The Annals of Statistics
2021, Vol. 49, No. 5, 2642–2671
https://doi.org/10.1214/21-AOS2049

SET STRUCTURED GLOBAL EMPIRICAL RISK MINIMIZERS ARE RATE OPTIMAL IN GENERAL DIMENSIONS

BY QIYANG HAN

Department of Statistics, Rutgers University, qh85@stat.rutgers.edu

Entropy integrals are widely used as a powerful empirical process tool to obtain upper bounds for the rates of convergence of global empirical risk minimizers (ERMs), in standard settings such as density estimation and regression. The upper bound for the convergence rates thus obtained typically matches the minimax lower bound when the entropy integral converges, but admits a strict gap compared to the lower bound when it diverges. Birgé and Massart (*Probab. Theory Related Fields* **97** (1993) 113–150) provided a striking example showing that such a gap is real with the entropy structure alone: for a variant of the natural Hölder class with low regularity, the global ERM actually converges at the rate predicted by the entropy integral that substantially deviates from the lower bound. The counter-example has spawned a long-standing negative position on the use of global ERMs in the regime where the entropy integral diverges, as they are heuristically believed to converge at a suboptimal rate in a variety of models.

The present paper demonstrates that this gap can be closed if the models admit certain degree of "set structures" in addition to the entropy structure. In other words, the global ERMs in such set structured models will indeed be rate-optimal, matching the lower bound even when the entropy integral diverges. The models with set structures we investigate include (i) image and edge estimation, (ii) binary classification, (iii) multiple isotonic regression, (iv) s-concave density estimation, all in general dimensions when the entropy integral diverges. Here, set structures are interpreted broadly in the sense that the complexity of the underlying models can be essentially captured by the size of the empirical process over certain class of measurable sets, for which matching upper and lower bounds are obtained to facilitate the derivation of sharp convergence rates for the associated global ERMs.

1. Introduction.

1.1. *Overview.* Empirical risk minimization (ERM) is one of the most widely used statistical procedures for the purpose of estimation and inference. Theoretical properties for various ERMs, in particular in terms of rates of convergence, have been intensively investigated by various authors [2–4, 24, 39–42, 44, 47, 48], in a number of by-now standard settings. To motivate our discussion, let us focus on the standard Gaussian regression setting: Let $X_1, \ldots, X_n$ be i.i.d. covariates taking value in $(\mathcal{X}, \mathcal{A})$ with law P, and the responses Y_i's are given by

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

where ξ_i's are i.i.d. $\mathcal{N}(0, 1)$, and f_0 belongs to a uniformly bounded class $\mathcal{F} \subset L_\infty(1)$. One canonical global ERM in the regression model (1.1) is the least squares estimator (LSE):

$$\widehat{f_n} \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2.$$

Received May 2019; revised September 2020.
MSC2020 subject classifications. Primary 60E15; secondary 62G05.
Key words and phrases. Empirical risk minimization, empirical process, classification, nonparametric regression, density estimation, non-Donsker.

The performance of $\widehat{f}_n$ is usually evaluated through the risk under squared L_2 loss $\mathbb{E}_{f_0}\|\widehat{f}_n - f_0\|^2_{L_2(P)}$, or its "probability" version.

The seminal work of Birgé and Massart [4] (and other references cited above) shows that an upper bound $\bar{r}_n^2$ for the risk $\mathbb{E}_{f_0}\|\widehat{f}_n - f_0\|^2_{L_2(P)}$ can be obtained by solving

$$\int_{c\bar{r}_n^2}^{\bar{r}_n} \sqrt{\log \mathcal{N}_{[\cdot]}\big(\varepsilon, \mathcal{F}, L_2(P)\big)}\,\mathrm{d}\varepsilon \asymp \sqrt{n}\cdot \bar{r}_n^2. \tag{1.2}$$

Here, $\mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}, L_2(P))$ is the ε-bracketing number of $\mathcal{F}$ under $L_2(P)$. On the other hand, a lower bound $\underline{r}_n^2$ for the risk, often evaluated in a minimax framework, that is, $\inf_{\widetilde{f}_n}\sup_{f_0\in\mathcal{F}}\mathbb{E}_{f_0}\|\widetilde{f}_n - f_0\|^2_{L_2(P)} \geq \underline{r}_n^2$, can be obtained (cf. [3, 49]) via a different equation

$$\underline{r}_n\sqrt{\log \mathcal{N}\big(\underline{r}_n, \mathcal{F}, L_2(P)\big)} \asymp \sqrt{n}\cdot \underline{r}_n^2, \tag{1.3}$$

where $\mathcal{N}(\varepsilon, \mathcal{F}, L_2(P))$ is the ε-covering number of $\mathcal{F}$ under $L_2(P)$. Note that the left-hand side of (1.2) is no smaller than the left-hand side of (1.3), so we always have $\underline{r}_n \lesssim \bar{r}_n$. Suppose for now that the difference in the covering and bracketing entropy can be ignored, and it holds for some $\alpha > 0$ that

$$\log \mathcal{N}\big(\varepsilon, \mathcal{F}, L_2(P)\big) \asymp \log \mathcal{N}_{[\cdot]}\big(\varepsilon, \mathcal{F}, L_2(P)\big) \asymp \varepsilon^{-2\alpha}. \tag{1.4}$$

The parameter $\alpha > 0$ measures the complexity of $\mathcal{F}$, and is closely related to the "smoothness" of $\mathcal{F}$, cf. [16, 44, 45]. Solving the equations (1.2) and (1.3) yields that

$$\underline{r}_n \asymp n^{-\frac{1}{2(1+\alpha)}}, \qquad \bar{r}_n \asymp \big(n^{-\frac{1}{2(1+\alpha)}} \vee n^{-\frac{1}{4\alpha}}\big)\sqrt{\log^{\mathbf{1}(\alpha=1)} n}. \tag{1.5}$$

Modulo the logarithmic factor in the boundary case $\alpha = 1$, we see a somewhat strange phase-transition phenomenon:

- If $\alpha \in (0, 1)$, $0 < \liminf_n \bar{r}_n/\underline{r}_n \leq \overline{\lim}_n \bar{r}_n/\underline{r}_n < \infty$. In this regime, $\mathcal{F}$ is *Donsker* since a central limit theorem in $\ell^\infty(\mathcal{F})$ holds for the empirical process due to the convergence of the bracketing entropy integral at 0; cf. [47], Section 2.5.2.
- If $\alpha > 1$, $\liminf_n \bar{r}_n/\underline{r}_n = \infty$. In this regime, $\mathcal{F}$ is *non-Donsker* since there does not exist a central limit theorem in $\ell^\infty(\mathcal{F})$ for the empirical process—the limiting Brownian bridge process indexed by $\mathcal{F}$ is not sample bounded.

Although at this point (1.2) only gives an upper bound for $\bar{r}_n$, Birgé and Massart [4] showed by a stunning example that in the regime $\alpha > 1$, $\bar{r}_n$ can actually be attained (up to logarithmic factors) for the global ERM (called "minimum contrast estimators" therein) over a slightly constructed $\mathcal{F}$ based on Hölder classes on $[0, 1]$ with smoothness less than $1/2$. Consequently, there is a genuine gap between the upper bound $\bar{r}_n$ and the lower bound $\underline{r}_n$ obtained from general empirical process techniques based on L_2 entropy structures (1.2) and (1.3) alone, in the regime $\alpha > 1$.

The counterexample in [4] results in a long-standing negative position on the use of global ERMs in the regime $\alpha > 1$, as they are heuristically believed to be rate-suboptimal in various problems falling into the non-Donsker regime, beyond the natural setting of Hölder-type smoothness classes; cf. [18, 23, 37, 44], just to name a few references. A common (but perhaps vague) heuristic is that when $\alpha > 1$, the class $\mathcal{F}$ is too "massive" for global ERMs to achieve the optimal rate.

It should be mentioned that the rate suboptimality phenomenon is due to the *global* nature of ERM that searches over the entire parameter space, since it is easy to construct a "theoretical" rate-optimal estimator by searching over certain maximal packing sets of $\mathcal{F}$ even in the regime $\alpha > 1$ (usually known as the "sieve" estimator [17, 28]). For instance, the LSE over a maximal $\underline{r}_n$-packing set of $\mathcal{F}$ typically leads to the desired optimal rate of convergence. Such

a theoretical construction often occurs in a minimax approach for a given statistical model; cf. [5, 18, 30].

At a deeper level from the perspective of empirical process theory, the upper bound (1.2) comes from the Dudley's entropy integral, and the lower bound (1.3) is inherited with Sudakov minorization. From the recent work [7, 22, 36, 43], it is now understood that the risk $r_n^2 \equiv \mathbb{E}_{f_0}\|\widehat{f}_n - f_0\|_{L_2(P)}^2$ can be *completely* characterized (at least in the simple Gaussian regression model with uniformly bounded $\mathcal{F}$), by the following (not fully rigorous but essential)[1] equation:

$$\mathbb{E} \sup_{f\in\mathcal{F}-f_0:\|f\|_{L_2(P)}\le r_n} |\mathbb{G}_n(f)| \asymp \sqrt{n}\cdot r_n^2. \tag{1.6}$$

Here, $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$ is the empirical process. Since Dudley's entropy integral provides an upper bound, while the Sudakov minorization gives a lower bound, for the empirical process in (1.6) as soon as it enters the "Gaussian domain" (= for n large in our case), the only possibility for which $\bar{r}_n$ and $\underline{r}_n$ do not match lies in situations where the entropy integral bound deviates substantially from the Sudakov minorization. This is indeed the case in the non-Donsker regime $\alpha > 1$: under a variant of the L_2 entropy condition (1.4) (see (2.7)), standard bounds lead to the estimates

$$n^{(\alpha-1)/2(\alpha+1)} \lesssim \mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)| \lesssim n^{(\alpha-1)/2\alpha}. \tag{1.7}$$

The upper and lower bounds above do not match, and neither of them can be improved without further structural assumptions. In particular, (1.7) leads to the discrepancy between $\bar{r}_n$ and $\underline{r}_n$ in (1.5) for non-Donsker $\mathcal{F}$'s.

Despite the strong suspicion in the literature (cited above) that the actual rate r_n of global ERMs will likely match $\bar{r}_n$ which has a strict gap compared to the minimax lower bound $\underline{r}_n$, there appears recently some surprising special cases in which global ERMs are proved to be rate optimal even in the regime $\alpha > 1$. One example is given by the multiple isotonic regression model studied by the author in [20]. When $d \ge 3$, by the entropy estimate in [14], the class of multiple isotonic functions is in the non-Donsker regime $\alpha > 1$, but interestingly [20] proved that the natural LSE (= global ERM) is still minimax rate-optimal (up to logarithmic factors) in L_2 loss. The proof techniques in [20] are rather intricate and somewhat indirect, so they unfortunately do not shed light on why the LSE must be rate-optimal (see Remark 5.5 for more technical details).

The purpose of the present paper is to demonstrate a general underlying mechanism for the rate-optimality phenomenon for global ERMs beyond the isotonic LSE in general dimensions as mentioned above. This amounts to the identification of a sub-family of $\mathcal{F}$'s satisfying the L_2 entropy condition (1.4) (or see its variant (2.7)), in which the the associated global ERMs remain rate-optimal. As one may expect, the key step is to close the gap between the upper and lower bounds in (1.7) under suitable structural assumptions on $\mathcal{F}$, and for the purpose of rate-optimality, one should aim at improving the upper bound $n^{(\alpha-1)/2\alpha}$ to match the lower bound $n^{(\alpha-1)/2(\alpha+1)}$ in (1.7). We show (cf. Theorem 2.1) that this is indeed possible for $\mathcal{F} \subset L_\infty(1)$ in the non-Donsker regime $\alpha > 1$, under a stronger L_1 entropy condition:

$$\log \mathcal{N}_{[\cdot]}\big(\varepsilon^2, \mathcal{F}, L_1(P)\big) \lesssim \varepsilon^{-2\alpha}. \tag{1.8}$$

An important class of $\mathcal{F}$ that verifies (1.8) is the class of indicators indexed by a class of measurable sets. More specifically, for a class of measurable sets $\mathscr{C}$, as the $L_1(P)$ size of

[1] Rigorously, r_n is determined by the location of the maxima of the map $r \mapsto \sup_{f\in\mathcal{F}-f_0:\|f\|_{L_2(P)}\le r} \mathbb{G}_n(\xi f - f^2) - \sqrt{n}r^2$ provided it exists uniquely; cf. [22, 43]. For Gaussian errors ξ_i's and uniformly bounded $\mathcal{F}$, the order of r_n can typically be obtained by matching upper and lower moment estimates for the LHS of (1.6).

any element $\mathbf{1}_C$ where $C \in \mathscr{C}$ is the same as its squared $L_2(P)$ size, (1.8) is automatically verified provided the L_2 entropy condition (1.4) (or (2.7)) is satisfied. Therefore, under the prescribed L_2 entropy condition alone, as long as the L_2-size of $\mathscr{C}$ is not too small, it holds for $\alpha \neq 1$ that

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} \left|\mathbb{G}_n(C)\right| \asymp \max\left\{\sigma^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\right\}. \tag{1.9}$$

Here, $\mathscr{C}(\sigma) \equiv \{C \in \mathscr{C} : P(C) \leq \sigma^2\}$, and for a measurable set C, $\mathbb{G}_n(C) \equiv \mathbb{G}_n(\mathbf{1}_C)$. For $\alpha > 1$, the empirical process (1.9) is in the non-Donsker regime, and our estimate (1.9) is still sharp in this challenging regime thanks to the improved estimate of the generic bound (1.7) due to the L_1 entropy condition (1.8).

In light of (1.9), we will show that in models with certain "set structures," the global ERMs will achieve the minimax optimal rates of convergence, that is, $\bar{r}_n \asymp r_n \asymp \underline{r}_n$ (up to logarithmic factors) even in the regime $\alpha > 1$ when the entropy integral diverges. Here, "set structures" are interpreted broadly in the sense that the size of the underlying empirical process (1.6) indexed by $\mathcal{F}$ can be characterized by an empirical process indexed by certain class of measurable sets for which an estimate of the type (1.9) is possible. This concept will be illustrated throughout a detailed study on the risk behavior (or rates of convergence) for the natural global ERMs in the following models:

1. Image and edge estimation;
2. Binary classification;
3. Multiple isotonic regression (revisited);
4. s-concave density estimation,

all of which will be considered in general dimensions, where the problems necessarily fall into the non-Donsker regime $\alpha > 1$. In the special case of the multiple isotonic regression model, our new techniques present a much easier and intuitive proof (compared to the previous work [20]) that explains the reason why the natural least squares estimator is indeed rate minimax (up to logarithmic factors) for $d \geq 3$—the complexity of the isotonic LSE is captured by that of the class of upper and lower sets that arise naturally in the min-max representation of the isotonic LSE; cf. [35].

