ROBUSTNESS OF SHAPE-RESTRICTED REGRESSION ESTIMATORS: AN ENVELOPE PERSPECTIVE

QIYANG HAN AND JON A. WELLNER

Abstract. Classical least squares estimators are well-known to be robust with respect to moment assumptions concerning the error distribution in a wide variety of finite-dimensional statistical problems; generally only a second moment assumption is required for least squares estimators to maintain the same rate of convergence that they would satisfy if the errors were assumed to be Gaussian. In this paper, we give a geometric characterization of the robustness of shape-restricted least squares estimators (LSEs) to error distributions with an $L_{2,1}$ moment, in terms of the 'localized envelopes' of the model.

This envelope perspective gives a systematic approach to proving oracle inequalities for the LSEs in shape-restricted regression problems in the random design setting, under a minimal $L_{2,1}$ moment assumption on the errors. The canonical isotonic and convex regression models, and a more challenging additive regression model with shape constraints are studied in detail. Strikingly enough, in the additive model both the adaptation and robustness properties of the LSE can be preserved, up to error distributions with an $L_{2,1}$ moment, for estimating the shape-constrained proxy of the marginal L_2 projection of the true regression function. This holds essentially regardless of whether or not the additive model structure is correctly specified.

The new envelope perspective goes beyond shape constrained models. Indeed, at a general level, the localized envelopes give a sharp characterization of the convergence rate of the L_2 loss of the LSE between the worst-case rate as suggested by the recent work of the authors [25], and the best possible parametric rate.

1. Introduction

1.1. **Overview.** [1] Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the regression model

$$Y_i = f_0(X_i) + \xi_i, \quad 1 \le i \le n. \tag{1.1}$$

Date: May 8, 2018.

2000 *Mathematics Subject Classification.* 60F17, 62E17.

Key words and phrases. robustness, shape-restricted regression, additive model, oracle inequality, localized envelope.

Supported in part by NSF Grant DMS-1566514, NI-AID grant R01 AI029168, and by Isaac Newton Institute for Mathematical Sciences, program *Statistical Scalability*, EPSRC Grant Number LNAG/036 RG91310.

[1] See Section 1.2 for notation.

where the X_i's are independent and identically distributed $\mathcal{X}$-valued covariates with law P, and the ξ_i's are mean-zero errors independent of X_i's. The goal is to recover the true signal f_0 based on the observed data $\{(X_i, Y_i)\}_{i=1}^n$.

In the canonical setting where the errors ξ_i's are Gaussian, perhaps the simplest estimation procedure for the regression model (1.1) is the *least squares estimator* (LSE) $\hat{f}_n$ defined by

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

where $\mathcal{F}$ is a model chosen by the user. The use of the LSE in the Gaussian regression model has been theoretically justified in the 1990s and the early 2000s, cf. [5, 6, 9, 27, 28, 33, 40, 43, 45]:

Theorem A. *Suppose that:*

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

(F) the model $\mathcal{F}$ satisfies an entropy condition with exponent $\alpha \in (0, 2)$[2]*.*

Then

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

Furthermore, the rate (1.3) is unimprovable under the entropy conditions (F) in a minimax sense, see e.g. [47].

Although the condition (F) is widely accepted in the literature as a complexity measurement of the model $\mathcal{F}$, it is far from clear if the light-tailed condition on the errors (E) is necessary for the theory. Recently, we showed [25] that the condition (E) is actually more than a mere technicality:

Theorem B. *Suppose that condition (E) in Theorem A is replaced by*

(E') the errors $\{\xi_i\}$ have a finite $L_{p,1}$ moment ($p \geq 1$)

and (F) holds. Then

$$\|\hat{f}_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2+\alpha}} \vee n^{-\frac{1}{2}+\frac{1}{2p}}\big). \tag{1.4}$$

We also showed [25] that the rate (1.4) cannot be improved under (F) alone. Comparing with (1.3), the rate in (1.4) clearly indicates that if the model $\mathcal{F}$ only satisfies (F), the best possible moment condition on the errors to guarantee the same rate of convergence of the LSE as in the case of Gaussian errors is $p \geq 1 + 2/\alpha$.

The starting point for this paper originates from a remarkable result due to Cun-Hui Zhang [48] in the context of isotonic regression. Zhang [48] showed that the L_2 loss of the isotonic LSE achieves the usual worst-case (minimax) $\mathcal{O}_{\mathbf{P}}(n^{-1/3})$ rate, and the adaptive rate $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$ if the true signal is, say, f_0 equals a constant, under only a second moment assumption on the errors.

[2] $\mathcal{F}$ satisfies an entropy condition with exponent $\alpha \in (0, 2)$ if either (i) $\sup_Q \log \mathcal{N}(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)) \lesssim \varepsilon^{-\alpha}$, where the supremum is over all finitely discrete measures Q on $(\mathcal{X}, \mathcal{A})$; or (ii) $\log \mathcal{N}_{[]}(\varepsilon, \mathcal{F}, L_2(P)) \lesssim \varepsilon^{-\alpha}$.

We view the first of these two properties established by Zhang as a "robustness property" of the LSE with respect to the distribution of the errors $\{\xi_i\}$. We formalize this with the following definition:

Definition 1. We will say that the estimator sequence $\{\hat{f}_n\}$ is *L_2-robust with respect to the errors $\{\xi_i\}$ in the model $\mathcal{F}$* (or just L_2-robust), if $\hat{f}_n$ converges to f_0 in $L_2(P)$ at the same rate for zero mean 0 errors with $\|\xi_i\|_2 < \infty$ as for errors $\{\xi_i\}$ that are Gaussian or sub-Gaussian. Similarly, if the same rate holds for zero mean errors with $\|\xi_i\|_{2,1} < \infty$, we say that $\{\hat{f}_n\}$ is *$L_{2,1}$-robust with respect to the errors $\{\xi_i\}$ in the model $\mathcal{F}$.*

Similarly, we view the second of the two properties established by Zhang as an "adaptation property" of the LSE with respect to the model $\mathcal{F}$:

Definition 2. We will say that the estimator sequence $\{\hat{f}_n\}$ is *adaptive to a subset $\mathcal{G}_m$ of the model $\mathcal{F}$* if it achieves a nearly (up to factors of $\log n$) parametric rate of convergence at all points $f \in \mathcal{G}_m$.

For the shape-constrained models we consider here the subsets $\mathcal{G}_m$ of $\mathcal{F}$ are natural subclasses of extreme points of the class $\mathcal{F}$: in the isotonic model $\mathcal{F}$ the collections $\mathcal{G}_m$ consisting of m constant non-decreasing pieces, and in the convex regression model $\mathcal{G}_m$ can be taken to be the piecewise linear (convex) functions with at most m linear pieces.

Zhang's work [48] has generated intensive research interest in further understanding the adaptation properties of the isotonic and other shape-restricted LSEs in recent years, cf. [7, 12, 13, 22, 23]. These papers share a common theme: the shape-restricted LSEs are adaptive to certain subsets $\{\mathcal{G}_m\}$ of the model $\mathcal{F}$ under a (sub-)gaussian assumption on the distribution of the errors in the regression model.

Despite substantial progress in the adaptation properties of various shape-restricted LSEs, there remains little progress in further understanding their L_2-robustness properties beyond the isotonic model studied by Zhang [48]. Indeed, the challenges involved here were noted in Guntuboyina and Sen [23] (page 30) as follows: "*.....However the existing proof techniques for these risk bounds strongly rely on the assumption of sub-Gaussianity. It will be very interesting to prove risk bounds in these problems without Gaussianity. We believe that new techniques will need to be developed for this*". One of the goals of this paper is to provide new approaches and insights concerning the L_2 (or $L_{2,1}$)-robustness of various shape-restricted LSEs.

Initially we had hoped to study this problem by appealing to the general Theorem B. However, the theory in Theorem B requires at least a third moment (note that here $\alpha = 1$ for the isotonic model). This implies that the isotonic shape constraint must contain more information than that provided by the entropic structure alone, so that Theorem B fails to fully capture the L_2-robustness of the isotonic LSE.

One particular useful feature of the isotonic model is an explicit min-max formula for the isotonic LSE in terms of partial sum processes; see e.g. [36].

Zhang's techniques [48] make full use of the min-max representation, and are therefore substantially of an analytic flavor. Similar techniques have also been used in [12, 17], but have apparently not yet successful in dealing with any other shape constrained models. The rigidity in this analytic approach naturally motivates the search for other 'softer' properties of the isotonic shape constrained model that explain the robustness of the LSE. These considerations lead to the following question.

Question 1. *What geometric aspects of the isotonic shape constrained model give rise to the L_2(or $L_{2,1}$)-robustness property of the LSE?*

To put this question into a more general setting, note that Theorem B implies that the LSE can converge as slowly as $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ for certain hard models when the errors only have a second moment, while in the aforementioned isotonic regression case, it is possible that the LSE converges at a nearly parametric rate $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$ for certain special isotonic functions. Therefore it seems more promising to search for a characterization of the convergence rate of the L_2 loss of the LSE in terms of some geometric feature of the model $\mathcal{F}$, when the errors have only an L_2(or $L_{2,1}$) moment.

The first main contribution of this paper is to shed light on Question 1 from an 'envelope' perspective at this general level. Roughly speaking, the size of the 'localized envelopes' of the model $\mathcal{F}$ determines the convergence rate of the L_2 loss of the LSE when the errors only have an $L_{2,1}$ moment. More specifically, let $F_0(\delta)$ be the envelope for $\mathcal{F}_0(\delta) \equiv \{f \in \mathcal{F}_0 : Pf^2 \le \delta^2\}$ where $\mathcal{F}_0 \equiv \mathcal{F} - f_0$. We show that (cf. Theorem 1), under a certain uniform entropy condition on the function class, if for some $0 \le \gamma \le 1$, the localized envelopes have the growth rate

$$\|F_0(\delta)\|_{L_2(P)} \sim \delta^\gamma : \tag{1.5}$$

then the convergence rate of the LSE in the L_2 loss is no worse than

$$\mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2(2-\gamma)}}\big). \tag{1.6}$$

Furthermore, the rate (1.6) cannot be improved under the condition (1.5), cf. Theorem 2. It is easily seen from (1.6) that, as the size of the localized envelopes increases, the rate of the L_2 loss of the LSE deteriorates from the parametric rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ to the worst-case rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ as suggested by Theorem B. For isotonic regression, we will see that the localized envelopes of the model are small in the sense that $\gamma \approx 1$ (up to logarithmic factors) when $f_0 = 0$, and hence the LSE converges at a nearly parametric rate under an $L_{2,1}$ moment assumption on the errors. For the hard models identified in [25] (cf. Example 4 below), the localized envelopes are big in the sense that $\gamma = 0$ so the LSE can only converge at the worst-case rate.

Addressing Question 1 from a geometric point of view is not only of interest in its own right, but also serves as an important step in better understanding the robustness properties of other shape constrained models. This is the context of the second main contribution of this paper: we aim

at improving our understanding of the $L_{2,1}$-robustness property of shape restricted LSEs, by providing a systematic approach to proving oracle inequalities in the random design regression setting for these LSEs under an $L_{2,1}$ moment condition on the errors. This goal is achieved by exploiting the idea of small envelopes from the solution to Question 1. The formulation of the oracle inequality follows its fixed-design counterparts that highlight the automatic rate-adaptive behavior of the LSE, cf. [7, 12]. More specifically, we first prove the following oracle inequality that holds for the canonical isotonic and convex LSEs in the simple regression models (cf. Theorem 3): Suppose that $\|f_0\|_\infty < \infty$ and the errors $\{\xi_i\}$ are i.i.d. mean-zero with $\|\xi_1\|_{2,1} < \infty$. Then for any $\delta \in (0,1)$, there exists some constant $c > 0$ such that with probability $1-\delta$,

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \le c \inf_{m\in\mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right), \tag{1.7}$$

where f_0^* is the $L_2(P)$-projection of f_0 onto the space of square ntegrable monotonic non-decreasing (resp. convex) functions, and $\mathcal{G}_m$ is the class of piecewise constant non-decreasing (resp. linear convex) functions on $[0,1]$ with at most m pieces in the isotonic (resp. convex) model. The oracle inequality (1.7) is further verified for the shape-restricted LSEs in the additive model (cf. Theorem 4), where now f_0 is the marginal L_2 projection of the true regression function. One striking message of the oracle inequality for the shape-restricted LSEs in the additive model is the following: both the adaptation and $L_{2,1}$-robustness properties of the LSE can be preserved, up to error distributions with an $L_{2,1}$ moment, for estimating the shape-constrained proxy of the marginal L_2 projection of the true regression function, *essentially regardless of whether or not the additive structure is correctly specified*.

The proofs in this paper rely heavily on the new empirical process tools and proof techniques developed in [25]. Although we will list relevant results, readers are referred to [25] for more discussion of the new tools. Along the way we also resolve the stochastic boundedness issue of convexity shape-restricted LSEs at the boundary, which may be of independent interest (this problem is in fact an open problem in the field, cf. [23]).

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

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

It is well known that $L_{p+\varepsilon} \subset L_{p,1} \subset L_p$ holds for any underlying probability measure, and hence a finite $L_{p,1}$ condition requires slightly more than a p-th moment, but no more than any $p+\varepsilon$ moment, see Chapter 10 of [29]. In this paper, we will primarily be concerned with the case $p = 2$.

For a real-valued measurable function f defined on $(\mathcal{X}, \mathcal{A}, P)$, $\|f\|_{L_p(P)} \equiv \big(P|f|^p\big)^{1/p}$ denotes the usual L_p-norm under P, and $\|f\|_\infty \equiv \|f\|_{L_\infty} \equiv \sup_{x\in\mathcal{X}}|f(x)|$. f is said to be P-centered if $Pf = 0$. $L_p(g, B)$ denotes the $L_p(P)$-ball centered at g with radius B. For simplicity we write $L_p(B) \equiv L_p(0, B)$.

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of the normed space of real functions $f : \mathcal{X} \to \mathbb{R}$. Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-covering number, and let $\mathcal{N}_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the ε-bracketing number; see page 83 of [45] for more details. To avoid unnecessary measurability digressions, we assume that $\mathcal{F}$ is countable throughout the article. As usual, for any $\phi : \mathcal{F} \to \mathbb{R}$, we write $\|\phi(f)\|_{\mathcal{F}}$ for $\sup_{f\in\mathcal{F}}|\phi(f)|$.

Throughout the article $\varepsilon_1, \ldots, \varepsilon_n$ will be i.i.d. Rademacher random variables independent of all other random variables. C_x will denote a generic constant that depends only on x, whose numeric value may change from line to line unless otherwise specified. $a \lesssim_x b$ and $a \gtrsim_x b$ mean $a \le C_x b$ and $a \ge C_x b$ respectively, and $a \asymp_x b$ means $a \lesssim_x b$ and $a \gtrsim_x b$ [$a \lesssim b$ means $a \le Cb$ for some absolute constant C]. For two real numbers a, b, $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. We slightly abuse notation by defining $\log(x) \equiv \log(x \vee e)$.

1.3. **Organization.** Section 2 is devoted to a treatment of the relationship between the size of the localized envelopes and the convergence rate of the L_2 loss of the least squares estimator. Section 3 is devoted to applications to shape-restricted regression problems. Proofs are deferred to Sections 4 and 5.

2. Convergence rate of the LSE: the envelope characterization

2.1. **Upper and lower bounds.** Our first main result is the following.

Theorem 1. *Suppose that $\xi_1, \ldots, \xi_n$ are i.i.d. mean-zero errors independent of i.i.d. covariates $X_1, \ldots, X_n$ with law P such that $\|\xi_1\|_{2,1} < \infty$. Further suppose that $\mathcal{F}_0 \equiv \mathcal{F} - f_0$ is a VC-subgraph class, and the envelopes $F_0(\delta)$ of $\mathcal{F}_0(\delta) \equiv \{f \in \mathcal{F}_0 : Pf^2 \le \delta^2\}$ satisfy the growth condition*

$$\|F_0(\delta)\|_{L_2(P)} \le c \cdot \delta^\gamma, \quad \text{for all } \delta > 0 \tag{2.1}$$

for some constants $0 \le \gamma \le 1$ and $c > 0$. If $\|\hat f_n - f_0\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$, then

$$\|\hat f_n - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2(2-\gamma)}}\big).$$

Remark 1. Some technical remarks are in order.

(1) If instead of $\|\hat f_n - f_0\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$ it is assumed that $\mathcal{F}_0 \subset L_\infty(1)$, then the conclusion of Theorem 1 can be strengthened to an expectation: $\mathbb{E}\|\hat f_n - f_0\|_{L_2(P)} = \mathcal{O}\big(n^{-\frac{1}{2(2-\gamma)}}\big)$.

(2) Condition (2.1) on the size of the localized envelopes can be modified to incorporate logarithmic factors. In particular, if

$$\|F_0(\delta)\|_{L_2(P)} \le c \cdot \delta^\gamma \log^\tau(1/\delta),$$

then we may slightly modify the proof of Theorem 1 to see that the convergence rate of the L_2 loss of the LSE is given by

$$\mathcal{O}_{\mathbf{P}}\big(n^{-\frac{1}{2(2-\gamma)}}\log^{\frac{\tau}{2-\gamma}} n\big).$$

(3) We assume that the errors are identically distributed for simplicity: the case of mean-zero, independent but not necessarily identically distributed errors follows from a minor modification of the proof.

Remark 2. Theorem 1 is actually proved for $\mathcal{F}_0$ under a more general *uniform VC-type* condition: $\mathcal{F}_0$ is said to be of uniform VC-type if there exists some $\alpha \in [0,2)$ and $\beta \in [0,\infty)$ [3] such that for any probability measure Q, and any $\varepsilon \in (0,1), \delta > 0$,

$$\log \mathcal{N}\left(\varepsilon \|F_0(\delta)\|_{L_2(Q)}, \mathcal{F}_0(\delta), L_2(Q)\right) \lesssim \varepsilon^{-\alpha}\log^{\beta}(1/\varepsilon). \tag{2.2}$$

The most significant examples for uniform VC-type classes are the VC-subgraph classes ($\alpha = 0, \beta = 1$). Other important examples include the VC-major classes, which satisfy (2.2) up to a logarithmic factor (cf. Lemma 8). As we will see in Section 3, the canonical examples of VC-major classes that satisfy (2.2) considered in this paper are the classes of bounded monotonic non-decreasing and convex functions on $[0,1]$.

Remark 3. From a purely probabilistic point of view, the condition (2.1) is related to Alexander's capacity function [1, 2, 3] defined for VC class of sets that gives relatively sharp asymptotic local moduli of weighted empirical processes indexed by such classes. Results in a similar vein can be found in [18] who generalized this notion to bounded VC-subgraph function classes.

