y - f_0(X))^2] \geq \|f - f_0\|^2,$$  \hspace{1cm} (19)

for any function $f \in F$\(^{16}\); see (31) in the proof of Theorem C.1. A natural next step is the study of the LSE when $f_0 \notin F$. LSEs under misspecification have received a lot of attention but most works assume restrictive conditions on $\epsilon$; see e.g., [7, 25, 31, 32, 55]. The techniques developed in this paper can be used to relax the assumptions on $\epsilon$ and allow for heteroscedastic and heavy-tailed errors.

If the true conditional expectation ($f_0$) does not belong to the class $F$ but $F$ is a convex set, then defining

$$\hat{f} := \arg \min_{f \in F} \mathbb{E}[(Y - f(X))^2],$$

we have that for any $f \in F$,

$$\frac{\partial}{\partial t} \mathbb{E}[(Y - (\hat{f} + t(f - \hat{f})))^2]|_{t=0} = 0 \quad \text{and thus} \quad \mathbb{E}[(Y - \hat{f}(X))(f - \hat{f})(X)] = 0.$$  

This further implies that $\mathbb{E}[(Y - f(X))^2] - \mathbb{E}[(Y - \hat{f}(X))^2] = \|f - \hat{f}\|^2$. Using this fact instead of (19), the results proved in previous sections imply the same rate bounds for $\|\hat{f} - f\|$. Thus, when $F$ is a convex set, the proofs of Theorems 2.1 and 3.1 go through by replacing $\epsilon$ with $\xi = Y - \hat{f}(X)$; note that $\mathbb{E}(\xi|X) = 0$. Most of the function classes considered in this paper are convex sets and hence our results do not require the well-specification assumption. This discussion concludes that an analogue\(^{17}\) of Theorem 5.1 of [32] holds even when the response has finite number of moments. The analysis, however, is different if $F$ is a non-convex set. Examples of non-convex function spaces include single or multiple index models and sparse linear or non-linear models; e.g., see Section 3.1. When $F$ is a non-convex set, the inequality (19) may not hold and finding upper rate bound for $\|\hat{f} - f\|$ requires different proof techniques. However, even in this case, the tools developed in this paper can be used, because the proofs for rate bounds under misspecification hinge on the control of an empirical process analogous to (3); see [3, Equations (1) and (2)] and [32, Theorem 5.2]. We leave the precise details of this argument for future research.

\(^{16}\)Recall that $f_0(x) = \mathbb{E}[Y|X = x]$.

\(^{17}\)The tail probabilities will decay polynomially and not exponentially.
6. Concluding remarks

Least squares estimators in nonparametric regression models are known to be minimax rate optimal when $\epsilon$ is sub-Gaussian when $F$ satisfies appropriate entropy assumptions. We show that in a wide variety of cases LSE attains the same rate of convergence even when $\epsilon$ is neither sub-Gaussian nor independent of $X$. We find sufficient moment conditions on $\epsilon$ under which the rate of convergence of the LSE under heavy-tailed errors matches the rate of the LSE under sub-Gaussian errors, i.e., the LSE is “robust” to heavy-tailed errors. Our sufficient conditions depend on the complexity ($\alpha$) and the local structure ($s$) of the function class $F$. The results justify the usage of LSE even under heteroscedastic and heavy-tailed errors. The necessity of our conditions is currently under investigation. In this paper, all our results focus on the squared error loss but our results can be easily generalized to other smooth loss functions.

Appendix A: Interpolation inequalities

In this section, we state and prove three interpolation inequalities that find the local envelope and $s$ (coefficient for the local structure) for the examples considered in the paper.

**Proposition A.1** (Local envelope for bounded convex function). Let

$$F := \{ f : [0, 1] \to [-\Phi, \Phi] \mid f \text{ is convex} \}$$

and $P_X$ be the uniform distribution on $[0, 1]$. Fix any $f_0 \in F$, then for any $x \in [0, 1]$,

$$F_\delta(x) \leq \min \left\{ 2(2\Phi)^{1/3}\delta^{2/3}\max\{x^{-1/3}, (1-x)^{-1/3}\}, 2\Phi \right\}.$$

Thus $\|F_\delta\|_\infty = 2\Phi$ and

$$\|F_\delta\|_3 \leq 4\Phi^{1/3}\delta^{2/3}\left[ \log(0.5\Phi^2/\delta^2) \right]^{1/3}. \tag{20}$$

**Proof.** A convex function $f$ bounded by $\Phi$ on $[0, 1]$ is Lipschitz on any sub-interval $[a, b]$ with Lipschitz constant $2\Phi/\min\{a, 1-b\}$. Fix any $x \in (0, 1/2)$. On any interval $[a, b] \subseteq [0, 1]$ containing $x$, $f$ and $f_0$ are both Lipschitz with Lipschitz constant $2\Phi/a$ (which implies that $f - f_0$ is Lipschitz with Lipschitz constant $4\Phi/a$). Using the interpolation for Lipschitz functions [13, Lemma 2], we have

$$|(f - f_0)(x)| \leq 2 \left( \int_a^b |f - f_0|^2(t)dt \right)^{1/3} \left( \frac{2\Phi}{a} \right)^{1/3}. \tag{21}$$

Hence if $F_\delta := \{ f : [0, 1] \to [-\Phi, \Phi] \mid \|f - f_0\| \leq \delta \text{ and } f \in F \}$, then for every $0 < x \leq 1/2$, we have

$$\sup_{f \in F_\delta} |(f - f_0)(x)| \leq 2(2\Phi)^{1/3}\delta^{2/3}x^{-1/3},$$

where we replaced $a$ in (21) by $x$ by taking limit $a \downarrow x$. Thus, by symmetry

$$F_\delta(x) \leq 2(2\Phi)^{1/3}\delta^{2/3}\max\{x^{-1/3}, (1-x)^{-1/3}\}.$$

However, $\|f\|_\infty \leq \Phi$ for every $f \in F$ thus

$$F_\delta(x) \leq \min \left\{ 2(2\Phi)^{1/3}\delta^{2/3}\max\{x^{-1/3}, (1-x)^{-1/3}\}, 2\Phi \right\}.$$

To prove (20), observe that

$$\int_0^{1/2} |F_\delta^3(x)|dx = \int_0^\eta |F_\delta^3(x)|dx + \int_\eta^{1/2} |F_\delta^3(x)|dx \leq 8\Phi^3\eta + (2(2\Phi)^{1/3}\delta^{2/3}\log(1/(2\eta))).$$

Taking $\eta = \Phi^{-2}\delta^2$ implies

$$\|F_\delta\|_3^3 \leq 16(2\Phi)\delta^2\log(0.5\Phi^2/\delta^2).$$

\qed
Proposition A.2 (Local envelope for additive Models). Suppose \( f : [0, 1]^d \to \mathbb{R} \) can be written as \( f(x) = f_1(x_1) + f_2(x_2) \) for some functions \( f_j : [0, 1]^{d_j} \to \mathbb{R} \), \( j = 1, 2 \), for every \( x = (x_1^T, x_2^T) \in [0, 1]^d \) with \( x_1 \in [0, 1]^{d_1}, x_2 \in [0, 1]^{d_2} \). If \( f_j \in \mathcal{F}_{\gamma_j, d_j}(L) \), \( j = 1, 2 \) where for any \( \gamma > 0 \) and dimension \( d \)
\[
\mathcal{F}_{\gamma, d}(L) := \{ f : \mathcal{X} \to \mathbb{R} : f/L \in \mathcal{F}_{\gamma, d} \},
\]
and \( P_X \) is the uniform distribution on \([0, 1]^d\). Then
\[
\|f\|_{\infty} \leq 5(\|f\|_2^c + \|f\|_2^b) \left(L^{1-c_1} + L^{1-c_2}\right),
\]
where \( c_j := 2\gamma_j/(2\gamma_1 + d_j) \), \( j = 1, 2 \). In particular, if \( f(x) = \sum_{j=1}^d f_j(x_j) \) for functions \( f_j : [0, 1] \to \mathbb{R} \) such that \( f_j \in \mathcal{F}_{\gamma, 1}(L) \), then
\[
\|f\|_{\infty} \leq 5d\|f\|_2 L^{1-c}, \quad \text{where} \quad c = 2\gamma/(2\gamma + 1).
\]

Proof. Consider \( f(x_1, x_2) = f_1(x_1) + f_2(x_2) \) for two functions \( f_1 : \mathbb{R}^{d_1} \to \mathbb{R}, f_2 : \mathbb{R}^{d_2} \to \mathbb{R} \). Define
\[
\tilde{f}_1(x_1) = f_1(x_1) - \int f_1(t) dt \quad \text{and} \quad \tilde{f}_2(x_2) = f_2(x_2) - \int f_2(t) dt.
\]
Because \( f_j \in \mathcal{F}_{\gamma_j, d_j}(L) \) are \( \gamma_j \)-smooth, \( \tilde{f}_j \), \( j = 1, 2 \) are also \( \gamma_j \)-smooth. Hence by [13, Lemma 2], we have
\[
\|\tilde{f}_1\|_{\infty} \leq 2\|\tilde{f}_1\|_2 L^{1-c_1} \quad \text{and} \quad \|\tilde{f}_2\|_{\infty} \leq 2\|\tilde{f}_2\|_2 L^{1-c_1},
\]
where \( c_j = 2\gamma_j/(2\gamma_1 + d_j) \), \( j = 1, 2 \). Observe now that
\[
\mathbb{E}[(\tilde{f}_1(X_1) + \tilde{f}_2(X_2))^2] = \|\tilde{f}_1\|_2^2 + \|\tilde{f}_2\|_2^2.
\]
Therefore,
\[
\sup_{(x_1, x_2) \in [0, 1]^{d_1+d_2}} |\tilde{f}_1(x_1) + \tilde{f}_2(x_2)| \leq \|\tilde{f}_1\|_{\infty} + \|\tilde{f}_2\|_{\infty}
\leq 2 \left[\|\tilde{f}_1\|_2 + \|\tilde{f}_2\|_2\right]^{c_1} + (\|\tilde{f}_1\|_2 + \|\tilde{f}_2\|_2) c_2 (L^{1-c_1} + L^{1-c_2}).
\]
Because
\[
\|f\|_1 \leq \|f_1\|_1 + \|f_2\|_1 \quad \text{and} \quad \|f\|_2^b = \|f_1\|_2^b + \|f_2\|_2^b + \|f_1\|_2^c + \|f_2\|_2^c,
\]
we get
\[
\|f\|_{\infty} \leq \int |f(x_1, x_2)| dx_1 dx_2 + 4\|f\|_2^c + \|f\|_2^b \left(L^{1-c_1} + L^{1-c_2}\right)
\leq \|f\|_2 + 4\|f\|_2^c + \|f\|_2^b \left(L^{1-c_1} + L^{1-c_2}\right).
\]
Furthermore, since \( \|f\|_{\infty} \leq L \), we get \( \|f\|_2^{1-c_1} \leq L^{1-c_1} \) and \( \|f\|_2^{1-c_2} \leq L^{1-c_2} \). Thus, we have
\[
\|f\|_{\infty} \leq 5\|f\|_2^{1-c_1} + \|f\|_2^{1-c_2},
\]
If \( f(x_1, \ldots, x_d) = \sum_{j=1}^d f_j(x_j) \) and \( f_j(\cdot) \) is \( \gamma \)-smooth for all \( 1 \leq j \leq d \), then
\[
\|f\|_{\infty} \leq 5d\|f\|_2 L^{1-c}, \quad \text{where} \quad c = 2\gamma/(2\gamma + 1). \quad \square
\]

Proposition A.3 (Local envelope for multiple index models). Suppose \( f(x) = m(Bx) - m_0(B_0x) \) for functions \( m, m_0 : \mathbb{R}^p \to \mathbb{R} \) satisfying \( m, m_0 \in \mathcal{F}_{\gamma, p}(L) \) (defined in (22)) and \( B, B_0 \in \mathbb{R}^{p \times d} \) with \( p < d \). If \( X \in \mathbb{R}^d \) is a random vector such that \((BX)^T, (B_0X)^T\) has a density with respect to the Lebesgue measure that is lower bounded by \( C > 0 \), then
\[
\|m \circ B - m_0 \circ B_0\|_{\infty} \leq 10C^{-c/2}\|m \circ B - m_0 \circ B_0\|_c L^{1-c},
\]
where \( \|m \circ B - m_0 \circ B_0\| := (\mathbb{E}[|m(BX) - m_0(B_0X)|^2])^{1/2} \) and \( c = 2\gamma/(2\gamma + p) \).
Proof. Define
\[ y := \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} := \begin{pmatrix} Bx \\ B_0x \end{pmatrix} \in \mathbb{R}^{2p}. \]

Then we can write \( f(y) = f_1(y_1) + f_2(y_2) \) where \( f_1(y_1) = m(y_1) \) and \( f_2(y_2) = -m_0(y_2) \). Observe that if \( Y = ((BX)\top, (B_0X)\top)\top \) has a density with respect to the Lebesgue measure bounded away from zero, that is, \( p_Y(y) \geq C \) with \( p_Y(\cdot) \) representing the pdf of \( Y \), then
\[
\|f\|_{2,Y} = (\mathbb{E}[f^2(Y)])^{1/2} \geq C^{1/2} \left( \int f^2(y) dy \right)^{1/2}.
\]

Applying Proposition A.2 with \( \gamma_1 = \gamma_2 = \gamma \) and \( d_1 = d_2 = p \) we get
\[
\|f\|_\infty \leq 10 \left[ \int f^2(y) dy \right]^{c/2} L^{1-c},
\]
where \( c = 2\gamma/(2\gamma + p) \). Hence from (23), we get
\[
\|f\|_\infty \leq 10C^{-c/2}\|f\|_{2,Y}^c L^{1-c},
\]
which implies the result. \( \square \)

Appendix B: A new maximal inequality for finite maximums

The following maximal inequality will be used in the proof of Theorem 3.1 but is also of interest.