1.2. *Related works.* Prior upper bounds for the empirical process (1.9) in the regime $\alpha > 1$ are obtained in, for example, [12], Theorem 2, or [13], Theorem 11.4, with additional logarithmic factors and in the weaker "in probability" form. This paper provides stronger matching upper and lower bounds for the expected supremum of the empirical process in (1.9), and in fact the bounds hold in much greater generality; see Theorem 2.1 for precise statements.

The major part of this work is based on Chapter 4 of the author's University of Washington Ph.D. thesis in 2018. During the preparation the paper, the author becomes aware of the very nice work [9] which derives, among other things, global risk bounds for the log-concave (= 0-concave) maximum likelihood estimators (MLEs) based on a reduction scheme of [6] and an upper bound similar to (1.9). Here, we prove that the rate-optimality in example (4) holds for the maximum regime of s in which the s-concave MLE exists. See Remark 3.9 for more technical remarks.

1.3. *Organization.* The rest of the paper is organized as follows. Section 2 is devoted to the new upper and lower bounds for the size of the empirical process indexed by a class of functions satisfying certain special entropy conditions that includes the class of measurable sets. Applications of these new bounds to the models mentioned above are detailed in Section 3. For clarity of presentation, proofs are deferred to Sections 4–5, and the Supplementary Material [19].

1.4. *Notation.* For a real-valued random variable ξ and $1 \le p < \infty$, let $\|\xi\|_p := (\mathbb{E}|\xi|^p)^{1/p}$ denote the ordinary p-norm.

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

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 $a \wedge b \equiv \min\{a, b\}$. Slightly abusing notation, we write $\log(x) \equiv \log(e \vee x)$, and $\log\log(x) \equiv \log(e \vee \log(e \vee x))$.

2. Empirical processes indexed by sets.

2.1. *Setup and assumptions.* Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution P on a sample space $(\mathcal{X}, \mathcal{A})$, and $\mathscr{C}$ be a collection of measurable sets contained in $\mathcal{X}$. To avoid measurability digressions, we assume that $\mathscr{C}$ is countable throughout the article. For any $\sigma > 0$, let $\mathscr{C}(\sigma) \equiv \{C \in \mathscr{C} : P(C) \le \sigma^2\}$.

Following the standard notation for covering and bracketing numbers (cf. [47], pp. 83), for a normed linear space $(\mathcal{F}, \|\cdot\|)$, let the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the minimum number of balls $\{g : \|g - f\| < \varepsilon\}$ of radius ε under $\|\cdot\|$ needed to cover $\mathcal{F}$. Let the bracketing number $\mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the minimum number of ε-brackets under $\|\cdot\|$ needed to cover $\mathcal{F}$, where an ε-bracket $[\ell, u] \equiv \{f : \ell \le f \le u\}$ with $\|u - \ell\| < \varepsilon$. Following the notation in [13], p. 270, (7.4), let $\mathcal{N}_I(\varepsilon, \mathscr{C}, P)$ be the ε-bracketing number for $\mathscr{C}$ under P, that is, $\mathcal{N}_I(\varepsilon, \mathscr{C}, P) \equiv \mathcal{N}_{[\cdot]}(\varepsilon, \mathbf{1}_\mathscr{C}, L_1(P))$ with $\mathbf{1}_\mathscr{C} \equiv \{\mathbf{1}_C : C \in \mathscr{C}\}$.

ASSUMPTION A. Fix $\alpha > 0$.

(E1) $\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \le L\varepsilon^{-\alpha}$.
(E2) $\log \mathcal{N}(\varepsilon/4, \mathscr{C}(\sqrt{\varepsilon}), P) \ge L^{-1}\varepsilon^{-\alpha}$.

For examples satisfying the above entropy conditions see, for example, [13], Sections 8.3/8.4, or Theorem 8.3.2, on the class of upper/lower sets and convex bodies (cf. [13], Theorem 8.4.1/Corollary 8.4.2). L will be a large enough absolute constant throughout the article, the dependence on which will not be explicitly stated in the theorems.

For $0 < \alpha < 1$, the bracketing condition in (E1) can also be replaced by a uniform entropy condition $\sup_Q \log \mathcal{N}(\varepsilon, \mathscr{C}, Q) \le L\varepsilon^{-\alpha}$, where the supremum is taken over all finitely discrete probability measures Q. Such a uniform entropy condition is satisfied if $\mathscr{C}$ is a VC-class (cf. [16], Section 3.6).[2] This case is essentially covered in [15].

2.2. *Upper and lower bounds.* We first state the general upper and lower bounds for empirical processes indexed by general function classes satisfying certain entropy conditions.

[2]Baraud [1] advocates the notion of weak VC-major class as a generalization of VC-major class. The class of indicators over a class of sets $\mathscr{C}$ is weakly VC-major if and only if $\mathscr{C}$ is VC (cf. [1], Definition 2.2).

THEOREM 2.1.

1. *Fix $p \geq 1$. Let $\mathcal{F} \subset L_\infty(1)$, and $\sup_{f\in\mathcal{F}}\|f\|_{L_p} \leq \sigma$. Suppose there exists some $\alpha > 0$ such that for all $\varepsilon > 0$, $\log \mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}, L_p(P)) \leq L\varepsilon^{-\alpha}$. Then*

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}|\mathbb{G}_n(f)| \\
&\lesssim_{\alpha,p} \inf_{0\leq\gamma\leq\sigma/2}\big\{(\sigma\mathbf{1}_{\alpha<p\wedge 2}+\gamma\mathbf{1}_{\alpha>p\wedge 2})^{\{(p\wedge 2)-\alpha\}/2}+n^{-1/2}\sigma^{-\alpha}+\sqrt{n}\gamma\big\} \qquad (2.1)\\
&\asymp \begin{cases} \sigma^{\{(p\wedge 2)-\alpha\}/2}+n^{-1/2}\sigma^{-\alpha}, & \alpha < p\wedge 2; \\ n^{\frac{\alpha-p\wedge 2}{2(\alpha+2-p\wedge 2)}}+\sigma^{-\frac{\alpha-p\wedge 2}{2}}+n^{-1/2}\sigma^{-\alpha}, & \alpha > p\wedge 2. \end{cases}
\end{aligned}$$

2. *Suppose that the following entropy estimate holds for some $\alpha > 0$:*

$$\log\mathcal{N}_{[\cdot]}\big(\varepsilon^2, \mathcal{F}, L_1(P)\mathbf{1}_{\alpha>1}+L_2^2(P)\mathbf{1}_{0<\alpha<1}\big) \leq L\varepsilon^{-2\alpha}. \qquad (2.2)$$

Then for $\sigma^2 \gtrsim n^{-1/(\alpha+1)}$, $\alpha \neq 1$, we have

$$\mathbb{E}\sup_{f\in\mathcal{F}(\sigma)}|\mathbb{G}_n(f)| \lesssim_\alpha \max\{\sigma^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}. \qquad (2.3)$$

3. *If in addition to (2.2), it holds that*

$$\log\mathcal{N}\big(\varepsilon/2, \mathcal{F}(\varepsilon), L_2(P)\big) \geq L^{-1}\varepsilon^{-2\alpha}. \qquad (2.4)$$

Then for $\sigma^2 \gtrsim n^{-1/(\alpha+1)}$, $\alpha \neq 1$, we have

$$\mathbb{E}\sup_{f\in\mathcal{F}(\sigma)}|\mathbb{G}_n(f)| \gtrsim_\alpha \max\{\sigma^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}. \qquad (2.5)$$

Here, $\mathcal{F}(\sigma) \equiv \{f \in \mathcal{F} : Pf^2 \leq \sigma^2\}$.

PROOF. See Section 4. □

REMARK 2.2 (Comparison to classical bounds). By the standard local maximal inequality for the empirical processes (cf. [47], Lemma 2.14.3), we have for $\sigma > 0$ not too small,

$$\mathbb{E}\sup_{f\in\mathcal{F}(\sigma)}|\mathbb{G}_n(f)| \lesssim \inf_{0\leq\gamma\leq\sigma/2}\left\{\sqrt{n}\gamma+\int_\gamma^\sigma\sqrt{\log\mathcal{N}_{[\cdot]}\big(\varepsilon, \mathcal{F}, L_2(P)\big)}\,\mathrm{d}\varepsilon\right\}. \qquad (2.6)$$

Suppose $\mathcal{F} \subset L_\infty(1)$ satisfies the L_2 entropy condition:

$$\varepsilon^{-2\alpha} \lesssim \log\mathcal{N}\big(\varepsilon, \mathcal{F}(2\varepsilon), L_2(P)\big) \leq \log\mathcal{N}_{[\cdot]}\big(\varepsilon, \mathcal{F}, L_2(P)\big) \lesssim \varepsilon^{-2\alpha}. \qquad (2.7)$$

In the Donsker regime $\alpha < 1$, standard upper bound (2.6) and the lower bound (2.5) already match each other under (2.7). In the non-Donsker regime $\alpha > 1$, these bounds lead to (1.7) whose upper and lower bounds do not match, both of which can be attained for certain instances of $\mathcal{F}$ satisfying (2.7). This means improvements for the generic bound (1.7) must require additional structural assumptions on $\mathcal{F}$. As the L_1 entropy condition (2.2) along with $\mathcal{F} \subset L_\infty(1)$ implies the rightmost part of (2.7), it follows that within the family of $\mathcal{F} \subset L_\infty(1)$'s satisfying the L_2 entropy condition (2.7) with $\alpha > 1$, if in addition the stronger L_1 entropy condition (2.2) is satisfied, then the upper bound in (1.7) can be improved from $n^{(\alpha-1)/2\alpha}$ to $n^{(\alpha-1)/2(\alpha+1)}$ that matches the lower bound.

An important family of $\mathcal{F}$ satisfying the L_1 entropy condition (2.2) is the class of indicators over measurable sets. We formalize the results below.

THEOREM 2.3. *Suppose* (*E*1) *holds and* $\sigma^2 \gtrsim n^{-1/(\alpha+1)}$, $\alpha \neq 1$. *Then*

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \lesssim_\alpha \max\{\sigma^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}.$$

If furthermore (*E*2) *holds*, *then*

$$\mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \gtrsim_\alpha \max\{\sigma^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}.$$

PROOF OF THEOREM 2.3. (2.2) is verified using the fact that for any measurable set C, with $f \equiv \mathbf{1}_C$ we have $Pf = Pf^2$. (2.4) can be verified by noting that $\mathcal{N}_I(\varepsilon/4, \mathscr{C}(\sqrt{\varepsilon'}), P) = \mathcal{N}_{[\cdot]}(\sqrt{\varepsilon}/2, \mathcal{F}(\sqrt{\varepsilon'}), L_2(P))$ holds with $\mathcal{F} \equiv \{\mathbf{1}_C : C \in \mathscr{C}\}$ and any $\varepsilon, \varepsilon' > 0$ (and similarly for the covering number). □

REMARK 2.4. Some remarks on the upper bounds in Theorems 2.1 and 2.3:

1. Roughly speaking, the improved estimates (2.1) compared to the classical bound (2.6) come from careful L_p chaining with bracketing that provides tighter controls at the finest resolution of the chaining step; see Section 4.1 for some heuristics and details.

2. It is also possible to consider the boundary case $\alpha = 1$ in Theorem 2.3. Then the upper bound deviates from the lower bound by a logarithmic factor. In particular, suppose (E1)–(E2) hold and $\sigma^2 \gtrsim n^{-1/2}$. Then $1 \lesssim \mathbb{E} \sup_{C \in \mathscr{C}(\sigma)} |\mathbb{G}_n(C)| \lesssim \log n$.

REMARK 2.5. Some remarks on the lower bounds in Theorems 2.1 and 2.3:

1. The proof for the lower bound (2.5) is based on Gaussian randomization followed by an application of the multiplier inequality derived in the author's previous work [22] that removes the effect of Gaussianization. This only requires some sharp upper bounds for the unconditional processes, as opposed to the approach of [15] using Rademacher minorization, which requires sharp upper bounds for *conditional* processes.

2. Condition (2.4) is also assumed in [15] under the name "α-fullness" (cf. Definition 3.3 therein). This condition is best verified on a case-by-case basis. For instance the α-Hölder class on $[0, 1]$ is α-full; cf. the proof of Lemma 6 in [22].

3. Rate-optimal global ERMs. In this section, we apply the new bounds derived in the previous section to several models including (i) image and edge estimation, (ii) binary classification, (iii) multiple isotonic regression, and (iv) s-concave density estimation, all in general dimensions. Global ERMs in these models are non-Donsker problems in general dimensions, but we will show that in each of these models, the underlying empirical process problem (1.6) can be essentially characterized by an empirical process indexed by certain class of measurable sets. The bounds in Theorem 2.3 can then be used to prove that these global ERMs converge at an optimal rate (up to logarithmic factors), rather than a strictly sub-optimal rate as predicted using the entropy integral (= (1.2)) in [4]. Interestingly, Theorem 2.3 is typically applied without the localization. This is viable as the size of the expected supremum of empirical process is already diverging in the non-Donsker regime, so localization is usually uninformative.