So far we have derived an upper bound for the convergence rate of the L_2 loss of the LSE under the condition (2.1). It is natural to wonder if such an upper bound is sharp in an appropriate sense.

Theorem 2. *Let P be the uniform distribution on $[0,1]$. For any $\gamma \in (0,1]$, there exists some uniformly bounded VC-subgraph class $\tilde{\mathcal{F}}$ on $[0,1]$ and some $f_0 \in \tilde{\mathcal{F}}$ such that $\tilde{\mathcal{F}}_0 \equiv \tilde{\mathcal{F}} - f_0$ satisfies (2.1), and the following property holds: for each $\varepsilon \in (0,1/2)$, there exist some constants $c_{\varepsilon,\gamma} > 0$, $\mathfrak{p} > 0$ and some law for ξ_1 with $\|\xi_1\|_{2(1-\varepsilon)} < \infty$ such that, for n large enough depending on ε, γ, there exists a LSE f_n^* whose L_2 loss satisfies*

$$\|f_n^* - f_0\|_{L_2(P)} \geq c_{\varepsilon,\gamma} \cdot n^{-\frac{1}{2(2-\gamma)} - c'_\gamma \varepsilon}$$

with probability at least $\mathfrak{p} > 0$. The constant c'_γ can be taken to be $2/\gamma$.

Theorem 2 shows that our upper bound Theorem 1 cannot be improved substantially under (2.1): the size of the localized envelopes drives the convergence rate of the L_2 loss of the LSE over VC-subgraph models (or more generally, models of uniform VC-type) in the heavy-tailed regression setting where the errors only admit (roughly) a second moment. Since the median

[3] We can also allow $\alpha = 2, \beta < -2$ but we are not aware of any such examples.

regression estimator over VC-subgraph models achieves a nearly parametric rate $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$ at least when the errors are symmetric and admit smooth densities; cf. Section 3.4.4 of [45], Theorem 2 illustrates a genuine deficiency of the LSE in VC-subgraph models when the envelopes of the model are not small. We remark that the case $\gamma = 0$ is excluded mainly for simplicity of presentation; similar conclusions hold under a slightly weaker formulation, cf. Theorem 5 of [25].

The proofs of Theorems 1 and 2 are based on recent developments on the *equivalence* between the convergence rate of the L_2 loss of the LSE and the size of the multiplier empirical process, cf. [11, 25, 44]. For the upper bound, our proofs rely heavily on a new multiplier inequality developed in [25]. The lower bound, on the other hand, is based on an explicit construction of $\tilde{\mathcal{F}}$ that witnesses the desired rate within uniformly bounded VC-subgraph classes satisfying (2.1).

2.2. **Examples.** In this section, we use Theorem 1 to examine the convergence rate of the L_2 loss of the LSE in several important examples.

Example 1 (Linear model)**.** Let $\mathcal{F} \equiv \{f_\beta(x) \equiv \beta^\top x : \beta \in \mathbb{R}^d\}$ and let P be the uniform distribution on $[0,1]^d$. This is the simplest linear regression model. A second moment assumption on the errors ξ_i's yields a closed-form LSE with a parametric convergence rate: $\|\hat{f}_n - f_0\|_{L_2(P)} \asymp \|\hat{\beta}_n - \beta_0\|_2 = \mathcal{O}_{\mathbf{P}}(n^{-1/2})$. This rate is obviously much faster than the worst-case rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ as suggested by Theorem B. Thus the LSE sequence $\{\hat{f}_n\}$ is L_2-robust for the model $\mathcal{F}$ by a direct argument while our Theorem 1 very nearly recovers this: it shows that $\{\hat{f}_n\}$ is $L_{2,1}$-robust for the model $\mathcal{F}$.

For simplicity of discussion, we assume $d = 1$ in the sequel. We may also restrict the model to be $\{f_\beta : \beta \in [-1,1]\}$; this is viable since the LSE *localizes* in the sense that $\|\hat{f}_n\|_\infty = |\hat{\beta}_n| = \mathcal{O}_{\mathbf{P}}(1)$. Moreover, it is clear that the model is a VC-subgraph class. For any $\delta > 0$, $\|f_\beta\|_{L_2(P)} \le \delta$ implies that $|\beta| \le \sqrt{3}\delta$, and thus

$$F(\delta)(x) = \sup_{\beta \in [-\sqrt{3}\delta, \sqrt{3}\delta]} |\beta x| = \sqrt{3}\delta|x|,$$

which in turn yields $\|F(\delta)\|_{L_2(P)} = \delta$. Hence Theorem 1 applies with $\gamma = 1$ to recover the usual parametric rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ for the L_2 loss of the LSE.

Our approach here should be compared with the common practice of using local entropy to recovery the exact parametric rate for parametric models—but the latter does not extend directly to the heavy-tailed regression setting, cf. pages 152-153 of [43].

Example 2 (Isotonic model)**.** Let $\mathcal{F}$ be the class of monotonic non-decreasing functions on $[0,1]$ and let P be the uniform distribution on $[0,1]$. It is shown in a related fixed design setting (cf. [12, 17, 48]) that a second moment condition on the errors ξ_i is sufficient for the isotonic LSE to achieve the nearly parametric adaptive rate $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$ in the discrete ℓ_2 loss, when the

true signal is $f_0 = 0$. This naturally suggests a similar rate for the L_2 loss of the isotonic LSE in the random design setting. Apparently, this (suggested) nearly parametric rate is far from the worst-case rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$.

In this model, since the univariate isotonic LSE localizes in L_∞ norm (cf. Lemma 5), we may assume without loss of generality that $\mathcal{F} \equiv \{f : \text{non-decreasing}, \|f\|_\infty \leq 1\}$. The entropy condition (2.2) can be verified using the VC-major property of $\mathcal{F}$ up to a logarithmic factor (cf. Lemma 8). On the other hand, for any $\delta > 0$, by monotonicity and the L_2 constraint, we can take

$$F(\delta)(x) \equiv \delta \cdot \left(x^{-1/2} \vee (1-x)^{-1/2}\right) \wedge 1.$$

Evaluating the integral we see that $\|F(\delta)\|_{L_2(P)} \lesssim \delta\sqrt{\log(1/\delta)}$. Then an application of Theorem 1 along with Remarks 1 (2) and 2, we see that the L_2 loss of the LSE $\hat{f}_n$ converges at a parametric rate up to logarithmic factors when the truth f_0 is a constant function and the errors are $L_{2,1}$. The observation concerning the role of the localized envelopes in the isotonic model here is the starting point for a systematic development of oracle inequalities for shape-restricted LSEs in Section 3.

Example 3 (Single change-point model)**.** Let $\mathcal{F} \equiv \{\mathbf{1}_{[a,1]} : a \in [0,1]\}$ be the model containing signals on $[0,1]$ with a single change point. Let P be the uniform distribution on $[0,1]$.

This model is contained in the isotonic model—from here we already know by Example 2 that the localized envelopes of $\mathcal{F}$ are small, and hence the LSE converges at a rate no worse than a nearly parametric rate under an $L_{2,1}$ moment assumption on the errors. We can do better: since the localized envelopes are exactly given by $F(\delta) = \mathbf{1}_{[1-\delta^2,1]}$, it follows that $\|F(\delta)\|_{L_2(P)} = \delta$, and hence by Theorem 1 with $\gamma = 1$ we see that the LSE converges exactly at the parametric rate $\mathcal{O}_{\mathbf{P}}(n^{-1/2})$ even if the errors only admit an $L_{2,1}$ moment. This is in stark contrast with the *multiple change-points model* detailed below.

Example 4 (Multiple change-points model)**.** Consider the following multiple change-points model:

$$\mathcal{F}_k \equiv \Big\{ \sum_{i=1}^{k} c_i \mathbf{1}_{[x_{i-1},x_i]} : |c_i| \leq 1, \\ 0 \leq x_0 < x_1 < \ldots < x_{k-1} < x_k \leq 1 \Big\}, k \geq 1.$$

It is shown in [25] that the L_2 loss of the LSE over (a subset of) $\mathcal{F}_k$ cannot converge at a rate faster than $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ for some errors ξ_i with only (roughly) a second moment. The LSE fails to be rate-optimal in this model: if the errors are Gaussian (or even bounded), the convergence rate of the L_2 loss of the LSE (over VC-subgraph classes) is no worse than $\mathcal{O}_{\mathbf{P}}(\sqrt{\log n/n})$.

Note that in this model, the localized envelopes are given by $F(\delta) \equiv 1$ for any $\delta > 0$ and hence $\|F(\delta)\|_{L_2(P)} = 1$. Applying Theorem 1 with $\gamma = 0$ recovers the correct rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ for the L_2 loss of the LSE in this model.

Example 5 (Unimodal model)**.** Let $\mathcal{F}$ contain all (bounded) unimodal functions on $[0,1]$, i.e. all $f : [0,1] \to \mathbb{R}$ such that there exists some $x^* \in [0,1]$ with $f|_{[0,x^*]}$ non-decreasing and $f|_{[x^*,1]}$ non-increasing. [13] and [7] considered the performance of the LSE in a fixed-design unimodal Gaussian regression setting, where similar adaptive behavior as in the isotonic case (cf. [48]) is derived. Since the class of (bounded) unimodal functions on $[0,1]$ contains the class of multiple change-points model $\mathcal{F}_1$ as studied in Example 4, our results here imply that *the unimodal shape constraint does not inherit* the L_2 (or $L_{2,1}$)-robustness property as in the isotonic shape constraint in Example 2: the worst-case $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ is attained by the LSE in the unimodal regression model for some errors ξ_i's with (roughly) a second moment.

3. Shape-restricted regression problems

As briefly mentioned in the Introduction, it is well-known that in the fixed design regression setting, the isotonic least squares estimator (LSE) only requires a second moment condition on the errors to enjoy an oracle inequality, cf. [12, 17, 48]. The proof techniques used therein rely crucially on (i) some form of representation of the isotonic LSE in terms of partial sum processes, and (ii) martingale inequalities. Unfortunately, such an explicit representation does not exist beyond the isotonic LSE, and hence these techniques do not readily extend to other problems.

Our goal here is to give a systematic treatment of the robustness properties of shape-restricted LSEs in a random design setting, up to error distributions with an $L_{2,1}$ moment. The examples we examine are (i) the canonical isotonic and convex regression models, and (ii) additive regression models with monotonicity and convexity shape constraints. As we will see, the 'smallness' of the localized envelopes, along with their special geometric properties, play a central role in our approach.

Henceforth, the isotonic (resp. convex) model refers to the regression model based on the class of monotonic non-decreasing (resp. convex) functions on $[0,1]$.

3.1. Prologue: the canonical problems.

We start by considering the 'canonical' problems in the area of shape restricted regression: the isotonic and convex regression problems. Note that a generic LSE $\hat{f}_n$ in (1.2) is only well-defined on the design points $X_1, \ldots, X_n$. Our results below hold for the *canonical LSEs*: for the isotonic (respectively convex) model, $\hat{f}_n$ is defined to be the unique left-continuous piecewise constant (resp. linear) function on $[0,1]$ with jumps (respectively kinks) at (potentially a subset of) $\{\hat{f}_n(X_i)\}_{i=1}^n$.

Some further notation: let $\mathcal{M}_m \equiv \mathcal{M}_m([0,1])$ (respectively $\mathcal{C}_m \equiv \mathcal{C}_m([0,1])$) be the class of all non-decreasing piecewise constant functions (respectively convex piecewise linear functions) on $[0,1]$ with at most m pieces. Let P denote the uniform distribution on $[0,1]$ for simplicity of exposition.

Theorem 3. *Consider the regression model (1.1). Let $\mathcal{F}$ be either the isotonic or convex model. Suppose that $\|f_0\|_\infty < \infty$, and the errors are i.i.d. mean-zero with $\|\xi_1\|_{2,1} < \infty$. Then for any $\delta \in (0,1)$, there exists $c \equiv c(\delta, \|\xi\|_{2,1}, \|f_0\|_\infty, \mathcal{F}) > 0$ such that with probability $1-\delta$, the canonical LSE $\hat{f}_n$ defined above satisfies*

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \le c \inf_{m \in \mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right),$$

where $f_0^* = \mathrm{argmin}_{g \in \mathcal{F} \cap L_2(P)} \|f_0 - g\|_{L_2(P)}$, and $\mathcal{G}_m = \mathcal{M}_m$ for the isotonic model and $\mathcal{G}_m = \mathcal{C}_m$ for the convex model.

The isotonic regression problem, included here mainly for sake of later development in the additive model, is a benchmark example in the family of shape-restricted regression problems. Even in this simplest case, the above oracle inequality in $L_2(P)$ loss seems new[4].

For the more interesting convex regression problem, our oracle inequality here confirms for the first time both the adaptation and robustness properties of the convex LSE up to error distributions with an $L_{2,1}$ moment. Previous oracle inequalities for the convex LSE exclusively focused on the fixed-design setting under a (sub-)Gaussian assumption on the errors [7, 12]; see also Section 3 of [23] for a review.

Remark 4. Two technical comments on the formulation of the oracle inequality in Theorem 3:

(1) The oracle inequality holds for the projection f_0^* of f_0 to $\mathcal{F} \cap L_2(P)$ and hence allows for model mis-specification: the only assumption on f_0 is boundedness: $\|f_0\|_\infty < \infty$. The same comment also applies to the oracle inequality in the additive model below.
(2) The oracle inequality cannot be strengthened to an expectation, in view of a counterexample discovered in [4] in the convex model: the convex LSE $\hat{f}_n$ has infinite L_2 risk in estimating $f_0 = 0$ even if the errors are bounded: $\mathbb{E}\|\hat{f}_n - 0\|_{L_2(P)} = \infty$.

3.1.1. *Proof strategy of Theorem 3.* The proof of Theorem 3 contains two major steps.

[4]An oracle inequality in $L_2(\mathbb{P}_n)$ loss follows immediately from [12] (with a second moment assumption on the errors) since the monotone cone does *not* change with the design points. See [24] for different techniques in the multivariate isotonic regression problem when the errors are Gaussian.

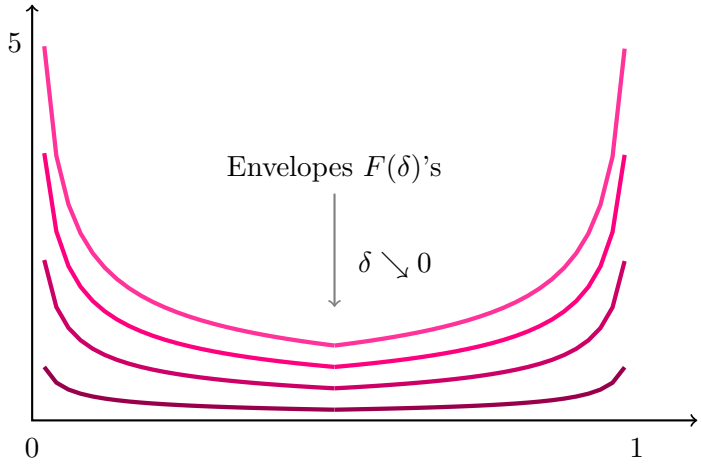

FIGURE 1. Envelopes for isotonic model with $c = 1$ in (3.2). From top to bottom: $\delta = 0.7, 0.5, 0.3, 0.1$.

(Step 1) We first localize the shape-restricted LSEs in L_∞ norm. This step requires some understanding of the boundary behavior of the shape-restricted LSEs under a second moment assumption on the errors. The case for isotonic regression is relatively straightforward, while the case for convex regression is much more difficult. Here we resolve this issue in Lemma 5.

(Step 2) After the localization in Step 1, the problem essentially reduces to controlling a multiplier empirical process of the form

$$\mathbb{E} \sup_{f\in\mathcal{F}: f-f_0^*\in L_2(\delta_n)\cap L_\infty(B)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (f - f_0^*)(X_i) \right|. \tag{3.1}$$

A sharp bound for (3.1) is inspired by the observation in Example 2, where the (untruncated) localized envelopes of the isotonic model take the form

$$F(\delta)(x) \equiv c\delta \cdot \left(x^{-1/2} \vee (1-x)^{-1/2} \right) \tag{3.2}$$

for some absolute constant $c > 0$. The envelopes for the convex model also take the same form (3.2), cf. Lemma 9. On the other hand, the localized envelopes (3.2) are centered at 0, while the multiplier empirical process (3.1) in question is centered at f_0^*. By exploiting the exact form of (3.2), we perform a 'change-of-center argument' on (3.1) by shifting f_0^* to an arbitrary piecewise simple signal $f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)$, cf. Lemma 6, thereby reducing the control of (3.1) to control of several multiplier empirical processes centered at 0. The effect of the heavy-tailed ξ_i's is then accounted for, via the multiplier inequality developed in [25], by a uniform estimate for

the corresponding empirical processes in terms of the L_2 size of the localized envelopes (3.2).

Remark 5. Currently our oracle inequality comes with a $\log^2 n$ term. It is known in (i) the fixed design isotonic model with a second moment assumption on the errors, and (ii) the fixed design convex model with sub-Gaussian errors, that the power of the logarithmic factor can be reduced to 1. The additional logarithmic factor in Theorem 3 occurs due to the use of VC-major property for the isotonic and convex models in the random design setting: the entropy estimate of bounded VC-major classes comes with logarithmic factors that involve the L_2 size of the envelopes (cf. Lemma 8).

3.2. **Additive regression model with shape constraints.** Consider fitting $(x, z) \mapsto \phi_0(x, z)$, the conditional mean of the regression model

$$Y_i = \phi_0(X_i, Z_i) + \xi_i, \quad 1 \le i \le n, \tag{3.3}$$

by additive models of the form $\{(x, z) \mapsto f(x) + h(z)\}_{f\in\mathcal{F},h\in\mathcal{H}}$, where $\mathcal{F}, \mathcal{H}$ are two function classes on $[0, 1]$. To capture the mathematical essence of the problem, we assume that the covariates $\{(X_i, Z_i)\}_{i=1}^n$ are i.i.d. from the uniform law P on $[0, 1]^2$ and are independent of the errors $\{\xi_i\}$. We use P_X, P_Z to denote the marginal distributions of P. For identifiability we assume that $\mathcal{H}$ is centered.

Additive models of the type have a long history; see e.g. [26, 38]. When the additive model is well specified (i.e. $\phi_0(x, z) = f_0(x) + h_0(z)$ with $f_0 \in \mathcal{F}, h_0 \in \mathcal{H}$), and the nonparametric components enjoy smoothness assumptions, standard methods such as iterative backfitting, e.g. [30] and penalized LSE (smooth spline), e.g. [46], can be used to estimate f_0 and h_0.