**Proposition B.1.** Let \( x_1, \ldots, x_n \) be independent random variables in \( \mathbb{R}^p \) with \( p \geq 2 \) and \( R_1, \ldots, R_n \) be i.i.d. Rademacher random variables (independent of \( x_1, \ldots, x_n \)). If for all \( i \in \{1, \ldots, n\} \),
\[
\mathbb{E}[\xi_i^q] < \infty \quad \text{where} \quad \xi_i := \max_{1 \leq j \leq p} |x_{i,j}|,
\]
where \( x_i := (x_{i,1}, \ldots, x_{i,p}) \). Then
\[
\mathbb{E} \left[ \sup_{1 \leq j \leq p} \sum_{i=1}^n R_i x_{i,j} \right] \leq \sqrt{\log p} \sup_{j} \mathbb{E}^{1/2} \left[ \sum_{i=1}^n x_{i,j}^2 \right] + (\log p)^{1-1/q} \left( \sum_{i=1}^n \mathbb{E}[\xi_i^q] \right)^{1/q}.
\]

**B.1. Example 4: An application of Proposition B.1**

Suppose we have \( n \) i.i.d. pairs \( (X_i, \epsilon_i), 1 \leq i \leq n \) such that \( \mathbb{E}[\epsilon_i|X_i] = 0 \) and \( \mathbb{E}[\epsilon_i^2|X_i] \leq \sigma^2 \) a.e. \( P_X \). Let \( \{f_1, \ldots, f_N\} \) be a collection of functions from \( \chi \) to \( \mathbb{R} \) and let \( F(\cdot) := \max_{1 \leq j \leq N} |f_j(\cdot)| \) denote their envelope. Then Proposition (B.1) (with \( q = 2 \)) yields
\[
\mathbb{E} \left[ \max_{1 \leq j \leq N} |G_n[f_j(X)]| \right] \lesssim \sqrt{\log N} \mathbb{E}^{1/2} \left[ \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 f_j^2(X_i) \right] + (\log N)^{1-1/q} \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\epsilon_i^q F^n(X_i)] \right)^{1/q},
\]
\[
\lesssim \sigma \sqrt{\log N} \mathbb{E}^{1/2} \left[ \max_{1 \leq j \leq N} |f_j| + \sigma \|F\| (\log N)^{1/2} \right] \quad \text{(taking} \quad q = 2 \text{)}
\]
\[
\lesssim \sigma \sqrt{\log N} \mathbb{E}^{1/2} \left[ \max_{1 \leq j \leq N} |f_j| + \|F\| \right],
\]
which implies that
\[
\max_{1 \leq j \leq N} |G_n(\epsilon f_j(X))| = O_p(\sqrt{\log N})
\] (25)
whenever \(\|F\| = O(1)\). In contrast, Lemma 8 of [14] implies
\[
E \left[ \max_{1 \leq j \leq N} |G_n(\epsilon f_j(X))| \right] \lesssim \sqrt{\log N} \max_j E^{1/2} \left[ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 f_j^2(X_i) \right] + \log N \sqrt{E} \max_{1 \leq i \leq n} |\epsilon_i|^2 F^2(X_i).
\]

Even when \(\|F\|_\infty \leq C < \infty\), under only the second moment assumption, \(E[\max_{1 \leq i \leq n} |\epsilon_i|^2] = O(n)\) and hence the second term on the right hand side will be of the order \(\log N\). Thus Lemma 8 of [14] will imply that \(\max_{1 \leq j \leq N} |G_n(\epsilon f_j(X))| = O_p(\log N)\). Thus (25) is a significant improvement, as the above calculation now implies that the lasso estimator is minimax rate optimal under just the conditional second moment assumption (CVar), when the covariates are coordinate-wise bounded; see Theorem 11.1 of [27]. Proposition B.1 can also be used in proving consistency of the multiplier bootstrap under finite moment assumptions; see e.g., Remark 5.2 of [33].

### B.2. Proof of Proposition B.1

It is clear that if \(R_1, \ldots, R_n\) are i.i.d. Rademacher random variables then
\[
E \left[ \sup_{1 \leq i \leq p} \sum_{i=1}^{n} R_i y_{i,j} \right] \leq \inf_{Y = Y^{(1)} + Y^{(2)}} \left\{ E \left[ \sup_{1 \leq i \leq p} \sum_{i=1}^{n} R_i y_{i,j}^{(1)} \right] + \sup_{1 \leq i \leq p} \sum_{i=1}^{n} |y_{i,j}^{(2)}| \right\},
\]
for any set of fixed numbers \(y_{i,j}, 1 \leq i \leq n, 1 \leq j \leq p\) and decomposition \(Y = Y^{(1)} + Y^{(2)}\) of \(Y = (y_{i,j})\). Recall the envelope
\[
\xi_i = \max_{1 \leq j \leq p} |x_{i,j}|.
\]

Now consider the decomposition
\[
x_{i,j} = x_{i,j} \mathbb{1}\{\xi_i \leq B\} + x_{i,j} \mathbb{1}\{\xi_i > B\} =: x_{i,j}^{(1)} + x_{i,j}^{(2)}.
\]

Then we get
\[
E \left[ \sup_{1 \leq i \leq p} \sum_{i=1}^{n} R_i x_{i,j} \right] \leq E \left[ \sup_{1 \leq i \leq p} \sum_{i=1}^{n} R_i x_{i,j} \mathbb{1}\{\xi_i \leq B\} \right] + E \left[ \sup_{1 \leq i \leq p} \sum_{i=1}^{n} |x_{i,j}| \mathbb{1}\{\xi_i > B\} \right]
\]
\[
\lesssim E \left[ \sup_{1 \leq i \leq p} \left( \sum_{i=1}^{n} x_{i,j}^2 \mathbb{1}\{\xi_i \leq B\} \right)^{1/2} \right] \sqrt{\log p} + \sum_{i=1}^{n} E[\xi_i^2 \mathbb{1}\{\xi_i > B\}]
\]
\[
\leq E \left[ \sup_{1 \leq i \leq p} \left( \sum_{i=1}^{n} x_{i,j}^2 \mathbb{1}\{\xi_i \leq B\} \right)^{1/2} \right] \sqrt{\log p} + B^{-q+1} \sum_{i=1}^{n} E[\xi_i^q].
\]

By Lemma 9 of [14], we get
\[
E \left[ \sup_{1 \leq i \leq p} \sum_{i=1}^{n} x_{i,j}^2 \mathbb{1}\{\xi_i \leq B\} \right] \leq \sup_{1 \leq j \leq p} \sum_{i=1}^{n} E[x_{i,j}^2] + B^2 \log p,
\]
and hence
\[ E \left[ \sup_{1 \leq j \leq p} \left( \sum_{i=1}^{n} x_{i,j}^2 \mathbb{I} \{ \xi_i \leq B \} \right)^{1/2} \right] \leq \sup_{1 \leq j \leq p} E^{1/2} \left[ \sum_{i=1}^{n} x_{i,j}^2 \right] + B \sqrt{\log p}. \]

Substituting this in the above bound for \( E[\sup_j \sum_{i=1}^{n} R_i x_{i,j}] \) yields
\[ E \left[ \sup_{1 \leq j \leq p} \sum_{i=1}^{n} R_i x_{i,j} \right] \leq \sqrt{\log p} \sup_j E^{1/2} \left[ \sum_{i=1}^{n} x_{i,j}^2 \right] + B \log p + B^{-q+1} \sum_{i=1}^{n} E[\xi_i^q]. \]

Taking \( B = (\log p)^{-1/q}(\sum_{i=1}^{n} E[\xi_i^q])^{1/q} \) we get
\[ E \left[ \sup_{1 \leq j \leq p} \sum_{i=1}^{n} R_i x_{i,j} \right] \leq \sqrt{\log p} \sup_j E^{1/2} \left[ \sum_{i=1}^{n} x_{i,j}^2 \right] + (\log p)^{1-1/q} \left( \sum_{i=1}^{n} E[\xi_i^q] \right)^{1/q}. \]

**Appendix C: A refined peeling result**

In this section, we prove a new peeling result. The result is a key component in the proofs of the rate results (Theorems 2.1, 3.1, and 4.1) in the paper. It is this refinement that helps us prove fast rates of convergence of the LSE in previously inaccessible cases. Before stating the result, we will introduce some notations. Let
\[ \hat{f} := \arg \min_{f \in \mathcal{F}} M_n(f) \quad \text{where} \quad M_n(f) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \]
and
\[ T(f; \epsilon, X) := 2\epsilon(f - f_0)(X) - (f - f_0)^2(X). \]
Furthermore, let \( U : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \) be such that
\[ \sup_{f: \|f - f_0\| \leq \delta} |T(f; \epsilon, X)| \leq U(\epsilon, X; \delta), \quad (27) \]
for all values of \( \epsilon, X, \) and \( \delta. \) If \( \Phi := \sup_{f \in \mathcal{F}} \|f\|_{\infty}, \) then a trivial choice is \( U(\epsilon, X; \delta) = 4|\epsilon| + \Phi \) (we use this choice in the proof of Theorem 2.1). Now for any \( B > 0, \) let
\[ T_B(f; \epsilon, X, \delta) := T(f; \epsilon, X) \mathbb{I} \{ U(\epsilon, X; \delta) \leq B \}. \]

Theorem C.1 below is useful because it provides tail bounds for \( \|\hat{f} - f_0\| \) in terms of upper bounds on a bounded (note that \( T(\cdot, \cdot, \cdot) \) is unbounded while \( T_B(\cdot, \cdot, \cdot) \) is bounded) empirical process and most existing maximal inequalities provide upper bounds for only bounded empirical processes.

**Theorem C.1.** (Peeling with truncation) Suppose \( f_0 \in \mathcal{F}, \Phi := \sup_{f \in \mathcal{F}} \|f\|_{\infty}, \) and there exists a real valued function \( \phi_n(\cdot; \cdot) \) such that
\[ \sup_{f: \|f - f_0\| \leq \delta} \mathbb{E} T_B^2(f; \epsilon, X, \delta) \leq 4(\sigma + \Phi)^2 \delta^2 \]
and
\[ \mathbb{E} \left[ \sup_{\delta/2 \leq \|f - f_0\| \leq \delta} \mathcal{G}_n(\mathbb{T}_B(f; \epsilon, X, \delta)) \right] \leq \phi_n(\delta; B), \quad (28) \]
for every \( n \) and any \( \delta, B > 0. \) Further, if there exists \( \gamma \geq 2 \) and \( s_\gamma(\cdot) \) such that
\[ \mathbb{E} [U^\gamma(\epsilon, X; \delta)] \leq s_\gamma(\delta). \]
Then for any positive $D, \varepsilon_n$, and $\{B_k\}_{k=1}^\infty$, $\beta \geq 1$, and $n \geq 1$ we have
\[
P \left( \| \hat{f} - f_0 \| \geq D\varepsilon_n \right) \leq \left( \frac{C}{\sqrt{n}(D\varepsilon_n)^2} \right)^{\beta} \sum_{k=1}^{\infty} \phi_n^{\beta}(2^k D\varepsilon_n, B_k) + \left( \frac{8C\sqrt{3}(\sigma + \Phi)}{\sqrt{n}(D\varepsilon_n)} \right)^{\beta} + \left( \frac{C\beta}{n(D\varepsilon_n)^2} \right)^{\beta} \sum_{k=1}^{\infty} \frac{B_k^\beta}{2^{2k\beta}} + \frac{16}{(D\varepsilon_n)^2} \sum_{k=0}^{\infty} s_n(2^k D\varepsilon_n),
\]
for some universal constant $C$.

**Remark C.1.** Shen and Wong [44] propose an iterative (non-dyadic) version of the dyadic peeling argument presented here. After a thorough investigation, we have found that their approach does not lead to better rates.