3.1. *Image estimation.* Let $X_1, \ldots, X_n$ be i.i.d. samples with law P on a sample space $(\mathcal{X}, \mathcal{A})$. In this subsection, we consider the regression model:

$$Y_i = \mathbf{1}_{C_0}(X_i) + \xi_i, \quad i = 1, \ldots, n. \tag{3.1}$$

This model has been considered by [25, 26] and more recently by [5] under the name "image estimation" (cf. [26], Section 3.1), where C_0 is considered as the "image," and $\mathcal{X} \setminus C_0$ is

considered as "background." We assume for simplicity that the ξ_i's are i.i.d. $\mathcal{N}(0,1)$ and are independent of X_i's. Let $\mathscr{C}$ be a collection of measurable sets in $\mathcal{X}$, and we will fit the regression model by $\{\mathbf{1}_C : C \in \mathscr{C}\}$. Our interest will be the behavior of the least squares estimator $\widehat{C}_n$ defined by

$$\widehat{C}_n \in \arg\min_{C\in\mathscr{C}} \sum_{i=1}^{n} \big(Y_i - \mathbf{1}_C(X_i)\big)^2. \tag{3.2}$$

We assume that $\widehat{C}_n$ is well defined without loss of generality. We will measure the quality of $\widehat{C}_n$ via the expected symmetric difference of $\widehat{C}_n$ and C_0 under P defined by

$$P|\widehat{C}_n \Delta C_0| = P(\mathbf{1}_{\widehat{C}_n} - \mathbf{1}_{C_0})^2 = \int (\mathbf{1}_{\widehat{C}_n} - \mathbf{1}_{C_0})^2 \,\mathrm{d}P. \tag{3.3}$$

By a relatively standard reduction (cf. Lemma 5.1), the risk of $\widehat{C}_n$ in symmetric difference can be related to the expected supremum

$$\mathbb{E} \sup_{C\in\mathscr{C}: P|C\Delta C_0|\le\delta^2} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i (\mathbf{1}_C - \mathbf{1}_{C_0})(X_i) \right|,$$

and, therefore, we may apply Theorem 2.3 (or the more general Theorem 2.1). We formalize the result below.

THEOREM 3.1. *Suppose that for some $\alpha \neq 1$, $\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \le L\varepsilon^{-\alpha}$. Then*

$$\sup_{C_0\in\mathscr{C}} \mathbb{E}_{C_0} P|\widehat{C}_n \Delta C_0| \lesssim n^{-1/(\alpha+1)}.$$

PROOF. See Section 5. □

By [49], the rate $n^{-1/(\alpha+1)}$ cannot be improved in a minimax sense if furthermore a lower bound on the metric entropy on the same order as that of the upper bound is available.

As a straightforward corollary of the above Theorem 3.1, let $\mathscr{C}_d$ be the collection of all convex bodies contained in the unit ball in $\mathbb{R}^d$ and P the uniform distribution on the unit ball.

COROLLARY 3.2. *Fix $d \ge 4$. Then*

$$\sup_{C_0\in\mathscr{C}_d} \mathbb{E}_{C_0} P|\widehat{C}_n \Delta C_0| \lesssim n^{-2/(d+1)}.$$

PROOF. The claim essentially follows from [13], Theorem 8.25, Corollary 8.26, asserting that we can take $\alpha = (d-1)/2$ in Theorem 3.1. □

The corollary shows that we can use a global least squares estimator rather than a sieved least squares estimator (cf. [5]) to achieve the optimal rate of convergence.

REMARK 3.3. It is possible to impose certain tail conditions on the density of P to extend the above corollary to a maximum risk bound over all convex sets in $\mathbb{R}^d$. In particular, the above result holds for any P with compact support in $\mathbb{R}^d$ with a bounded Lebesgue density. A proof in this vein is carried out in the context of s-concave density estimation in $\mathbb{R}^d$ to be detailed ahead.

3.2. *Edge estimation*. In this subsection, we consider the regression model studied in [26, 30]:

$$Y_i = f_{C_0}(X_i)\eta_i, \tag{3.4}$$

where $f_{C_0}(x) = 2\mathbf{1}_{C_0}(x) - 1$ and η_i's are i.i.d. random variables such that $\mathbb{P}(\eta_i = 1) = 1/2 + a$ and $\mathbb{P}(\eta_i = -1) = 1/2 - a$ for some known constant $a \in (0, 1/2)$. Such a model is motivated by estimation of sets in multidimensional "black and white" pictures, where $Y_i = 1$ is interpreted as observing black, and $Y_i = -1$ is white. We refer the reader to [30] for more motivation for this model. The model (3.4) can be rewritten as

$$Y_i = 2af_{C_0}(X_i) + \xi_i, \tag{3.5}$$

where $\xi_i = f_{C_0}(X_i)(\eta_i - 2a)$'s are bounded errors. An important property for these errors is that $\mathbb{E}[\xi_i | X_i] = 0$ for all $i = 1, \dots, n$. Note here ξ_i is *not* independent of X_i and hence a different analysis is needed. Now consider the least squares estimator

$$\widehat{C}_n \equiv \arg\min_{C \in \mathscr{C}} \sum_{i=1}^{n} \big(Y_i - 2af_C(X_i)\big)^2. \tag{3.6}$$

A careful analysis to be detailed in Lemma 5.2 ahead shows that the risk of $\widehat{C}_n$ in symmetric difference (3.3) can still be related to the expected supremum of empirical process, so Theorem 2.3 is applicable in this setting as well. Formally, we have the following.

THEOREM 3.4. *Suppose that for some* $\alpha \neq 1$, $\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \leq L\varepsilon^{-\alpha}$. *Then*

$$\sup_{C_0 \in \mathscr{C}} \mathbb{E}_{C_0} P|\widehat{C}_n \Delta C_0| \lesssim n^{-1/(\alpha+1)}.$$

PROOF. See Section 5. □

Compared to [30], Theorem 4.1, we use an unsieved least squares estimator to achieve the optimal rate, rather than their theoretical "sieved" estimator. This provides another example for which the simple least squares estimator can be rate-optimal for non-Donsker function classes in a natural setting.

3.3. *Binary classification*: *Excess risk bounds*. In this subsection, we consider the binary classification problem in the learning theory; cf. [32, 38]. Suppose one observes i.i.d. $(X_1, Y_1), \dots, (X_n, Y_n)$ with law P, where X_i's take values in $\mathcal{X}$, and the responses $Y_i \in \{0, 1\}$. A classifier $g : \mathcal{X} \to \{0, 1\}$ over a class $\mathcal{G}$ has a generalization error $P(Y \neq g(X))$. The excess risk for a classifier g over $\mathcal{G}$ under law P is given by

$$\mathcal{E}_P(g) \equiv P\big(Y \neq g(X)\big) - \inf_{g' \in \mathcal{G}} P\big(Y \neq g'(X)\big).$$

It is known that for a given law P on (X, Y), the minimal generalized error is attained by a Bayes classifier $g_0(x) \equiv \mathbf{1}_{\eta(x) \geq 1/2}$ where $\eta(x) \equiv \mathbb{E}[Y|X = x]$; cf. [10]. It is then natural to consider an estimator of g_0 by minimizing the empirical training error:

$$\widehat{g}_n \equiv \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Y_i \neq g(X_i)}. \tag{3.7}$$

We assume $g_0 \in \mathcal{G}$ for simplicity. The global ERM $\widehat{g}_n$ is previously studied in, for example, [38], pp. 136, [32], p. 2327, [24], pp. 2627–2629, [15], pp. 1211–1213. The quality of the estimator $\widehat{g}_n$ is measured by the excess risk:

$$\mathcal{E}_P(\widehat{g}_n) \equiv P\big(Y \neq \widehat{g}_n(X)\big) - P\big(Y \neq g_0(X)\big).$$

Let Π be the marginal distribution of X under P. We assume the following "Tsybakov's margin(low noise) condition" (cf. [31, 38]): there exists some $c > 0$ such that for all $g \in \mathcal{G}$,

$$\mathcal{E}_P(g) \geq c\big(\Pi\big(g(X) \neq g_0(X)\big)\big) = c\|g - g_0\|_{L_2(P)}^2. \tag{3.8}$$

Here, we have assumed that the margin condition holds with $\kappa = 1$. Although faster rates are possible under more general margin condition $\kappa \geq 1$ (cf. [31, 38]), we do not go into this direction to avoid distraction from our main points.

Below is the main result in this subsection, the formulation of which follows that of [15, 24].

THEOREM 3.5. *Suppose $\mathcal{G} \equiv \{\mathbf{1}_C : C \in \mathscr{C}\}$ satisfies the following entropy condition: there exists some $\alpha \neq 1$ such that for all $\varepsilon > 0$, $\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P) \leq L\varepsilon^{-\alpha}$. If $r_n^2 \geq Kn^{-1/(\alpha+1)}$ for a large enough constant $K > 0$, then*

$$\mathbb{P}\big(\mathcal{E}_P(\widehat{g}_n) \geq r_n^2\big) \leq K' \exp\big(-nr_n^2/K'\big)$$

holds for some constant $K' > 0$.

PROOF. See Section 5. □

Roughly speaking, the key to prove the above theorem is a control for the random variable

$$\max_{1 \leq j \leq \ell} \frac{\sup_{f \in \mathcal{F}_j} |\mathbb{P}_n(f) - P(f)|}{r_n^2 2^j},$$

where ℓ is the smallest integer such that $r_n^2 2^\ell \geq 1$, and $\mathcal{F}_j \equiv \{\mathbf{1}_{y \neq g_1(x)} - \mathbf{1}_{y \neq g_2(x)} : \mathcal{E}_P(g_1) \vee \mathcal{E}_P(g_2) \leq r_n^2 2^j\}$. A sharp estimate for the above random variable is achieved by an application of Theorem 2.3 and Talagrand's inequality (cf. Appendix C).

Examples of $\mathcal{G}$ that satisfy the prescribed entropy conditions in the above theorem can be found in the comments after [31], Theorem 1, p. 1813. To put the above results in the literature, [38] considered the same problem under the working assumption $\alpha \in (0, 1)$ (cf. [38], Assumption A2, p. 140). Massart and Nédélec [32] used ratio-type empirical process techniques to give a more unified treatment of deriving risk bounds for this problem, when the class of classifiers satisfies a Donsker bracketing entropy condition (i.e., $0 < \alpha < 1$), or a Donsker uniform entropy condition. Giné and Koltchinskii [15] further improved the result of [32] in the Donsker regime under a uniform entropy condition, by taking into account the size of the localized envelopes. See also [24], p. 2618, [29], p. 1706, for similar Donsker conditions. To the best knowledge of the author, our Theorem 3.5 gives a first result for the global ERM $\widehat{g}_n$ in (3.7) to be rate-optimal in the non-Donsker regime $\alpha > 1$ in the classification problem.

3.4. *Multiple isotonic regression.* Let $X_1, \ldots, X_n$ be i.i.d. with law P on $[0, 1]^d$. For simplicity we assume that P is the uniform distribution on $[0, 1]^d$. Consider the multiple isotonic regression model

$$Y_i = f_0(X_i) + \xi_i, \quad i = 1, \ldots, n, \tag{3.9}$$

where ξ_i's are i.i.d. Gaussian errors $\mathcal{N}(0, 1)$, and $f_0 \in \mathcal{M}_d \equiv \{f : [0, 1]^d \to \mathbb{R}, f(x) \leq f(y)$ for any $x \leq y\}$. Consider the isotonic least squares regression estimator $\widehat{f}_n$ defined via

$$\widehat{f}_n \equiv \operatorname*{arg\,min}_{f \in \mathcal{M}_d} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

The performance of $\widehat{f}_n$ in the multivariate setting is examined by [8] for $d = 2$ and [20] for $d \geq 3$. By the entropy estimate for uniformly bounded multiple isotonic functions in [14], $\mathcal{M}_d \cap L_\infty(1)$ is in the non-Donsker regime when $d \geq 3$, which is the main interesting case here.

THEOREM 3.6. *Let $d \geq 2$. Then with $\gamma_{d,iso} = (2)\mathbf{1}_{d=2} + \mathbf{1}_{d\geq 3}$,*

$$\sup_{f_0 \in \mathcal{M}_d \cap L_\infty(1)} \mathbb{E}_{f_0} \|\widehat{f}_n - f_0\|_{L_2(P)}^2 \lesssim_d n^{-1/d} \log^{\gamma_{d,iso}} n.$$

PROOF. See Section 5. □

Compared to [20], Theorem 4, the above result gives improvements over logarithmic factors. The logarithmic factors in $d \geq 3$ are due to boundary behavior of $\widehat{f}_n$. For instance, if the errors are bounded, then we may remove these logarithmic factors to get a sharp rate $n^{-1/d}$ for $d \geq 3$. These logarithmic factors cannot be removed by the proof techniques in [20] even if the errors are bounded. The rate $n^{-1/d}$ is shown to be minimax optimal for squared L_2 loss in [20].

The proof of Theorem 3.6 contains two inter-related steps:

1. First, we show that with high enough probability,

$$\|\widehat{f}_n - f_0\|_\infty = \mathcal{O}(\sqrt{\log n})$$

under the Gaussian noise assumption.

2. Second, using the first step, we show that

$$\mathbb{E} \sup_{f \in \mathcal{M}_d \cap L_\infty(C\sqrt{\log n})} |\mathbb{G}_n(f - f_0)| \lesssim \sqrt{\log n} \cdot \mathbb{E} \sup_{C \in \mathcal{L}_d} |\mathbb{G}_n(C)|,$$

where $\mathcal{L}_d$ is the collection of all upper and lower sets contained in $[0, 1]^d$ (precise definition see the paragraph before the proof of Theorem 3.6 in Section 5). The expected supremum on the right-hand side of the above display can be controlled using Theorem 2.3. Finally, the claim follows by a standard reduction for the risk of LSE to expected supremum of empirical process.

Compared to the proof of [20], Theorem 4, the proof strategy described above is more informative by making a clear connection to the class $\mathcal{L}_d$ that drives the minimax rates of convergence for the multiple isotonic LSE. See also Remark 5.5 for a detailed technical comparison.

The approach outlined above can also be adapted to the problem of multivariate convex regression modulo technical difficulties due to unsolved boundary behavior of the convex LSE. See Remark 5.6 for some details.

3.5. *s-concave density estimation in $\mathbb{R}^d$.* We first introduce the class of s-concave densities on $\mathbb{R}^d$. The exposition follows that of [21]. Let

$$M_s(a, b; \theta) \equiv \begin{cases} \big((1-\theta)a^s + \theta b^s\big)^{1/s}, & s \neq 0, a, b > 0, \\ 0, & s < 0, ab = 0, \\ a^{1-\theta} b^\theta, & s = 0, \\ a \wedge b, & s = -\infty. \end{cases}$$

A density p on $\mathbb{R}^d$ is called s-concave, that is, $p \in \mathcal{P}_s$ if and only if for all $x_0, x_1 \in \mathbb{R}^d$ and $\theta \in (0, 1)$, $p((1-\theta)x_0 + \theta x_1) \geq M_s(p(x_0), p(x_1); \theta)$. It is easy to see that the densities p

have the form $p = \varphi_+^{1/s}$ for some concave function φ if $s > 0$, $p = \exp(\varphi)$ for some concave φ if $s = 0$, and $p = \varphi_+^{1/s}$ for some convex φ if $s < 0$. The function classes $\mathcal{P}_s$ are nested in s in that for every $r > 0 > s$, we have $\mathcal{P}_r \subset \mathcal{P}_0 \subset \mathcal{P}_s \subset \mathcal{P}_{-\infty}$.