Instead of computational issues, we will be interested here in certain structural aspects of the additive LSE $\hat{f}_n$ defined via:

$$(\hat{f}_n, \hat{h}_n) \in \operatorname*{argmin}_{(f,h)\in\mathcal{F}\times\mathcal{H}} \sum_{i=1}^n \big(Y_i - f(X_i) - h(Z_i)\big)^2. \tag{3.4}$$

Since the true regression function ϕ_0 need not have an additive structure, one may naturally expect that $\hat{f}_n$ and $\hat{h}_n$ estimate the marginal L_2 projections $x \mapsto f_0(x) \equiv P_Z\phi_0(x, Z)$ and $z \mapsto h_0(z) \equiv P_X\phi_0(X, z) - P\phi_0$ (cf. Appendix 4, page 439 of [8]). Our primary *structural question* on the behavior of the additive LSE $\hat{f}_n$ concerns the situation in which the model $\mathcal{F}$ involves shape constraints:

Question 2. *Does the additive LSE $\hat{f}_n$ over the shape constrained model $\mathcal{F}$ enjoy similar robustness and adaptation properties as in the univariate case (treated in Theorem 3)?*

The next theorem gives an affirmative answer to Question 2.

Theorem 4. *Suppose that (X_i, Z_i, Y_i), $i = 1, \dots, n$, are i.i.d. with values in $[0, 1] \times [0, 1] \times \mathbb{R}$ and satisfy (3.3) where $\|\phi_0\|_\infty < \infty$, and the errors $\{\xi_i\}$ are*

i.i.d. mean zero with $\|\xi_1\|_{2,1} < \infty$. *Let* $\mathcal{F}$ *be either the isotonic or convex model. Further suppose that* $\mathcal{H} \subset L_\infty(2\|\phi_0\|_\infty)$ *satisfies the following* L_∞ *covering bound: for some* $\gamma \in (0,2)$

$$\log \mathcal{N}(\varepsilon, \mathcal{H}, L_\infty) \lesssim \varepsilon^{-\gamma}, \text{ for all } \varepsilon \in (0,1). \tag{3.5}$$

Then for any $\delta \in (0,1)$, *there exists* $c \equiv c(\delta, \|\xi\|_{2,1}, \|\phi_0\|_\infty, \mathcal{F}, \mathcal{H}) > 0$ *such that with probability* $1-\delta$, *the canonical LSE* $\hat{f}_n$ *in (3.4) satisfies*

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \le c \inf_{m \in \mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right),$$

where $f_0^* = \operatorname{argmin}_{g \in \mathcal{F} \cap L_2(P)} \|f_0 - g\|_{L_2(P)}$ *with* $f_0 = P_Z \phi_0(\cdot, Z)$, *and* $\mathcal{G}_m = \mathcal{M}_m$ *for the isotonic model and* $\mathcal{G}_m = \mathcal{C}_m$ *for the convex model.*

There is very limited theoretical understanding of the properties of shape-restricted estimators when additive models are used. [34] investigated identifiability issue for the additive LSE in the fixed design setting. [31] considered *pointwise* performance of the LSE where both $\mathcal{F}$ and $\mathcal{H}$ are monotonic with errors admitting exponential moments. [15] gives an extension to a semi-parametric setting assuming the same moment condition on the errors, still considering pointwise performance of the LSEs for the isotonic components. [14] proved consistency of the MLEs for a generalized class of additive and index models with shape constraints, without rate considerations. A common feature of all these works is that the model is required to be well-specified.

To the best knowledge of the authors, Theorem 4 is the first oracle inequality for shape-restricted LSEs in regression using an additive model, and moreover, allowing for model mis-specification: not only the regression function class $\mathcal{F}$ can be mis-specified, but the additive model itself may also be mis-specified. Our result here therefore gives a strong positive answer to Question 2: both the adaptation and robustness properties of additive shape-restricted LSEs can be preserved in estimating the shape constrained proxy of the marginal L_2 projection of the true regression function, up to error distributions with an $L_{2,1}$ moment, *essentially regardless of whether or not the additive structure is correctly specified.*

3.2.1. *Examples under correct specification of the additive structure.* Now we consider the important situation when ϕ_0 has an additive structure:

$$\phi_0(x,z) \equiv f_0(x) + h_0(z).$$

In such a scenario, our result here is related to the recent work [42], who asserted that the rate optimality nature of the (penalized) LSE over $\mathcal{F}$ in the Gaussian regression setting can be preserved regardless of the smoothness level of $\mathcal{H}$. Our Theorem 4 reveals a further structural property of the LSEs: the robustness and adaptation merits due to shape constraints can also be preserved, regardless of the choice of $\mathcal{H}$ under the entropy condition (3.5).

To further illustrate this point, we consider some examples.

- (*Parametric model*) $\mathcal{H} \equiv \{f_\beta(z) \equiv \beta(z - 1/2) : \beta \in [-1, 1]\}$. In this case (3.3) becomes the semiparametric partially linear model.
- (*Smooth model*) $\mathcal{H}$ is the class of centered uniformly bounded α-Hölder ($\alpha > 1/2$) continuous functions on $[0, 1]$ with uniformly bounded derivatives (cf. Theorem 2.7.1 of [45]).
- (*Shape constrained model*) $\mathcal{H}$ is the class of centered uniformly Lipschitz convex functions on $[0, 1]$ (cf. Corollary 2.7.10 of [45]).

3.2.2. *Proof strategy of Theorem 4.* The basic strategy in our proof of Theorem 4 is similar to that of Theorem 3. First, we need to localize the LSEs in L_∞ norm under a second moment assumption on the errors and $P_Z H^2 < \infty$, cf. Lemma 13. Next, in addition to the multiplier empirical process (3.1), the major additional empirical process we need to control is

(3.6)
$$\mathbb{E} \sup_{\substack{f \in \mathcal{F}: f - f_0^* \in L_2(\delta_n) \cap L_\infty(B) \\ h \in \mathcal{H}}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f - f_0^*)(X_i)(h - (\phi_0 - f_0))(X_i, Z_i) \right|.$$

where the ε_i's are i.i.d. Rademacher random variables. One notable feature in (3.6) is that the supremum over $\mathcal{H}$ need *not* be localized when the interest is in the behavior of $\hat{f}_n$, cf. Proposition 4. In other words, no apriori information on the behavior of $\hat{h}_n$ (other than the assumption (3.5)) is needed in order to understand the behavior of $\hat{f}_n$.

The entropy condition (3.5) serves as a sufficient condition for a sharp estimate for (3.6) (and thereby for the oracle inequality in Theorem 4), but is apparently not necessary; we make such a choice here to cover the above common examples. A case-by-case study is possible as long as (3.6) can be well-controlled. For instance, it is not hard to verify a similar bound for (3.6) as in Lemma 10 (and hence the oracle inequality for shape-restricted LSEs $\hat{f}_n$) when the additive structure is correctly specified, and $\mathcal{H}$ is the class of centered indicator functions over closed intervals on $[0, 1]$ and $h_0 = 0$ (note that this class fails to satisfy (3.5) since $\mathcal{H}$ is not totally bounded in L_∞). This is a difficult case: although the L_2 loss of the LSE $\hat{h}_n$ is known to converge at a worst-case rate $\mathcal{O}_{\mathbf{P}}(n^{-1/4})$ (cf. Example 4), Theorem 4 tells us that the bad behavior of $\hat{h}_n$ has no effect on the good (robust and adaptive) performance of $\hat{f}_n$, at least under reasonable assumption on the distribution of the covariates (X, Z).

4. Proofs of the main results

In this section we outline the main steps in proving the main results of the paper, namely:

(1) Theorems 1 and 2 characterizing the geometric feature of the model that determines the actual convergence rate of the L_2 loss of the least squares estimator, and

(2) Theorems 3 and 4 highlighting oracle inequalities in shape restricted regression models with a $L_{2,1}$ moment assumption on the errors.

Proofs of many technical intermediate results will be deferred to Section 5.

4.1. **Preliminaries.** In this subsection we collect the empirical process tools that will be needed in the proofs to follow. Our first ingredient is a sharp multiplier inequality proved in [25].

Lemma 1 (Theorem 1 in [25]). *Suppose that $\xi_1, \dots, \xi_n$ are i.i.d. mean-zero random variables independent of i.i.d. $X_1, \dots, X_n$. Let $\mathcal{F}_1 \supset \cdots \supset \mathcal{F}_n$ be a non-increasing sequence of function classes. Assume further that there exist non-decreasing concave functions $\{\psi_n\} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\psi_n(0) = 0$ such that*

$$\mathbb{E}\left\| \sum_{i=1}^{k} \varepsilon_i f(X_i) \right\|_{\mathcal{F}_k} \leq \psi_n(k) \tag{4.1}$$

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

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

Lemma 1 controls the first moment of the multiplier empirical process. For higher moments, the following moment inequality is useful.

Lemma 2 (Proposition 3.1 of [19]). *Suppose $X_1, \dots, X_n$ are i.i.d. with law P and $\xi_1, \dots, \xi_n$ are i.i.d. mean-zero random variables with $\|\xi_1\|_2 < \infty$. Let $\mathcal{F}$ be a class of measurable functions such that $\sup_{f \in \mathcal{F}} Pf^2 \leq \sigma^2$. Then for any $q \geq 1$,*

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

Here $K > 0$ is a universal constant.

To use Lemma 1, we need to control the size of the empirical process. Let

$$J(\delta, \mathcal{F}, L_2) \equiv \int_0^\delta \sup_Q \sqrt{1 + \log \mathcal{N}\big(\varepsilon \|F\|_{L_2(Q)}, \mathcal{F}, L_2(Q)\big)}\, \mathrm{d}\varepsilon \tag{4.2}$$

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

We will frequently use the following Koltchinskii-Pollard maximal inequality.

Lemma 3 (Theorem 2.14.1 of [45])**.** *Let $\mathcal{F}$ be a class of measurable functions with measurable envelope F, and $X_1, \ldots, X_n$ are i.i.d. random variables with law P. Then*

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

Our last technical ingredient is Talagrand's concentration inequality [39] for the empirical process in the form given by [32]:

Lemma 4. *Let $\mathcal{F}$ be a class of measurable functions such that $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq b$. Then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\mathbb{G}_n f| \geq 2\mathbb{E} \sup_{f\in\mathcal{F}} |\mathbb{G}_n f| + \sqrt{8\sigma^2 x} + 34.5 b \frac{x}{\sqrt{n}} \right) \leq e^{-x},$$

where $\sigma^2 \equiv \sup_{f\in\mathcal{F}} \mathrm{Var}_P f$, and $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$.

4.2. **Proof of Theorem 1.**

Proof of Theorem 1. We only prove the case $\mathcal{F}_0 \subset L_\infty(1)$ as in Remark 1 (1). The proof for the case $\|\hat{f}_n - f_0\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$ follows with only minor modifications. We also work with the more general uniform VC-type condition as in Remark 2. Let $\delta_n \equiv n^{-\frac{1}{2(2-\gamma)}}$. By the proof of Proposition 2 of [25], we only need to estimate for each $t \geq 1$, with $\mathcal{F}_0(r) = \{f \in \mathcal{F} - f_0 : \|f\|_{L_2(P)} \leq r\}$,

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}_0(2^j t\delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i f(X_i) \right| \right)^2, \quad \mathbb{E}\left(\sup_{f\in\mathcal{F}_0(2^j t\delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f^2(X_i) \right| \right)^2.$$

By the contraction principle for Rademacher processes and the moment inequality Lemma 2, we only need to estimate the sum of

(4.3)
$$(I) \equiv \left(\mathbb{E} \sup_{f\in\mathcal{F}_0(2^j t\delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i f(X_i) \right| \right)^2 + \left(\mathbb{E} \sup_{f\in\mathcal{F}_0(2^j t\delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right)^2$$

and

(4.4)
$$(II) \equiv \left(2^j t\delta_n (\|\xi_1\|_2 \vee 1)\right)^2 + n^{-1} \cdot \mathbb{E} \max_{1\leq i\leq n} (|\xi_i| \vee 1)^2 \cdot \|F_0(2^j t\delta_n)\|^2_{L_2(P)}.$$

For the first summand (4.3), by the Koltchinskii-Pollard maximal inequality for empirical processes (cf. Lemma 3), since $\mathcal{F}$ is of uniform VC-type, it follows that

$$\max_{1\leq k\leq n} \mathbb{E} \sup_{f\in\mathcal{F}_0(2^j t\delta_n)} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \varepsilon_i f(X_i) \right| \leq C_{\mathcal{F}} \|F_0(2^j t\delta_n)\|_{L_2(P)} \leq C'_{\mathcal{F}} (2^j t)^\gamma \delta_n^\gamma.$$

We may apply the multiplier inequality Lemma 1 with $\psi_n(k) \equiv \sqrt{k} C'_{\mathcal{F}} (2^j t)^\gamma \delta_n^\gamma$ to see that

$$\mathbb{E} \sup_{f \in \mathcal{F}_0(2^j t \delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f(X_i) \right| \le 4 C'_{\mathcal{F}} (2^j t)^\gamma \|\xi_1\|_{2,1} \delta_n^\gamma.$$

Hence,

$$(4.3) \le C_{\mathcal{F},\xi} (2^j t \delta_n)^{2\gamma}. \tag{4.5}$$

(4.4) is easy to handle by noting that $\mathbb{E} \max_{1 \le i \le n} (|\xi_i| \vee 1)^2 \lesssim n$ under the assumption that $\|\xi_1\|_2 < \infty$, which entails that

$$(4.4) \le C_\xi \left((2^j t \delta_n)^2 + (2^j t \delta_n)^{2\gamma} \right). \tag{4.6}$$

Combining (4.5) and (4.6) and the arguments in the proof of Proposition 2 of [25], we have

$$\begin{aligned} \mathbb{P}\big(\|\hat{f}_n - f_0\|_{L_2(P)} \ge t \delta_n \big) &\le C_{\mathcal{F},\xi} \sum_{j \ge 0 : 2^j t \delta_n \le 2} \frac{(2^j t \delta_n)^2 + (2^j t \delta_n)^{2\gamma}.}{\left(2^{2j} t^2 \sqrt{n} \delta_n^2\right)^2} \\ &\le C'_{\mathcal{F},\xi} \big(n \delta_n^{2(2-\gamma)} \big)^{-1} \sum_{j \ge 0} \frac{1}{(2^j t)^{4-2\gamma}} \le C''_{\mathcal{F},\xi} t^{-2}, \end{aligned}$$

where the last inequality follows from the choice of δ_n. Now the claim of the theorem (in the form of Remark 1 (1) and under the more general condition as in Remark 2) follows by integrating the above tail estimate. □

4.3. **Proof of Theorem 2.** The basic device we will use to derive a lower bound for the risk of the least squares estimator is the following.

Proposition 1 (Proposition 6 of [25])**.** *Let*

$$F_n(\delta) \equiv \sup_{f \in \mathcal{F} - f_0 : P f^2 \le \delta^2} (\mathbb{P}_n - P)(2\xi f - f^2) - \delta^2 \equiv E_n(\delta) - \delta^2.$$

Suppose that $0 < \delta_1 < \delta_2$ are such that $E_n(\delta_1) < F_n(\delta_2)$. Then there exists a LSE f_n^* such that $\|f_n^* - f_0\|_{L_2(P)} \ge \delta_1$.

The key ingredient in applying the above device is the following.

Proposition 2. *For any $\gamma \in (0,1]$, there exists some VC-subgraph class $\tilde{\mathcal{F}}$ satisfying (2.1) with the following property: for each $\varepsilon \in (0, 1/2)$, there exists some law for ξ_1 with $\|\xi_1\|_{2(1-\varepsilon)} < \infty$ such that*

(1) for any $\vartheta \ge 4$, there exists some $\mathfrak{p} > 0$, with $\delta_2 \equiv \vartheta n^{-\frac{1}{2(2-\gamma)}}$,

$$\mathbb{P}\left(F_n(\delta_2) \ge \frac{1}{2} c_1 \vartheta^\gamma n^{-\frac{1}{2-\gamma}} \tau_n(\varepsilon, \gamma) \right) \ge 2\mathfrak{p},$$

holds for n large enough depending on $\varepsilon, \vartheta, \gamma$. Here c_1 depends on ε, γ, and $\tau_n(\varepsilon, \gamma) \equiv n^{\frac{1-\gamma}{2-\gamma} \cdot \frac{\varepsilon}{2-\varepsilon}}$.

(2) for any $\rho > 0$, with $\delta_1 \equiv \rho n^{-\frac{1}{2(2-\gamma)}-\beta_\varepsilon}$,

$$\mathbb{P}\left(E_n(\delta_1) \le \mathfrak{p}^{-1} C_{\varepsilon,\xi} \rho^\gamma n^{-\frac{1}{2-\gamma}} \omega_n(\varepsilon,\gamma)\right) \ge 1 - \mathfrak{p}.$$

Here $\omega_n(\varepsilon,\gamma) = n^{-\gamma\beta_\varepsilon + \frac{\varepsilon}{2(1-\varepsilon)}}$.

In (1)-(2) above, $F_n(\delta) \equiv \sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta^2}(\mathbb{P}_n - P)(2\xi f - f^2) - \delta^2 \equiv E_n(\delta) - \delta^2$.

The proof of Proposition 2 relies on a delicate construction of a tree-structured $\tilde{\mathcal{F}}$, and a sequence of technical arguments including concentration of empirical processes, the Paley-Zygmund moment argument, and an exact characterization of the size of the maxima of summations. To ease reading, a formal proof of Proposition 2 will be given in Section 5.

Proof of Theorem 2. Let $f_0 = 0$. In order to apply Proposition 1, we only need to require an order in the exponent of $\tau_n(\cdot,\cdot)$ and $\omega_n(\cdot,\cdot)$ in Proposition 2, by making a good choice of β_ε. To this end, it suffices to require

$$-\gamma\beta_\varepsilon + \frac{\varepsilon}{2(1-\varepsilon)} < \frac{1-\gamma}{2-\gamma}\frac{\varepsilon}{2-\varepsilon} \Leftrightarrow \beta_\varepsilon > \frac{\varepsilon}{\gamma}\left[\frac{2-\varepsilon\gamma}{(2-\varepsilon)(2-\gamma)(2-2\varepsilon)}\right].$$

Since $\varepsilon \in (0, 1/2)$ and $\gamma \in (0, 1]$, we may choose $\beta_\varepsilon = (2/\gamma)\cdot\varepsilon$, along with any $\vartheta \ge 4$ and $\rho > 0$ small enough to conclude. □

4.4. **Proof of Theorem 3.** The proof of Theorem 3 follows from a more principled oracle inequality presented below—it captures the essential geometric property in the model that accounts for both the adaptation and robustness property of the shape-restricted LSE up to error distributions with an $L_{2,1}$ moment.