**Proof.** The proof follows along the standard peeling argument [55, Theorem 3.2.5]. But the crucial observation here is that the empirical processes involved here are not bounded and thus to be able to apply the rich literature of maximal inequalities we truncate the empirical process involved in the peeling step. The proof is split into 3 main steps.

**Step 1: Peeling and truncation.** From the definition of $\hat{f}$, it follows that
\[
P \left( \| \hat{f} - f_0 \| \geq D\varepsilon_n \right) \leq \mathbb{P} \left( \sup_{\| f - f_0 \| \geq D\varepsilon_n} \mathbb{M}_n(f_0) - \mathbb{M}_n(f) \geq 0 \right)
\leq \mathbb{P} \left( \bigcup_{k=0}^{\infty} \left\{ \sup_{f \in A_k} \mathbb{M}_n(f_0) - \mathbb{M}_n(f) \geq 0 \right\} \right),
\]
where
\[A_k = \{ f : 2^{k-1} D\varepsilon_n \leq \| f - f_0 \| \leq 2^k D\varepsilon_n \}.
\]
From the definition of $f_0$, we obtain
\[
\mathbb{E} \left[ \mathbb{M}_n(f) - \mathbb{M}_n(f_0) \right] = \mathbb{E} \left[ \hat{f}^2(X) - 2\hat{Y}(f(X) - f_0(X)) - f^2_0(X) \right] = \| f - f_0 \|^2,
\]
(31)
as $\mathbb{E}(Y|X) = f_0(X)$. Thus
\[
P \left( \| \hat{f} - f_0 \| \geq D\varepsilon_n \right) \leq \mathbb{P} \left( \bigcup_{k=1}^{\infty} \left\{ \sup_{f \in A_k} \mathbb{G}_n(T(f; \epsilon, X)) \geq \sqrt{n}(2^{k-1} D\varepsilon_n)^2 \right\} \right).
\]
Observe that $T$ is unbounded. To control the tail probabilities, we will truncate $T$ at a sequence $\{B_k\}_{k=1}^\infty$. Thus
\[
P \left( \| \hat{f} - f_0 \| \geq D\varepsilon_n \right) \leq \mathbb{P} \left( \bigcup_{k=1}^{\infty} \left\{ \sup_{f \in A_k} \mathbb{G}_n(T_B(f; \epsilon, X, 2^k D\varepsilon_n)) \geq \sqrt{n}(2^{k-1} D\varepsilon_n)^2/2 \right\} \right)
+ \mathbb{P} \left( \bigcup_{k=1}^{\infty} \left\{ \sup_{f \in A_k} \mathbb{G}_n(T(f; \epsilon, X) - T_B(f; \epsilon, X, 2^k D\varepsilon_n)) \geq \sqrt{n}(2^{k-1} D\varepsilon_n)^2/2 \right\} \right)
=: \mathbf{P}_1 + \mathbf{P}_2.
\]
Observe that $\mathbf{P}_1$ corresponds to the bounded part and $\mathbf{P}_2$ corresponds to the unbounded part.

**Step 2: The unbounded part.** To bound $\mathbf{P}_2$, observe that by definition of $U(\cdot, \cdot; \cdot)$
\[
\sup_{f \in A_k} | T(f; \epsilon, X) | \leq U(\epsilon, X; 2^k D\varepsilon_n).
\]
Therefore, for any $f \in A_k$, we have
\[
|G_n(T(f; \epsilon, X) - T_{B_k}(f; \epsilon, X))| \leq \sqrt{n}(P_n + P)|T(f; \epsilon, X)|1\{U(\epsilon, X; 2^kD\epsilon_n) \geq B_k\}
\leq \sqrt{n}(P_n + P)U(\epsilon, X; 2^kD\epsilon_n)1\{U(\epsilon, X; 2^kD\epsilon_n) \geq B_k\},
\]
and hence,
\[
P_2 = P\left(\bigcup_{k=1}^{\infty} \left\{ \sup_{f \in A_k} G_n(T(f; \epsilon, X) - T_{B_k}(f; \epsilon, X, 2^kD\epsilon_n)) \geq \sqrt{n}(2^{k-1}D\epsilon_n)^2/2 \right\} \right)
\leq \sum_{k=1}^{\infty} P\left( \sup_{f \in A_k} G_n(T(f; \epsilon, X) - T_{B_k}(f; \epsilon, X, 2^kD\epsilon_n)) \geq \sqrt{n}(2^{k-1}D\epsilon_n)^2/2 \right)
\leq \sum_{k=1}^{\infty} \frac{1}{\sqrt{n}(2^{k-1}D\epsilon_n)^2/2} \left( \sup_{f \in A_k} G_n(T(f; \epsilon, X) - T_{B_k}(f; \epsilon, X, 2^kD\epsilon_n)) \right)
\leq \sum_{k=1}^{\infty} \frac{1}{\sqrt{n}(2^{k-1}D\epsilon_n)^2/2} \left( \sqrt{n}(P_n + P)U(\epsilon, X; 2^kD\epsilon_n)1\{U(\epsilon, X; 2^kD\epsilon_n) \geq B_k\} \right)
\leq \sum_{k=1}^{\infty} 4E \left[ U(\epsilon, X; 2^kD\epsilon_n)1\{U(\epsilon, X; 2^kD\epsilon_n) \geq B_k\} \right] / (2^{k-1}D\epsilon_n)^2.
\]

However, by assumption (29), we have
\[
E[U(\epsilon, X; 2^kD\epsilon_n)1\{U(\epsilon, X; 2^kD\epsilon_n) \geq B_k\}] \leq \frac{E[U^\gamma(\epsilon, X; 2^kD\epsilon_n)]}{B_k^{\gamma-1}} \leq \frac{s_\gamma(2^kD\epsilon_n)}{B_k^{\gamma-1}}.
\]

Thus
\[
P_2 \leq \frac{16}{(D\epsilon_n)^2} \sum_{k=1}^{\infty} s_\gamma(2^kD\epsilon_n) / 2^{2k}B_k^{\gamma-1}.
\]

**Step 3: The bounded part.** Now to bound $P_1$, observe that
\[
P_1 \leq \sum_{k=1}^{\infty} P\left( \sup_{f \in A_k} G_nT_{B_k}(f; \epsilon, X, 2^kD\epsilon_n) \geq \sqrt{n}(2^{k-1}D\epsilon_n)^2/2 \right).
\]
We will use Markov’s inequality to bound each of the term on the right in the above display. For any $\beta \geq 1$, we have
\[
P\left( \sup_{f \in A_k} G_nT_{B_k}(f; \epsilon, X, 2^kD\epsilon_n) \geq \sqrt{n}(2^{k-1}D\epsilon_n)^2/2 \right) \leq \frac{8^\beta E \left[ \sup_{f \in A_k} |G_nT_{B_k}(f; \epsilon, X, 2^kD\epsilon_n)|^\beta \right]}{(\sqrt{n}(2^kD\epsilon_n)^2)^{\beta/2}}.
\]
Using Proposition 3.1 of [20], for every $\beta \geq 1$ we get
\[
E \left[ \sup_{f \in A_k} |G_nT_{B_k}(f; \epsilon, X, 2^kD\epsilon_n)|^\beta \right] \leq C^\beta \phi_n(2^kD\epsilon_n, B_k) + C^\beta \beta^{\beta/2} \sup_{f \in A_k} \left( E \left[ T_{B_k}(f; \epsilon, X, 2^kD\epsilon_n) \right] \right)^{\beta/2}
\]
\[
+ C^\beta \beta^{\beta/2} E \left[ \max_{1 \leq i \leq n} \sup_{f \in A_k} |T_{B_k}(f; \epsilon_i, X_i, 2^kD\epsilon_n)|^\beta \right]
\leq C^\beta \left[ \phi_n(2^kD\epsilon_n, B_k) + \beta^{\beta/2}(\sigma + M)^{\beta}(2^kD\epsilon_n)^\beta + \frac{\beta B_k^\beta}{n^{\beta/2}} \right],
\]

\[\text{We get non trivial bound for every } \beta > 0 \text{ because } T_B(\cdot, \cdot, \cdot) \text{ uniformly bounded.}\]
Lemma 3.4.2 of [55]. Recall that

The proof proof of Theorem 2.1 will be split into two steps. In the first step, we find \( \phi \), and in the second step we apply Theorem C.1 with an appropriate choice of \( \beta, U \), and \( \gamma \).

Finally, from the calculations of Section 3.4.3.2 of [55] it follows that for any \( \delta > 0 \),

where \( C \) is a universal constant. This is because by assumption of the theorem, we have

\[
\sup_{f \in A_n} \mathbb{E} \left[ T_{B_k}^2(f; \epsilon, X, 2^k D \varepsilon_n) \right] \leq 4(\sigma + \Phi)^2 (2^k D \varepsilon_n)^2
\]

and \( |T_{B_k}(f; \epsilon, X, 2^k D \varepsilon_n)| \leq B_k \). Thus the above three displays combined with (32), we get

\[
\mathbb{P} \left( \| \hat{f} - f_0 \| \geq D \varepsilon_n \right) \leq \left( \frac{C}{\sqrt{n}(D \varepsilon_n)^2} \right)^{\frac{\beta}{2}} \sum_{k=1}^{\infty} \phi_n^k(2^k D \varepsilon_n, B_k) + \left( \frac{8C \beta(2D \varepsilon_n)}{\sqrt{n}(D \varepsilon_n)} \right)^{\beta} \sum_{k=1}^{\infty} \frac{1}{2^k \beta} + \frac{16}{(D \varepsilon_n)^2} \sum_{k=1}^{\infty} s_{\gamma}(2^k D \varepsilon_n) \right)^{\frac{1}{2k \beta}}.
\]

\[\square\]

Appendix D: Proof of Theorem 2.1

The proof proof of Theorem 2.1 will be split into two steps. In the first step, we find \( \phi_n(\delta, B) \) satisfying the assumptions of Theorem C.1. In the second step we apply Theorem C.1 with an appropriate choice of \( \beta, U \), and \( s_{\gamma} \).

**Finding \( \phi_n(\delta, B) \):** We will find a choice for \( \phi_n(\delta, B) \) that satisfies (28). To find \( \phi_n(\cdot, \cdot) \), we will apply Lemma 3.4.2 of [55]. Recall that

\[ T(f; \epsilon, X) = 2\epsilon(f - f_0)(X) - (f - f_0)^2(X), \]

Observe from the definition, we can choose the local envelope to be

\[ U(\epsilon, X, \delta) := 2(\epsilon + 2\Phi) F_\delta(X) \]

since it satisfies (27). Recall that \( T_B(f; \epsilon, X, \delta) = T(f; \epsilon, X) 1 \{ U(\epsilon, X, \delta) \leq B \} \), where \( T \) is defined in (26). Observe that by definition

\[ \| T_B(f; \cdot, \cdot, \delta) \|_\infty \leq B, \]

and for \( f \) satisfying \( \| f - f_0 \| \leq \delta \),

\[ P \left( T_B^2(f; \cdot, \cdot, \delta) \right) \leq P(T^2(f; \cdot, \cdot, \delta)) \leq 4\sigma^2 \delta^2 + 4\Phi^2 \delta^2 = (2\sigma + 2\Phi)^2 \delta^2. \]

Finally, from the calculations of Section 3.4.3.2 of [55] it follows that for any \( \delta > 0 \), we have

\[ N[1](\eta, \{ T_B(f; \epsilon, X, \delta) : f \in F \}, \| \cdot \| \leq N[1](\eta/(2\sigma + 2\Phi), F, \| \cdot \|). \]

From Lemma 3.4.2 of [55], we get

\[
\mathbb{E} \left[ \sup_{\delta/2 \leq \| f - f_0 \| \leq \delta} \mathbb{C}_n(T_B(f; \epsilon, X, \delta)) \right] \leq C_J[1](\sqrt{2}(2\sigma + 2\Phi) \delta) \left( 1 + \frac{J[1](\sqrt{2}(2\sigma + 2\Phi) \delta) B}{\sqrt{n(2\sigma + 2\Phi)^2 2^k \beta}} \right),
\]

where by \( (L_2) \),

\[
J[1](\beta) = \int_0^\beta \sqrt{\log N[1](\eta, \{ T_B(f; \epsilon, X) : f \in F \}, \| \cdot \|)} d\eta \\
\leq \int_0^\beta \sqrt{A(\eta/(2\sigma + 2\Phi))^{-1} \eta} d\eta \\
= A^{1/2}(2\sigma + 2\Phi)^{\alpha/2} \beta^{1-\alpha/2}/(1-\alpha/2).
\]
Therefore $J_{||}(\sqrt{2}(2\sigma + 2\Phi)\delta) \leq A^{1/2}2^{3/2-\alpha/4}(2\sigma + 2\Phi)\delta^{1-\alpha/2}/(2-\alpha)$ and

$$
\mathbb{E} \left[ \sup_{\|f \|=\delta} \mathbb{G}_n(T_B(f; \epsilon, X, \delta)) \right] \leq C \frac{A^{1/2}(\sigma + \Phi)\delta^{1-\alpha/2}}{2-\alpha} + C \frac{A\delta^{2-\alpha}B}{(2-\alpha)^2 \sqrt{n} \delta^2} \\
\leq C\frac{A^{1/2}(\sigma + \Phi)\delta^{1-\alpha/2}}{2-\alpha} + CAB\frac{\delta^{-\alpha}}{(2-\alpha)^2 \sqrt{n}} \\
:= \phi_n(\delta, B).
$$

With $\phi_n(\cdot, \cdot)$ defined, we will apply Theorem C.1 to find the tail probability bound.