Maximum likelihood estimation over $\mathcal{P}_s$ is proposed in [37], where existence and consistency of the MLE $\widehat{p}_n$ is proved. Global rates of convergence of the MLE $\widehat{p}_n$ over $\mathcal{P}_s$ is primarily studied in the special case $s = 0$, also known as the log-concave MLE; cf. [23]. For general s-concave MLEs, the only result concerning global convergence rates is due to [11], who studied the univariate case $d = 1$, $s > -1$, showing that $h^2(\widehat{p}_n, p_0) = \mathcal{O}_{\mathbf{P}}(n^{-4/5})$, where $h(\cdot, \cdot)$ is the Hellinger distance. Here, we will be interested in general s-concave MLEs in general dimensions.

THEOREM 3.7. *Suppose $s > -1/d$ and $d \geq 2$. Then*

$$h^2(\widehat{p}_n, p_0) = \mathcal{O}_{\mathbf{P}}\big(n^{-2/(d+1)} \log^{\gamma_{d,s}} n\big),$$

where $\gamma_{d,s} = (2/3)\mathbf{1}_{d=2} + (2)\mathbf{1}_{d=3} + \mathbf{1}_{d\geq 4}$.

PROOF. See Section 5. □

The most interesting regime here is $d \geq 4$ when the entropy integral for the class of s-concave densities diverges. Modulo logarithmic factors, the rates of convergence for the s-concave MLE $\widehat{p}_n$ in squared Hellinger distance is $\mathcal{O}_{\mathbf{P}}(n^{-2/(d+1)})$, which matches the minimax lower bound for the smaller log-concave (= 0-concave) class; cf. [23]. During the preparation the paper, the author becomes aware of the very nice work [9] which derives, among other things, global risk bounds for the log-concave (i.e., $s = 0$) MLEs in the Hellinger distance. The techniques used in both papers in this example share certain common features, while our general setting brings about further technical challenges. See Remark 3.9 below for more technical comments on the proof of Theorem 3.7.

The integrability restriction $s > -1/d$ is very natural in this setting: if $s < -1/d$, then there exists a family of s-concave densities with singularities so that the MLE does not exist. The following proposition makes this precise.

PROPOSITION 3.8. *The s-concave MLE does not exist for $s < -1/d$.*

PROOF. For $a \in \mathbb{R}^d$, $b > 0$, let $\widetilde{\varphi}_{a,b}(x) \equiv \|x - a\|\mathbf{1}_{\|x-a\|\leq b} + \infty \mathbf{1}_{\|x-a\|>b}$. Since $c_b \equiv \int \widetilde{\varphi}_{a,b}^{1/s} = \int \widetilde{\varphi}_{0,b}^{1/s} < \infty$ for $s < -1/d$, $p_{a,b} \equiv \widetilde{\varphi}_{a,b}^{1/s}/c_b$ is an s-concave density. The log likelihood function for observed $X_1, \ldots, X_n$ is $\ell(a, b) \equiv \log \prod_{i=1}^n p_{a,b}(X_i) = \sum_{i=1}^n [(1/s) \times \log(\|X_i - a\|) - \log c_b]$ for (a, b) such that $\max_i \|X_i - a\| \leq b$ and $X_i \neq a$ for $i = 1, \ldots, n$. For b large enough and a approaches any of X_i's, $\ell(a, b) \uparrow \infty$, so the MLE does not exist. □

The univariate case $d = 1$ for the above proposition can also be found in [11].

REMARK 3.9. The proof of Theorem 3.7 relies on the following reduction scheme:

$$h^2(\widehat{p}_n, p_0) \lesssim \log n \cdot \mathbb{E} \sup_{C\in\mathscr{C}_d} \big|(\mathbb{P}_n - P_0)(C)\big| + \mathcal{O}_{\mathbf{P}}\big(n^{-1/2}\big), \tag{3.10}$$

where $\mathscr{C}_d$ is the class of convex bodies on $\mathbb{R}^d$.

The above reduction scheme (3.10) for $s = 0$ is essentially achieved in [6], but a sharp bound for the expected supremum of the empirical process on the right- hand side of the above display is not available therein. Here, we show that the reduction (3.10) holds for the

maximum regime in which the s-concave MLE exists. Once (3.10) is proven, the expected supremum on its right-hand side can be controlled by Theorem 2.3 combined with a standard technique of "domain extension" (cf. [46] or [47], Corollary 2.7.4) under the envelope control (5.7). See the proof of Theorem 3.7 for more details.

4. Proofs for Section 2. We will prove Theorem 2.1 in this section. As (2) is a direct consequence of (1), we will prove (1) and (3) only.

4.1. *Proof of Theorem* 2.1(1). A heuristic way of seeing the bound (2.1) is the following. This requires some understanding for the proof of the classical maximal inequality (2.6):

- The entropy integral term $\int_\gamma^\sigma \sqrt{\log \mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}, L_2(P)}\,\mathrm{d}\varepsilon$ in (2.6) comes from L_2 chaining with bracketing in the "Gaussian regime" using Bernstein's inequality, starting from σ down to γ.
- The residual term $\sqrt{n}\gamma$ in (2.6) comes from the bound $\|f\|_{L_1(P)} \le \|f\|_{L_2(P)}$ toward the end level γ of the L_2 chaining.

Now we wish to carry out L_p chaining with bracketing for, say, $p \in [1,2]$. Clearly, $\|f\|_{L_2(P)} \le \|f\|_{L_p(P)}^{p/2}$ and $\|f\|_{L_1(P)} \le \|f\|_{L_p(P)}$. This naturally hints the following conjecture: for $\sigma > 0$ not "too small,"

$$
\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}(\sigma)} |\mathbb{G}_n(f)| &\lesssim \inf_{0\le\gamma\le\sigma/2}\biggl\{\sqrt{n}\gamma + \int_{\gamma^{p/2}}^{\sigma^{p/2}} \sqrt{\log \mathcal{N}_{[\cdot]}\bigl(\varepsilon^{2/p}, \mathcal{F}, L_p(P)\bigr)}\,\mathrm{d}\varepsilon\biggr\}\\
&\asymp \inf_{0\le\gamma\le\sigma/2}\biggl\{\sqrt{n}\gamma + \int_{\gamma}^{\sigma} u^{p/2-1}\sqrt{\log \mathcal{N}_{[\cdot]}\bigl(u, \mathcal{F}, L_p(P)\bigr)}\,\mathrm{d}u\biggr\}\\
&\lesssim \inf_{0\le\gamma\le\sigma/2}\biggl\{\sqrt{n}\gamma + \int_{\gamma}^{\sigma} u^{(p-\alpha)/2-1}\,\mathrm{d}u\biggr\}\\
&\sim \inf_{0\le\gamma\le\sigma/2}\bigl\{\sqrt{n}\gamma + |u^{(p-\alpha)/2}|_{u=\gamma}^{u=\sigma}|\bigr\}\\
&\sim \begin{cases} \sigma^{\frac{p-\alpha}{2}}, & \alpha < p,\\ \inf\limits_{0\le\gamma\le\sigma/2}\bigl\{\sqrt{n}\gamma + \gamma^{-(\alpha-p)/2}\bigr\}, & \alpha > p, \end{cases}\\
&\sim \begin{cases} \sigma^{\frac{p-\alpha}{2}}, & \alpha < p,\\ n^{\frac{\alpha-p}{2(\alpha+2-p)}}, & \alpha > p. \end{cases}
\end{aligned}
$$

Below we implement this heuristic program rigorously and identify the regime of $\sigma > 0$ in which the above bound holds.

PROOF OF THEOREM 2.1(1). The proof is inspired by the proof of [47], Lemma 2.14.3, that is originated in [33]. We use the same notation for convenience of the readers. Without loss of generality, we assume $\sigma = 2^{-q_0}$ for some $q_0 \in \mathbb{N}$. By the assumption, there exist nested partitions $\{\mathcal{F} = \bigcup_{i=1}^{N_q} \mathcal{F}_{q,i}\}_{q=q_0}^\infty$ such that for all $q \ge q_0$, (i) $\max_{1\le i\le N_q} \|\sup_{f,g\in\mathcal{F}_{q,i}} |f - g|\|_{L_p} \le 2^{-q}$, and (ii) $\log N_q \lesssim_{L,\alpha} 2^{q\alpha}$. [Such nested partitions can be constructed as follows. First, taking unnested partitions $\{\mathcal{F} = \bigcup_{i=1}^{\bar N_q} \bar{\mathcal{F}}_{q,i}\}_{q=q_0}^\infty$ with $\bar N_q \le \mathcal{N}_{[\cdot]}(2^{-q}, \mathcal{F}, L_p) \le \exp(L\cdot 2^{q\alpha})$. Then at level q, we use the partition consisting of all intersections of form $\{\bigcup_{r=q_0}^q \bar{\mathcal{F}}_{r,i_r} : 1 \le i_r \le \bar N_r, q_0 \le r \le q\}$. The first property above is obvious. The second one follows as $N_q \le \prod_{r=q_0}^q \bar N_r$.].

Pick any $f_{q,i} \in \mathcal{F}_{q,i}$. For any $f \in \mathcal{F}$, let $\pi_q f \equiv f_{q,i}$ and $\Delta_q f \equiv \sup_{f,g\in\mathcal{F}_{q,i}} |f - g|$ if $f \in \mathcal{F}_{q,i}$. Let $a_q \equiv 2^{-q\{(p\wedge 2)+\alpha\}/2}$. Now define the indicator functions $A_{q-1} f \equiv \mathbf{1}_{\Delta_{q_0} f \le \sqrt{n} a_{q_0}, \ldots, \Delta_{q-1} f \le \sqrt{n} a_{q-1}}$, $B_q f \equiv \mathbf{1}_{\Delta_{q_0} f \le \sqrt{n} a_{q_0}, \ldots, \Delta_{q-1} f \le \sqrt{n} a_{q-1}, \Delta_q f > \sqrt{n} a_q}$, and $B_{q_0} f \equiv \mathbf{1}_{\Delta_{q_0} f > \sqrt{n} a_{q_0}}$. In words, $A_{q-1} f$ indicates the region for which "Gaussian estimates" are valid up to level $q-1$ for f, while B_q indicates the region for which the "Gaussian estimate" is first violated at level q for f.

By the last display in [47], p. 241, the following chaining holds for any $q_1 > q_0$:

$$
\begin{aligned}
f - \pi_{q_0} f = (f - \pi_{q_0} f) B_{q_0} f + \sum_{q=q_0+1}^{q_1} (f - \pi_q f) B_q f \\
+ \sum_{q=q_0+1}^{q_1} (\pi_q f - \pi_{q-1} f) A_{q-1} f + (f - \pi_{q_1} f) A_{q_1} f.
\end{aligned} \tag{4.1}
$$

We bound the expectation of the above four terms when applied with the empirical process, and name them (I)–(IV). Roughly speaking, (III) is the term with "Gaussian behavior" due to the construction of $A_{q-1} f$, while for the terms (I)–(II), the Gaussian estimate fails on $B_q f$ at level q. This will be compensated by the observation that the Gaussian estimate still holds at level $q-1$. The single term (IV) terminates the chaining and will be handled via an L_p control. Below we implement this rough idea precisely.

For (I) and (II), note that

$$
\begin{aligned}
&\mathbb{E} \sup_{f\in\mathcal{F}} \left|\mathbb{G}_n (f - \pi_q f) B_q f\right| \\
&\le \mathbb{E} \sup_{f\in\mathcal{F}} \sqrt{n} \mathbb{P}_n |f - \pi_q f| B_q f + \sqrt{n} \sup_{f\in\mathcal{F}} P |f - \pi_q f| B_q f \\
&\le \mathbb{E} \sup_{f\in\mathcal{F}} \sqrt{n} \mathbb{P}_n \Delta_q f B_q f + \sqrt{n} \sup_{f\in\mathcal{F}} P \Delta_q f B_q f \\
&\le \mathbb{E} \sup_{f\in\mathcal{F}} \left|\mathbb{G}_n (\Delta_q f B_q f)\right| + 2\sqrt{n} \sup_{f\in\mathcal{F}} P \Delta_q f B_q f.
\end{aligned} \tag{4.2}
$$

Hence

$$
\begin{aligned}
(I) + (II) &= \sum_{q=q_0}^{q_1} \mathbb{E} \sup_{f\in\mathcal{F}} \left|\mathbb{G}_n (f - \pi_q f) B_q f\right| \\
&\le \sum_{q=q_0}^{q_1} \mathbb{E} \sup_{f\in\mathcal{F}} \left|\mathbb{G}_n (\Delta_q f B_q f)\right| + 2\sqrt{n} \sum_{q=q_0}^{q_1} \sup_{f\in\mathcal{F}} P \Delta_q f B_q f.
\end{aligned} \tag{4.3}
$$

For $q = q_0$, we use the trivial bound $\Delta_{q_0} f B_{q_0} f \le 1$. For $q \ge q_0 + 1$, we will however implement control at level $q-1$: As the partitions are nested, $\Delta_q f B_q f \le \Delta_{q-1} f B_q f \le \sqrt{n} a_{q-1}$ for all $q \ge q_0 + 1$. In summary, for $q \ge q_0$,

$$
\Delta_{q_0} f B_{q_0} f \le 1 \wedge \sqrt{n} a_{q-1}.
$$

On the other hand,

$$
\sup_{f\in\mathcal{F}} P (\Delta_q f B_q f)^2 \le \sup_{f\in\mathcal{F}} \left(P (\Delta_q f B_q f)^p\right)^{1\wedge(2/p)} \le 2^{-q(p\wedge 2)}.
$$

By Bernstein inequality, for any $q \ge q_0$,

$$
\mathbb{E} \sup_{f\in\mathcal{F}} \left|\mathbb{G}_n (\Delta_q f B_q f)\right|
$$

$$\begin{aligned}(4.4)\qquad &\lesssim 2^{-q(p\wedge 2)/2}\sqrt{\log N_q}+n^{-1/2}(1\wedge\sqrt{n}a_{q-1})\log N_q\\ &\lesssim 2^{-q\{(p\wedge 2)-\alpha\}/2}+\big(n^{-1/2}\wedge 2^{-q\{(p\wedge 2)+\alpha\}/2}\big)\cdot 2^{q\alpha}\asymp 2^{-q\{(p\wedge 2)-\alpha\}/2}.\end{aligned}$$