4.4.1. *The general oracle inequality.* First some definitions.

Definition 3. $\mathcal{F}$ is said to satisfy a *convexity-based shape constraint* (under P) if $\mathcal{F}$ is convex, and $\mathcal{F}(\delta) = \{f \in \mathcal{F} : Pf^2 \le \delta^2\}$ admits a convex envelope $F(\delta)$.

Definition 4. $\mathcal{G} \subset \mathcal{F}$ is said to be a *basic adaptive subset* of $\mathcal{F}$ if $\mathcal{F} - \mathcal{G} \subset \mathcal{F}$. $\mathcal{G}_m$ is said to be an *m-th order adaptive subset* of $\mathcal{F}$ if for any $g_m \in \mathcal{G}_m$, there is an interval partition $\{I_j\}_{j=1}^m$ of $\mathcal{X} = [0,1]$ and elements $\tilde{g}_j \in \mathcal{G}$ such that $g_m = \sum_{i=1}^m \mathbf{1}_{I_j}\tilde{g}_j \in \mathcal{F}$.

Before stating the general oracle inequality, recall that a function class $\mathcal{F}$ defined on $\mathcal{X} = [0,1]$ is called VC-major if the sets $\{x \in \mathcal{X} : f(x) \ge t\}$ with f ranging over $\mathcal{F}$ and t over $\mathbb{R}$ form a VC-class of sets.

Theorem 5. *Consider the regression model (1.1) and the LSE $\hat{f}_n$ in (1.2). Suppose that $\|f_0\|_\infty \vee \|f_0^*\|_\infty < \infty$, and that $\xi_1, \dots, \xi_n$ are mean zero errors independent of i.i.d. covariates X_i's with $\|\xi_1\|_{2,1} < \infty$. Further assume that: (i) $\mathcal{F}$ satisfies a convexity-based shape constraint, and $\mathcal{F} \cap L_\infty(B)$ is a VC-major class for any $B > 0$, and (ii) $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$. Then for any*

$\delta \in (0,1)$, there exists $c \equiv c(\delta, \|\xi\|_{2,1}, \mathcal{F}, \|f_0\|_\infty, \|f_0^*\|_\infty) > 0$ such that with probability $1-\delta$,

$$\|\hat{f}_n - f_0^*\|_{L_2(P)}^2 \le c \inf_{m \in \mathbb{N}} \left(\inf_{f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)} \|f_m - f_0^*\|_{L_2(P)}^2 + \frac{m}{n} \cdot \log^2 n \right),$$

where $f_0^* = \mathrm{argmin}_{g \in \mathcal{F} \cap L_2(P)} \|f_0 - g\|_{L_2(P)}$, and $\mathcal{G}_m$ is an m-th order adaptive subset of $\mathcal{F}$.

The proof of Theorem 5 will be deferred to the next subsection. We first use it to prove Theorem 3. To this end, we only need to check: (i) the convexity-based shape constraint and VC-major condition of the isotonic and convex models; and (ii) the stochastic boundedness condition for the corresponding LSEs $\hat{f}_n$.

Proof of Theorem 3. For the isotonic model $\mathcal{F}$, $\mathcal{F}$ is clearly convex, and (3.2) is an envelope for $\mathcal{F}(\delta)$ by the L_2 constraint and monotonicity of the function class. Furthermore, it is clear by definition that $\mathcal{F} \cap L_\infty(B)$ is VC-major. Similarly we can verify that the convex model satisfies both the convexity-based shape constraint with the envelope (3.2) (cf. Lemma 9) and the VC-major condition.

The stochastic boundedness of the isotonic and convex LSEs is established in the following lemma:

Lemma 5. *If $\|f_0\|_\infty < \infty$ and $\|\xi_1\|_2 < \infty$, then both the canonical isotonic and convex LSEs are stochastically bounded: $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$.*

For the isotonic LSE, we use an explicit min-max representation (cf. [36]) to prove this lemma, while for the convex LSE, the explicit characterization of the convex LSE derived in [21] plays a crucial role. The details of the proof of this lemma can be found in Section 5. Now the claim of Theorem 3 follows from Theorem 5, by noting that $\|f_0^*\|_\infty < \infty$ under $\|f_0\|_\infty < \infty$, and that $\inf_{f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)} \|f_m - f_0^*\|_{L_2(P)}^2 = \inf_{f_m \in \mathcal{G}_m} \|f_m - f_0^*\|_{L_2(P)}^2$ for isotonic model, and the same holds for the convex model when $L_\infty(\|f_0^*\|_\infty$ is replaced by $L_\infty(C\|f_0^*\|_\infty)$ for some large enough $C > 0$. □

4.4.2. *Proof of Theorem 5.* The first ingredient of the proof is the following proposition relating the convergence rate of $\hat{f}_n$ to the size of localized empirical processes.

Proposition 3. *Consider the regression model (1.1) and the least squares estimator $\hat{f}_n$ in (1.2). Suppose that $\xi_1, \ldots, \xi_n$ are mean-zero random variables independent of $X_1, \ldots, X_n$, and $\mathcal{F}$ is convex with $\mathcal{F} - f_0^* \subset L_\infty(1)$. Further assume that*

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

$$\mathbb{E} \sup_{f \in \mathcal{F}: \|f - f_0^*\|_{L_2(P)} \le \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f - f_0^*)(X_i) \right| \lesssim \phi_n(\delta), \tag{4.7}$$

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

hold for some ϕ_n such that $\delta\mapsto\phi_n(\delta)/\delta$ is non-increasing. Then $\|\hat{f}_n-f_0^*\|_{L_2(P)}=\mathcal{O}_{\mathbf{P}}(\delta_n)$ holds for any δ_n such that $\phi_n(\delta_n)\le\sqrt{n}\delta_n^2$.

Proof. This is a special case of Proposition 4, the proof of which will be given therein. □

By Proposition 3, we only need to control the size of the empirical processes (4.7) centered at f_0^*. The following lemma will be useful in this regard by approximating f_0^* via arbitrary $f_m\in\mathcal{G}_m$.

Lemma 6. *Suppose that the hypotheses of Theorem 5 hold. Let $\{\delta_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers such that $\delta_n\ge 1/n$. Then for any $f_m\in\mathcal{G}_m\cap L_\infty(\|f_0^*\|_\infty)$ and $B>0$,*

$$\begin{aligned}\max\Bigg\{&\mathbb{E}\sup_{f\in\mathcal{F}:f-f_0^*\in L_2(\delta_n)\cap L_\infty(B)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i(f-f_0^*)(X_i)\right|,\\ &\mathbb{E}\sup_{f\in\mathcal{F}:f-f_0^*\in L_2(\delta_n)\cap L_\infty(B)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i(f-f_0^*)(X_i)\right|,\\ &\mathbb{E}\sup_{f\in\mathcal{F}:f-f_0^*\in L_2(\delta_n)\cap L_\infty(B)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i(f-f_0^*)(X_i)(f_0-f_0^*)(X_i)\right|\Bigg\}\\ &\le C_{\mathcal{F},\|f_0\|_\infty,\|f_0^*\|_\infty,B}\cdot\|\xi_1\|_{2,1}\sqrt{\log(1/\delta_n)}\bar{L}_n\cdot(\delta_n\vee\|f_m-f_0^*\|_{L_2(P)})\sqrt{m},\end{aligned}$$

where $\bar{L}_n\equiv\sqrt{\log n}$.

To prove Lemma 6, we need the following form of a multiplier inequality proved in Proposition 1 of [25].

Lemma 7. *Suppose that $\xi_1,\dots,\xi_n$ are i.i.d. mean-zero random variables independent of i.i.d. $X_1,\dots,X_n$. Then for any function class $\mathcal{F}$,*

$$\mathbb{E}\left\|\sum_{i=1}^n\xi_i f(X_i)\right\|_{\mathcal{F}}\le\mathbb{E}\left[\sum_{k=1}^n(|\eta_{(k)}|-|\eta_{(k+1)}|)\mathbb{E}\left\|\sum_{i=1}^k\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right]\tag{4.8}$$

where $|\eta_{(1)}|\ge\cdots\ge|\eta_{(n)}|\ge|\eta_{(n+1)}|\equiv 0$ are the reversed order statistics for $\{|\xi_i-\xi_i'|\}_{i=1}^n$ with $\{\xi_i'\}$ being an independent copy of $\{\xi_i\}$.

The following entropy estimate for bounded VC-major classes will be useful.

Lemma 8. *Let $\mathcal{F}_0\subset L_\infty(1)$ be a VC-major class defined on $\mathcal{X}$. Then there exists some constant $C\equiv C_{\mathcal{F}_0}>0$ such that for any $\mathcal{F}\subset\mathcal{F}_0$, and any probability measure Q, the entropy estimate*

$$\log\mathcal{N}\big(\varepsilon\|F\|_{L_2(Q)},\mathcal{F},L_2(Q)\big)\le\frac{C}{\varepsilon}\log\left(\frac{C}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon\|F\|_{L_2(Q)}}\right),\ \text{for all }\varepsilon\in(0,1)$$

holds for any envelope F of $\mathcal{F}$.

The proof of this lemma essentially follows from page 1171-1172 of [18] with a minor modification. We include some details in Section 5 for the convenience of the reader.

We also need the following lemma concerning the envelope of a convex function given constraints on its L_2 size. The proof can be found in Lemma 7.3 of [22].

Lemma 9. *If f is a convex function on $[0,1]$ with $\int_0^1 |f(x)|^2 \, \mathrm{d}x \le 1$, then $|f(x)| \le 2\sqrt{3}\big(x^{-1/2} \vee (1-x)^{-1/2}\big)$ for all $x \in (0,1)$.*

Proof of Lemma 6. In the proof we omit the dependence on $L_\infty(B)$ if there is no confusion. All three empirical processes can be handled in essentially the same way so we focus on the most difficult first one (with ξ_i's only admitting a $L_{2,1}$ moment). We will apply Lemma 7 in the following form:

$$\mathbb{E} \sup_{f\in\mathcal{F}: f-f_0^* \in L_2(\delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (f-f_0^*)(X_i) \right| \le 3\|\xi_1\|_{2,1} \max_{1\le k\le n} \mathbb{E} \sup_{f\in\mathcal{F}: f-f_0^* \in L_2(\delta_n)} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i (f-f_0^*)(X_i) \right|. \tag{4.9}$$

To see this, note that the right hand side of (4.8) can be bounded by

$$\mathbb{E}\bigg[\sum_{k=1}^n \sqrt{k}(|\eta_{(k)}| - |\eta_{(k+1)}|) \bigg] \cdot \max_{1\le k\le n} \mathbb{E} \bigg\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i f(X_i) \bigg\|_{\mathcal{F}}$$

where $\mathbb{E}\big[\sum_{k=1}^n \sqrt{k}(|\eta_{(k)}| - |\eta_{(k+1)}|)\big] \le \sqrt{n}\|\eta_1\|_{2,1} \le 3\sqrt{n}\|\xi_1\|_{2,1}$. The first inequality follows from similar lines as in the proof of Theorem 1 of [25] and the second inequality uses Problem 2 on page 186 of [45]. This proves (4.9). Note that any $f_m \in \mathcal{G}_m$ has a representation $f_m = \sum_{j=1}^m g_j \mathbf{1}_{I_j}$, where $\{I_j = [x_j, x_{j+1}]\}_{j=1}^m$ is a partition of $\mathcal{X} = [0,1]$ with $x_1 = 0, x_{m+1} = 1$ and $g_j \in \mathcal{G}$. Then for any $f_m \in \mathcal{G}_m$, the empirical process localized at f_0^* can be controlled via

$$\mathbb{E} \sup_{f\in\mathcal{F}: f-f_0^* \in L_2(\delta_n)\cap L_\infty(B)} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i (f-f_0^*)(X_i) \right| \le \mathbb{E} \sup_{\substack{f\in\mathcal{F}: \|f-f_m\|_{L_2(P)} \le \delta_n + \|f_m - f_0^*\|_{L_2(P)}, \\ \|f\|_\infty \le B + \|f_0^*\|_\infty}} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i (f-f_m)(X_i) \right| + \|f_0^* - f_m\|_{L_2(P)}, \tag{4.10}$$

where the second term holds because the collection $\{f_0^* - f_m\}$ consists of just one element. The first term in the above term can be further bounded

by

(4.11)
$$\sum_{j=1}^{m} \mathbb{E}\Bigg[\frac{\sqrt{k_j}}{\sqrt{k}} \mathbb{E}\Bigg[\sup_{\substack{f\in\mathcal{F}: \|f-f_m\|_{L_2(P)} \le \delta_n + \|f_m - f_0^*\|_{L_2(P)}, \\ \|f\|_\infty \le B + \|f_0^*\|_\infty}} \left| \frac{1}{\sqrt{k_j}} \sum_{X_i \in I_j} \varepsilon_i (f - g_j)(X_i) \right| \Bigg| k_j(\boldsymbol{X}) = k_j \Bigg] \Bigg]$$
$$\le \sum_{j=1}^{m} \mathbb{E}\Bigg[\frac{\sqrt{k_j}}{\sqrt{k}} \mathbb{E}\Bigg[\sup_{\substack{f|_{I_j} \in \mathcal{F}|_{I_j}: \\ \|f\|_\infty \le B + 2\|f_0^*\|_\infty, \\ Pf^2 \le (\delta_n + \|f_m - f_0^*\|_{L_2(P)})^2}} \left| \frac{1}{\sqrt{k_j}} \sum_{X_i \in I_j} \varepsilon_i f|_{I_j}(X_i) \right| \Bigg| k_j(\boldsymbol{X}) = k_j \Bigg] \Bigg]$$

where $k_j(\boldsymbol{X}) = \sum_{i=1}^{k} \mathbf{1}_{I_j}(X_i)$, and in the second line we used the definition of a basic adaptive subset (cf. Definition 4). From now on we write $\tilde{\delta}_n \equiv \delta_n + \|f_m - f_0^*\|_{L_2(P)}$ and $B_0 \equiv B + 2\|f_0^*\|_\infty$ for notational convenience. Since $\big(\mathcal{F} \cap L_\infty(B_0)\big)|_{I_j}$ is VC-major, so is its subset $\mathcal{F}_{I_j}(\tilde{\delta}_n) \equiv \{f|_{I_j} \in \big(\mathcal{F} \cap L_\infty(B_0)\big)|_{I_j} : Pf^2 \le \tilde{\delta}_n^2\}$. It follows by Lemma 8 that there exists some $C \equiv C_{\mathcal{F},B_0} > 0$ such that for any probability measure Q on I_j, and any $\varepsilon \in (0,1)$,

$$\log \mathcal{N}\Big(\varepsilon \|F_{I_j}(\tilde{\delta}_n)\|_{L_2(Q)}, \mathcal{F}_{I_j}(\tilde{\delta}_n), L_2(Q)\Big) \le \frac{C}{\varepsilon} \log\left(\frac{C}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \|F_{I_j}(\tilde{\delta}_n)\|_{L_2(Q)}}\right),$$

where $F_{I_j}(\delta)$ is any envelope for $\mathcal{F}_{I_j}(\delta)$. This enables us to apply the Koltchinskii-Pollard maximal inequality to see that the summand (=conditional expectation) in the second line of (4.11) can be bounded by (further conditioning on which X_i's lie in the interval I_j, each case corresponds to i.i.d. uniforms on I_j)

$$\int_0^1 \sqrt{\frac{C}{\varepsilon} \log\left(\frac{C}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \inf_Q \|F_{I_j}(\tilde{\delta}_n)\|_{L_2(Q)}}\right)}\, \mathrm{d}\varepsilon \cdot \sqrt{P_{I_j} F_{I_j}^2(\tilde{\delta}_n)}, \tag{4.12}$$

where P_{I_j} is the uniform distribution on I_j.

In order to evaluate (4.12), note that by the definition of convexity-based shape constraint and Lemma 9, the envelopes $F_{I_j}(\delta)$'s can be taken as the restrictions of the global envelope

$$F(\delta)(x) \equiv \left(\frac{\delta}{\sqrt{x}} \vee \frac{\delta}{\sqrt{1-x}}\right) \wedge B_0$$

to the I_j's. Without loss of generality we assume: (i) $B_0 = 1$, (ii) $\tilde{\delta}_n^2 < 1/2$ and (iii) $\tilde{\delta}_n^2$ and $1 - \tilde{\delta}_n^2$ are one of the endpoints of some intervals in $\{I_j\}$ (otherwise, we may take an alternative representation of $f_m \in \mathcal{G}_{m+2}$ by adding these two points).