**Application of Theorem C.1:** To apply Theorem C.1, we need $\beta, \gamma$, and $s_\gamma(\delta)$. From assumption (6), we can choose

$$
\gamma = q \quad \text{and} \quad s_\gamma(\delta) = \mathbb{E}(\|U(\epsilon, X, \delta)|^q) \leq 4^q \|\|e| + \Phi\|F_\delta\|^q \leq 4^q C^q \Phi^{2q} \delta^q.
$$

Then $s_\gamma(\delta)$ satisfies (29). We will choose $\beta$ later. Theorem C.1 now implies that

$$
\mathbb{P}\left(\|\hat{f} - f_0\| \geq \delta \varepsilon_n\right) \leq \left(\frac{C}{\sqrt{n}(\delta \varepsilon_n)^2}\right)^\beta \sum_{k=1}^\infty \phi_n^\beta(2kD \varepsilon_n, B_k) + \left(8C \sqrt{\beta(\sigma + \Phi)}\right)^\beta \\
+ \left(\frac{C \beta}{n(D \varepsilon_n)^2}\right)^\beta \sum_{k=1}^\infty \frac{B_k^\beta}{2^{2k\beta}} + \frac{16}{(D \varepsilon_n)^2} \sum_{k=0}^\infty s_\gamma(2kD \varepsilon_n) \frac{2^{2k}B_k^{q-1}}{2^{2k}B_k^{q}} \\
\leq \left(\frac{C}{\sqrt{n}(D \varepsilon_n)^2}\right)^\beta \sum_{k=1}^\infty \phi_n^\beta(2kD \varepsilon_n, B_k) + \left(8C \sqrt{\beta(\sigma + \Phi)}\right)^\beta \\
+ \left(\frac{C \beta}{n(D \varepsilon_n)^2}\right)^\beta \sum_{k=1}^\infty \frac{B_k^\beta}{2^{2k\beta}} + \frac{4^{q+2}C^q \Phi^{2q}}{(D \varepsilon_n)^2} \sum_{k=1}^\infty \left(\frac{2kD \varepsilon_n}{2^{2k}B_k^{q-1}}\right)^{q}.
$$

From the definition of $\phi_n(\cdot, \cdot)$, write $\phi_n(\cdot, \cdot) = A_1(\delta) + BA_2(\delta)$, where

$$
A_1(\delta) := \frac{KA^{1/2}(\sigma + \Phi)\delta^{1-\alpha/2}}{(2-\alpha)^2}, \quad \text{and} \quad A_2(\delta) := \frac{KA\delta^{-\alpha}}{n^{1/2}(2-\alpha)^2}.
$$

This implies

$$
\phi_n^\beta(2kD \varepsilon_n; B_k) \leq 2^\beta A_1^\beta(2kD \varepsilon_n) + 2^\beta B_k^\beta A_2^\beta(2kD \varepsilon_n).
$$

Hence

$$
\mathbb{P}\left(\|\hat{f} - f_0\| \geq \delta \varepsilon_n\right) \leq \sum_{k=1}^\infty \left(\frac{C A_1^\beta(2kD \varepsilon_n)}{2^{2k} \sqrt{n}(D \varepsilon_n)^2}\right)^\beta + \left(\frac{8C \sqrt{\beta(\sigma + \Phi)}\right)^\beta \\
+ \sum_{k=1}^\infty \frac{C \beta \beta}{2^{2k\beta}} \left(\frac{\beta}{n(D \varepsilon_n)^2} + \frac{A_2^\beta(2kD \varepsilon_n)}{\sqrt{n}(D \varepsilon_n)^2}\right)^\beta \\
+ \frac{4^{q+2}C^q \Phi^{2q}}{(D \varepsilon_n)^2} \sum_{k=1}^\infty \left(\frac{2kD \varepsilon_n}{2^{2k}B_k^{q-1}}\right)^{q},
$$

for a universal constant $C$. We now choose $B_k$ to balance the summands of the last two terms

$$
B_k \left(\frac{C \beta}{n(2^kD \varepsilon_n)^2} + \frac{C A_1^\beta(2kD \varepsilon_n)}{\sqrt{n}(2^kD \varepsilon_n)^2}\right) = \left(\frac{4^{q+2}C^q \Phi^{2q}}{(2^kD \varepsilon_n)^{2-q}B_k^{q-1}}\right)^{1/\beta}.
$$
Equivalently,

\[
B_k = \left( \frac{C\beta}{n(2kD\varepsilon_n)^2} + \frac{CA_2(2kD\varepsilon_n)}{\sqrt{n}(2kD\varepsilon_n)^2} \right)^{-\beta/(\beta+q-1)} \left( \frac{4^{q+2}Cq2^q}{(2kD\varepsilon_n)^2-q^2} \right)^{1/(\beta+q-1)}. \tag{36}
\]

Hence the last two terms in (35) become

\[
\begin{aligned}
\sum_{k=1}^{\infty} \frac{4^{q+2}Cq2^q}{(2kD\varepsilon_n)^2-q^2} &\left( \frac{C\beta}{n(2kD\varepsilon_n)^2} + \frac{CA_2(2kD\varepsilon_n)}{\sqrt{n}(2kD\varepsilon_n)^2} \right)^{\beta/(\beta+q-1)} \\
\sum_{k=1}^{\infty} \frac{4^{q+2}Cq2^q}{(2kD\varepsilon_n)^2-q^2} &\left( \frac{C\beta}{n(2kD\varepsilon_n)^2} + \frac{CA_2(2kD\varepsilon_n)}{\sqrt{n}(2kD\varepsilon_n)^2} \right)^{\beta/(\beta+q-1)} \\
\sum_{k=1}^{\infty} \frac{4^{q+2}Cq2^q}{(2kD\varepsilon_n)^2-q^2} &\left( \frac{C\beta}{n(2kD\varepsilon_n)^2} + \frac{CA_2(2kD\varepsilon_n)}{\sqrt{n}(2kD\varepsilon_n)^2} \right)^{\beta/(\beta+q-1)}.
\end{aligned}
\tag{37}
\]

where the last equality follows from the definition of \( A_2(\cdot) \) in (34). Substituting this in (35) and using the definition of \( A_1(\cdot) \), we get

\[
\begin{aligned}
P(\|\hat{f} - f_0\| \geq D\varepsilon_n) \\
\leq \sum_{k=1}^{\infty} \left( \frac{CKA^{1/2}(\sigma + \Phi)}{\sqrt{n}(2kD\varepsilon_n)^{1+\alpha/2}(2 - \alpha)} \right)^{\beta} + \left( \frac{8C\sqrt{3}(\sigma + \Phi)}{\sqrt{n}(D\varepsilon_n)} \right)^{\beta/2} \\
+ \sum_{k=1}^{\infty} \left( \frac{Ck^{q-1}4q^{2q}2^q}{(2kD\varepsilon_n)^{2-q^2}} \right)^{\beta/(\beta+q-1)} \left( \frac{C\beta}{n(2kD\varepsilon_n)^2} + \frac{CA_2(2kD\varepsilon_n)}{\sqrt{n}(2kD\varepsilon_n)^2} \right)^{(q-1)} \\
\leq \left( \frac{CKA^{1/2}(\sigma + \Phi)}{\sqrt{n}(D\varepsilon_n)^{1+\alpha/2}(2 - \alpha)} \right)^{\beta} + \left( \frac{8C\sqrt{3}(\sigma + \Phi)}{\sqrt{n}(D\varepsilon_n)} \right)^{\beta/2} \\
+ \left( \frac{Ck^{q-1}4q^{2q}2^q}{(2kD\varepsilon_n)^{2-q^2}} \right)^{\beta/(\beta+q-1)} \left( \frac{CA_2(2kD\varepsilon_n)}{n(2kD\varepsilon_n)^2+2(2kD\varepsilon_n)} \right)^{(q-1)}.
\end{aligned}
\tag{38}
\]

In the inequalities above the constant \( C \) could be different in different lines. Now choose \( \varepsilon_n \) so that the following inequalities are satisfied:

\[
\frac{A^{1/2}(\sigma + \Phi)}{\sqrt{n}\varepsilon_n^{1+\alpha/2}} \leq 1 \iff \varepsilon_n \geq A^{1/(2+\alpha)}(\sigma + \Phi)^{2/(2+\alpha)}n^{-1/(2+\alpha)},
\]

\[
\frac{(\sigma + \Phi)}{\sqrt{n}\varepsilon_n} \leq 1 \iff \varepsilon_n \geq (\sigma + \Phi)n^{-1/2},
\]

\[
\frac{Cq2^q}{n^{q-1}\varepsilon_n^{2-q^2}} \leq 1 \iff \varepsilon_n \geq C^{1/(2-s)}\varepsilon_n^{2-q^2}n^{-(q-1)/(q(2-s))},
\]

\[
\frac{A^{q-1}Cq2^q}{n^{q-1}\varepsilon_n^{2-q^2+(2+\alpha)(q-1)}} \leq 1 \iff \varepsilon_n \geq (A^{q-1}Cq2^q)^{1/(2-q^2+(2+\alpha)(q-1))}n^{-(2+\alpha+(2-q^2)/(q-1))}. 
\]

Take

\[
\varepsilon_n := \max \left\{ \frac{(\sigma + \Phi)^{2/(2+\alpha)}n^{\alpha/2}}{A^{1/(2+\alpha)}}, \frac{(\sigma + \Phi)n^{-1/2}}{C^{1/(2-s)}\varepsilon_n^{2-q^2+(2+\alpha)(q-1)}}, \frac{Cq2^q}{n^{q-1}/(q(2-s))}, \frac{A^{q-1}Cq2^q}{n^{1/(2+\alpha+(2-q^2)/(q-1))}} \right\},
\]
for which, the tail bound becomes

\[
\mathbb{P}(\|\hat{f} - f_0\| \geq D\varepsilon_n) \leq \left( \frac{CK}{D^{1+\alpha/2}(2-\alpha)} \right)^\beta + \left( \frac{8C\sqrt{\beta}}{D} \right)^\beta + \left( \frac{Cq^{-1}q^{-1}(q+2)^2}{D^{q(q-2)}} \right)^{\beta/(\beta+q-1)} + \left( \frac{CK(2-\alpha)q^{-1}(q+2)^2}{(2-\alpha)^2D^{q+2(q-1)}} \right)^{\beta/(\beta+q-1)}.
\]

Take \( \beta \) such that \( \beta \geq q \) and \( \beta q/(\beta + q - 1) \geq q \) or equivalently \( \beta \geq \max\{q, (q-1)/(1-s)\} \).

In case \( s = 1 \), fix any \( \eta > 0 \) and take \( \beta \) such that \( \beta \geq q \) and \( \beta q/(\beta + q - 1) = q - \eta \) or equivalently \( \beta \geq \max\{q, (q-1)(q-\eta)/\eta\} \). This would imply that for all \( D > 0 \),

\[
\mathbb{P}(\|\hat{f} - f_0\| \geq D\varepsilon_n) \leq CD^{-\eta+\eta1(s=1)},
\]

for a constant \( C > 0 \) depending only on \( q, s, \alpha, \) and \( \eta \). Because \( A \geq 1, \varepsilon_n^{-1} \) is equal to \( r_n \) in (8).