Furthermore,

$$\begin{aligned}(4.5)\qquad P\Delta_q f B_q f &\le P\Delta_q f\mathbf{1}_{\Delta_q f>\sqrt{n}a_q}\le(\sqrt{n}a_q)^{-1}P(\Delta_q f)^2\\ &\le n^{-1/2}a_q^{-1}2^{-q(p\wedge 2)}=n^{-1/2}2^{-q\{(p\wedge 2)-\alpha\}/2}.\end{aligned}$$

Combining (4.3)–(4.5), we have

$$\begin{aligned}(4.6)\qquad (I)+(II)&\lesssim\sum_{q=q_0}^{q_1}2^{-q\{(p\wedge 2)-\alpha\}/2}\\ &\asymp_{p,\alpha}2^{-q_0\{(p\wedge 2)-\alpha\}/2}\mathbf{1}_{\alpha<p\wedge 2}+2^{q_1\{\alpha-(p\wedge 2)\}/2}\mathbf{1}_{\alpha>p\wedge 2}.\end{aligned}$$

For (III), we have Gaussian estimates as follows. Note that

$$\begin{aligned}&|\pi_q f-\pi_{q-1}f|A_{q-1}f\le 1\wedge\Delta_{q-1}fA_{q-1}f\le 1\wedge\sqrt{n}a_{q-1},\\ &P\big(|\pi_q f-\pi_{q-1}f|A_{q-1}f\big)^2\le P(\Delta_{q-1}f)^2\le 2^{-(q-1)(p\wedge 2)}.\end{aligned}$$

As the cardinality of $\{(\pi_q f-\pi_{q-1}f)A_{q-1}f:f\in\mathcal{F}\}$ is at most $N_qN_{q-1}\le N_q^2$, by Bernstein inequality and a similar argument to (4.4),

$$\begin{aligned}(III)&=\sum_{q=q_0+1}^{q_1}\mathbb{E}\sup_{f\in\mathcal{F}}\big|\mathbb{G}_n\big((\pi_q f-\pi_{q-1}f)A_{q-1}f\big)\big|\\ &\lesssim\sum_{q=q_0+1}^{q_1}2^{-q\{(p\wedge 2)-\alpha\}/2}\qquad(4.7)\\ &\asymp_{p,\alpha}2^{-q_0\{(p\wedge 2)-\alpha\}/2}\mathbf{1}_{\alpha<p/2}+2^{q_1\{\alpha-(p\wedge 2)\}/2}\mathbf{1}_{\alpha>p\wedge 2}.\end{aligned}$$

For (IV), using similar arguments as in (4.2),

$$\begin{aligned}(IV)&=\mathbb{E}\sup_{f\in\mathcal{F}}\big|\mathbb{G}_n(f-\pi_{q_1}f)A_{q_1}f\big|\\ &\le\mathbb{E}\sup_{f\in\mathcal{F}}\big|\mathbb{G}_n(\Delta_{q_1}fA_{q_1}f)\big|+2\sqrt{n}\sup_{f\in\mathcal{F}}P\Delta_{q_1}fA_{q_1}f\qquad(4.8)\\ &\lesssim 2^{-q_1\{(p\wedge 2)-\alpha\}/2}+\sqrt{n}2^{-q_1}.\end{aligned}$$

Here, in the last inequality we used $P\Delta_{q_1}fA_{q_1}f\le\|\Delta_{q_1}fA_{q_1}f\|_{L_p}\le 2^{-q_1}$. Finally, using Bernstein's inequality again,

$$(4.9)\qquad \mathbb{E}\sup_{f\in\mathcal{F}}\big|\mathbb{G}_n(\pi_{q_0}f)\big|\lesssim 2^{-q_0\{(p\wedge 2)-\alpha\}/2}+n^{-1/2}2^{q_0\alpha}.$$

Combining (4.1), (4.6)–(4.9), we obtain

$$\mathbb{E}\sup_{f\in\mathcal{F}}\big|\mathbb{G}_n(f)\big|\lesssim\begin{cases}2^{-q_0\{(p\wedge 2)-\alpha\}/2}+n^{-1/2}2^{q_0\alpha}+\sqrt{n}2^{-q_1}, & \alpha<p\wedge 2,\\ 2^{q_1\{\alpha-(p\wedge 2)\}/2}+n^{-1/2}2^{q_0\alpha}+\sqrt{n}2^{-q_1}, & \alpha>p\wedge 2.\end{cases}$$

The $\lesssim$ does not depend on q_1. Now optimize over $q_1>q_0$ to conclude. $\square$

REMARK 4.1. Pollard [34] suggested the following key "recursive equality" (cf. [34], p. 1043, last display)

$$R_iT_i = R_{i+1}T_{i+1} - R_{i+1}T_i^cT_{i+1} + (R_i - R_{i+1})T_iT_{i+1} + R_iT_iT_{i+1}^c \tag{4.10}$$

for the chaining method of [33]. See [34] for definitions of the above notation. Interestingly, (4.10) can also be viewed as a one-step chaining of (4.1) from i to $i+1$, as we may regard $f - \pi_i f = R_i$, $A_i f = T_i$, $A_{i+1} f = T_iT_{i+1}$ and $B_{i+1} f = T_iT_{i+1}^c$ by translating the notation used here to those in [34]. Some elementary algebra then reduces (4.1) to (4.10).

4.2. *Proof of Theorem* 2.1(3). The proof of (2.5) is divided into two steps:

1. First, we establish a lower bound for the *Gaussianized* empirical process; see Proposition 4.2. This can be done roughly via Sudakov minimization in the Gaussian regime, that is, when σ is not too small.

2. Second, we will control the Gaussianized empirical process by the standard symmetrized empirical process from above. This step requires several subtle estimates as a naive bound would incur additional undesirable logarithmic factors. This is done via the help of a multiplier inequality derived in [22] (cf. Appendix C). Roughly speaking, the logarithmic factors can be removed for the Gaussianized empirical process at sample size n as long as the empirical process has size no smaller than Gaussian maxima along the "entire path" from 1 to n. Details see the proofs of Propositions 4.4 and 4.6.

We first prove the lower bound for Gaussianized empirical process.

PROPOSITION 4.2. *Let $\mathcal{F} \subset L_\infty(1)$. For any $\sigma \geq 50n^{-1/2}$ such that*

$$\log \mathcal{N}\big(\sigma/4, \mathcal{F}(\sigma), L_2(P)\big) \leq n\sigma^2/4000,$$

we have

$$\mathbb{E} \sup_{f\in\mathcal{F}(\sigma)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i f(X_i) \right| \gtrsim \sigma \sqrt{\log \mathcal{N}\big(\sigma/2, \mathcal{F}(\sigma), L_2(P)\big)}.$$

Here, $g_1, \ldots, g_n$ are i.i.d. $\mathcal{N}(0,1)$.

PROOF OF PROPOSITION 4.2. By Sudakov minorization (cf. Lemma C.2), for any $\sigma > 0$,

$$\mathbb{E} \sup_{f\in\mathcal{F}(\sigma)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i f(X_i) \right| \gtrsim \mathbb{E}\sigma \sqrt{\log \mathcal{N}\big(\sigma/10, \mathcal{F}(\sigma), L_2(\mathbb{P}_n)\big)}. \tag{4.11}$$

We claim that for any $\sigma > 0$ such that $\log \mathcal{N}(\sigma/4, \mathcal{F}(\sigma), L_2(P)) \leq n\sigma^2/4000$,

$$\mathbb{P}\big(\mathcal{N}\big(\sigma/10, \mathcal{F}(\sigma), L_2(\mathbb{P}_n)\big) \geq \mathcal{N}\big(\sigma/2, \mathcal{F}(\sigma), L_2(P)\big)\big) \geq 1 - e^{-n\sigma^2/2000}. \tag{4.12}$$

To see this, let $f_1, \ldots, f_N$ be a maximal $\sigma/2$-packing set of $\mathcal{F}(\sigma)$ in the $L_2(P)$ metric, that is, for $i \neq j$, $P(f_i - f_j)^2 \geq \sigma^2/4$. Since $P(f_i - f_j)^4 \leq 4P(f_i - f_j)^2 \leq 16\sigma^2$, we apply Bernstein's inequality followed by a union bound to see that with probability at least $1 - N^2\exp(-t)$,

$$\max_{1\leq i\neq j\leq N} \left(nP(f_i - f_j)^2 - \sum_{k=1}^n (f_i - f_j)^2(X_k) \right) \leq \frac{2t}{3} + \sqrt{32tn\sigma^2}.$$

With $t = cn\sigma^2$ for a constant $c > 0$ to be specified below, we obtain

$$\mathbb{P}\left(\min_{1\le i\ne j\le N}\frac{1}{n}\sum_{k=1}^{n}(f_i - f_j)^2(X_k) \ge \sigma^2(1/4 - 2c/3 - \sqrt{32c})\right)$$
$$\ge 1 - e^{2\log\mathcal{D}(\sigma/2,\mathcal{F}(\sigma),L_2(P))-cn\sigma^2} \ge 1 - e^{2\log\mathcal{N}(\sigma/4,\mathcal{F}(\sigma),L_2(P))-cn\sigma^2},$$

where $\mathcal{D}(\cdot,\cdot,\cdot)$ stands for the packing number. By choosing $c = 1/10^3$ and $\log\mathcal{N}(\sigma/4, \mathcal{F}(\sigma), L_2(P)) \le n\sigma^2/4000$, we have

$$\mathbb{P}\left(\min_{1\le i\ne j\le N}\frac{1}{n}\sum_{k=1}^{n}(f_i - f_j)^2(X_k) \ge 0.04\sigma^2\right) \ge 1 - \exp\bigl(-n\sigma^2/2000\bigr).$$

This entails that $\mathcal{D}(\sigma/5, \mathcal{F}(\sigma), L_2(\mathbb{P}_n)) \ge N \equiv \mathcal{D}(\sigma/2, \mathcal{F}(\sigma), L_2(P))$ with the above probability. Hence for any $\sigma > 0$ such that $\log\mathcal{N}(\sigma/4, \mathcal{F}(\sigma), L_2(P)) \le n\sigma^2/4000$, with probability at least $1 - e^{-n\sigma^2/2000}$,

$$\mathcal{N}\bigl(\sigma/10, \mathcal{F}(\sigma), L_2(\mathbb{P}_n)\bigr) \ge \mathcal{D}\bigl(\sigma/5, \mathcal{F}(\sigma), L_2(\mathbb{P}_n)\bigr)$$
$$\ge \mathcal{D}\bigl(\sigma/2, \mathcal{F}(\sigma), L_2(P)\bigr) \ge \mathcal{N}\bigl(\sigma/2, \mathcal{F}(\sigma), L_2(P)\bigr),$$

completing the proof of (4.12). Hence for any $\sigma \ge 50n^{-1/2}$ such that the entropy $\log\mathcal{N}(\sigma/4, \mathcal{F}(\sigma), L_2(P)) \le n\sigma^2/4000$, the claim of the proposition follows from (4.11) and (4.12). □

Next, we eliminate the effect of the Gaussian multiplier. We need a technical lemma.

LEMMA 4.3. *Let $g_1, \dots, g_n$ be i.i.d. $\mathcal{N}(0,1)$, and $|g_{(n)}| \le \cdots \le |g_{(1)}|$ be reversed order statistics of $\{|g_1|, \dots, |g_n|\}$. Then there exists an absolute constant $K > 0$ such that for any $c \in (0, K^{-1})$, and $0 \le t \le K^{-1}\sqrt{\log(1/c)}$, we have $\mathbb{P}(|g_{(\lfloor cn\rfloor)}| \le t) \le e^{-c^2n/K}$.*

PROOF. For notational convenience, we assume that $cn \in \mathbb{N}$. Let $\phi(t) \equiv \mathbb{P}(|g_1| > t)$. By [16], (2.23), $\sqrt{2/\pi}\cdot\frac{t}{t^2+1}e^{-t^2/2} \le \phi(t) \le \min\{1, \sqrt{2/\pi}\cdot t^{-1}\}e^{-t^2/2}$. Let $t_c > 0$ be such that $\phi(t_c) = 2c$. Then $t_c \le \sqrt{2\log(1/2c)}$. By Bernstein's inequality, for $0 \le t \le t_c$, $2c \le \phi(t) \le 1$, so

$$\mathbb{P}\bigl(|g_{(cn)}| \le t\bigr) \le \mathbb{P}\left(\sum_{i=1}^{n}\mathbf{1}_{|g_i|>t} \le cn\right)$$
$$= \mathbb{P}\left(\sum_{i=1}^{n}\bigl(\mathbf{1}_{|g_i|>t} - \phi(t)\bigr) \le -\bigl(\phi(t) - c\bigr)n\right)$$
$$\le \exp\left(-\frac{(\phi(t) - c)^2n^2}{2n\phi(t) + 4(\phi(t) - c)n/3}\right) \le e^{-c^2n/K},$$

proving the claim. □

PROPOSITION 4.4. *Let the conditions in Theorem 2.1(3) hold for some $\alpha \in (0,1)$ and $L > 0$ large enough. Then for $\sigma_n^2 \ge cn^{-1/(\alpha+1)}$ with some constant $c > 0$, $\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}|\mathbb{G}_n(f)| \gtrsim_\alpha \sigma_n^{1-\alpha}$.*

PROOF. By Proposition 4.2, the Gaussianized empirical process satisfies

$$\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_i f(X_i)\right| \gtrsim \sigma_n\sqrt{\log\mathcal{N}\bigl(\sigma_n/2, \mathcal{F}(\sigma_n), L_2(P)\bigr)} \ge C_1^{-1}\sigma_n^{1-\alpha}.$$

Suppose that $\sigma_n^2 \le c$. Without loss of generality, we assume that $\sigma_n^2 \equiv \sigma_n(\gamma)^2 = cn^{-\gamma}$ for some $0 \le \gamma \le 1/(\alpha+1)$ and define $\sigma_k^2 \equiv ck^{-\gamma}$. We first prove the following claim: there exists some $c_1 \equiv c_1(c,\alpha) > 0$ such that for any $0 \le \gamma \le 1/(\alpha+1)$,

$$\mathbb{E} \sup_{f \in \mathcal{F}(\sigma_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \ge c_1 \sigma_n^{1-\alpha}. \tag{4.13}$$