Note that $\inf_Q \|F_{I_j}(\tilde{\delta}_n)\|_{L_2(Q)} \geq \sqrt{2}\tilde{\delta}_n > 1/n$ by the assumption $\delta_n \geq 1/n$, and hence the integral term in (4.12) can be bounded by

$$\int_0^1 \sqrt{\frac{C}{\varepsilon} \log\left(\frac{C}{\varepsilon}\right) \log\left(\frac{n}{\varepsilon}\right)}\, \mathrm{d}\varepsilon \lesssim \sqrt{\log n} \equiv \bar{L}_n.$$

To handle the $\sqrt{P_{I_j} F_{I_j}^2(\tilde{\delta}_n)}$ term in (4.12), define the index sets $\mathcal{J}_1 \equiv \{1 \leq j \leq m : I_j \subset [0, \tilde{\delta}_n^2]\}, \mathcal{J}_2 \equiv \{1 \leq j \leq m : I_j \subset [\tilde{\delta}_n^2, 1 - \tilde{\delta}_n^2]\}$ and $\mathcal{J}_3 \equiv \{1 \leq j \leq m : I_j \subset [1 - \tilde{\delta}_n^2, 1]\}$. It is easy to see that $\mathcal{J}_1 \cup \mathcal{J}_2 \cup \mathcal{J}_3 = \{1, \ldots, m\}$. Clearly for $j \in \mathcal{J}_1 \cup \mathcal{J}_3$,

$$P_{I_j} F_{I_j}^2(\tilde{\delta}_n) = |I_j|^{-1} \int_{I_j} F_{I_j}^2(\tilde{\delta}_n)(x)\, \mathrm{d}x \leq 1,$$

and for $j \in \mathcal{J}_2$,

$$\begin{aligned} P_{I_j} F_{I_j}^2(\tilde{\delta}_n) &\leq |I_j|^{-1} \tilde{\delta}_n^2 \int_{x_j}^{x_{j+1}} \left(\frac{1}{x} \vee \frac{1}{1-x}\right) \mathrm{d}x \\ &\leq |I_j|^{-1} \tilde{\delta}_n^2 \left[\log\left(\frac{x_{j+1}}{x_j}\right) \vee \log\left(\frac{1-x_j}{1-x_{j+1}}\right)\right]. \end{aligned}$$

Summarizing the above discussion shows that we can further bound (4.11) by a $\mathcal{O}(\bar{L}_n)$ multiple of

(4.13)

$$\begin{aligned} &\sum_{j \in \mathcal{J}_1 \cup \mathcal{J}_3} \mathbb{E}\left[\sqrt{\frac{k_j}{k}} \cdot 1\right] + \sum_{j \in \mathcal{J}_2} \tilde{\delta}_n \cdot \mathbb{E}\left[\sqrt{\frac{k_j}{k}} \cdot \sqrt{\frac{\log(x_{j+1}) - \log(x_j)}{x_{j+1} - x_j}}\right] \\ &\qquad + \sum_{j \in \mathcal{J}_2} \tilde{\delta}_n \cdot \mathbb{E}\left[\sqrt{\frac{k_j}{k}} \cdot \sqrt{\frac{\log(1-x_j) - \log(1-x_{j+1})}{(1-x_j) - (1-x_{j+1})}}\right] \\ &\equiv (I) + (II) + (III). \end{aligned}$$

The first term of (4.13) is easy to handle: by the Cauchy-Schwarz inequality,

$$(I) \leq \sqrt{k^{-1}\left(\mathbb{E} \sum_{j \in \mathcal{J}_1 \cup \mathcal{J}_3} k_j(\boldsymbol{X})\right) \cdot |\mathcal{J}_1 \cup \mathcal{J}_3|} \leq \sqrt{\sum_{j \in \mathcal{J}_1 \cup \mathcal{J}_3} |I_j|} \cdot \sqrt{m} \lesssim \tilde{\delta}_n \sqrt{m}.$$

The second and third terms of (4.13) can be handled in a similar fashion; we only consider the second term of (4.13). Again by the Cauchy-Schwarz inequality,

$$\begin{aligned} (II) &\leq \tilde{\delta}_n \sqrt{m} \cdot \sqrt{\mathbb{E}\left[\sum_{j \in \mathcal{J}_2} \frac{k_j(\boldsymbol{X})}{k} \cdot \frac{\log(x_{j+1}) - \log(x_j)}{x_{j+1} - x_j}\right]} \\ &= \tilde{\delta}_n \sqrt{m} \sqrt{\sum_{j \in \mathcal{J}_2} \big(\log(x_{j+1}) - \log(x_j)\big)} \lesssim \sqrt{m} \cdot \tilde{\delta}_n \sqrt{\log(1/\tilde{\delta}_n)}. \end{aligned}$$

Collecting the above estimates, we see that (4.11) can be bounded by a constant multiple of $\sqrt{m}\cdot\tilde{\delta}_n\sqrt{\log(1/\tilde{\delta}_n)}\bar{L}_n$. Thus, (4.10) yields that

$$\max_{1\le k\le n}\mathbb{E}\sup_{f\in\mathcal{F}:f-f_0^*\in L_2(\delta_n)}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\varepsilon_i(f-f_0^*)(X_i)\right|\le C'\sqrt{m}\cdot\tilde{\delta}_n\sqrt{\log(1/\tilde{\delta}_n)}\bar{L}_n.$$

Combined with (4.9), the claim of the lemma follows. □

Proof of Theorem 5. The proof follows easily from the reduction scheme Proposition 4 and Lemma 6 by solving a quadratic inequality. We provide some details below. Abusing notation, we let $f_m\in\operatorname{argmin}_{g_m\in\mathcal{G}_m}\|g_m-f_0^*\|_{L_2(P)}$ and m be the index attaining the infimum of the oracle inequality in the statement of the theorem. We only need to choose δ_n such that

$$\sqrt{m}(\delta_n+\|f_m-f_0^*\|_{L_2(P)})\sqrt{\log(1/\delta_n)}\bar{L}_n\le c_{\delta,\mathcal{F},\|f_0^*\|_\infty,\|\xi\|_{2,1}}\sqrt{n}\delta_n^2.$$

Suppose $\log(1/\delta_n)\lesssim\log n$. Then we can easily solve for the zeros for quadratic forms to see that the inequality in the last display holds if

$$\delta_n^2\gtrsim\frac{m\bar{L}_n^2\log n}{n}+\sqrt{\frac{m\bar{L}_n^2\log n}{n}}\|f_m-f_0^*\|_{L_2(P)}.$$

The assumption $\log(1/\delta_n)\lesssim\log n$ apparently holds. The right hand side of the above display can be further bounded up to a constant by $\frac{m\bar{L}_n^2\log n}{n}+\|f_m-f_0^*\|_{L_2(P)}^2$ by the basic inequality $ab\le(a^2+b^2)/2$, thereby completing the proof of Theorem 5. □

4.5. **Proof of Theorem 4.** The proof of Theorem 4 follows a similar strategy as that of Theorem 5. First we need the following reduction scheme.

Proposition 4. *Consider the additive model (3.3) and the least squares estimator $\hat{f}_n$ in (3.4). Suppose that $\xi_1,\dots,\xi_n$ are mean-zero random variables independent of $(X_1,Z_1),\dots,(X_n,Z_n)$, and $\mathcal{F}$ is convex with $\mathcal{F}-f_0^*\subset L_\infty(1)$. Further assume that all three parts of (4.7) and*

(4.14)

$$\mathbb{E}\sup_{\substack{f\in\mathcal{F}:\|f-f_0^*\|_{L_2(P)}\le\delta\\ h\in\mathcal{H}}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(f-f_0^*)(X_i)(h-(\phi_0-f_0))(X_i,Z_i)\right|\lesssim\phi_n(\delta),$$

hold for some ϕ_n such that $\delta\mapsto\phi_n(\delta)/\delta$ is non-increasing. Then $\|\hat{f}_n-f_0^*\|_{L_2(P)}=\mathcal{O}_{\mathbf{P}}(\delta_n)$ holds for any δ_n such that $\phi_n(\delta_n)\le\sqrt{n}\delta_n^2$.

Proof. Recall that $f_0=P_Z\phi_0(\cdot,Z)$. By the definition of the LSE,

$$\begin{aligned}
&\mathbb{P}_n(\phi_0+\xi-\hat{f}_n-\hat{h}_n)^2\le\mathbb{P}_n(\phi_0+\xi-f_0^*-\hat{h}_n)^2\\
\Leftrightarrow\quad&\mathbb{P}_n(f_0^*-\hat{f}_n)\big(2\phi_0+2\xi-\hat{f}_n-f_0^*-2\hat{h}_n\big)\le0\\
\Leftrightarrow\quad&\mathbb{P}_n(f_0^*-\hat{f}_n)^2+2\mathbb{P}_n(f_0^*-\hat{f}_n)\big(\phi_0+\xi-f_0^*-\hat{h}_n\big)\le0\\
\Leftrightarrow\quad&-\mathbb{P}_n(f_0^*-\hat{f}_n)^2-2\mathbb{P}_n(f_0^*-\hat{f}_n)\xi-2\mathbb{P}_n(f_0^*-\hat{f}_n)(f_0-f_0^*)
\end{aligned}$$

$$
\begin{aligned}
&\quad - 2\mathbb{P}_n(f_0^* - \hat{f}_n)(\phi_0 - f_0 - \hat{h}_n) \geq 0 \\
\Leftrightarrow \quad & -(\mathbb{P}_n - P)\Big[(f_0^* - \hat{f}_n)^2 - 2\xi(f_0^* - \hat{f}_n)\Big] - P(f_0^* - \hat{f}_n)^2 \\
&\quad - 2(\mathbb{P}_n - P)(f_0^* - \hat{f}_n)(f_0 - f_0^*) - 2P(f_0^* - \hat{f}_n)(f_0 - f_0^*) \\
&\quad - 2(\mathbb{P}_n - P)(f_0^* - \hat{f}_n)(\phi_0 - f_0 - \hat{h}_n) \geq 0.
\end{aligned}
$$

The last equivalence holds since

$$
\begin{aligned}
&P(f_0^* - \hat{f}_n)(X)(\phi_0 - f_0 - \hat{h}_n)(X, Z) \\
&= P\Big[(f_0^* - \hat{f}_n)(X) P\big[(\phi_0 - f_0 - \hat{h}_n)(X, Z)\big|X\big]\Big] \\
&= P\Big[(f_0^* - \hat{f}_n)(X)\big(P\big[\phi_0(X, Z)\big|X\big] - f_0(X) - P\hat{h}_n(Z)\big)\Big] = 0,
\end{aligned}
$$

where we used (i) $P[\phi_0(X, Z)|X] = f_0(X)$, and (ii) $Ph = 0$ for all $h \in \mathcal{H}$. Now since $f_0^* \in \operatorname{argmin}_{g \in \mathcal{F} \cap L_2(P)} \|f_0 - g\|_{L_2(P)}$, it follows from the convexity of $\mathcal{F}$ that $P(f_0^* - \hat{f}_n)(f_0 - f_0^*) \geq 0$ [more specifically, for each $\varepsilon > 0$, since $(1 - \varepsilon)f_0^* + \varepsilon \hat{f}_n^* \in \mathcal{F} \cap L_2(P)$ by convexity of $\mathcal{F}$, the definition of f_0^* yields that $P(f_0 - f_0^*)^2 \leq P(f_0 - (1 - \varepsilon)f_0^* - \varepsilon \hat{f}_n)^2 = P(f_0 - f_0^* + \varepsilon(f_0^* - \hat{f}_n))^2$. The claim follows by expanding the square and taking $\varepsilon \to 0$]. This implies that, with $S_j(\delta_n) \equiv \{f \in \mathcal{F} : 2^{j-1}\delta_n < \|f - f_0^*\|_{L_2(P)} \leq 2^j \delta_n\}$, on the event $\{2^{j-1}\delta_n < \|\hat{f}_n - f_0^*\|_{L_2(P)} \leq 2^j \delta_n\}$, it holds that

$$
\begin{aligned}
&\sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)(f - f_0^*)^2| + 2 \sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)\xi(f - f_0^*)| \\
&\quad + 2 \sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)(f - f_0^*)(f_0 - f_0^*)| \\
&\quad + 2 \sup_{f \in S_j(\delta_n), h \in \mathcal{H}} |(\mathbb{P}_n - P)(f - f_0^*)(h - (\phi_0 - f_0))| \\
&\geq -(\mathbb{P}_n - P)\Big[(f_0^* - \hat{f}_n)^2 - 2\xi(f_0^* - \hat{f}_n)\Big] \\
&\quad - 2(\mathbb{P}_n - P)(f_0^* - \hat{f}_n)(f_0 - f_0^*) - 2(\mathbb{P}_n - P)(f_0^* - \hat{f}_n)(\phi_0 - f_0 - \hat{h}_n) \\
&\geq 2^{2j-2}\delta_n^2.
\end{aligned}
$$

Hence by symmetrization, the contraction principle for Rademacher processes and the assumptions we see that

$$
\begin{aligned}
&\mathbb{P}\big(\|\hat{f}_n - f_0^*\|_{L_2(P)} > 2^{M-1}\delta_n\big) \\
&\leq \sum_{j \geq M} \mathbb{P}\Big(\sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)(f - f_0^*)^2| + 2 \sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)\xi(f - f_0^*)| \\
&\qquad + 2 \sup_{f \in S_j(\delta_n)} |(\mathbb{P}_n - P)(f - f_0^*)(f_0 - f_0^*)|
\end{aligned}
$$

$$+ 2 \sup_{f\in S_j(\delta_n), h\in\mathcal{H}} |(\mathbb{P}_n - P)(f - f_0^*)(h - (\phi_0 - f_0))| \geq 2^{2j-2}\delta_n^2\Big)$$

$$\lesssim \sum_{j\geq M} \big(2^{2j}\sqrt{n}\delta_n^2\big)^{-1}\Big(\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_0(2^j\delta_n)} \vee \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_0(2^j\delta_n)\otimes\xi}$$

$$\vee\, \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_0(2^j\delta_n)\otimes(f_0-f_0^*)} \vee \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_0(2^j\delta_n)\otimes(\mathcal{H}-(\phi_0-f_0))}\Big)$$

$$\leq C\sum_{j\geq M} \frac{\phi_n(2^j\delta_n)}{2^{2j}\sqrt{n}\delta_n^2} \leq C\sum_{j\geq M} \frac{\phi_n(\delta_n)}{2^j\sqrt{n}\delta_n^2} \lesssim \sum_{j\geq M} 2^{-j} \to 0$$

as $M \to \infty$. Here we denote $\mathcal{F}_0 \equiv \mathcal{F} - f_0^*$, and in the last sequence of inequalities we used the assumption that $\delta \mapsto \phi_n(\delta)/\delta$ is non-decreasing and the definition of δ_n. This completes the proof. □

By Proposition 4, apart from the empirical processes in Lemma 6, we also need to control the empirical process (4.14) indexed by a suitably localized subset of $\mathcal{F}\otimes(\mathcal{H}-(\phi_0-f_0)) \equiv \{f(x)\big(h(z)-\phi_0(x,z)-f_0(x)\big) : f\in\mathcal{F}, h\in\mathcal{H}\}$. In a related work, [41] derived bounds for similar empirical processes under L_∞-type entropy conditions for both $\mathcal{F}$ and $\mathcal{H}$ (cf. Theorem 3.1 of [41]), which apparently fail for shape constrained classes.

Lemma 10. *Suppose that the hypotheses of Theorem 4 hold. Let $\{\delta_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers such that $\delta_n \geq 1/n$. Then for any $f_m \in \mathcal{G}_m \cap L_\infty(\|f_0^*\|_\infty)$, and $B > 0$,*

$$\mathbb{E} \sup_{\substack{f\in\mathcal{F}: f-f_0^*\in L_2(\delta_n)\cap L_\infty(B)\\ h\in\mathcal{H}}} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i(f-f_0^*)(X_i)(h-h_0)(X_i,Z_i)\right|$$

$$\leq C_{\mathcal{H},\mathcal{F},\|\phi_0\|_\infty,\|f_0^*\|_\infty,B}\cdot\sqrt{\log(1/\delta_n)}\bar{L}_n\cdot(\delta_n \vee \|f_m - f_0^*\|_{L_2(P)})\sqrt{m}.$$

Here $\bar{L}_n \equiv \sqrt{\log n}$.

We need some technical lemmas. Recall P_X, P_Z are the marginal probability distributions of (X,Z), i.e. uniform distribution on $[0,1]$.

Lemma 11. *Let $\mathcal{H}$ be a class of measurable functions defined on $[0,1]$, and let $f \in L_2(P_X), g \in L_2(P)$. Then for any probability measure Q on $[0,1]^2$,*

$$\mathcal{N}\big(\varepsilon\|f\otimes 1\|_{L_2(Q)}, f\otimes\big(\mathcal{H}-g\big), L_2(Q)\big) \leq \mathcal{N}\big(\varepsilon,\mathcal{H},L_\infty\big).$$

Lemma 12. *Suppose the conditions on $\mathcal{H}$ in Theorem 4 hold and $\mathcal{F}$ is the class of monotonic non-decreasing or convex functions on $[0,1]$. Then for any $\mathcal{F}'\subset\mathcal{F}\cap L_\infty(1)$ and any probability measure Q on $[0,1]^2$, the entropy estimate*

$$\log\mathcal{N}\big(\varepsilon\|F'\otimes 1\|_{L_2(Q)}, \mathcal{F}'\otimes(\mathcal{H}-(\phi_0-f_0)), L_2(Q)\big)$$

$$\lesssim \frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon\|F'\otimes 1\|_{L_2(Q)}}\right)\vee\varepsilon^{-\gamma},\ \textit{for all } \varepsilon\in(0,1)$$

holds for any envelope F' of $\mathcal{F}'$. The constant in the above estimate does not depend on the choice of $\mathcal{F}'$ or Q.

The proofs of Lemmas 11 and 12 are standard. We include the details in Section 5 for completeness.