### D.1. Additional log factors in (6)

Suppose \( F_\delta \) and \( \epsilon \) do not satisfy (6) but satisfy

\[
\|(a| + \Phi)F_\delta(X)\|_q \leq C\Phi^2\delta^r \log^{r'}(1/\delta).
\]

Then, by modifying \( s_q \) in (33) and incorporating the changes in (35), (36), and (37), it can be shown that \( \hat{f} \) will satisfy the following modification of (38),

\[
\mathbb{P}\left(\|\hat{f} - f_0\| \geq D\varepsilon_n\right) \leq \left( \frac{CKA^{1/2}(\sigma + \Phi)}{\sqrt{n}(D\varepsilon_n)^{1+\alpha/2}(2-\alpha)} \right)^\beta + \left( \frac{8C\sqrt{\beta}(\sigma + \Phi)}{\sqrt{n}(D\varepsilon_n)} \right)^\beta + \left( \frac{Cq^{-1}q^{-1}q+2Cq^{2}(\log(1/D\varepsilon_n))^{1+\nu}}{n^{q-1}(D\varepsilon_n)^{2q+2(q-1)}} \right)^{\beta/(\beta+q-1)} + \left( \frac{CK(2-\alpha)q^{-1}(q+2)^2}{(2-\alpha)^2D^{q+2(q-1)}} \right)^{\beta/(\beta+q-1)}.
\]

We will now chose \( \varepsilon_n \) that satisfies the following inequalities

\[
\frac{A^{1/2}(\sigma + \Phi)}{n^{1+\alpha/2}} \leq 1 \iff \varepsilon_n \geq A^{1/(2+\alpha)}(\sigma + \Phi)^{2/(2+\alpha)}n^{-1/(2+\alpha)},
\]

\[
\frac{\sigma + \Phi}{\sqrt{n}\varepsilon_n} \leq 1 \iff \varepsilon_n \geq (\sigma + \Phi)n^{-1/2},
\]

\[
\frac{Cq^{2}(\log(1/\varepsilon_n))^{1+\nu}}{n^{q-1}(2-s)^{2q+2(q-1)}} \leq 1 \iff \varepsilon_n \geq C^{1/(2-s)}(\Phi^{2/(2-s)}n^{-(q-1)/4(q-2-s)})(\log n)^{\nu}/(q(2-s)),
\]

\[
\frac{A^{q-1}Cq^{2}(\log(1/\varepsilon_n))^{1+\nu}}{n^{q-1}(2-s)^{2q+2(q-1)}} \leq 1 \iff \varepsilon_n \geq (A^{q-1}Cq^{2}(\log n)^{1/2q+2q+2(q-1)}n^{-(q-1)/2+q+2(q-1)})(q(2-s))/4(q-2-s).
\]

Thus by choosing the \( \beta \) as before, we have that \( \hat{f} \) will satisfy (39) with

\[
\varepsilon_n := \max \left\{ \left( \frac{(\sigma + \Phi)^{1/(2+\alpha)}}{nA^{1/(2+\alpha)}}, \frac{C^{1/(2-s)}n^{1/(2+\alpha)}}{(q(2-s))^{1/(2q+2q+2q+2(q-1))}}, \frac{(A^{q-1}Cq^{2}(\log n)^{1/2q+2q+2q+2(q-1)}n^{-(q-1)/2+q+2(q-1)})^{1/2q+2q+2q+2(q-1)}}{n^{q-1}} \right\}.
\]
Appendix E: Proof of Theorem 3.1

As in the proof of Theorem 2.1, we will first find an appropriate \( \phi_n(\delta, B) \). We will use the maximal inequality derived in Proposition B.1 in conjunction with generic chaining [48] to find this. Recall that \( T(f, \epsilon, X) := 2\epsilon(f-f_0)(X)-(f-f_0)^2(X) \). By symmetrization and contraction (Theorem 3.1.21 and Corollary 3.2.2 of [21], respectively), we get that

\[
\phi_n(\delta, B) \leq \mathbb{E} \left( \sup_{f \in \mathcal{F}_n} G_nT(f; \epsilon, X, \delta) \right) \leq \mathbb{E} \left( \sup_{f \in \mathcal{F}_n} G_n(\epsilon + 16\Phi)(f - f_0) \right).
\]

Define a stochastic process \( X(\cdot) \) on \( \mathcal{F} \) as

\[
X(f) := G_n(\epsilon + 16\Phi)(f - f_0).
\]

We will now bound \( \mathbb{E}(\sup_{f \in \mathcal{F}_n} X(f)) \). Let \( \{f_0\} = S_0 \subset S_1 \subset \cdots \subset S_m \subset \cdots \) be a sequence of incremental subsets of \( \mathcal{F}_3 \). We take these sets \( S_i \) so that \( \log |S_{i+1}| \leq 2^{i+1} \). Let \( \epsilon_{2,i} \) denote the smallest \( \epsilon \) so that \( \log N(\epsilon_{2,i}, \mathcal{F}_3, \|\cdot\|) \leq 2^i \) and \( \epsilon_{\infty,i} \) denote the smallest \( \epsilon \) so that \( \log N(\epsilon_{\infty,i}, \mathcal{F}_3, \|\cdot\|_\infty) \leq 2^i \).

Now set \( A_i = \{f_1, \ldots, f_{2^{2^i}}\} \) as the \( \epsilon_{2,i} \)-net of \( \mathcal{F}_3 \) with respect to \( \|\cdot\| \) and \( B_i = \{g_1, \ldots, g_{2^{2^i}}\} \) as the \( \epsilon_{\infty,i} \)-net of \( \mathcal{F}_3 \) with respect to \( \|\cdot\|_\infty \). Define the partition of \( \mathcal{F}_3 \) by

\[
\{B_2(f_j, \epsilon_{2,i}) \cap B_{\infty}(g_k, \epsilon_{\infty,i}) : 1 \leq j, k \leq 2^{2^i}\}.
\]

The number of sets in this partition is bounded above by \( (2^{2^i})^2 = 2^{2^{i+1}} \). Take \( S_{i+1} \subset \mathcal{F}_3 \) of cardinality at most \( 2^{2^{i+1}} \) by taking one element in each of \( B_2(f_j, \epsilon_{2,i}) \cap B_{\infty}(g_k, \epsilon_{\infty,i}) \) for all \( 1 \leq j, k \leq 2^{2^i} \). For any \( f \in \mathcal{F}_3 \), we let \( \pi_if \) be the element in \( S_i \) that is closest to \( f \) so that

\[
\max_{f \in \mathcal{F}_3} \|f - \pi_if\| \leq \epsilon_{2,i-1} \quad \text{and} \quad \max_{f \in \mathcal{F}_3} \|f - \pi_if\|_\infty \leq \epsilon_{\infty,i-1}.
\]

Using these \( \pi_if \), we can write

\[
X(f) - X(f_0) = \sum_{i \geq 1} \{X(\pi_if) - X(\pi_{i-1}f)\}.
\]

Thus

\[
\mathbb{E} \left[ \sup_{f \in \mathcal{F}_3} X(f) \right] = \mathbb{E} \left[ \sup_{f \in \mathcal{F}_3} \{X(f) - X(f_0)\} \right] \leq \sum_{t \geq 1} \mathbb{E} \left[ \sup_{f \in \mathcal{F}_3} \{X(\pi_tf) - X(\pi_{t-1}f)\} \right].
\]

By symmetrization, for every \( t \geq 1 \) we have

\[
\mathbb{E} \left[ \sup_{f \in \mathcal{F}_3} \{X(\pi_tf) - X(\pi_{t-1}f)\} \right] = \mathbb{E} \left[ \sup_{f \in \mathcal{F}_3} \{G_n(\epsilon + 16\Phi)[\pi_tf - \pi_{t-1}f]\} \right] \leq 2 \frac{1}{\sqrt{n}} \mathbb{E} \left[ \sup_{f \in \mathcal{F}_3} \left| \sum_{i=1}^{n} R_i(\epsilon_i + 16\Phi)[\pi_tf(X_i) - \pi_{t-1}f(X_i)] \right| \right].
\]

Observe that the number of possible pairs \( (\pi_tf, \pi_{t-1}f) \) is bounded by \( |S_t||S_{t-1}| \). Let \( N_{t+1} = 2^{2^{i+1}} \) and \( \{(g_j, h_j)\}_{j=1}^{N_t} \) denote all such pairs. Thus in (41), the supremum is over a finite set. Thus

\[
\frac{1}{\sqrt{n}} \mathbb{E} \left[ \sup_{f \in \mathcal{F}_3} \left| \sum_{i=1}^{n} R_i(\epsilon_i + 16\Phi)[\pi_tf(X_i) - \pi_{t-1}f(X_i)] \right| \right] \leq \frac{1}{\sqrt{n}} \mathbb{E} \left[ \max_{1 \leq j \leq N_{t+1}} \left| \sum_{i=1}^{n} R_i(\epsilon_i + 16\Phi)[g_j(X_i) - h_j(X_i)] \right| \right].
\]
To bound the above expectation, we will use the maximal inequality in Proposition B.1 with
\[ x_{i,j} = (\epsilon_i + 16\Phi)[g_j(X_i) - h_j(X_i)] \quad \text{and} \quad \xi_i := \max_{1 \leq j \leq N_{t+1}} |(\epsilon_i + 16\Phi)[g_j(X_i) - h_j(X_i)]|. \]

Observe that
\[
\mathbb{E} \left[ \sum_{i=1}^{n} x_{i,j}^2 \right] \leq \mathbb{E} \left[ \sum_{i=1}^{n} (\epsilon_i + 16\Phi)[g_j(X_i) - h_j(X_i)]^2 \right] \leq (\sigma + 16\Phi)^2 n \|g_j - h_j\| \]

and
\[
\mathbb{E}[\xi_i^q] = \mathbb{E} \left[ \max_{1 \leq j \leq N_{t+1}} \left\{ (\epsilon_i + 16\Phi)^q [g_j(X_i) - h_j(X_i)] \right\}^q \right] \leq (\|\epsilon\|_q + 16\Phi)^q \max_{1 \leq j \leq N_{t+1}} \|g_j - h_j\|_q.
\]

Thus by Proposition B.1, we have that
\[
\frac{1}{\sqrt{n}} \mathbb{E} \left[ \max_{1 \leq j \leq N_{t+1}} \left\{ R_i(\epsilon_i + 16\Phi)[g_j(X_i) - h_j(X_i)] \right\} \right] \leq (\sigma + 16\Phi) \sqrt{\log N_{t+1}} \max_{1 \leq j \leq N_{t+1}} \|g_j - h_j\| (43)
\]

We will now bound \[\max_{1 \leq j \leq N_{t+1}} \|g_j - h_j\|\] and \[\max_{1 \leq j \leq N_{t+1}} \|g_j - h_j\|_\infty.\] By definition of \{(g_j, h_j)\}_{j=1}^{N_t}, we have that
\[
\max_{1 \leq j \leq N_{t+1}} \|g_j - h_j\| = \sup_{f \in F_t} \|f - \pi_t f\| \leq \sup_{f \in F_t} \|f - \pi_t f\| + \sup_{f \in F_t} \|f - \pi_{t-1} f\| \leq 2\varepsilon_{2,t-2}.
\]

Similarly, we also have
\[
\max_{1 \leq j \leq N_{t+1}} \|g_j - h_j\|_\infty \leq 2\varepsilon_{\infty,t-2}.
\]

Thus combining (41), (42), and (43), we have that
\[
\sum_{i \geq 1} \mathbb{E} \left[ \sup_{f \in F_t} |X(\pi_i f) - X(\pi_{i-1} f)| \right] \lesssim (\sigma + \Phi) \sum_{t \geq 1} 2^{t/2} \varepsilon_{2,t-2} + \varepsilon_{\infty,t-2} 2^{t/2} X(\pi_{i-1} f) \]
\[
\leq (\sigma + \Phi) \sum_{t \geq 1} 2^{t/2} \varepsilon_{2,t-2} + \frac{\|\epsilon\|_q + \Phi}{n^{1/2 - 1/q}} \sum_{t \geq 1} 2^{t(1 - 1/q)} \varepsilon_{\infty,t-2},
\]

where \[\varepsilon_{2,-1} = \delta\] and \[\varepsilon_{2,-1} = \|F_0\|_\infty.\] Thus
\[
\mathbb{E} \left[ \sup_{f \in F_t} \mathbb{G}_n T(f; \epsilon, X, \delta) \right] \leq \sum_{i \geq 1} \mathbb{E} \left[ \sup_{f \in F_t} |X(\pi_i f) - X(\pi_{i-1} f)| \right] \]
\[
\lesssim (\sigma + \Phi) \left( \delta + \sum_{t \geq 2} 2^{t/2} \varepsilon_{2,t-2} + \frac{\|\epsilon\|_q + \Phi}{n^{1/2 - 1/q}} 2^{t(1 - 1/q)} \varepsilon_{\infty,t-2} \right) \]
\[
\lesssim (\sigma + \Phi) \left( \delta + \sum_{t \geq 2} 2^{t/2} \varepsilon_{2,t} + \frac{\|\epsilon\|_q + \Phi}{n^{1/2 - 1/q}} 2^{t(1 - 1/q)} \varepsilon_{\infty,t} \right).
\]
By (2.37) of [49, Page 22], we have that
\[
\sum_{t \geq 0} 2^{t/2} \varepsilon_{2,t} \leq \int_0^\delta \sqrt{\log N(\eta, F_\delta, \| \cdot \|)} \, d\eta \quad \text{and} \quad \sum_{t \geq 0} 2^{t/2} \varepsilon_{\infty,t} \leq \int_0^{\|F_\delta\|_{\infty}} \{\log N(\eta, F_\delta, \| \cdot \|_{\infty})\}^{1-1/q} \, d\eta.
\]