To this end, let $a_n \equiv (\sigma_n^{1-\alpha})^{-1} \mathbb{E} \sup_{f \in \mathcal{F}(\sigma_n)} |\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i)|$. Then by the local maximal inequality (cf. Theorem 2.1-(1)), we see that $\sup_{k \in \mathbb{N}} a_k \le C_2 \equiv C_2(\alpha)$. Since $\sqrt{n}\sigma_n^{1-\alpha} = c^{(1-\alpha)/2} n^\beta$ where $\beta \equiv \beta(\alpha,\gamma) \equiv \frac{1}{2}(1-(1-\alpha)\gamma) \in [\alpha/(1+\alpha), 1/2]$, we have by Lemma C.4 that for a constant $c' > 0$ to be determined later,

$$\begin{aligned}
C_1^{-1} n^\beta &\le c^{-(1-\alpha)/2} \mathbb{E} \sup_{f \in \mathcal{F}(\sigma_n)} \left| \sum_{i=1}^n g_i f(X_i) \right| \\
&\le c^{-(1-\alpha)/2} \mathbb{E} \left[\sum_{k=1}^n \big(|g_{(k)}| - |g_{(k+1)}| \big) \mathbb{E} \sup_{f \in \mathcal{F}(\sigma_k)} \left| \sum_{i=1}^k \varepsilon_i f(X_i) \right| \right] \\
&\le \mathbb{E} \left[\sum_{k=1}^{\lfloor c'n \rfloor - 1} \big(|g_{(k)}| - |g_{(k+1)}| \big) a_k k^\beta \right] \\
&\quad + \mathbb{E} \left[\sum_{k=\lfloor c'n \rfloor}^n \big(|g_{(k)}| - |g_{(k+1)}| \big) a_k k^\beta \right] \equiv (I) + (II).
\end{aligned} \tag{4.14}$$

For (I) in (4.14), using the same notation as in the proof of Lemma 4.3,

$$\begin{aligned}
(I) &\le C_2 \cdot \mathbb{E} \left[\sum_{k=1}^{\lfloor c'n \rfloor - 1} \big(|g_{(k)}| - |g_{(k+1)}| \big) k^\beta \right] \\
&\le C_2 \cdot \mathbb{E} \left[\sum_{k=1}^{\lfloor c'n \rfloor - 1} \int_{|g_{(k+1)}|}^{|g_{(k)}|} k^\beta \, \mathrm{d}t \right] \\
&\le C_2 \cdot \mathbb{E} \int_0^\infty \left(\sum_{i=1}^n \mathbf{1}_{|g_i| \ge t} \mathbf{1}_{|g_{(\lfloor c'n \rfloor)}| \le t \le |g_{(1)}|} \right)^\beta \mathrm{d}t \\
&\le C_2 n^\beta \cdot \int_0^\infty \big(\mathbb{P}(|g_1| > t) \mathbb{P}\big(|g_{(\lfloor c'n \rfloor)}| \le t \le |g_{(1)}| \big) \big)^{\beta/2} \, \mathrm{d}t \\
&\le C_2 n^\beta \left(\int_0^{K^{-1}\sqrt{\log(1/c')}} \phi(t)^{\beta/2} \mathbb{P}\big(|g_{(\lfloor c'n \rfloor)}| \le t \big)^{\beta/2} \, \mathrm{d}t + \int_{K^{-1}\sqrt{\log(1/c')}}^\infty \phi(t)^{\beta/2} \, \mathrm{d}t \right) \\
&\le C_2 n^\beta \left(\int_0^{K^{-1}\sqrt{\log(1/c')}} e^{-\beta t^2/4 - \beta (c')^2 n/2K} \, \mathrm{d}t + \int_{K^{-1}\sqrt{\log(1/c')}}^\infty e^{-\beta t^2/4} \, \mathrm{d}t \right) \\
&\le C_3 n^\beta \big(e^{-(c')^2 n/C_3} + e^{-\log(1/c')/C_3} \big) \le \big(C_1^{-1}/2 \big) n^\beta,
\end{aligned}$$

by choosing $c' \equiv \exp(-C_3 \log(4C_1C_3))$ and $n \ge C_3 \log(4C_1C_3)/(c')^2$. On the other hand, for (II) in (4.14), we have

$$(II) \le \Big(\max_{\lfloor c'n \rfloor \le k \le n} a_k \Big) \mathbb{E} \left[\sum_{k=1}^n \big(|g_{(k)}| - |g_{(k+1)}| \big) k^\beta \right] \le \Big(\max_{\lfloor c'n \rfloor \le k \le n} a_k \Big) G_\alpha n^\beta,$$

where $G_\alpha = \int_0^\infty (\mathbb{P}(|g_1| > t))^{\alpha/(1+\alpha)}\,\mathrm{d}t < \infty$ since Gaussian random variables have finite moments of any order, and the last inequality follows from Jensen's inequality. Combining the above displays, we see that

$$1/(2C_1G_a) \le \max_{\lfloor c'n\rfloor \le k \le n} a_k \le \max_{\lfloor c'n\rfloor \le k \le n} \bigl(\sigma_k^{1-\alpha}\bigr)^{-1}\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_k)}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^k \varepsilon_i f(X_i)\right|$$
$$\le \sigma_n^{-(1-\alpha)}\cdot\sqrt{1/\bigl(c'-1/n\bigr)}\cdot\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_{\lfloor c'n\rfloor})}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right|,$$

where in the last inequality we used Jensen's inequality. This proves our claim (4.13) by adjusting the constant and choosing $n \ge 2/c'$. Now by desymmetrization inequality (cf. [47], Lemma 2.3.6), we have that

$$\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right| \le 2\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\bigl|\mathbb{G}_n(f)\bigr| + 2\sigma_n. \tag{4.15}$$

For $\sigma_n \le (c_1/4)^{1/\alpha}\wedge c^{1/2}$, the claim of the proposition follows from (4.13) and (4.15). On the other hand, the claim is trivial for $\sigma_n > (c_1/4)^{1/\alpha}\wedge c^{1/2}$. □

REMARK 4.5. From the proof of Proposition 4.4, the bracketing entropy upper bound is only used to prove $\sup_{k\in\mathbb{N}} a_k \le C_2 \equiv C_2(\alpha)$. This means that we may impose instead a uniform entropy upper bound condition as in [15] in the regime $0 < \alpha < 1$.

PROPOSITION 4.6. *Let the conditions in Theorem 2.1(3) hold for some $\alpha > 1$ and $L > 0$ large enough. Then for $\sigma_n^2 \equiv cn^{-1/(\alpha+1)}$ with some constant $c > 0$, $\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}|\mathbb{G}_n(f)| \gtrsim n^{(\alpha-1)/2(\alpha+1)}$.*

PROOF. Proposition 4.2 shows that

$$\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i f(X_i)\right| \gtrsim \sigma_n\sqrt{\log\mathcal{N}\bigl(\sigma_n/2,\mathcal{F}(\sigma_n),L_2(P)\bigr)} \gtrsim n^{(\alpha-1)/2(\alpha+1)}.$$

Now applying Lemma C.4 in the following form:

$$\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i f(X_i)\right| \lesssim \max_{1\le k\le n}\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^k \varepsilon_i f(X_i)\right|,$$

we see that for some $K > 0$,

$$\max_{1\le k\le n}\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\bigl|\mathbb{G}_k(f)\bigr| \ge K^{-1}n^{(\alpha-1)/2(\alpha+1)}.$$

On the other hand, by enlarging K if necessary, Theorem 2.1(1) entails that $\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_k)}|\mathbb{G}_k(f)| \le K\cdot k^{(\alpha-1)/2(\alpha+1)}$, and hence

$$\max_{1\le k\le n}\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_k)}\bigl|\mathbb{G}_k(f)\bigr| \le Kn^{(\alpha-1)/2(\alpha+1)}$$

by the assumption $\alpha > 1$. Combining the upper and lower estimates, we see that

$$K^{-1}n^{(\alpha-1)/2(\alpha+1)} \le \max_{1\le k\le n}\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\bigl|\mathbb{G}_k(f)\bigr|$$
$$\le \max_{1\le k\le n}\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_k)}\bigl|\mathbb{G}_k(f)\bigr| \le Kn^{(\alpha-1)/2(\alpha+1)}.$$

Now we will argue that the max operator can be "eliminated." To this end, let $a_k \equiv \mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}|\mathbb{G}_k(f)|$ and $\beta \equiv (\alpha-1)/2(\alpha+1)$ for notational convenience. Let $k_n \equiv \arg\max_{1\le k\le n} a_k$. We claim that $k_n \in [cn, n]$ where $c = K^{-2/\beta} \in (0,1)$. To see this, we only need to note $K^{-1}n^\beta \le \max_{1\le k\le n} a_k = a_{k_n} \le K k_n^\beta$, which entails $k_n^\beta \ge K^{-2}n^\beta$. Hence

$$K^{-1}n^{(\alpha-1)/2(\alpha+1)} \le \mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\left|\mathbb{G}_{k_n}(f)\right| \le \frac{1}{\sqrt{c}}\mathbb{E}\sup_{f\in\mathcal{F}(\sigma_n)}\left|\mathbb{G}_n(f)\right|,$$

where the last inequality follows from Jensen's inequality, proving the claim. □

PROOF OF THEOREM 2.1(3). The claims follow by combining Propositions 4.4 and 4.6. □

5. Proofs for Section 3.

5.1. *Proof of Theorem 3.1.* Before the proof of Theorem 3.1, we need the following.

LEMMA 5.1. *Consider the regression model (3.1) and the least squares estimator $\widehat{C}_n$ in (3.2). Suppose that $(\xi_1, X_1), \dots, (\xi_n, X_n)$ are i.i.d. random vectors with $\mathbb{E}[\xi_1|X_1] = 0$ and $\mathbb{E}[\xi_1^2|X_1] \lesssim 1 \vee \|\xi_1\|_2^2$ almost surely. Further assume that*

$$\begin{aligned}(5.1)\qquad &\mathbb{E}\sup_{C\in\mathscr{C}:P|C\Delta C_0|\le\delta^2}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i(\mathbf{1}_C-\mathbf{1}_{C_0})(X_i)\right| \\ &\vee \mathbb{E}\sup_{C\in\mathscr{C}:P|C\Delta C_0|\le\delta^2}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i(\mathbf{1}_C-\mathbf{1}_{C_0})(X_i)\right| \lesssim \phi_n(\delta),\end{aligned}$$

hold for some ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta$ is nonincreasing. Then $\mathbb{E}_{C_0}P|\widehat{C}_n\Delta C_0| = \mathcal{O}(\delta_n^2)$ holds for any $\delta_n \ge n^{-1/2}\max\{1, \|\xi_1\|_2, \mathbb{E}^{1/8}\max_{1\le i\le n}|\xi_i|^4\}$ such that $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$, where the constant in $\mathcal{O}$ only depends on the constants in (5.1).

PROOF. See Appendix B. □

PROOF OF THEOREM 3.1. By Lemma 5.1, the risk of the least squares estimator

$$\delta_n^2 \equiv \sup_{C_0\in\mathscr{C}}\mathbb{E}_{C_0}P|\widehat{C}_n\Delta C_0| = \sup_{C_0\in\mathscr{C}}\mathbb{E}_{C_0}\int(\mathbf{1}_{\widehat{C}_n}-\mathbf{1}_{C_0})^2\,\mathrm{d}P$$

can be solved by estimating the empirical processes in (5.1). Since the global entropy estimate is translation invariant (i.e., the metric entropy of $\{\mathbf{1}_C - \mathbf{1}_{C_0} : C\in\mathscr{C}\}$ is the same as that of $\{\mathbf{1}_C : C\in\mathscr{C}\}$), by Theorem 2.3, we obtain an estimate for the Rademacher randomized empirical process:

$$\sup_{C_0\in\mathscr{C}}\mathbb{E}\sup_{C\in\mathscr{C}:P|C\Delta C_0|\le\delta_n^2}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i(\mathbf{1}_C-\mathbf{1}_{C_0})(X_i)\right| \lesssim \max\{\delta_n^{1-\alpha}, n^{(\alpha-1)/2(\alpha+1)}\}.$$

It is now easy to see that the choice $\delta_n^2 \asymp n^{-1/(\alpha+1)}$ leads to an upper bound of the above display on the desired order $\sqrt{n}\delta_n^2$. Note that the bound continues to hold when the left-hand side of the above display is replaced with expected supremum without localization over $C\in\mathscr{C}$ for $P|C\Delta C_0| \le \delta_n^2$. The Gaussian randomized empirical process can be handled via the multiplier inequality Lemma C.3 by letting $\psi_n(t) \equiv \psi(t) \equiv t^{\alpha/(\alpha+1)}$, whence

$$\sup_{C_0\in\mathscr{C}}\mathbb{E}\sup_{C\in\mathscr{C}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i(\mathbf{1}_C-\mathbf{1}_{C_0})(X_i)\right| \lesssim n^{(\alpha-1)/2(\alpha+1)},$$

completing the proof. □

5.2. *Proof of Theorem 3.4.* We need the following analogy of Lemma 5.1 before proving Theorem 3.4.

LEMMA 5.2. *Consider the regression model (3.5) and the least squares estimator $\widehat{C}_n$ in (3.6). Further assume that*

$$\mathbb{E}\sup_{C\in\mathscr{C}:P|C\Delta C_0|\le\delta^2}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i(\mathbf{1}_C-\mathbf{1}_{C_0})(X_i)\right| \vee \mathbb{E}\sup_{C\in\mathscr{C}:P|C\Delta C_0|\le\delta^2}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i(\mathbf{1}_{C\cap C_0}-\mathbf{1}_{C_0})(X_i)\right| \lesssim \phi_n(\delta),$$

holds for some ϕ_n such that $\delta\mapsto\phi_n(\delta)/\delta$ is nonincreasing. Then $\mathbb{E}_{C_0}P|\widehat{C}_n\Delta C_0| = \mathcal{O}(\delta_n^2)$ holds for any $\delta_n \ge n^{-1/2}\max\{1, \|\xi_1\|_2, \mathbb{E}^{1/8}\max_{1\le i\le n}|\xi_i|^4\}$ such that $\phi_n(\delta_n)\le\sqrt{n}\delta_n^2$, where the constant in $\mathcal{O}$ only depends on the constants in the above inequality.

PROOF. See Appendix B. □

PROOF OF THEOREM 3.4. The proof follows by Lemma 5.2 and similar arguments as in the proof of Theorem 3.1. □

5.3. *Proof of Theorem 3.5.* We need some further notation to implement this program. For any $g\in\mathcal{G}$, write $f_g(x,y)\equiv\mathbf{1}_{y\ne g(x)}$. Let $\mathcal{G}(\delta)\equiv\{g\in\mathcal{G}:\mathcal{E}_P(g)\le\delta\}$. Let ℓ be the smallest integer such that $r_n^2 2^\ell\ge 1$, and for any $1\le j\le\ell$, let $\mathcal{F}_j\equiv\{f_{g_1}-f_{g_2}: g_1,g_2\in\mathcal{G}(r_n^2 2^j)\}$.