Proof of Lemma 10. The proof follows the same strategy as that of Lemma 6. We only prove the isotonic case $\mathcal{G}_m = \mathcal{M}_m$; the convex case follows by similar arguments. As in the proof of Lemma 6, we will omit the explicit dependence on $L_\infty(B)$ if no confusion arises. Note that

(4.15)

$$\begin{aligned}
&\mathbb{E} \sup_{f\in\mathcal{F}: f-f_0^*\in L_2(\delta_n), h\in\mathcal{H}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f-f_0^*)(X_i)(h-(\phi_0-f_0))(X_i,Z_i) \right| \\
&\le \mathbb{E} \sup_{\substack{f\in\mathcal{F}: \|f-f_m\|_{L_2(P)}\le \delta_n + \|f_m - f_0^*\|_{L_2(P)}, \\ h\in\mathcal{H}}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f-f_m)(X_i)(h-(\phi_0-f_0))(X_i,Z_i) \right| \\
&\qquad + \mathbb{E} \sup_{h\in\mathcal{H}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f_m-f_0^*)(X_i)(h-(\phi_0-f_0))(X_i,Z_i) \right| \equiv (I)+(II).
\end{aligned}$$

We first handle (II) in (4.15). The entropy assumption on $\mathcal{H}$ coupled with Lemma 11 entails that the uniform entropy integral for the class $(f_m - f_0^*)\otimes \big(\mathcal{H} - (\phi_0 - f_0)\big)$ converges. By Theorem 2.14.1 of [45], we have the following estimate

$$(II) \le C_{\mathcal{H}} \|f_m - f_0^*\|_{L_2(P)}.$$

For the first term (I) in (4.15), we mimic the proof strategy in Lemma 6: any piecewise constant $f_m \in \mathcal{M}_m$ has a representation $f_m = \sum_{j=1}^m g_j \mathbf{1}_{I_j}$, where $\{I_j = [x_j, x_{j+1}]\}_{j=1}^m$ is a partition of $[0,1]$ with $x_1 = 0, x_{m+1} = 1$ and g_j takes constant values on the intervals I_j. Then for such $f_m \in \mathcal{M}_m$, write $\tilde{I}_j = I_j \times [0,1]$, we have

$$\begin{aligned}
&\sum_{j=1}^m \mathbb{E}\bigg[\frac{\sqrt{n_j}}{\sqrt{n}} \mathbb{E}\bigg[\sup_{\substack{f\in\mathcal{F}: f-f_m\in L_2(\tilde{\delta}_n) \\ h\in\mathcal{H}}} \bigg| \frac{1}{\sqrt{n_j}} \sum_{(X_i,Z_i)\in\tilde{I}_j} \varepsilon_i (f-g_j)(X_i)(h-(\phi_0-f_0))(X_i,Z_i) \bigg| \\
&\qquad\qquad \bigg| n_j(\boldsymbol{X},\boldsymbol{Z}) = n_j \bigg] \bigg] \\
&\le \sum_{j=1}^m \mathbb{E}\bigg[\frac{\sqrt{n_j}}{\sqrt{n}} \mathbb{E}\bigg[\sup_{\substack{f|_{I_j}\in\mathcal{F}|_{I_j}: \\ \|f\|_\infty \le B+2\|f_0^*\|_\infty, \\ P_X f^2 \le \delta_n^2, h\in\mathcal{H}}} \bigg| \frac{1}{\sqrt{n_j}} \sum_{(X_i,Z_I)\in\tilde{I}_j} \varepsilon_i f\otimes(h-(\phi_0-f_0))|_{\tilde{I}_j}(X_i,Z_i) \bigg| \\
&\qquad\qquad \bigg| n_j(\boldsymbol{X},\boldsymbol{Z}) = n_j \bigg] \bigg]
\end{aligned}$$

where $\tilde{\delta}_n \equiv \delta_n + \|f_m - f_0^*\|_{L_2(P)}$. Here $n_j(\boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^n \mathbf{1}_{\tilde{I}_j}(X_i, Z_i)$ and in the second line we used the fact that $(f - f_m)|_{I_j} \in \mathcal{F}|_{I_j}$. By Lemma 12 and the Koltchinskii-Pollard maximal inequality, each summand of the above display can be bounded up to a constant (depending on $\mathcal{F}, \mathcal{H}, \|\phi_0\|_\infty$) by

$$\int_0^1 \sqrt{\frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)\log\left(\frac{1}{\varepsilon\|F_{I_j}(\tilde{\delta}_n)\otimes 1\|_{L_2(Q)}}\right)\vee \varepsilon^{-\gamma}}\,\mathrm{d}\varepsilon \times \left(P_{\tilde{I}_j}F_{I_j}^2(\tilde{\delta}_n)\right)^{1/2} \lesssim \bar{L}_n \cdot \sqrt{P_{I_j}F_{I_j}^2(\tilde{\delta}_n)},$$

where $P_{\tilde{I}_j}$ is the uniform distribution on $\tilde{I}_j$ and $F_{I_j}(\delta)$ is the envelope for $\big(\mathcal{F} \cap L_\infty(B + 2\|f_0^*\|_\infty) \cap L_2(\delta)\big)|_{I_j}$, and the inequality in the above display follows from similar arguments as in the proof of Lemma 6. From here the proof proceeds along the same lines as that of the proof for Lemma 6. □

Proof of Theorem 4. The proof of Theorem 4 follows the arguments of the proof of Theorem 5 by using Proposition 4 along with Lemmas 6 and 10, combined with the stochastic boundedness of the LSE:

Lemma 13. *Suppose that the hypotheses of Theorem 4 hold (except that $\mathcal{H}$ is only required to have a continuously square integrable envelope $P_Z H^2 < \infty$). Then both the canonical isotonic and convex LSEs in the additive regression model (3.4) are stochastically bounded: $\|\hat{f}_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$.*

The proof of this lemma will be detailed in Section 5, and hence completes the proof of Theorem 4. □

5. Proofs of technical results

In this section, we collect the proofs for technical results in three groups:

(1) the key Proposition 2 used in the proof of Theorem 2;
(2) entropy results in Lemmas 8, 11 and 12;
(3) stochastic boundedness for shape-restricted LSEs in Lemmas 5 and 13.

5.1. **Proof of Proposition 2.**

In the next few subsections, we will prove Proposition 2 step by step.

5.1.1. *Construction of $\tilde{\mathcal{F}}$.* First consider the case $\gamma \in (0, 1)$. We will do the construction iteratively. For $l = 1$, since $[0, 1]$ contains $\left\lfloor 2^{\frac{1}{1-\gamma}} \right\rfloor$ many equal-length intervals (with length $\left(2^{\frac{1}{1-\gamma}}\right)^{-1}$), we can pick 2 intervals among them; this is denoted $\tilde{\mathscr{C}}_1$. For $l = 2$, each interval in $\tilde{\mathscr{C}}_1$ contains $\left\lfloor 2^{\frac{1}{1-\gamma}} \right\rfloor$ many equal-length subintervals with length $\left(2^{\frac{1}{1-\gamma}}\right)^{-2}$, we can pick 2 subintervals among each of the interval; this is denoted $\tilde{\mathscr{C}}_2$. In this way we can define iteratively $\tilde{\mathscr{C}}_l$ for any $l \in \mathbb{N}$. Let $\tilde{\mathcal{F}}_l \equiv \{\mathbf{1}_I : I \in \tilde{\mathscr{C}}_l\}$. Clearly $|\tilde{\mathcal{F}}_l| = 2^l$

and contains indicators over intervals in $[0,1]$ with length $\left(2^{\frac{1}{1-\gamma}}\right)^{-l}$. Now let $\tilde{\mathcal{F}} \equiv \cup_{l\in\mathbb{N}}\tilde{\mathcal{F}}_l \cup \{\mathbf{0}\}$ where $\mathbf{0}$ denotes a mapping taking identical value 0. Next, for $\gamma = 1$, let $\tilde{\mathcal{F}} \equiv \{\mathbf{1}_{[0,\delta]} : 0 \le \delta \le 1\}$.

We show that the constructed $\tilde{\mathcal{F}}$ satisfies the desired growth condition (2.1). Recall P is the uniform distribution on $[0,1]$.

Lemma 14. *It holds that*

$$\|\tilde{F}(\delta)\|_{L_2(P)} \le \sqrt{2}\delta^{\gamma},$$

where $\tilde{F}(\delta)$ denotes the envelope for $\tilde{\mathcal{F}}(\delta)$.

Proof. The claim is trivial for $\gamma = 1$. For $\gamma \in (0,1)$, since each element in $\tilde{\mathscr{C}}_{l+1}$ is contained in some element in $\tilde{\mathscr{C}}_l$, we only need to count the number of intervals for the smallest level $l(\delta)$ such that the length of intervals in $\tilde{\mathcal{F}}_{l(\delta)}$ is no more than δ^2. In other words, $l(\delta)$ is the integer for which

$$\left(2^{\frac{1}{1-\gamma}}\right)^{-l(\delta)} \le \delta^2, \quad \left(2^{\frac{1}{1-\gamma}}\right)^{-l(\delta)+1} > \delta^2.$$

Hence the number of intervals in $\tilde{\mathcal{F}}_{l(\delta)}$ is $N(\delta) = 2^{l(\delta)} \in [\delta^{-(2-2\gamma)}, 2\delta^{-(2-2\gamma)}]$, from which the claim of the lemma holds. □

5.1.2. *Proof of claim (1) of Proposition 2.* The following standard Paley-Zygmund lower bound will be used.

Lemma 15 (Paley-Zygmund)**.** *Let Z be any non-negative random variable. Then for any $\varepsilon > 0$, $\mathbb{P}(Z > \varepsilon\mathbb{E}Z) \ge \left(\frac{(1-\varepsilon)\mathbb{E}Z}{(\mathbb{E}Z^q)^{1/q}}\right)^{q'}$, where $q, q' \in (1,\infty)$ are conjugate indices: $1/q + 1/q' = 1$.*

We need the following exact characterization concerning the size of maxima of a sequence of independent random variables due to [20], see also Corollary 1.4.2 of [16].

Lemma 16. *Let $\xi_1, \ldots, \xi_n$ be a sequence of independent non-negative random variables such that $\|\xi_i\|_r < \infty$ for all $1 \le i \le n$. For $\lambda > 0$, set $\delta_0(\lambda) \equiv \inf\{t > 0 : \sum_{i=1}^n \mathbb{P}(\xi_i > t) \le \lambda\}$. Then*

$$\frac{1}{1+\lambda}\sum_{i=1}^n \mathbb{E}\xi_i^r \mathbf{1}_{\xi_i > \delta_0} \le \mathbb{E}\max_{1\le i\le n}\xi_i^r \le \frac{1}{1\wedge\lambda}\sum_{i=1}^n \mathbb{E}\xi_i^r \mathbf{1}_{\xi_i > \delta_0}.$$

Proof of Proposition 2, claim (1). **(Case 1: $0 < \gamma < 1$).** Recall $\delta_2 \equiv \vartheta n^{-\frac{1}{2(2-\gamma)}}$. Then by the proof of Lemma 14, we see that there exists some level $l(\delta_2) \in \mathbb{N}$ such that the $N(\delta_2)$ many intervals $\{I_l\}_{l=1}^{N(\delta_2)}$ in $\tilde{\mathcal{F}}_{l(\delta_2)}$ have length at most δ_2^2 and at least $2^{-1/(1-\gamma)}\delta_2^2$, while the number of intervals satisfies $\vartheta^{-(2-2\gamma)}n^{\frac{1-\gamma}{2-\gamma}} \le N(\delta_2) \le 2\vartheta^{-(2-2\gamma)}n^{\frac{1-\gamma}{2-\gamma}}$. Let $\mathcal{E}_n$ be the event that all intervals $\{I_l\}_{l=1}^{N(\delta_2)}$ contain at least $2^{-\frac{2-\gamma}{1-\gamma}}\vartheta^2 n^{\frac{1-\gamma}{2-\gamma}}$ of the X_i's and at most

$\frac{5}{4}\vartheta^2 n^{\frac{1-\gamma}{2-\gamma}}$ of the X_i's. Then by a union bound and Bernstein's inequality (cf. (2.10) of [10]),

$$\mathbb{P}(\mathcal{E}_n^c) \le \mathbb{P}\bigg(\max_{1\le l\le N(\delta_2)} \bigg| \sum_{i=1}^n \mathbf{1}_{I_l}(X_i) - n|I_l| \bigg| > 2^{-\frac{2-\gamma}{1-\gamma}} \vartheta^2 n^{\frac{1-\gamma}{2-\gamma}} \bigg) \le 2\vartheta^{-(2-2\gamma)} n^{\frac{1-\gamma}{2-\gamma}} \exp\big(- c_\gamma \vartheta^2 n^{\frac{1-\gamma}{2-\gamma}} \big). \tag{5.1}$$

Let $\mathcal{I}_l \equiv \{X_i \in I_l\}$ for $1 \le l \le N(\delta_2)$ and $\{\xi_i^{(l)}\}_{i,l\ge 1}$ be i.i.d. random variables with the same law as ξ_1. Then for some $t_n > 0$ to be determined later,

$$\mathbb{P}\bigg(\sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_2^2} \bigg| \sum_{i=1}^n \xi_i f(X_i) \bigg| \ge t_n \bigg) \ge \mathbb{E}_{\boldsymbol{X}} \bigg[\mathbb{P}_{\boldsymbol{\xi}} \bigg(\max_{1\le l\le N(\delta_2)} \bigg| \sum_{i=1}^n \xi_i \mathbf{1}_{I_l}(X_i) \bigg| \ge t_n \bigg) \mathbf{1}_{\mathcal{E}_n} \bigg] = \mathbb{E}_{\boldsymbol{X}} \bigg[\mathbb{P}_{\boldsymbol{\xi}} \bigg(\max_{1\le l\le N(\delta_2)} \bigg| \sum_{i=1}^{|\mathcal{I}_l|} \xi_i^{(l)} \bigg| \ge t_n \bigg) \mathbf{1}_{\mathcal{E}_n} \bigg]. \tag{5.2}$$

Our goal now is to make a good choice of the law for $\xi_\cdot^{(\cdot)}$'s so that we may obtain a good estimate for t_n and thereby using the Paley-Zygmund argument. Let ξ_1 be distributed according to the symmetric $\alpha_\varepsilon \equiv 2-\varepsilon$ stable law, i.e. the characteristic function of ξ_1 is $\varphi_{\xi_1}(t) = \exp(-|t|^{\alpha_\varepsilon})$. Apparently, $k^{-1/\alpha_\varepsilon} \sum_{i=1}^k \xi_i^{(l)}$ has the same law as that of ξ_1, and hence we can take

$$t_n = \frac{1}{2} \bigg(2^{-\frac{2-\gamma}{1-\gamma}} \vartheta^2 n^{\frac{1-\gamma}{2-\gamma}} \bigg)^{1/\alpha_\varepsilon} \mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|. \tag{5.3}$$

Then the conditional probability in the last line of (5.2) can be bounded from below by

$$\mathbb{P}_{\boldsymbol{\xi}} \bigg(\max_{1\le l\le N(\delta_2)} |\xi_l| \ge \frac{1}{2} \mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l| \bigg) \ge \bigg(\frac{\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|}{2 \big(\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|^r \big)^{1/r}} \bigg)^{r'} \tag{5.4}$$

for some conjugate indices $(r, r') \in (1,\infty)^2$. (5.2) and (5.4) suggest that we need to derive a lower bound for $\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|$ and an upper bound for $\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|^r$. This can be done via the help of Lemma 16: since $\mathbb{P}(|\xi_1| > t) \asymp \frac{C_\varepsilon}{1+t^{\alpha_\varepsilon}}$ (cf. Property 1.2.15, page 16 of [37]), we can choose $\lambda \equiv 1$ and $\delta_0 \asymp_\varepsilon N(\delta_2)^{1/\alpha_\varepsilon}$ to see that

$$\mathbb{E}_{\boldsymbol{\xi}} \max_{1\le l\le N(\delta_2)} |\xi_l|^r \asymp \sum_{l=1}^{N(\delta_2)} \mathbb{E}|\xi_l|^r \mathbf{1}_{\xi_l > \delta_0} = N(\delta_2) \bigg(\mathbb{P}\left(|\xi_1| > \delta_0\right) \int_0^{\delta_0} r u^{r-1} \, \mathrm{d}u$$

$$+\int_{\delta_0}^{\infty} ru^{r-1}\mathbb{P}\left(|\xi_1|>u\right)\,\mathrm{d}u\Big)$$
$$\asymp_{\varepsilon,r} N(\delta_2)^{r/\alpha_\varepsilon}.$$

Now as long as $\varepsilon<1/2$, we may choose $r>1$ close enough to 1, e.g. $r=1.1$, to conclude that there exists $\mathfrak{p}_1\in(0,1/8)$ that only depends on ε such that

$$\text{Left hand side of (5.4)} \geq 8\mathfrak{p}_1. \tag{5.5}$$

Combining (5.1), (5.2) and (5.5), and the fact that $t_n=c_1\big(\vartheta^\gamma n^{\frac{1-\gamma}{2-\gamma}}\big)^{2/\alpha_\varepsilon}$ for some constant c_1 depending on ε,γ only, we have that for n large enough depending on ϑ,γ,

$$\mathbb{P}\bigg(\sup_{f\in\tilde{\mathcal{F}}:Pf^2\leq\delta_2^2}\Big|\sum_{i=1}^n\xi_i f(X_i)\Big|\geq c_1\big(\vartheta^\gamma n^{\frac{1-\gamma}{2-\gamma}}\big)^{2/\alpha_\varepsilon}\bigg)\geq 4\mathfrak{p}_1. \tag{5.6}$$

On the other hand, by Talagrand's concentration inequality (cf. Lemma 4) and the contraction principle for Rademacher processes, we have with probability at least $1-2\mathfrak{p}_1$,

(5.7)

$$\begin{aligned}\sup_{f\in\tilde{\mathcal{F}}:Pf^2\leq\delta_2^2}|\mathbb{G}_n(f^2)| &\leq C\big(\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}:Pf^2\leq\delta_2^2}|\mathbb{G}_n f|+\delta_2\sqrt{\log(1/2\mathfrak{p}_1)}+\log(1/2\mathfrak{p}_1)/\sqrt{n}\big)\\ &\leq C_\varepsilon\cdot\delta_2\sqrt{\log(1/\delta_2)}\leq C_{\varepsilon,\gamma}\vartheta n^{-\frac{1}{2(2-\gamma)}}\sqrt{\log n}.\end{aligned}$$

Combining (5.6)-(5.7), we see that with probability at least $2\mathfrak{p}_1$,

$$\begin{aligned}&\sup_{f\in\tilde{\mathcal{F}}:Pf^2\leq\delta_2^2}(\mathbb{P}_n-P)(2\xi f-f^2)\\ &\geq 2c_1\big(\vartheta^\gamma n^{\frac{1-\gamma}{2-\gamma}}\big)^{2/\alpha_\varepsilon}\cdot n^{-1}-C_{\varepsilon,\gamma}\vartheta n^{-\frac{1}{2(2-\gamma)}-\frac12}\sqrt{\log n}\\ &\geq 2c_1\vartheta^\gamma n^{-\frac{1}{2-\gamma}}\cdot\tau_n(\varepsilon,\gamma)-C_{\varepsilon,\gamma}\vartheta n^{-\frac{(3-\gamma)/2}{(2-\gamma)}}\sqrt{\log n}\geq c_1\vartheta^\gamma n^{-\frac{1}{2-\gamma}}\cdot\tau_n(\varepsilon,\gamma)\end{aligned}$$

for n large enough depending on $\varepsilon,\vartheta,\gamma$, where $\tau_n(\varepsilon,\gamma)\equiv n^{\frac{1-\gamma}{2-\gamma}\cdot\frac{\varepsilon}{2-\varepsilon}}$. Hence with the same probability estimate,

$$\begin{aligned}F_n(\delta_2)&=\sup_{f\in\tilde{\mathcal{F}}:Pf^2\leq\delta_2^2}(\mathbb{P}_n-P)(2\xi f-f^2)-\delta_2^2\\ &\geq c_1\vartheta^\gamma n^{-\frac{1}{2-\gamma}}\tau_n(\varepsilon,\gamma)-\vartheta^2 n^{-\frac{1}{2-\gamma}}\geq\frac12 c_1\vartheta^\gamma n^{-\frac{1}{2-\gamma}}\tau_n(\varepsilon,\gamma)\end{aligned}$$

holds for n large enough depending on $\varepsilon,\vartheta,\gamma$, completing the proof for the claim for $0<\gamma<1$.