Thus if \( \log N(\eta, F_\delta, \| \cdot \|_{\infty}) \leq A\eta^{-\alpha} \) for some \( \alpha \in [0, 2) \). Then
\[
\sum_{t \geq 0} 2^{t/2} \varepsilon_{2,t} \leq A^{1/2} \delta^{1-\alpha/2} / (1 - \alpha/2)
\]
and if \( \alpha(1 - 1/q) < 1 \) then
\[
\sum_{t \geq 0} 2^{t/2} \varepsilon_{\infty,t} \leq A^{1-1/q} \|F_\delta\|_{\infty}^{1-\alpha(1-1/q)}.
\]

Thus if \( \alpha(1 - 1/q) < 1 \), then
\[
\Phi_n(\delta, B) \leq C(\sigma + \Phi) \left( \delta + \frac{A^{1/2} \delta^{1-\alpha/2}}{1 - \alpha/2} \right) + C \|\epsilon\|_q + \Phi \left( \|F_\delta\|_{\infty} + A^{1-1/q} \|F_\delta\|_{\infty}^{1-\alpha(1-1/q)} \right)
\]
(44)
\[
\leq C(\sigma + \Phi) \left( \delta + \frac{A^{1/2} \delta^{1-\alpha/2}}{1 - \alpha/2} \right) + C \|\epsilon\|_q + \Phi \left( \Phi^{1-s} \delta^s + A^{1-1/q} \Phi^{1-s} \delta^s (1-\alpha(1-1/q)) \right).
\]

**Application of Theorem C.1:** To apply Theorem C.1, we need \( \beta, \gamma, \) and \( s_r(\delta) \). From assumption (10), we can choose
\[
\gamma = q \quad \text{and} \quad s_r(\delta) = E(\|U(\epsilon, X, \delta)\|^q) \leq 4^q \|\epsilon\|_q + \Phi \|F_\delta\|_{\infty}^q \leq C\Phi^{q+1-s} \delta^q.
\]
then \( s_q(\delta) \) satisfies (29). We will choose \( \beta \) later. Thus by (30), we have that
\[
P\left( \|\hat{f} - f_0\| \geq D \varepsilon_n \right) \leq \left( \frac{C}{\sqrt{n}(D \varepsilon_n)^2} \right)^\beta \sum_{k=1}^\infty \phi_n^\beta(2^k D \varepsilon_n, B_k) \left( 8C \sqrt{3}(\sigma + \Phi) \right)^\beta
\]
\[
+ \left( \frac{C}{n(D \varepsilon_n)^2} \right)^\beta \sum_{k=1}^\infty \frac{B_k^\beta}{2^{2k\beta}} \frac{16}{(D \varepsilon_n)^2} \sum_{k=1}^\infty \frac{s_q(2^k D \varepsilon_n)}{2^{2k} B_k^q}
\]
(45)
\[
\leq \left( \frac{C}{\sqrt{n}(D \varepsilon_n)^2} \right)^\beta \sum_{k=1}^\infty \phi_n^\beta(2^k D \varepsilon_n, B_k) \left( 8C \sqrt{3}(\sigma + \Phi) \right)^\beta
\]
\[
+ \left( \frac{C}{n(D \varepsilon_n)^2} \right)^\beta \sum_{k=1}^\infty \frac{B_k^\beta}{2^{2k\beta}} \frac{4^{q+2} C^q \Phi^{2q}}{(D \varepsilon_n)^2} \sum_{k=1}^\infty \frac{(2^k D \varepsilon_n)^q}{2^{2k} B_k^{q-1}}.
\]

We will now choose \( B_k \) to minimize the upper bound and then choose the smallest \( \varepsilon_n \) such that the right hand side is a does not depend on \( n \) and goes to zero as \( D \) increases to infinity. Substituting (44) in the probability bound (45), we get
\[
P\left( \|\hat{f} - f_0\| \geq D \varepsilon_n \right) \leq \sum_{k=1}^\infty \left( \frac{C A_1(2^k D \varepsilon_n)}{2^{2k} \sqrt{n}(D \varepsilon_n)^2} \right)^\beta \left( 8C \sqrt{3}(\sigma + \Phi) \right)^\beta
\]
\[
+ \sum_{k=1}^\infty C^\beta B_k^\beta \left( \frac{\beta}{n(D \varepsilon_n)^2} \right)^\beta + 4^{q+2} C^q \Phi^{2q} \sum_{k=1}^\infty \frac{(2^k D \varepsilon_n)^q}{2^{2k} B_k^{q-1}},
\]
(46)
for some universal constant \( C \) and
\[
A_1(\delta) := C(\sigma + \Phi) \left( \delta + \frac{A^{1/2} \delta^{1-\alpha/2}}{1 - \alpha/2} \right) + C \|\epsilon\|_q + \Phi \left( \Phi^{1-s} \delta^s + A^{1-1/q} \Phi^{1-s} \delta^s (1-\alpha(1-1/q)) \right).
\]
We now see that the choice of $B_k$ that balances the summands of the last two terms is

$$B_k = \left(\frac{C\beta}{n(2^k D\varepsilon_n)^2}\right)^{-\beta/((\beta+q-1))} \left(\frac{4^q C\Phi 2^q}{(2^k D\varepsilon_n)^2-q^s}\right)^{1/(\beta+q-1)}.$$

Hence the last two terms in (46) become

$$\sum_{k=1}^{\infty} \frac{A_{q+2} C\Phi 2^q}{(2^k D\varepsilon_n)^2-q^s} \left(\frac{C\beta}{n(2^k D\varepsilon_n)^2}\right)^{\beta(q-1)/(\beta+q-1)}$$

$$+ \sum_{k=1}^{\infty} \frac{A_{q+2} C\Phi 2^q}{(2^k D\varepsilon_n)^2-q^s} \left(\frac{C\beta}{n(2^k D\varepsilon_n)^2}\right)^{\beta(q-1)/(\beta+q-1)}.$$

Substituting this in the probability bound in (46) and simplifying, we obtain

$$\mathbb{P}\left(\|\hat{f} - f_0\| \geq D\varepsilon_n\right) \leq \sum_{k=1}^{\infty} \left(\frac{CA_1 (2^k D\varepsilon_n)^2}{2^k \sqrt{n(D\varepsilon_n)^2}}\right)^{\beta} + \left(\frac{8C\sqrt{\beta(\sigma + \Phi)}}{\sqrt{n(D\varepsilon_n)^2}}\right)^{\beta}$$

$$+ \sum_{k=1}^{\infty} \left(\frac{A_{q+2} C\Phi 2^q}{(2^k D\varepsilon_n)^2-q^s}\right)^{\beta/((\beta+q-1))} \left(\frac{C\beta}{n(2^k D\varepsilon_n)^2}\right)^{\beta(q-1)/(\beta+q-1)}.$$

Now we use the definitions of $A_1(\cdot)$ in (47), to get

$$\mathbb{P}\left(\|\hat{f} - f_0\| \geq D\varepsilon_n\right)$$

$$\leq \sum_{k=1}^{\infty} \left(\frac{C(\sigma + \Phi)}{\sqrt{n(2^k D\varepsilon_n)^2}}\right)^{\beta} + \sum_{k=1}^{\infty} \left(\frac{C\sqrt{\beta(\sigma + \Phi)}}{(1-\alpha/2)\sqrt{n(2^k D\varepsilon_n)^2}}\right)^{\beta} + \sum_{k=1}^{\infty} \left(\frac{\Phi^{1-\alpha} (\|\epsilon\|_q + \Phi)}{n^{1-\alpha}(2^k D\varepsilon_n)^2-s}\right)^{\beta}$$

$$+ \sum_{k=1}^{\infty} \left(\frac{A_{q+2} C\Phi 2^q}{(2^k D\varepsilon_n)^2-q^s}\right)^{\beta} \left(\frac{\Phi 2^q}{n^{2-q}(2^k D\varepsilon_n)^2-q^s}\right)^{\beta/(\beta+q-1)}.$$

We now choose $\varepsilon_n$ so that the following inequalities are satisfied:

$$\frac{A^{1/2}(\sigma + \Phi)}{n^{1/2} \varepsilon_n^{1/2}} \leq 1 \iff \varepsilon_n \geq \frac{A^{1/2}(\sigma + \Phi) 2^{1/2\alpha}}{n^{1/(2\alpha)}}$$

$$\frac{\Phi^{1-\alpha} (\|\epsilon\|_q + \Phi)}{n^{1-\alpha}(2^k D\varepsilon_n)^2-s} \leq 1 \iff \varepsilon_n \geq \frac{\Phi}{n^{(q-1)/(2-s)}}$$

$$\frac{A^{1-\alpha} (\sigma + \Phi)(1-\alpha(1-q)/2)}{n^{1-\alpha}(2^k D\varepsilon_n)^2} \leq 1 \iff \varepsilon_n \geq \frac{\Phi 2^{q-\alpha}}{n^{(q-1)/(2-s)}}$$

$$\frac{C\Phi 2^q}{n^{2-q} \varepsilon_n^{2-q}} \leq 1 \iff \varepsilon_n \geq \frac{C\Phi 2^{q-2}}{n^{(q-1)/(2-s)}}.$$
Define
\[ \epsilon_n := \max \left\{ \left( \frac{(\sigma + \Phi)^{2/(2+\alpha)}}{(nA^{-1})^{1/(2+\alpha)}} \right)^{1/(2-s)} \cdot \left( \frac{\phi^{2/(2-s)}}{n^{1/2}} \cdot \left( \frac{1}{n} \phi^{2/(2-s)} \right)^{1/(q(2-s)+\alpha s(q-1))} \right)^{1/(2-s)} \cdot \left( \frac{\beta(\alpha s(q-1) - \alpha(q-1))}{n^{q-1}} \right) \right\}. \]

From the definition of \( \epsilon_n \), the probability bound simplifies to
\[ \mathbb{P}\left( \| \hat{f} - f_0 \| \geq D\epsilon_n \right) \leq D^{-\beta(1+\alpha/2)} + D^{-\beta(2-s)} + D^{-(2-s+\alpha s(1-1/\eta))} + \beta\epsilon D^{-\beta} + \frac{\beta^2(q-1)/(\beta + q-1)}{D(2q-qs)} \]

Choose \( \beta \) such that \( \beta \geq q \) and \( \{(2q-qs)(\beta + q-1) \geq q \) or equivalently \( \beta \geq \max\{q, (q-1)/(1-s)\} \). If \( s = 1 \) choose \( \beta \) such that \( \beta \geq q - \eta \) and \( (2q-qs)(\beta + q-1)/(\beta + q-1) \geq q - \eta \) or equivalently \( \beta = \max\{q - \eta, (q-1)(q - \eta)/\eta\} \). This choice of \( \beta \) implies that
\[ \mathbb{P}(\| \hat{f} - \beta q \| \geq D\epsilon_n) \leq C^q D^{-q+\eta \{s=1\}}, \]

for a constant \( C > 0 \) depending only on \( q, s, \alpha, \) and \( \eta \). Because \( A \geq 1, r_n = \epsilon_n^{-1} \).

**Appendix F: Proof of Theorem 4.1**

Just as in the proofs of the earlier theorems, we will first find an appropriate \( \phi_n(\cdot, \cdot) \) and then apply Theorem C.1 with the right choice of \( \beta \) and \( s_2(\cdot); \gamma = 2 \) here. Note that by definition of \( \Phi \) and \( F_\delta \), we have that
\[ U(\epsilon, X, \delta) := 2(|\epsilon| + \Phi)F_\delta \]
satisfies (27). Recall that for this theorem, we have assumed that \( \epsilon \) has only two moments and \( \mathbb{E}(|\epsilon|^2) \leq \sigma^2 \). The following lemma proved in Section F.1 finds the bound \( \phi_n(\cdot, \cdot) \) for various values of \( \delta \) and \( B \).