LEMMA 5.3. *Suppose $\mathcal{G}\equiv\{\mathbf{1}_C: C\in\mathscr{C}\}$ satisfies the same entropy condition as in Theorem 3.5. Then*

$$\mathbb{P}\left(\max_{1\le j\le\ell}\frac{\sup_{f\in\mathcal{F}_j}|\mathbb{P}_n(f)-P(f)|}{r_n^2 2^j}\ge c\left(\frac{1}{4}+K\sqrt{\frac{s}{nr_n^2}}+K\frac{s}{nr_n^2}\right)\right)\le K'\exp(-s/K')$$

holds for some constants $K, K'>0$ provided $r_n^2\cdot n^{1/(\alpha+1)}\ge K''$ for a large enough constant $K''>0$ depending on $c>0$ in (3.8) only.

PROOF OF LEMMA 5.3. By Talagrand's concentration inequality (cf. Appendix C), with $\sigma_j^2\equiv\sup_{f\in\mathcal{F}_j}\|f\|_{L_2(P)}^2$,

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}_j}|\mathbb{G}_n(f)|\ge K\left(\mathbb{E}\sup_{f\in\mathcal{F}_j}|\mathbb{G}_n(f)|+\sqrt{\sigma_j^2 s_j}+\frac{s_j}{\sqrt{n}}\right)\right]\le K\exp(-s_j/K).$$

Let $\mathscr{S}\equiv\{S: f_g=\mathbf{1}_S, g\in\mathcal{G}\}$. Note that for $g_1=\mathbf{1}_{C_1}, g_2=\mathbf{1}_{C_2}\in\mathcal{G}$, where $C_1, C_2\in\mathscr{C}$, we have $f_{g_1}=\mathbf{1}_{S_1}$, $f_{g_2}=\mathbf{1}_{S_2}$, and hence

$$P(S_1\Delta S_2)=P(f_{g_1}-f_{g_2})^2\le P(g_1-g_2)^2=P(C_1\Delta C_2).$$

This shows that $\mathcal{N}_I(\varepsilon,\mathscr{S},P)\le\mathcal{N}_I(\varepsilon,\mathscr{C},P)$. Furthermore, for any $g\in\mathcal{G}(r_n^2 2^j)$, let $S\in\mathscr{S}$ be such that $f_g=\mathbf{1}_S$. Then similar to the above display, we have

$$P(S\Delta S_0)\le\|g-g_0\|_{L_2(P)}^2\le c^{-1}r_n^2 2^j,$$

where the last inequality follows from the margin condition. Now by Theorem 2.3, we obtain

$$\mathbb{E}\sup_{f\in\mathcal{F}_j}|\mathbb{G}_n(f)|\lesssim\mathbb{E}\sup_{g\in\mathcal{G}(r_n^2 2^j)}|\mathbb{G}_n(f_g)|\le\mathbb{E}\sup_{S\in\mathscr{S}:P(S\Delta S_0)\le c^{-1}r_n^2 2^j}|\mathbb{G}_n(S)|$$
$$\lesssim\max\{(r_n^2 2^j)^{(1-\alpha)/2}, n^{(\alpha-1)/2(\alpha+1)}\}.$$

On the other hand,

$$\sigma_j^2 \equiv \sup_{f\in\mathcal{F}_j}\|f\|_{L_2(P)}^2 = \sup_{g_1,g_2\in\mathcal{G}(r_n^2 2^j)}\|f_{g_1}-f_{g_2}\|_{L_2(P)}^2$$
$$\le 4\sup_{g\in\mathcal{G}(r_n^2 2^j)}\|g-g_0\|_{L_2(P)}^2 \le 4c^{-1}\sup_{g\in\mathcal{G}(r_n^2 2^j)}\mathcal{E}_P(g) \le 4c^{-1}r_n^2 2^j.$$

This implies that with $s_j = s2^j$,

$$\mathbb{P}\Bigg[\frac{\sup_{f\in\mathcal{F}_j}|\mathbb{P}_n(f)-P(f)|}{r_n^2 2^j}$$
$$\ge K_c r_n^{-2}2^{-j}\big(\max\{n^{-1/2}(r_n^2 2^j)^{(1-\alpha)/2}, n^{-1/(\alpha+1)}\}$$
$$+ n^{-1/2}(r_n^2 2^j)^{1/2}\sqrt{s}2^{j/2} + n^{-1}s2^j\big)\Bigg] \le K\exp(-s2^j/K).$$

Note that

$$r_n^{-2}2^{-j}\big(\max\{n^{-1/2}(r_n^2 2^j)^{(1-\alpha)/2}, n^{-1/(\alpha+1)}\} + n^{-1/2}(r_n^2 2^j)^{1/2}\sqrt{s}2^{j/2} + n^{-1}s2^j\big)$$
$$\le \max\left\{\frac{1}{\sqrt{n}r_n^{\alpha+1}}, \frac{1}{r_n^2 n^{1/(\alpha+1)}}\right\} + \sqrt{\frac{s}{nr_n^2}} + \frac{s}{nr_n^2} \le \frac{c}{4K_c} + \sqrt{\frac{s}{nr_n^2}} + \frac{s}{nr_n^2}$$

under the assumption. Now a union bound leads to the desired claim. □

PROOF OF THEOREM 3.5. Given the estimate in Lemma 5.3, the proof of the theorem closely follows that of [15], Theorem 7.1. We provide some details for the convenience of the reader. On the event

$$E \equiv \left\{\max_{1\le j\le \ell}\frac{\sup_{f\in\mathcal{F}_j}|\mathbb{P}_n(f)-P(f)|}{r_n^2 2^j} \le c\left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right)\right\},$$

we have for any $g\in\mathcal{G}(r_n^2 2^j)\setminus\mathcal{G}(r_n^2 2^{j-1})$ and $g'\in\mathcal{G}(\sigma)$ for some $0<\sigma<r_n^2 2^j$,

$$\mathcal{E}_P(g) = P(f_g - f_{g'}) + \big[P(f_{g'}) - Pf_{g_0}\big] \le P(f_g - f_{g'}) + \sigma$$
$$\le \mathbb{P}_n(f_g - f_{g'}) + \sigma + \sup_{f\in\mathcal{F}_j}\big|(\mathbb{P}_n - P)(f)\big|$$
$$\le \mathcal{E}_{\mathbb{P}_n}(g) + \sigma + c\left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right)r_n^2 2^j$$
$$\le \mathcal{E}_{\mathbb{P}_n}(g) + \sigma + \left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right)2\mathcal{E}_P(g).$$

Since $\sigma>0$ is taken arbitrarily, we see that on the event E, it holds that

$$\frac{\mathcal{E}_{\mathbb{P}_n}(g)}{\mathcal{E}_P(g)} \ge 1 - \left(\frac{1}{2} + 2K\sqrt{\frac{s}{nr_n^2}} + 2K\frac{s}{nr_n^2}\right) \tag{5.2}$$

for all $g\in\mathcal{G}$ such that $\mathcal{E}_P(g)\ge r_n^2$. Furthermore, the above display entails that on the event E, we necessarily have $\mathcal{E}_P(\widehat{g}_n) < r_n^2$ for n large enough. Hence for any $g\in\mathcal{G}(r_n^2 2^j)\setminus\mathcal{G}(r_n^2 2^{j-1})$, we have

$$\mathcal{E}_{\mathbb{P}_n}(g) = \mathbb{P}_n(f_g) - \mathbb{P}_n(f_{\widehat{g}_n}) \le Pf_g - Pf_{\widehat{g}_n} + \sup_{f\in\mathcal{F}_j}\big|(\mathbb{P}_n - P)(f)\big|$$
$$\le \mathcal{E}_P(g) + \left(\frac{1}{4} + K\sqrt{\frac{s}{nr_n^2}} + K\frac{s}{nr_n^2}\right)2\mathcal{E}_P(g).$$

This entails that

$$\frac{\mathcal{E}_{\mathbb{P}_n}(g)}{\mathcal{E}_P(g)} \le 1 + \Big(\frac{1}{2} + 2K\sqrt{\frac{s}{nr_n^2}} + 2K\frac{s}{nr_n^2}\Big). \tag{5.3}$$

The proof of the claim is complete by combining (5.2)–(5.3) along with Lemma 5.3. □

5.4. *Proof of Theorem 3.6.*

LEMMA 5.4. *It holds for $C > 0$ large enough that*

$$\mathbb{E}_{f_0}\|\widehat{f}_n - f_0\|_{L_2(P)}^2 \le \mathbb{E}_{f_0}\|\widehat{f}_n - f_0\|_{L_2(P)}^2 \mathbf{1}_{\|\widehat{f}_n - f_0\|_\infty \le C\sqrt{\log n}} + \mathcal{O}(n^{-1}).$$

The $\mathcal{O}$ term is uniform in $f_0 \in \mathcal{M}_d \cap L_\infty(1)$.

PROOF. Fix $f_0 \in \mathcal{M}_d \cap L_\infty(1)$. By [20], Lemma 10, supplement, $\sup_{x\in[0,1]^d}(\widehat{f}_n - f_0)(x) \le \max_{1\le i\le n} Y_i + \|f_0\|_\infty \le 2 + \max_{1\le i\le n}\xi_i$. Hence with $Z_n \equiv \|\widehat{f}_n - f_0\|_\infty$, for u large, $\mathbb{P}(Z_n > u\sqrt{\log n}) \le e^{-cu^2\log n}$ for some $c > 0$. In particular, $\mathbb{E}Z_n^4 \lesssim \log^2 n$. Now the claim of the lemma follows by noting that

$$\begin{aligned}
&\mathbb{E}_{f_0}\|\widehat{f}_n - f_0\|_{L_2(P)}^2 \mathbf{1}_{\|\widehat{f}_n - f_0\|_\infty > C\sqrt{\log n}} \\
&\quad\le \mathbb{E}Z_n^2 \mathbf{1}_{Z_n > C\sqrt{\log n}} \\
&\quad\le \sqrt{\mathbb{E}Z_n^4}\cdot\sqrt{\mathbb{P}(Z_n > C\sqrt{\log n})} \le \log n \cdot e^{-cC^2\log n/2} = \mathcal{O}(n^{-1})
\end{aligned}$$

for $C > 0$ large. □

Recall that $B \subset \mathbb{R}^d$ is a lower (resp., upper) set if and only if for all $x \in B$, $y \in \mathbb{R}^d$ with $y_i \le x_i$ (resp., $y_i \ge x_i$), $i = 1, \dots, d$, we have $y \in B$. Let $\mathcal{LL}_d$ be the collection of all upper and lower sets in $\mathbb{R}^d$ and $\mathcal{L}_d = \{B \cap [0,1]^d : B \in \mathcal{LL}_d\}$.

PROOF OF THEOREM 3.6. First, consider $d \ge 3$. By Lemma 5.4, we only need to compute an upper bound for $\mathbb{E}_{f_0}\|\widehat{f}_n - f_0\|_{L_2(P)}^2 \mathbf{1}_{\|\widehat{f}_n - f_0\|_\infty \le C\sqrt{\log n}} \le \bar{r}_n^2$. Similar to the proof of Lemma 5.1, this can be done by evaluating the size of two empirical processes

$$\begin{aligned}
&\mathbb{E}\sup_{\substack{f\in\mathcal{M}_d\cap L_\infty(C\sqrt{\log n}):\\ \|f-f_0\|_{L_2(P)}\le\bar{r}_n}} \big|\mathbb{G}_n(\xi(f - f_0))\big| \\
&\vee \mathbb{E}\sup_{\substack{f\in\mathcal{M}_d\cap L_\infty(C\sqrt{\log n}):\\ \|f-f_0\|_{L_2(P)}\le\bar{r}_n}} \big|\mathbb{G}_n\big((f - f_0)^2\big)\big| \lesssim \sqrt{n}\bar{r}_n^2.
\end{aligned} \tag{5.4}$$

Note that for any $f \in \mathcal{M}_d$,

$$\begin{aligned}
\big|(\mathbb{P}_n - P)f\big| &= \big|\mathbb{E}_{\mathbb{P}_n} f(X) - \mathbb{E}_P f(X)\big| \\
&\le \big|\mathbb{E}_{\mathbb{P}_n} f_+(X) - \mathbb{E}_P f_+(X)\big| + \big|\mathbb{E}_{\mathbb{P}_n} f_-(X) - \mathbb{E}_P f_-(X)\big| \\
&\le \left|\int_0^\infty \big(\mathbb{P}_{\mathbb{P}_n}\big(f_+(X) > t\big) - \mathbb{P}_P\big(f_+(X) > t\big)\big)\,\mathrm{d}t\right| \\
&\quad + \left|\int_0^\infty \big(\mathbb{P}_{\mathbb{P}_n}\big(f_-(X) > t\big) - \mathbb{P}_P\big(f_-(X) > t\big)\big)\,\mathrm{d}t\right| \\
&\le 2\|f\|_\infty \sup_{C\in\mathcal{L}_d} \big|(\mathbb{P}_n - P)(C)\big|.
\end{aligned}$$

Here, $\mathcal{L}_d$ is the class of all upper and lower sets in $[0,1]^d$. The last inequality follows since for any $f \in \mathcal{M}_d$, $\{f_+(x) > t\} = \{f(x) \vee 0 > t\} \in \mathcal{L}_d$ and $\{f_-(x) > t\} = \{-(f(x) \wedge 0) > t\} = \{f(x) \wedge 0 < -t\} \in \mathcal{L}_d$. Hence by [13], Theorem 8.22, we may apply Theorem 2.3 with $\alpha = d-1$ to see that

$$\mathbb{E} \sup_{f \in \mathcal{M}_d \cap L_\infty(C\sqrt{\log n})} |\mathbb{G}_n(f - f_0)| \lesssim \sqrt{\log n} \cdot \mathbb{E} \sup_{C \in \mathcal{L}_d} |\mathbb{G}_n(C)| + 1 \lesssim \sqrt{\log n} \cdot n^{\frac{d-2}{2d}}.$$

Using the multiplier inequality (cf. Lemma C.3) and contraction principle for empirical processes (cf. [20], Lemma 6, supplement), we may further bound the two empirical processes in (5.4) by

$$\mathbb{E} \sup_{\substack{f \in \mathcal{M}_d \cap L_\infty(C\sqrt{\log n}): \\ \|f - f_0\|_{L_2(P)} \le \bar{r}_n}} |\mathbb{G}_n(\xi(f - f_0))|$$

$$\vee\, \mathbb{E} \sup_{\substack{f \in \mathcal{M}_d \cap L_\infty(C\sqrt{\log n}): \\ \|f - f_0\|_{L_2(P)} \le \bar{r}_n}} |\mathbb{G}_n((f - f_0)^2)| \lesssim n^{\frac{d-2}{2d}} \log n.$$

Solving (5.4) using the above inequality, we obtain the rate $\bar{r}_n$ for $d \ge 3$.