(Case 2: $\gamma=1$). Recall $\delta_2=\vartheta n^{-1/2}$, and there exists one interval I with length δ_2^2. It is easy to see that $\mathbb{P}\big(|\sum_{i=1}^n\mathbf{1}_I(X_i)-\vartheta^2|>\vartheta^2/2\big)\leq 2\exp(-\vartheta^2/10)$. For $\vartheta\geq 4$, we see that with probability at least 0.5, there

are $\mathcal{O}(1)$ points $X_i \in I$. Denote this event $\mathcal{E}_1$. Let

$$Z_n \equiv \sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_2^2} \left(2\sum_{i=1}^n \xi_i f(X_i) - n\left(\mathbb{P}_n - P\right)(f^2)\right).$$

Note we can use the absolute value in the suprema in the above display. Since $\mathbb{E}|(\mathbb{P}_n - P)(\mathbf{1}_I)|^2 \le \vartheta^2 n^{-2}$, we see that on an event with probability at least 0.96, $|n(\mathbb{P}_n - P)(\mathbf{1}_I)| \le 25\vartheta$. Denote this event by $\mathcal{E}_2$. Then for any ξ such that $\mathbb{E}|\xi| \ge 25\vartheta$, let $t = \mathbb{E}|\xi| - 25\vartheta$, and $N_I \equiv \sum_{i=1}^n \mathbf{1}_I(X_i)$,

$$\begin{aligned}
\mathbb{P}\left(Z_n \ge t\right) &\ge \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{P}_{\boldsymbol{\xi}}\left(\left|2\sum_{i=1}^n \xi_i \mathbf{1}_I(X_i) - n\left(\mathbb{P}_n - P\right)(\mathbf{1}_I)\right| \ge t\right)\mathbf{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \\
&\ge \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{P}_{\boldsymbol{\xi}}\left(\left|\sum_{i=1}^{N_I}\xi_i\right| > (t+25\vartheta)/2\right)\mathbf{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right] \\
&\ge \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{P}_{\boldsymbol{\xi}}\left(\left|\sum_{i=1}^{N_I}\xi_i\right| > \frac{1}{2}\mathbb{E}_{\boldsymbol{\xi}}\left|\sum_{i=1}^{N_I}\xi_i\right|\right)\mathbf{1}_{\mathcal{E}_1\cap\mathcal{E}_2}\right]
\end{aligned}$$

where in the last inequality we used Jensen's inequality. Let η be a symmetric random variable given by $\mathbb{P}(|\eta| > t) = 1/(1+t^2)$, then it is easy to calculate that $\mathbb{E}|\eta| = \pi/2$, and $\mathbb{E}|\eta|^r \equiv c_r < \infty$ for $r < 2$. Let $\xi \equiv 50\vartheta^2 \cdot \eta$. Then $\mathbb{E}|\xi| = 25\pi\vartheta^2 > 25\vartheta$, and hence choosing $r > 1$ close enough to 1 in the Paley-Zygmund Lemma 15 yields that

$$\mathbb{P}\left(Z_n \ge 25\pi\vartheta^2 - 25\vartheta\right) \ge 2\mathfrak{p}_2$$

for some constant $\mathfrak{p}_2 > 0$ depending only on ϑ (through the estimate on N_I on the event $\mathcal{E}_1$). Hence with probability at least $2\mathfrak{p}_2$,

$$F_n(\delta_2) \ge (25\pi\vartheta^2 - 25\vartheta)n^{-1} - \vartheta^2 n^{-1} \ge 285\vartheta n^{-1}.$$

This completes the proof. □

5.1.3. *Proof of claim (2) of Proposition 2.*

Proof of Proposition 2, claim (2). Recall $\delta_1 \equiv \rho n^{-\frac{1}{2(2-\gamma)}-\beta_\varepsilon}$. Note that by Koltchinskii-Pollard maximal inequality for empirical processes (cf. Theorem 2.14.1 of [45]), we have

$$\max_{1\le k\le n} \mathbb{E}\sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_1^2}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^k \varepsilon_i f(X_i)\right| \lesssim \|\tilde{F}(\delta_1)\|_{L_2(P)} \le C_1\delta_1^\gamma.$$

Hence we may take $\psi_n(k) \equiv C_1 k^{1/(2-2\varepsilon)}\delta_1^\gamma$ in the multiplier inequality Lemma 1 to see that

$$\begin{aligned}
\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}: Pf^2\le\delta_1^2}\left|\sum_{i=1}^n \xi_i f(X_i)\right| &\le 4\int_0^\infty \psi_n\big(n\mathbb{P}(|\xi_1| > t)\big)\,\mathrm{d}t \\
&\le 4C_1\delta_1^\gamma n^{1/2(1-\varepsilon)}\|\xi_1\|_{2(1-\varepsilon),1}.
\end{aligned}$$

On the other hand, again by the Koltchinskii-Pollard maximal inequality and the contraction principle for Rademacher processes,

$$\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}:Pf^2\le\delta_1^2}|\mathbb{G}_n(f^2)|\lesssim\mathbb{E}\sup_{f\in\tilde{\mathcal{F}}:Pf^2\le\delta_1^2}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i f(X_i)\right|\lesssim\delta_1^\gamma.$$

Combining the above estimates, we arrive at

$$\begin{aligned}\mathbb{E}E_n(\delta_1)&\le 8C_1\delta_1^\gamma n^{-1}\cdot n^{1/2(1-\varepsilon)}\|\xi_1\|_{2(1-\varepsilon),1}+C_2n^{-1/2}\delta_1^\gamma\\&\le C_{\varepsilon,\xi}\rho^\gamma n^{-\frac{1}{2-\gamma}}\omega_n(\varepsilon,\gamma).\end{aligned}$$

The claim (2) of Proposition 2 now follows from Markov's inequality and hence the proof of Theorem 2 is complete. □

5.2. **Proof of entropy results.**

5.2.1. *Proof of Lemma 8.*

Proof of Lemma 8. Let $t_j\equiv(1+\varepsilon)^{-j}$ and $m(\varepsilon)$ be the smallest integer j such that $t_j\le\varepsilon\|F\|_{L_2(Q)}$. Now for any $f\in\mathcal{F}$, define

$$f_\varepsilon\equiv\sum_{j=1}^{m(\varepsilon)}\big(t_j\mathbf{1}_{t_j<f\le t_{j-1}}+(-t_{j-1})\mathbf{1}_{-t_{j-1}<f\le -t_j}\big).$$

Then if $x\in\mathcal{X}$ is such that

(1) $t_j<f(x)\le t_{j-1}$ for some $j\le m(\varepsilon)$,

$$0\le f(x)-f_\varepsilon(x)\le t_{j-1}-t_j\le\varepsilon t_j\le\varepsilon f(x)\le\varepsilon F(x).$$

(2) $-t_{j-1}<f(x)\le -t_j$ for some $j\le m(\varepsilon)$,

$$0\le f(x)-f_\varepsilon(x)\le -t_j-(-t_{j-1})\le\varepsilon t_j\le\varepsilon\big(-f(x)\big)\le\varepsilon F(x).$$

(3) $-t_{m(\varepsilon)}<f(x)\le t_{m(\varepsilon)}$,

$$|f(x)-f_\varepsilon(x)|\le t_{m(\varepsilon)}\le\varepsilon\|F\|_{L_2(Q)}.$$

Combining the above discussion we arrive at $\|f-f_\varepsilon\|_{L_2(Q)}^2\le 3\varepsilon^2\|F\|_{L_2(Q)}^2$. Let $\mathcal{F}_\varepsilon\equiv\{f_\varepsilon:f\in\mathcal{F}\}$. Then since the sets

$$\begin{aligned}\big\{(x,t):f_\varepsilon(x)\ge t\big\}=\cup_{j=1}^{m(\varepsilon)}\big\{x:f(x)\ge t_j\big\}\times(t_j,t_{j-1}]\\\bigcup\cup_{j=1}^{m(\varepsilon)}\big\{x:f(x)\ge -t_{j-1}\big\}\times(-t_{j-1},-t_j]\\\bigcup\big\{x:f(x)\ge -t_{m(\varepsilon)}\big\}\times(-t_{m(\varepsilon)},t_{m(\varepsilon)}]\end{aligned}$$

as f_ε ranges over $\mathcal{F}_\varepsilon$ is the union of at most $2m(\varepsilon)+1$ VC-classes with disjoint supports, and hence the VC-dimension of $\mathcal{F}_\varepsilon$ is no larger than $Vm(\varepsilon)$, where $V\in(0,\infty)$ only depends on $\mathcal{F}_0$. The rest of the proof proceeds along the same lines as in page 1172 of [18]. □

5.2.2. *Proof of Lemma 11.*

Proof of Lemma 11. Let $\{h_i\}_{i=1}^N$ be a minimal ε-covering set of $\mathcal{H}$ under L_∞. For any probability measure Q on $[0,1]^2$, and any $f \otimes (h-g) \in f \otimes \big(\mathcal{H} - g\big)$, take h_i such that $\|h - h_i\|_\infty \le \varepsilon$. Then

$$\|f \otimes (h_i - g) - f \otimes (h-g)\|_{L_2(Q)}^2 \le \|f\|_{L_2(Q)}^2 \varepsilon^2 = \|f \otimes 1\|_{L_2(Q)}^2 \varepsilon^2.$$

completing the proof. □

5.2.3. *Proof of Lemma 12.*

Proof of Lemma 12. Since $\mathcal{F}' \subset \mathcal{F} \cap L_\infty(1)$ is VC-major, Lemma 8 yields that for any probability measure Q_x on $[0,1]$ and any $\varepsilon > 0$,

$$\log \mathcal{N}\big(\varepsilon \|F'\|_{L_2(Q_x)}, \mathcal{F}', L_2(Q_x)\big) \le \frac{C}{\varepsilon} \log\left(\frac{C}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \|F'\|_{L_2(Q_x)}}\right).$$

Now for any discrete probability measure $Q = n^{-1}\sum_{i=1}^n \delta_{(x_i,z_i)}$ on $[0,1]^2$, let $Q_x \equiv n^{-1}\sum_{i=1}^n \delta_{x_i}$ be the (marginal) probability measure on $[0,1]$. Take a minimal $\varepsilon\|F'\|_{L_2(Q_x)}$-cover of $\mathcal{F}'$ under $L_2(Q_x)$, namely $\{f_k\}$, the log-cardinality of which is no more than

$$\frac{C}{\varepsilon} \log\left(\frac{C}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \|F'\|_{L_2(Q_x)}}\right).$$

Further take a minimal ε-cover of $\mathcal{H}$ under L_∞, namely $\{h_l\}$, the log-cardinality of which is at most a constant multiple of $\varepsilon^{-\gamma}$. Consider the set $\{f_k \otimes h_l\}$, the log-cardinality of which is at most a constant multiple of

$$\frac{1}{\varepsilon} \log\left(\frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon \|F' \otimes 1\|_{L_2(Q)}}\right) \vee \varepsilon^{-\gamma}.$$

For every $f \otimes (h - (\phi_0 - f_0)) \in \mathcal{F}' \otimes (\mathcal{H} - (\phi_0 - f_0))$, let $\tilde{f}_k, \tilde{h}_l$ be such that $\|f - \tilde{f}_k\|_{L_2(Q_x)} \le \varepsilon \|F'\|_{L_2(Q_x)}$ and $\|h - \tilde{h}_l\|_\infty \le \varepsilon$. Then

$$\begin{aligned}
&\|f \otimes (h - (\phi_0 - f_0)) - \tilde{f}_k \otimes (\tilde{h}_l - (\phi_0 - f_0))\|_{L_2(Q)}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Big(f(x_i)\big(h(z_i) - (\phi_0(x_i,z_i) - f_0(x_i))\big) \\
&\qquad\qquad - \tilde{f}_k(x_i)\big(\tilde{h}_l(z_i) - (\phi_0(x_i,z_i) - f_0(x_i))\big)\Big)^2 \\
&\lesssim \|h - \tilde{h}_l\|_\infty^2 \|F'\|_{L_2(Q_x)}^2 + \|\phi_0\|_\infty^2 \|f - \tilde{f}_k\|_{L_2(Q_x)}^2 \\
&\lesssim (1 \vee \|\phi_0\|_\infty^2)\varepsilon^2 \|F'\|_{L_2(Q_x)}^2 = (1 \vee \|\phi_0\|_\infty^2)\varepsilon^2 \|F' \otimes 1\|_{L_2(Q)}^2,
\end{aligned}$$

as desired. □

5.3. **Proof of stochastic boundedness of shape-restricted LSEs.**

5.3.1. *Proof of Lemma 5.*

Proof of Lemma 5, isotonic case. The isotonic least squares estimator $\hat{f}_n$ has a well-known min-max representation [36]:

$$\hat{f}_n(X_j) = \min_{v \geq j} \max_{u \leq j} \frac{1}{v-u+1} \sum_{i=u}^{v} Y_i \tag{5.8}$$

where we slightly abuse the notation X_i's so that $X_1 \leq \cdots \leq X_n$ denote the ordered covariates and Y_i denotes the corresponding observed response at X_i. Since $\hat{f}_n$ is non-decreasing, we only need to consider

$$\alpha_n \equiv \hat{f}_n(X_1) = \min_{v \geq 1} \frac{1}{v} \sum_{i=1}^{v} Y_i, \quad \beta_n \equiv \hat{f}_n(X_n) = \max_{u \leq n} \frac{1}{n-u+1} \sum_{i=u}^{n} Y_i.$$

Note that

$$\mathbb{E}|\alpha_n| \vee \mathbb{E}|\beta_n| \leq \mathbb{E} \max_{k \leq n} \left| \frac{1}{k} \sum_{i=1}^{k} \xi_i \right| + \|f_0\|_\infty.$$

The first term is $\mathcal{O}(1)$ by a simple blocking argument and a Lévy-type maximal inequality due to Montgomery-Smith [35] (see also Theorem 1.1.5 of [16]); we include some details for the convenience of the reader: suppose without loss of generality that $\log_2 n$ is an integer, then for any $t \geq 1$,

$$\begin{aligned}
\mathbb{P}\left(\max_{1 \leq k \leq n} \left| \frac{1}{k} \sum_{i=1}^{k} \xi_i \right| > t \right) &\leq \sum_{j=1}^{\log_2 n} \mathbb{P}\left(\max_{2^{j-1} \leq k < 2^j} \left| \frac{1}{k} \sum_{i=1}^{k} \xi_i \right| > t \right) + \mathbb{P}\left(\left| \sum_{i=1}^{n} \xi_i \right| > nt \right) \\
&\leq \sum_{j=1}^{\log_2 n} \mathbb{P}\left(\max_{2^{j-1} \leq k < 2^j} \left| \sum_{i=1}^{k} \xi_i \right| > 2^{j-1} t \right) + \frac{\|\xi_1\|_2^2}{nt^2} \\
&\leq 9 \sum_{j=1}^{\log_2 n} \mathbb{P}\left(\left| \sum_{i=1}^{2^j} \xi_i \right| > 2^{j-1} t/30 \right) + \frac{\|\xi_1\|_2^2}{nt^2} \\
&\leq C \|\xi_1\|_2^2 \left(\sum_{j=1}^{\log_2 n} \frac{1}{2^j t^2} + \frac{1}{nt^2} \right) \leq C' \|\xi_1\|_2^2 t^{-2},
\end{aligned}$$

completing the proof. □

The proof of stochastic boundedness of the convex least squares estimator crucially uses the characterization developed in Lemma 2.6 of [21]. Note that the characterization is purely deterministic.

Lemma 17. *$\hat{f}_n$ is a convex least squares estimator if and only if for all $j = 2, \ldots, n$,*

$$\sum_{k=1}^{j-1} R_k (X_{k+1} - X_k) \geq \sum_{k=1}^{j-1} S_k (X_{k+1} - X_k),$$

with inequality holds if and only if $\hat{f}_n$ has a kink at X_j. Here $R_k = \sum_{i=1}^k \hat{f}_n(X_i)$ and $S_k = \sum_{i=1}^k Y_i$, where we abuse the notation X_i's for the ordered covariates such that $X_1 \le \ldots \le X_n$, and Y_i's are the corresponding observed responses at X_i.

Proof of Lemma 5, convex case. By symmetry we only consider the behavior of $\hat{f}_n(0)$. Let τ_n denote the first kink of $\hat{f}_n$ away from 0. Then it follows from the characterization Lemma 17 that

$$\sum_{k=1}^{\tau_n-2} R_k(X_{k+1} - X_k) \ge \sum_{k=1}^{\tau_n-2} S_k(X_{k+1} - X_k),$$
$$\sum_{k=1}^{\tau_n-1} R_k(X_{k+1} - X_k) = \sum_{k=1}^{\tau_n-1} S_k(X_{k+1} - X_k).$$

The above two (in)equalities necessarily entail that

$$R_{\tau_n-1}(X_{\tau_n} - X_{\tau_n-1}) \le S_{\tau_n-1}(X_{\tau_n} - X_{\tau_n-1}).$$

Hence with probability 1 we have $R_{\tau_n-1} \le S_{\tau_n-1}$, i.e.

$$\sum_{i=1}^{\tau_n-1} \hat{f}_n(X_i) \le \sum_{i=1}^{\tau_n-1} Y_i. \tag{5.9}$$

Since $\hat{f}_n$ is linear on $[0, X_{\tau_n}]$, we can write

$$\hat{f}_n(x) = \left(1 - \frac{x}{X_{\tau_n}}\right)\hat{f}_n(0) + \frac{x}{X_{\tau_n}}\hat{f}_n(X_{\tau_n}). \tag{5.10}$$

Combining (5.9) and (5.10) we see that

$$\left[\sum_{i=1}^{\tau_n-1}\left(1 - \frac{X_i}{X_{\tau_n}}\right)\right]\hat{f}_n(0) + \left[\sum_{i=1}^{\tau_n-1}\frac{X_i}{X_{\tau_n}}\right]\hat{f}_n(X_{\tau_n}) \le \sum_{i=1}^{\tau_n-1} Y_i,$$

and hence

$$\hat{f}_n(0) \le \left(\frac{1}{1-\beta_{\tau_n}}\right)\cdot\frac{\sum_{i=1}^{\tau_n-1} Y_i}{\tau_n - 1} + \frac{\beta_{\tau_n}}{1-\beta_{\tau_n}}\big|\inf_{x\in[0,1]}\hat{f}_n(x)\big|, \tag{5.11}$$

where

$$\beta_k = \left(\frac{1}{k-1}\sum_{i=1}^{k-1} X_i\right)\cdot\frac{1}{X_k}.$$

By (5.11), we need to handle three terms:

(i) $(1-\beta_{\tau_n})^{-1}$,
(ii) $\frac{\sum_{i=1}^{\tau_n-1} Y_i}{\tau_n - 1}$, and
(iii) $|\inf_{x\in[0,1]}\hat{f}_n(x)|$.