**Lemma F.1 (Bound on \( \phi_n(\cdot, \cdot) \)).** Let \( \Phi(\cdot, \cdot) \) be as defined in (28). If \( \epsilon \) and \( F \) satisfy the assumptions of Theorem 4.1, then there exists a constant depending only on \( \alpha \geq 0 \) and \( \beta > 0 \) such that

1. If \( \alpha \in [0, 2) \) and \( \beta \geq 0 \), then
\[ \phi_n(\delta, B) \leq C A^{1/2}(\sigma + \Phi)\Phi^\alpha, \text{ for every } \delta, B > 0. \]
2. If \( \alpha = 2 \) and \( \beta = 0 \), then
\[ \phi_n(\delta, B) \leq C A^{1/2}(\sigma + \Phi)\Phi^\alpha \log(A^{-1}n), \text{ for all } n \geq A \text{ and every } \delta, B > 0. \]
3. If \( \alpha > 2 \) and \( \beta = 0 \), then
\[ \phi_n(\delta, B) \leq C^{1/2} A^{1/2-1/\alpha}(\sigma + \Phi)\Phi^\alpha, \text{ for all } n \geq A \text{ and every } \delta, B > 0. \]

**Application of Theorem C.1:** To apply Theorem C.1, we need \( \beta, \gamma, \) and \( s_\gamma(\delta) \). If we choose
\[ \beta = \gamma = 2 \quad \text{and} \quad s_2(\delta) = \mathbb{E}(|U(\epsilon, X, \delta)|^2) = 2^2\sigma^2\Phi^2\Phi^2, \]
then \( s_2(\delta) \) satisfies (29). Thus by (30), we have that
\[ \mathbb{P}\left( \| \hat{f} - f_0 \| \geq D\epsilon_n \right) \leq \left( \frac{C}{\sqrt{n}(D\epsilon_n)^2} \right)^2 \sum_{k=1}^{\infty} \frac{\phi_n^2(2^k D\epsilon_n, B_k)}{2^{2k}} + \left( \frac{C\sqrt{2}(\sigma + \Phi)}{\sqrt{n}D\epsilon_n} \right)^2 \sum_{k=1}^{\infty} \frac{1}{2^{2k}} \]
\[ + \left( \frac{2C}{n(D\epsilon_n)^2} \right)^2 \sum_{k=1}^{\infty} \frac{B_k^2}{2^{4k}} + \frac{16}{(D\epsilon_n)^2} \sum_{k=0}^{\infty} \frac{s_2(2^k D\epsilon_n)}{2^{2k} B_k}. \]
We will now choose \( \{B_k\} \) so as to minimize the right hand side. In contrast to the earlier proofs the \( \phi_n(\delta, B) \) obtained in Lemma F.1 do not depend on \( B \). Thus the choice of \( \{B_k\} \) that minimizes the bound in (51) is the value for which the last two terms are equal order for every \( k \geq 0 \), i.e.,

\[
\left( \frac{1}{n(D\epsilon_n)^2} \right)^{2} \frac{B_k^2}{24k^k} = \frac{\sigma^2 \Phi^2}{(D\epsilon_n)^2} \frac{(2^k D\epsilon_n)^{2s}}{2^{2k} B_k}.
\]

Thus, we choose

\[
B_k = \left[ n^2(D\epsilon_n)^{2(2+s)} \sigma^2 \Phi^2 2^{2k(1+2s)} \right]^{1/3}.
\]

Then

\[
\left( \frac{2C}{n(D\epsilon_n)^2} \right)^{2} \sum_{k=1}^{\infty} \frac{B_k^2}{24k} = \left( \frac{2C}{n(D\epsilon_n)^2} \right)^{2} \left[ n^2(D\epsilon_n)^{2(2+s)} \sigma^2 \Phi^2 \right]^{2/3} \sum_{k=1}^{\infty} \frac{1}{2^{8k(1+s)/3}} = \frac{4C^2 \sigma^3 \Phi^3}{n^{2/3}(D\epsilon_n)^{4(2-s)/3}}.
\]

Substituting this in (51), we get

\[
\mathbb{P} \left( \| \hat{f} - f_0 \| \geq D\epsilon_n \right) \\
\leq \left( \frac{C}{\sqrt{n}(D\epsilon_n)^2} \right)^{2} \sum_{k=1}^{\infty} \frac{\phi_n^2(2^k D\epsilon_n, B_k)}{24k} + \frac{\left( C \sqrt{2(\sigma + \Phi)} \right)^{2}}{\sqrt{n}D\epsilon_n} + C \frac{4C^2 \sigma^3 \Phi^3}{n^{2/3}(D\epsilon_n)^{4(2-s)/3}}
\]

\[
\leq C \left[ \left( \frac{1}{\sqrt{n}(D\epsilon_n)^2} \right)^{2} \sum_{k=1}^{\infty} \frac{\phi_n^2(2^k D\epsilon_n, B_k)}{24k} + \frac{\left( (\sigma + \Phi) \right)^{2}}{\sqrt{n}D\epsilon_n} + \frac{\sigma^3 \Phi^3}{n^{2/3}(D\epsilon_n)^{4(2-s)/3}} \right],
\]

where \( C \) is a universal constant. We will now compute the bound in the (52) for all the three cases in Lemma F.1 by substituting the appropriate values of \( \phi_n(\delta, \cdot) \). To aid in this calculation observe that each of the bound on \( \Phi_n(\delta, B) \) in Lemma F.1 is of the form \( h(A, \Phi, n, \alpha, \beta)\delta^s \), where \( h(A, \Phi, n, \alpha, \beta) \) is a function of it inputs. Thus (52) can be rewritten as

\[
\mathbb{P} \left( \| \hat{f} - f_0 \| \geq D\epsilon_n \right) \\
\leq C \left[ \left( \frac{1}{\sqrt{n}(D\epsilon_n)^2} \right)^{2} \sum_{k=1}^{\infty} \frac{\phi_n^2(2^k D\epsilon_n, B_k)}{24k} + \frac{\left( (\sigma + \Phi) \right)^{2}}{\sqrt{n}D\epsilon_n} + \frac{\sigma^3 \Phi^3}{n^{2/3}(D\epsilon_n)^{4(2-s)/3}} \right]
\]

\[
\leq C \left[ \frac{h^2(A, \Phi, n, \alpha, \beta)}{n(D\epsilon_n)^{2(2-s)}} + \frac{\left( (\sigma + \Phi) \right)^{2}}{\sqrt{n}D\epsilon_n} + \frac{\sigma^3 \Phi^3}{n^{2/3}(D\epsilon_n)^{4(2-s)/3}} \right],
\]

where the last inequality is true as \( s \leq 1 \). We will find a bound on the tail probability of the LSE and the upper bound on the rate of the LSE by substituting the appropriate \( h(A, \Phi, n, \alpha, \beta) \) for each of the cases. We do this below.

**Case 1:** \( \alpha \in [0, 2) \) and \( \beta \geq 0 \) Because we will allow \( A \) and \( \Phi \) to depend on \( n \) and assume \( \sigma, \alpha, \) and \( \beta \) to be constants. We will assume that without loss of generality that \( \sigma \leq \Phi \). Substituting (48)
in (53), we have that
\[
P \left( \| \hat{f} - f_0 \| \geq D \varepsilon_n \right) 
\leq C \left[ \frac{A^{1/2} \Gamma(1 + \beta/2)(1 - \alpha/2)^{1 + \beta/2}(\sigma + \Phi)}{n(D \varepsilon_n)^{2(2-s)}} + \left( \frac{(\sigma + \Phi)}{\sqrt{n}D \varepsilon_n} \right)^2 + \frac{\sigma^3 \Phi^3}{n^{2/3}(D \varepsilon_n)^{4(2-s)/3}} \right]
\]
where \( C \) is a constant depending only \( \alpha, \beta, \) and \( \sigma \). Now choose,
\[
\varepsilon_n := \max \left\{ \left[ \frac{A \Phi^2}{n} \right]^{1/2(2-s)}, \frac{\Phi}{\sqrt{n}}, \left[ \frac{\Phi^{9/2}}{n} \right]^{1/2(2-s)} \right\}
\]
With the above choice of \( \varepsilon \), we have that
\[
P \left( \| \hat{f} - f_0 \| \geq D \varepsilon_n \right) \leq \frac{3C}{D^{4(2-s)/3}}.
\] (55)

If we fix \( A \) and \( \Phi \) then it is clear that the rate of convergence of the LSE is no worse than \( n^{1/(2(2-s))} \).

**Case 2:** \([\alpha = 2 \text{ and } \beta = 0]\) Because we will allow \( A \) and \( \Phi \) to depend on \( n \) and assume \( \sigma \) to be a constant. We will assume that without loss of generality that \( \sigma \leq \Phi \). Substituting (49) in (53), we have that
\[
P \left( \| \hat{f} - f_0 \| \geq D \varepsilon_n \right) 
\leq C \left[ \frac{\{ A^{1/2}(\sigma + \Phi) \log (A^{-1} n) \}^2}{n(D \varepsilon_n)^{2(2-s)}} + \left( \frac{(\sigma + \Phi)}{\sqrt{n}D \varepsilon_n} \right)^2 + \frac{\sigma^3 \Phi^3}{n^{2/3}(D \varepsilon_n)^{4(2-s)/3}} \right]
\]
where \( C \) is a constant depending only \( \alpha, \beta, \) and \( \sigma \). Then it is easy to see that \( \hat{f} \) satisfies (55) if
\[
\varepsilon_n := \max \left\{ \left[ \frac{A \Phi^2 \log^2 (A^{-1} n)}{n} \right]^{1/2(2-s)}, \frac{\Phi}{\sqrt{n}}, \left[ \frac{\Phi^{9/2}}{n} \right]^{1/2(2-s)} \right\}
\] (56)
If we fix \( A \) and \( \Phi \) then it is clear that the rate of convergence of the LSE is no worse than \((\sqrt{n} / \log n)^{1/(2-s)}\).

**Case 3:** \([\alpha > 2 \text{ and } \beta = 0]\) Because we will allow \( A \) and \( \Phi \) to depend on \( n \) and assume \( \sigma \) to be a constant. We will assume that without loss of generality that \( \sigma \leq \Phi \). Substituting (50) in (53), we have that
\[
P \left( \| \hat{f} - f_0 \| \geq D \varepsilon_n \right) 
\leq C \left[ \frac{\{ A^{1/\alpha}(\sigma + \Phi)n^{1/2 - 1/\alpha} \}^2}{n(D \varepsilon_n)^{2(2-s)}} + \left( \frac{(\sigma + \Phi)}{\sqrt{n}D \varepsilon_n} \right)^2 + \frac{\sigma^3 \Phi^3}{n^{2/3}(D \varepsilon_n)^{4(2-s)/3}} \right]
\]
where \( C \) is a constant depending only \( \alpha, \beta, \) and \( \sigma \). Then it is easy to see that \( \hat{f} \) satisfies (55) if
\[
\varepsilon_n := \max \left\{ \left[ \frac{A \Phi^2 n^{1-2/\alpha}}{n(D \varepsilon_n)^{2(2-s)}} \right]^{1/2(2-s)}, \frac{\Phi}{\sqrt{n}}, \left[ \frac{\Phi^{9/2}}{n} \right]^{1/2(2-s)} \right\}
\] (56)
where $C$ is a constant depending only on $\alpha, \beta,$ and $\sigma$. Then it is easy to see that $\hat{f}$ satisfies (55) if

$$\varepsilon_n := \max \left\{ \frac{A^{2/\alpha} \Phi^2}{n^{2/\alpha}}, \frac{\Phi}{\sqrt{n}}, \frac{\Phi^{2/\alpha}}{n^{1/\alpha(2-s)}} \right\}.$$

(57)

As in the previous cases, if we fix $A$ and $\Phi$, the rate of convergence of the LSE is no worse than $n^{1/\alpha(2-s)}$.

Finally, for each of the above cases $s \leq 1$, thus we have that $\varepsilon_n^{-1} E\|\hat{f} - f_0\| = O(1)$.

### F.1. Proof of Lemma F.1

The proof is based on the use of symmetrization by Rademacher variables followed by application of the sub-Gaussian maximal inequality given by Corollary 2.2.8 of [55] conditionally on $\{\epsilon_i, X_i\}, 1 \leq i \leq n$. Recall that

$$\phi_n(\delta; B) = E \left[ \sup_{\delta/2 \leq \|f - f_0\| \leq \delta} \mathcal{G}_n(T_B(f; \epsilon, X, \delta)) \right].$$

By Symmetrization (Corollary 3.2.2 of [21]), we get

$$\phi_n(\delta; B) \leq 2E \left[ \sup_{f \in \mathcal{F}_3} \frac{1}{\sqrt{n}} \sum_{i=1}^n R_i T_B(f; \epsilon_i, X_i, \delta) \right],$$

where $R_1, \ldots, R_n$ are i.i.d. Rademacher random variables independent of $(\epsilon_1, X_1), \ldots, (\epsilon_n, X_n)$. Now by Lemma the A.1 of [45] (also see [11, Theorem 3.2]), we have

$$\phi_n(\delta; B) \leq E \left[ \inf_{\gamma \geq 0} 4n^{1/2} \gamma + 10 \int_{\gamma}^{\eta_n} \sqrt{\log(\eta, \mathcal{G}_n(T_B(f; \epsilon, X, \delta))}, \right],$$

(58)

where $\|g\|^2_n := n^{-1} \sum_{i=1}^n g^2(\epsilon_i, X_i)$, for $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, and $\eta_n := \sup_{f \in \mathcal{F}_3} \|T_B(f)\|_n$. It is clear that

$$|T_B(f_1; \epsilon, X) - T_B(f_2; \epsilon, X)| \leq (2|\epsilon| + 4M)(|f_1(X) - f_2(X)| \mathbb{I}\{U(\epsilon, X, \delta) \leq B\}).$$

(59)

Define a measure $Q$ on $\{X_1, \ldots, X_n\}$ as

$$Q(\{X_i\}) = \frac{(2|\epsilon_i| + 4M)^2 \mathbb{I}\{U(\epsilon_i, X_i; \delta) \leq B\}}{\sum_{j=1}^n (2|\epsilon_j| + 4M)^2 \mathbb{I}\{U(\epsilon_j, X_j; \delta) \leq B\}}, \quad 1 \leq i \leq n.$$

From inequality (59), we get

$$\|T_B(f_1; \epsilon, X) - T_B(f_2; \epsilon, X)\|_n \leq \|(2|\epsilon| + 4M) \mathbb{I}\{U(\epsilon, X, \delta) \leq B\}\|_n \|f_1 - f_2\|_{2,Q}.$$  

Thus it follows that for any $\eta > 0$,

$$\log N(\|(2|\epsilon| + 4M) \mathbb{I}\{U(\epsilon, X, \delta) \leq B\}\|_n, \mathcal{G}_n(T_B(f))), \| \cdot \|_n)
\leq \log N(\eta, \mathcal{F}_3, L_2(Q)) \leq A \left( \frac{\eta}{\|F_3\|_{2,Q}} \right)^{-\alpha} \log^\beta \left( \frac{\|F_3\|_{2,Q}}{\eta} \right) \quad \text{for all } \eta > 0.$$

(60)

Hence using the fact $\|U(\epsilon, X, \delta) \mathbb{I}\{U(\epsilon, X, \delta) \leq B\}\|_n = \|(2|\epsilon| + 4M) \mathbb{I}\{U(\epsilon, X, \delta) \leq B\}\|_n \|F_3\|_{2,Q}$, (60) yields

$$\log N(\|(2|\epsilon| + 4M) \mathbb{I}\{U(\epsilon, X, \delta) \leq B\}\|_n, \mathcal{F}_3, \| \cdot \|_n) \leq A \eta^{-\alpha} \log^\beta (1/\eta) \quad \text{for all } \eta > 0.$$
Substituting this bound in (58), we get

$$
\phi_n(\delta; B) \leq \mathbb{E} \left[ \|U(\epsilon, X; \delta)\|_{\text{n}} \inf_{\gamma \geq 0} 4\sqrt{n} \gamma + 10A^{1/2} \int_{\gamma}^{\Theta_n} \eta^{-\alpha/2} \log^{\beta/2}(1/\eta) d\eta \right],
$$

where $\Theta_n := \sup_{f \in F} \|T_B(f; \epsilon, \delta)\|_{\text{n}}/\|U(\epsilon, X; \delta)\|_{\text{n}}$. Because $\Theta_n \leq 1$ and the infimum is a non-random quantity, we obtain

$$
\phi_n(\delta; B) \leq \mathbb{E} \left[ \|U(\epsilon, X; \delta)\|_{\text{n}} \inf_{\gamma \geq 0} 4\sqrt{n} \gamma + 10A^{1/2} \int_{\gamma}^{1} \eta^{-\alpha/2} \log^{\beta/2}(1/\eta) d\eta \right]
$$

$$
\leq \|U(\epsilon, X; \delta)\| \inf_{\gamma \geq 0} 4\sqrt{n} \gamma + 10A^{1/2} \int_{\gamma}^{1} \eta^{-\alpha} \log^{\beta/2}(1/\eta) d\eta
$$

$$
\leq C(\sigma + \Phi)\Phi \delta^s G_n,
$$

where $C$ is a universal constant and

$$
G_n := \inf_{\gamma \geq 0} \sqrt{n} \gamma + A^{1/2} \int_{\gamma}^{1} \eta^{-\alpha/2} \log^{\beta/2}(1/\eta) d\eta.
$$

We now complete the proof by bounding $G_n$ separately in each of the three cases.

Proof of (48): If $\alpha < 2$ and $\beta > 0$, then we can take $\gamma = 0$ in the infimum of $G(\cdot)$ and using $\int_{0}^{1} \eta^{-\alpha/2} \log^{\beta/2}(1/\eta) d\eta < \infty$, we get

$$
\phi_n(\delta; B) \leq CA^{1/2}(\sigma + \Phi)\Phi \delta^s,
$$

for a constant $C > 0$ depending only on $\alpha, \beta$.

Proof of (49): If $\alpha = 2$ and $\beta = 0$, taking $\gamma = (A^{-1}n)^{-1/2}$ yields

$$
G_n \leq A^{1/2} + A^{1/2} \int_{(A^{-1}n)^{-1/2}}^{1} \eta^{-1} d\eta = 1 + A^{1/2} \log(A^{-1/2}n^{1/2}) \leq A^{1/2} \log(A^{-1}n),
$$

and hence $\phi_n(\delta; B) \leq CA^{1/2}(\sigma + \Phi)\Phi \delta^s \log(A^{-1}n)$.

Proof of (50): If $\alpha > 2$ and $\beta = 0$, taking $\gamma = (A^{-1}n)^{-1/\alpha}$ yields

$$
G_n \leq A^{1/\alpha} n^{1/2 - 1/\alpha} + A^{1/2} \int_{A^{1/\alpha} n^{-1/\alpha}}^{1} \eta^{-\alpha/2} d\eta
$$

$$
= A^{1/\alpha} n^{1/2 - 1/\alpha} + \frac{A^{1/2}}{\alpha/2 - 1} [(A^{-1}n)^{1/2 - 1/\alpha} - 1] \leq CA^{1/\alpha} n^{1/2 - 1/\alpha},
$$

for some constant $C > 0$ depending only on $\alpha$. Hence

$$
\phi_n(\delta; B) \leq CA^{1/\alpha} n^{1/2 - 1/\alpha}(\sigma + \Phi)\Phi \delta^s.
$$

References

[1] Agmon, S. (2010). Lectures on elliptic boundary value problems. AMS Chelsea Publishing, Providence, RI. Prepared for publication by B. Frank Jones, Jr. with the assistance of George W. Batten, Jr., Revised edition of the 1965 original.

[2] Audibert, J.-Y. and O. Catoni (2011). Robust linear least squares regression. The Annals of Statistics 39(5), 2766–2794.

[3] Bartlett, P. L. (2008). Fast rates for estimation error and oracle inequalities for model selection. Econometric Theory 24(2), 545–552.
[4] Beliakov, G. and M. Kohler (2005). Estimation of regression functions by Lipschitz continuous functions. Unpublished manuscript.
[5] Bélanger, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. The Annals of Statistics 46(2), 745–780.
[6] Birgé, L. and P. Massart (1993). Rates of convergence for minimum contrast estimators. Probability Theory and Related Fields 97(1-2), 113–150.
[7] Birgé, L. and P. Massart (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. Bernoulli 4(3), 329–375.
[8] Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. Ann. Statist. 17(2), 453–555.
[9] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. The Annals of Statistics 42(6), 2340–2381.
[10] Chatterjee, S., A. Guntuboyina, and B. Sen (2015). On risk bounds in isotonic and other shape restricted regression problems. The Annals of Statistics 43(4), 1774–1800.
[11] Chatterjee, S., A. Guntuboyina, and B. Sen (2018). On matrix estimation under monotonicity constraints. Bernoulli 24(2), 1072–1100.
[12] Chatterjee, S. and J. Lafferty (2015). Adaptive risk bounds in unimodal regression. arXiv preprint arXiv:1512.02956.
[13] Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. Econometrica, 289–314.
[14] Chernozhukov, V., D. Chetverikov, and K. Kato (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. Probab. Theory Related Fields 162(1-2), 47–70.
[15] de la Peña, V. H. and E. Giné (1999). Decoupling. Probability and its Applications (New York). Springer-Verlag, New York. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.
[16] Doss, C. R. (2015). Bracketing Numbers of Convex Functions on Polytopes. arXiv preprint arXiv:1506.00034.
[17] Friedman, J. H. and W. Stuetzle (1981). Projection pursuit regression. Journal of the American Statistical Association 76(376), 817–823.
[18] Gao, C., F. Han, and C.-H. Zhang (2017). On Estimation of Isotonic Piecewise Constant Signals. arXiv preprint arXiv:1705.06386.
[19] Giné, E. and V. Koltchinskii (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. Ann. Probab. 34(3), 1143–1216.
[20] Giné, E., R. Latała, and J. Zinn (2000). Exponential and moment inequalities for U-statistics. In High dimensional probability, II (Seattle, WA, 1999), Volume 47 of Progr. Probab., pp. 13–38. Birkhäuser Boston, Boston, MA.
[21] Giné, E. and R. Nickl (2016). Mathematical foundations of infinite-dimensional statistical models. Cambridge Series in Statistical and Probabilistic Mathematics, 40. Cambridge University Press, New York.
[22] Guntuboyina, A. and B. Sen (2015). Global risk bounds and adaptation in univariate convex regression. Probability Theory and Related Fields 163(1-2), 379–411.
[23] Guntuboyina, A. and B. Sen (2018). Nonparametric shape-restricted regression. Statistical Science 33(4), 568–594.
[24] Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). A distribution-free theory of nonparametric regression. Springer Series in Statistics. Springer-Verlag, New York.
[25] Han, Q. and J. A. Wellner (2018). Robustness of shape-restricted regression estimators: an envelope perspective. ArXiv preprints arxiv:1805.02542.
[26] Han, Q. and J. A. Wellner (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. Ann. Statist. 47, 2286 – 2319.
[27] Hastie, T., R. Tibshirani, and M. Wainwright (2015). Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC.
[28] Holloway, C. A. (1979). On the estimation of convex functions. *Operations Research* 27(2), 401–407.
[29] Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* 29(6), 1537–1566.
[30] Kolmogorov, A. N. (1949). On inequalities between the upper bounds of the successive derivatives of an arbitrary function on an infinite interval. *American Mathematical Society Translations* (1-2), 233–243.
[31] Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics* 34(6), 2593–2656.
[32] Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*, Volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d’Été de Probabilités de Saint-Flour.
[33] Kuchibhotla, A. K. and A. Chakrabortty (2018). Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*.
[34] Kuchibhotla, A. K., R. K. Patra, and B. Sen (2017). Least Squares Estimation in a Single Index Model with Convex Lipschitz link. *arXiv preprints arXiv:1708.00145v3 [math.ST]*.
[35] Lugosi, G. and S. Mendelson (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*.
[36] Massart, P. (2007). *Concentration inequalities and model selection*, Volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
[37] Mendelson, S. (2015). ‘Local’ vs. ‘global’ parameters – breaking the gaussian complexity barrier. *ArXiv preprints arXiv:1504.02191*.
[38] Mendelson, S. (2016). Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications* 126(12), 3652–3680.
[39] Nickl, R. and B. M. Pötscher (2007). Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability* 20(2), 177–199.
[40] Nirenberg, L. (2011). On elliptic partial differential equations. In *The principle of minimum and its applications to functional equations*, pp. 1–48. Springer.
[41] Pollard, D. (1984). *Convergence of stochastic processes*. Springer Series in Statistics. New York: Springer-Verlag.
[42] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* 7(2), 186–199.
[43] Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Chichester: John Wiley & Sons Ltd.
[44] Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, 580–615.
[45] Srebro, N., K. Sridharan, and A. Tewari (2010). Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*.
[46] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* 22(1), 118–184. With discussion by Andreas Buja and Trevor Hastie and a rejoinder by the author.
[47] Stout, Q. F. (2008). Unimodal regression via prefix isotonic regression. *Computational Statistics & Data Analysis* 53(2), 289–297.
[48] Talagrand, M. (1996). Majorizing measures: the generic chaining. *Ann. Probab.* 24(3), 1049–1103.
[49] Talagrand, M. (2014). *Upper and lower bounds for stochastic processes*, Volume 60 of *A Series of
Modern Surveys in Mathematics. Springer, Heidelberg. Modern methods and classical problems.

[50] Tsirelson, B. (1986). A geometric approach to maximum likelihood estimation for infinite-dimensional gaussian location. ii. Theory of Probability & Its Applications 30(4), 820–828.

[51] van de Geer, S. (1990). Estimating a regression function. The Annals of Statistics, 907–924.

[52] van de Geer, S. and M. J. Wainwright (2017). On concentration for (regularized) empirical risk minimization. Sankhya A 79(2), 159–200.

[53] van de Geer, S. and M. Wegkamp (1996). Consistency for the least squares estimator in nonparametric regression. The Annals of Statistics, 2513–2523.

[54] van de Geer, S. A. (2000). Applications of empirical process theory, Volume 6 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

[55] van der Vaart, A. W. and J. A. Wellner (1996). Weak convergence and empirical processes. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

[56] Zhang, C.-H. (2002). Risk bounds in isotonic regression. The Annals of Statistics 30(2), 528–555.