We first handle term (i). We claim that for some universal constant $C > 0$, it holds that

$$\mathbb{P}\big(\max_{2\le k\le n} (1-\beta_k)^{-1} \ge t\big) \le Ct^{-1}. \tag{5.12}$$

To see this, note that for each $k \le n$, conditional on X_k, $X_1/X_k, \ldots, X_{k-1}/X_k$ are distributed as the order statistics for $k-1$ uniform random variables on $[0,1]$. Let $U_1, \ldots, U_n$ be an i.i.d. sequence of uniformly distributed random variables on $[0,1]$, and $0 \le U^n_{(1)} \le \ldots \le U^n_{(n)} \le 1$ be their associated order statistics. Then by using a union bound, the probability in (5.12) is bounded by

$$\sum_{k=2}^{n} \mathbb{P}\bigg(\frac{1}{k-1}\sum_{i=1}^{k-1} \frac{X_i}{X_k} \ge 1 - t^{-1}\bigg) \le \sum_{k=1}^{n-1} \mathbb{E}\bigg[\mathbb{P}\bigg(\frac{1}{k}\sum_{j=1}^{k} U^k_{(j)} \ge 1 - t^{-1}\bigg) \bigg| X_{k+1}\bigg].$$

For $t \ge 3$, the probability in the bracket equals $\mathbb{P}\big(\sum_{j=1}^k U_j \le kt^{-1}\big) = \frac{(kt^{-1})^k}{k!}$ by volume computation: $|\{\sum_{j=1}^k x_j \le a\}| = a^k/k!$. Now combining the probability estimates we arrive at

$$\mathbb{P}\big(\max_{2\le k\le n} (1-\beta_k)^{-1} \ge t\big) \le \sum_{k\ge 1} \frac{(kt^{-1})^k}{k!} \le \sum_{k\ge 1} \frac{(kt^{-1})^k}{(k/e)^k} \le Ct^{-1},$$

proving the claim (5.12) for $t \ge 3$. For $t < 3$, it suffices to increase C.

The second term (ii) can be handled along the same lines as in the proof for the isotonic model, assuming $\|f_0\|_\infty < \infty$ and $\|\xi_1\|_2 < \infty$.

Finally we consider the third term (iii) $|\inf_{x\in[0,1]} \hat f_n(x)|$. We claim that with probability 1,

$$\limsup_{n\to\infty} \sup_{x\in[1/4,3/4]} |\hat f_n(x)| \le C_{\xi,f_0}. \tag{5.13}$$

The claim will be verified in the proof of Lemma 13 below in a more general setting. In particular, (5.13) implies that $\sup_{x\in[1/4,3/4]}|\hat f_n(x)| = \mathcal{O}_{\mathbf{P}}(1)$. Hence for any $\varepsilon > 0$, there exists a constant $K_\varepsilon > 0$ such that for all n large enough, with probability at least $1-\varepsilon$, $\sup_{x\in[1/4,3/4]}|\hat f_n(x)| \le K_\varepsilon$. This event is denoted $\mathcal{E}_\varepsilon$. Now by convexity of $\hat f_n$, it follows that $|\inf_{x\in[0,1]} \hat f_n(x)| \le 2K_\varepsilon$ on $\mathcal{E}_\varepsilon$. To see this, we only need to consider the case where the minimum of $\hat f_n$ is attained in, say, $[0,1/4]$: then the line connecting $(1/4, \hat f_n(1/4))$ and $(3/4, \hat f_n(3/4))$ minorizes $\hat f_n$ on $[0,1/4]$, which is bounded from below by $-2K_\varepsilon$ and hence the same lower bound holds for $\inf_{x\in[0,1]} \hat f_n(x)$ on the event $\mathcal{E}_\varepsilon$. An upper bound for $\inf_{x\in[0,1]} \hat f_n(x)$ is trivial: $\inf_{x\in[0,1]} \hat f_n(x) \le \sup_{x\in[1/4,3/4]} \hat f_n(x) \le K_\varepsilon$ on $\mathcal{E}_\varepsilon$. These arguments complete the proof for $|\inf_{x\in[0,1]} \hat f_n(x)| = \mathcal{O}_{\mathbf{P}}(1)$.

The claim that $\|\hat f_n\|_\infty = \mathcal{O}_{\mathbf{P}}(1)$ follows by combining the discussion of the three terms above and (5.11) which proved $|\hat f_n(0)| \vee |\hat f_n(1)| = \mathcal{O}_{\mathbf{P}}(1)$ and $|\inf_{x\in[0,1]} \hat f_n(x)| = \mathcal{O}_{\mathbf{P}}(1)$. □

5.3.2. *Proof of Lemma 13.*

Proof of Lemma 13, isotonic case. The proof essentially follows the isotonic case of Lemma 5 by noting that the least squares estimator $\hat{f}_n$ for $\mathcal{F}$ in the additive model has the following representation:

$$\hat{f}_n(X_j) = \min_{v \geq j} \max_{u \leq j} \frac{1}{v-u+1} \sum_{i=u}^{v} \big(Y_i - \hat{h}_n(Z_i)\big)$$

where $X_1 \leq \cdots \leq X_n$ denote the ordered X_i's, Y_i's are the observed responses at the corresponding X_i's, and Z_i's are the corresponding Z_i's following the ordering of the X_i's. The rest of the proof proceeds along the same lines as in the isotonic case of Lemma 5 by noting that

$$\begin{aligned} &\max_{1 \leq k \leq n} \sup_{h \in \mathcal{H}} \left| \frac{1}{k} \sum_{i=1}^{k} \big(\phi_0(X_i, Z_i) - h(Z_i)\right| \\ &\leq \|\phi_0\|_\infty + \max_{1 \leq k \leq n} \left(\frac{1}{k} \sum_{i=1}^{k} H(Z_i) \right) = \mathcal{O}_{\mathbf{P}}(1), \end{aligned} \tag{5.14}$$

where the stochastic boundedness follows from the same arguments using Lévy-type maximal inequality as in the isotonic case of Lemma 5, since we have assumed $P_Z H^2 < \infty$. □

Proof of Lemma 13, convex case. We use the same strategy as the convex case of Lemma 5 by replacing Y_i with $Y_i - \hat{h}_n(Z_i)$, and handling terms (i), (ii) and (iii) as in the proof of the convex case of Lemma 5 . Term (i) can be handled using the same arguments as in the proof of the convex case of Lemma 5 ; term (ii) can be handled similar to (5.14). Hence it remains to handle (iii). Let $\hat{\phi}_n(x,z) \equiv \hat{f}_n(x) + \hat{h}_n(z)$. We claim that there exists some $M > 0$ such that

$$\mathbb{P}\left(\inf_{(x,z) \in [1/4,3/4]^2} |\hat{\phi}_n(x,z) - \phi_0(x,z)| > M \text{ i.o.} \right) = 0. \tag{5.15}$$

Once (5.15) is proved, the event $\mathcal{E} \equiv \cup_{m \geq 1} \cap_{n \geq m} \{\inf_{x \in [1/4,3/4]} |\hat{f}_n(x)| \leq \bar{M}\}$ happens with probability 1, where $\bar{M} \equiv M + \sup_{(x,z) \in [1/4,3/4]^2} H(z) + \|\phi_0\|_\infty < \infty$. Let $x_n \in \operatorname{argmin}_{x \in [1/4,3/4]} \hat{f}_n(x)$ and $M_n \equiv |\hat{f}_n(x_n)|$. On the event $\mathcal{E}$, for all n large enough, there exists $x_n^* \in [1/4, 3/4]$ such that $|\hat{f}_n(x_n^*)| \leq 2\bar{M}$. The key observation is the following: if $M_n > 10\bar{M}$, then

$$\inf_{x \in [1/16,1/8]} \hat{f}_n(x) \vee \inf_{x \in [7/8,15/16]} \hat{f}_n(x) \geq \frac{1}{4}\big(M_n - 10\bar{M}\big). \tag{5.16}$$

To see this, we only consider the case $1/4 \leq x_n < x_n^* \leq 3/4$, and derive a lower bound for $\inf_{x \in [7/8,15/16]} \hat{f}_n(x)$; the other case follows from similar arguments. Note that the line L connecting $(x_n, \hat{f}_n(x_n))$ and $(x_n^*, \hat{f}_n(x_n^*))$ minorizes $\hat{f}_n$ on $[7/8, 15/16]$. Since $M_n > 10\bar{M} > 2\bar{M}$, $\hat{f}_n(x_n) < 0$ and hence

the line L has a positive slope s_L bounded below by $(M_n-2\bar{M})/(3/4-1/4) = 2(M_n-2\bar{M})$. This implies that for any $x \in [7/8, 15/16]$,

$$\begin{aligned}\hat{f}_n(x) \ge \hat{f}_n(7/8) \ge L(7/8) &= L(x_n^*) + s_L(7/8 - x_n^*)\\ &\ge \hat{f}_n(x_n^*) + 2(M_n - 2\bar{M}) \cdot (7/8 - 3/4)\\ &\ge (-2\bar{M}) + \frac{1}{4}(M_n - 2\bar{M}) = \frac{1}{4}\big(M_n - 10\bar{M}\big),\end{aligned}$$

proving (5.16). Now we assume without loss of generality that $\inf_{x\in[1/16,1/8]} \hat{f}_n(x) \ge (M_n - 10\bar{M})/4$. Let $I \equiv [1/16, 1/8] \times [0,1]$. Since

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \big(Y_i - \hat{\phi}_n(X_i, Z_i)\big)^2\\ &= \frac{1}{n}\sum_{i=1}^n \big(\xi_i + \phi_0(X_i, Z_i) - \hat{h}_n(Z_i) - \hat{f}_n(X_i)\big)^2\\ &\ge \frac{1}{n}\sum_{(X_i,Z_i)\in I} \big(\hat{f}_n(X_i) - \big(H(Z_i) + \|\phi_0\|_\infty + |\xi_i|\big)\big)_+^2\\ &\ge \frac{1}{2n}\sum_{(X_i,Z_i)\in I} \hat{f}_n^2(X_i) - \frac{1}{n}\sum_{(X_i,Z_i)\in I} \big(3H^2(Z_i) + 3\|\phi_0\|_\infty^2 + 3\xi_i^2\big)\\ &\ge \left(\frac{(M_n - 10\bar{M})^2}{32} - 3\|\phi_0\|_\infty^2\right)\frac{|\{i\in[1:n] : (X_i,Z_i)\in I\}|}{n}\\ &\qquad - \frac{3}{n}\sum_{(X_i,Z_i)\in I} H^2(Z_i) - \frac{3}{n}\sum_{(X_i,Z_i)\in I}\xi_i^2.\end{aligned}$$

Hence by the law of large numbers, on an event with probability 1, if $M_n > 10\bar{M}$,

(5.17)

$$\begin{aligned}&\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \big(Y_i - \hat{\phi}_n(X_i, Z_i)\big)^2\\ &\qquad \ge \frac{(\limsup_{n\to\infty} M_n - 10\bar{M})^2}{16\cdot 32} - \frac{3}{16}\big(\|\phi_0\|_\infty^2 + P_Z H^2 + \mathbb{E}\xi_1^2\big).\end{aligned}$$

On the other hand, since $\hat{\phi}_n$ is the least squares estimator, for any $h' \in \mathcal{H}$,

$$\begin{aligned}&\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \big(Y_i - \hat{\phi}_n(X_i, Z_i)\big)^2\\ &\le \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \big(Y_i - h'(Z_i)\big)^2 \le 3\mathbb{E}\xi_1^2 + 3\|\phi_0\|_\infty^2 + 3P_Z H^2.\end{aligned} \tag{5.18}$$

Combining (5.17) and (5.18), it follows that on an event with probability 1,

$$\limsup_{n\to\infty} M_n \le C\big(\|\xi_1\|_2 + \|H\|_{L_2(P_Z)} + \|\phi_0\|_\infty + \bar{M}\big),$$

holds for some absolute constant $C > 0$, thus proving that with probability 1,

$$\limsup_{n\to\infty} \Big| \inf_{x\in[1/4,3/4]} \hat{f}_n(x) \Big| \le C_{\xi,H,\phi_0,M}.$$

That

$$\limsup_{n\to\infty} \Big| \sup_{x\in[1/4,3/4]} \hat{f}_n(x) \Big| \le C'_{\xi,H,\phi_0,M}$$

with probability 1 can be proved in a completely similar manner by noting that the supremum of $\hat{f}_n$ over $[1/4, 3/4]$ is taken either at $1/4$ or $3/4$. These claims show that with probability 1,

$$\limsup_{n\to\infty} \sup_{x\in[1/4,3/4]} |\hat{f}_n(x)| \le C''_{\xi,H,\phi_0,M}.$$

Note that we have also verified the announced claim (5.13) in the convex case of Lemma 5 by taking $\phi_0(x, z) \equiv f_0(x)$ and $\mathcal{H} \equiv \{0\}$. The rest of proof for handling term (iii) proceeds along the same lines as in the proof of the convex case of Lemma 5, modulo the unproved claim (5.15). Below we prove that (5.15) holds for $M > \sqrt{32\big(\|\xi_1\|_2^2 + \|\phi_0\|_\infty^2 + P_Z H^2\big)}$. To this end, first we prove

$$\mathbb{P}\bigg(\mathcal{E}_1 \equiv \Big\{ \inf_{(x,z)\in[1/4,3/4]^2} \big(\hat{\phi}_n(x,z) - \phi_0(x,z)\big) > M \text{ i.o.} \Big\}\bigg) = 0. \tag{5.19}$$

On the event $\mathcal{E}_1$ intersecting a probability-one event, there exists a subsequence $\{n_k\}_{k\ge1}$ such that

$$\begin{aligned} &\liminf_{k\to\infty} \frac{1}{n_k} \sum_{i=1}^{n_k} \big(Y_i - \hat{\phi}_{n_k}(X_i, Z_i)\big)^2 \\ &\ge \liminf_{k\to\infty} \frac{1}{2n_k} \sum_{(X_i,Z_i)\in[1/4,3/4]^2} \big(\phi_0 - \hat{\phi}_{n_k}\big)^2(X_i, Z_i) - \lim_{k\to\infty} \frac{1}{n_k} \sum_{i=1}^{n_k} \xi_i^2 \\ &\ge M^2/8 - \mathbb{E}\xi_1^2, \end{aligned} \tag{5.20}$$

and thus by (5.18), $M^2 \le 32\big(\|\xi_1\|_2^2 + \|\phi_0\|_\infty^2 + P_Z H^2\big)$. Hence $\mathcal{E}_1$ must be a probability-zero event, which proves (5.19). Using the same arguments we can prove

$$\mathbb{P}\bigg(\sup_{(x,z)\in[1/4,3/4]^2} \big(\hat{\phi}_n(x,z) - \phi_0(x,z)\big) < -M \text{ i.o.} \bigg) = 0. \tag{5.21}$$

The claim (5.15) now follows from (5.19) and (5.21). This completes the proof. □

Acknowledgements

We thank Tengyao Wang for his generous help in the proof of Lemma 5.

References

[1] K. S. Alexander. Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 475–493. Wadsworth, Belmont, CA, 1985.

[2] K. S. Alexander. The central limit theorem for weighted empirical processes indexed by sets. *J. Multivariate Anal.*, 22(2):313–339, 1987.

[3] K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probab. Theory Related Fields*, 75(3):379–423, 1987.

[4] G. Balázs, A. György, and C. Szepesvári. Near-optimal max-affine estimators for convex regression. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 56–64, 2015.

[5] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[6] P. L. Bartlett and S. Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

[7] P. C. Bellec. Sharp oracle inequalities for Least Squares estimators in shape restricted regression. *Ann. Statist.*, 46(2):745–780, 2018.

[8] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York, 1998. Reprint of the 1993 original.

[9] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150, 1993.

[10] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[11] S. Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 2014.

[12] S. Chatterjee, A. Guntuboyina, and B. Sen. On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800, 2015.

[13] S. Chatterjee and J. Lafferty. Adaptive risk bounds in unimodal regression. *arXiv preprint arXiv:1512.02956*, 2015.

[14] Y. Chen and R. J. Samworth. Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(4):729–754, 2016.

[15] G. Cheng. Semiparametric additive isotonic regression. *J. Statist. Plann. Inference*, 139(6):1980–1991, 2009.

[16] V. H. de la Peña and E. Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

[17] C. Gao, F. Han, and C.-H. Zhang. Minimax risk bounds for piecewise constant models. *arXiv preprint arXiv:1705.06386*, 2017.

[18] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216, 2006.

[19] E. Giné, R. Latała, and J. Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.

[20] E. Giné and J. Zinn. Central limit theorems and weak laws of large numbers in certain Banach spaces. *Z. Wahrsch. Verw. Gebiete*, 62(3):323–354, 1983.

[21] P. Groeneboom, G. Jongbloed, and J. A. Wellner. Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, 29(6):1653–1698, 2001.

[22] A. Guntuboyina and B. Sen. Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields*, 163(1-2):379–411, 2015.

[23] A. Guntuboyina and B. Sen. Nonparametric shape-restricted regression. *arXiv preprint arXiv:1709.05707*, 2017.
[24] Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth. Isotonic regression in general dimensions. *arXiv preprint arXiv:1708.09468*, 2017.
[25] Q. Han and J. A. Wellner. A sharp multiplier inequality with applications to heavy-tailed regression problems. *arXiv preprint arXiv:1706.02410*, 2017.
[26] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London, 1990.
[27] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.
[28] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 443–457. Birkhäuser Boston, Boston, MA, 2000.
[29] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.
[30] E. Mammen, O. Linton, and J. Nielsen. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, 27(5):1443–1490, 1999.
[31] E. Mammen and K. Yu. Additive isotone regression. In *Asymptotics: particles, processes and inverse problems*, volume 55 of *IMS Lecture Notes Monogr. Ser.*, pages 179–195. Inst. Math. Statist., Beachwood, OH, 2007.
[32] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000.
[33] P. Massart and E. Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.
[34] M. C. Meyer. Semi-parametric additive constrained regression. *J. Nonparametr. Stat.*, 25(3):715–730, 2013.
[35] S. J. Montgomery-Smith. Comparison of sums of independent identically distributed random vectors. *Probab. Math. Statist.*, 14(2):281–285 (1994), 1993.
[36] T. Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1988.
[37] G. Samorodnitsky and M. S. Taqqu. *Stable non-Gaussian random processes*. Stochastic Modeling. Chapman & Hall, New York, 1994. Stochastic models with infinite variance.
[38] C. J. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13(2):689–705, 1985.
[39] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.
[40] S. van de Geer. Estimating a regression function. *Ann. Statist.*, 18(2):907–924, 1990.
[41] S. van de Geer. On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electron. J. Stat.*, 8(1):543–574, 2014.
[42] S. van de Geer and A. Muro. Penalized least squares estimation in the additive model with different smoothness for the components. *J. Statist. Plann. Inference*, 162:43–61, 2015.
[43] S. A. van de Geer. *Applications of Empirical Process Theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.
[44] S. A. van de Geer and M. Wainwright. On concentration for (regularized) empirical risk minimization. *arXiv preprint arXiv:1512.00677*, 2015.
[45] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[46] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.
[47] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.
[48] C.-H. Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, 2002.

(Q. Han) Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, USA.
E-mail address: royhan@uw.edu

(J. A. Wellner) Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, USA.
E-mail address: jaw@stat.washington.edu