For $d = 2$, we may estimate the empirical process with an additional $\log n$:

$$\mathbb{E} \sup_{f \in \mathcal{M}_d \cap L_\infty(C\sqrt{\log n})} |\mathbb{G}_n(f - f_0)| \lesssim \log^{3/2} n,$$

and the rate can be obtained similarly as above. □

REMARK 5.5. The proof for the analogue of Theorem 3.6 in [20], that is, [20], Theorem 4, uses a completely different strategy. A rough argument is as follows. Han et al. [20] first consider the problem $f_0 = 0$, where it is shown in Proposition 9 therein that for $\delta_n > 0$ not too small,

$$\mathbb{E} \sup_{f \in \mathcal{M}_d \cap L_\infty(\sqrt{C \log n}): \|f\|_{L_2(P)} \le \delta_n} |\mathbb{G}_n(f)| \lesssim \delta_n \cdot n^{1/2 - 1/d} \log^\gamma n. \tag{5.5}$$

Then by a simple triangle inequality, if $d \ge 2$,

$$\begin{aligned} &\mathbb{E} \sup_{\substack{f \in \mathcal{M}_d \cap L_\infty(C\sqrt{\log n}): \\ \|f - f_0\|_{L_2(P)} \le \delta_n}} |\mathbb{G}_n(f - f_0)| \\ &\le \mathbb{E} \sup_{\substack{f \in \mathcal{M}_d \cap L_\infty(C\sqrt{\log n}): \\ \|f\|_{L_2(P)} \le \delta_n + \|f_0\|_\infty}} |\mathbb{G}_n(f)| + \mathbb{E}|\mathbb{G}_n(f_0)| \lesssim (\delta_n + \|f_0\|_\infty) n^{1/2 - 1/d} \log^\gamma n. \end{aligned} \tag{5.6}$$

Using the above inequality and (5.4), we obtain $\bar{r}_n^2 \lesssim n^{-1/d}$ up to logarithmic factors. It is clear from the sketch here that the property of isotonic regression functions is only used in (5.5) where the problem is $f_0 = 0$. The proof for general $f_0 \in L_\infty(1)$ in (5.6) is not very informative in the sense that the method of (5.6) is valid for any problem as long as one could solve the risk problem (= empirical process problem (5.5)) for one particular f_0. In contrast, the proof of Theorem 3.6 here shows that it is the complexity of the class of upper and lower sets $\mathcal{L}_d$ that leads to the minimax rate of convergence for the multiple isotonic LSE.

REMARK 5.6. It is possible to adapt the present approach to the problem of multivariate convex regression. The major difficulty here is to understand the boundary behavior for the convex LSE $\widehat{f}_n^{\mathrm{cvx}}$. In particular, if we can prove that the convex LSE $\widehat{f}_n^{\mathrm{cvx}}$ satisfies $\|\widehat{f}_n^{\mathrm{cvx}} - f_0\|_\infty = \mathcal{O}_{\mathbf{P}}(L_n)$ for some slowly growing L_n (in similar spirit to Lemma 5.4 for

the isotonic LSE), then using similar arguments as in the proof of Theorem 3.6, we may conclude $\|\widehat{f}_n^{\mathrm{cvx}} - f_0\|_{L_2(P)} = \mathcal{O}_{\mathbf{P}}(n^{-1/(d+1)}\bar{L}_n)$ for some slowly growing $\bar{L}_n$. Interestingly, recently [27] proved that under a fixed lattice design in a polytopal domain, the rate of the convex LSE can be improved to $n^{-2/(d+4)}$ for $d \le 4$ and $n^{-1/d}$ for $d \ge 5$ (up to logarithmic factors) and these rates cannot be further improved in the worst case. These improved rates are due to reduced complexity of the class of convex functions on polytopal domains than those on smooth domains. Similar rates are obtained in [27] for bounded and Lipschitz convex LSEs under random designs. It remains open whether the unconstrained convex LSEs attain these rates over polytopal and more general domains under random designs, where the boundary behavior of $\widehat{f}_n^{\mathrm{cvx}}$ may play a crucial role.

5.5. *Proof of Theorem 3.7.*

PROOF. We only provide the proof for the most difficult case $-1/d < s < 0$; the other cases are similar or simpler.

(*Case* 1: $d \ge 4$). We will relate the squared Hellinger distance $h^2(p_0, \widehat{p}_n)$ to that of the expected supremum of empirical process over the class of convex sets. The proof for this reduction is largely inspired by the idea of [6].

Using same arguments as in Step 1 of the proof of [11], Theorem 4.3, we may assume without loss of generality that $p_0 \in \mathcal{P}_{s,M/2}$ and $\widehat{p}_n$ belongs to

$$\mathcal{P}_{s,M} \equiv \Big\{ p \in \mathcal{P}_s : \sup_{x \in \mathbb{R}^d} p(x) \le M, \inf_{x \in B(0,1)} p(x) \ge 1/M \Big\}$$

for some large M with high probability. By the proof of [21], Lemma F.7 (especially (F.3) therein),

$$\sup_{p \in \mathcal{P}_{s,M}} p(x) \le C_M \big(1 + \|x\|\big)^{1/s} \le C_{M,d} \bigg(1 + \prod_{k=1}^{d} |x_k|^{1/d} \bigg)^{1/s}. \tag{5.7}$$

Furthermore, it is not hard to see that $\widehat{p}_n$ is supported in the convex hull of $X_1, \dots, X_n$. By (5.7), $\kappa_q \equiv \mathbb{E}_{X \sim p_0}(1/p_0(X))^q = \int p_0^{1-q} < \infty$ for $q \in (0, 1 + sd)$. This means that $\mathbb{E}_{X \sim p_0} \max_{1 \le i \le n} (1/p_0(X_i))^q \le n \cdot \kappa_q$, so $\log \max_i (1/p_0(X_i)) \le C_1 \log n$ with high probability for large $C_1 > 0$. Hence with $c_n \equiv n^{-C_1}$, $X_1, \dots, X_n \in \{p_0 \ge c_n\}$ with high probability. Let $\widetilde{p}_n \equiv (\widehat{p}_n \vee c_n)\mathbf{1}_{p_0 \ge c_n} / \int (\widehat{p}_n \vee c_n)\mathbf{1}_{p_0 \ge c_n}$. Then with $b_n \equiv \int (\widehat{p}_n \vee c_n)\mathbf{1}_{p_0 \ge c_n}$, it follows that with high probability

$$b_n^{-1} = b_n^{-1} \int_{p_0 \ge c_n} \widehat{p}_n \le b_n^{-1} \int_{p_0 \ge c_n} (\widehat{p}_n \vee c_n) = 1,$$

$$b_n - 1 = \int_{p_0 \ge c_n} |(\widehat{p}_n \vee c_n) - \widehat{p}_n| \le c_n |\{p_0 \ge c_n\}| \lesssim c_n^{1+sd}.$$

The last inequality in the second line of the above display follows as $\{x : \|x\| > (c_n/C_M)^s\} \subset \{x : p_0(x) < c_n\}$ by (5.7) and, therefore, $|\{p_0 \ge c_n\}| \subset |\{x : \|x\| \le (c_n/C_M)^s\}| \asymp_{d,M} c_n^{sd}$. As $s > -1/d$, by choosing $C_1 > 0$ large, we have $0 \le b_n - 1 \le \mathcal{O}(n^{-1})$. This implies with high probability,

$$\begin{aligned} h^2(\widehat{p}_n, \widetilde{p}_n) &\lesssim \int |\widehat{p}_n - \widetilde{p}_n| = \int |\widehat{p}_n - (\widehat{p}_n \vee c_n)\mathbf{1}_{p_0 \ge c_n}/b_n| \\ &\le |1 - b_n^{-1}| \int \widehat{p}_n + b_n^{-1} \int_{p_0 \ge c_n} |\widehat{p}_n - (\widehat{p}_n \vee c_n)| \\ &= |1 - b_n^{-1}| + b_n^{-1}(b_n - 1) = \mathcal{O}(n^{-1}). \end{aligned} \tag{5.8}$$

On the other hand, let $\widetilde{p}_0 \equiv p_0 \mathbf{1}_{p_0 \geq c_n} / \int p_0 \mathbf{1}_{p_0 \geq c_n}$ and $b_0 \equiv \int p_0 \mathbf{1}_{p_0 \geq c_n}$. As $s > -1/d$, with $q \in (0, 1 + sd)$,

$$\begin{aligned} b_0 &= \int p_0 \mathbf{1}_{p_0 \geq c_n} \leq 1, \\ 1 - b_0 &= \int p_0 \mathbf{1}_{p_0 < c_n} \lesssim c_n^q \int (1 + \|x\|)^{(1-q)/s} \, \mathrm{d}x = \mathcal{O}(c_n^q). \end{aligned}$$

This means by choosing $C_1 > 0$ large, we have $0 \leq 1 - b_0 \leq \mathcal{O}(n^{-1})$. Hence

$$\begin{aligned} h^2(p_0, \widetilde{p}_0) &\asymp \int p_0 \mathbf{1}_{p_0 < c_n} + \int (\sqrt{p_0} - \sqrt{p_0/b_0})^2 \mathbf{1}_{p_0 \geq c_n} \\ &= \mathcal{O}(c_n^q) + (1 - b_0^{-1/2})^2 \int p_0 \mathbf{1}_{p_0 \geq c_n} = \mathcal{O}(n^{-1}), \end{aligned} \tag{5.9}$$

and

$$\begin{aligned} &\left| \int p_0 \log p_0 - \int \widetilde{p}_0 \log \widetilde{p}_0 \right| \\ &\leq \left| \int p_0 \log p_0 \mathbf{1}_{p_0 < c_n} \right| + \left| \int_{p_0 \geq c_n} p_0 \log p_0 - (p_0/b_0) \log(p_0/b_0) \right| \\ &\leq \mathcal{O}(c_n^q) + |1 - b_0^{-1}| \int_{p_0 \geq c_n} |p_0 \log p_0| + |\log b_0 / b_0| \int_{p_0 \geq c_n} p_0 \\ &= \mathcal{O}(n^{-1}), \end{aligned} \tag{5.10}$$

and

$$\int |p_0 - \widetilde{p}_0| \leq \int p_0 \mathbf{1}_{p_0 < c_n} + |1 - b_0^{-1}| \int_{p_0 \geq c_n} p_0 = \mathcal{O}(n^{-1}). \tag{5.11}$$

Now by the integrability $\mathbb{E}_{P_0} \log^2 p_0 < \infty$, with P_0, $\widetilde{P}_0$ denoting the distributions of p_0, $\widetilde{p}_0$, it follows that with high probability,

$$\begin{aligned} h^2(p_0, \widehat{p}_n) &\lesssim h^2(p_0, \widetilde{p}_0) + h^2(\widetilde{p}_0, \widetilde{p}_n) + h^2(\widetilde{p}_n, \widehat{p}_n) \\ &\leq \mathbb{E}_{\widetilde{P}_0} \log(\widetilde{p}_0 / \widetilde{p}_n) + \mathcal{O}(n^{-1}) \text{(by (5.8) and (5.9))} \\ &\leq \mathbb{E}_{P_0} \log p_0 - \mathbb{E}_{\widetilde{P}_0} \log \widetilde{p}_n + \mathcal{O}(n^{-1}) \text{(by (5.10))} \\ &\leq \mathbb{E}_{\mathbb{P}_n} \log p_0 - \mathbb{E}_{\widetilde{P}_0} \log \widetilde{p}_n + \mathcal{O}_{\mathbf{P}}(n^{-1/2}) \text{(by integrability of } \log p_0) \\ &\leq \mathbb{E}_{\mathbb{P}_n} \log \widehat{p}_n - \mathbb{E}_{\widetilde{P}_0} \log \widetilde{p}_n + \mathcal{O}_{\mathbf{P}}(n^{-1/2}) \text{(as } \widehat{p}_n \text{ is the MLE)} \\ &\leq \mathbb{E}_{\mathbb{P}_n} \log[b_n^{-1}(\widehat{p}_n \vee c_n) \mathbf{1}_{p_0 \geq c_n}] - \mathbb{E}_{\widetilde{P}_0} \log \widetilde{p}_n + \log b_n + \mathcal{O}_{\mathbf{P}}(n^{-1/2}) \\ &\leq |(\mathbb{P}_n - \widetilde{P}_0) \log \widetilde{p}_n| + \mathcal{O}_{\mathbf{P}}(n^{-1/2} \vee |1 - b_n|) \\ &\leq \left| \int_0^\infty (\mathbb{P}_{\mathbb{P}_n}((\log \widetilde{p}_n)_+(X) \geq t) - \mathbb{P}_{\widetilde{P}_0}((\log \widetilde{p}_n)_+(X) \geq t)) \, \mathrm{d}t \right| \\ &\quad + \left| \int_0^\infty (\mathbb{P}_{\mathbb{P}_n}((\log \widetilde{p}_n)_-(X) \leq t) - \mathbb{P}_{\widetilde{P}_0}((\log \widetilde{p}_n)_-(X) \leq t)) \, \mathrm{d}t \right| \\ &\quad + \mathcal{O}_{\mathbf{P}}(n^{-1/2}) \\ &\overset{(*)}{\lesssim} \log n \cdot \mathbb{E} \sup_{C \in \mathscr{C}_d} |(\mathbb{P}_n - \widetilde{P}_0)(C)| + \mathcal{O}_{\mathbf{P}}(n^{-1/2}) \end{aligned} \tag{5.12}$$

$$\le \log n \cdot \Big[\mathbb{E}\sup_{C\in\mathscr{C}_d}|(\mathbb{P}_n - P_0)(C)| + \sup_{C\in\mathscr{C}_d}\int_C |p_0 - \widetilde{p}_0|\Big] + \mathcal{O}_{\mathbf{P}}(n^{-1/2})$$

$$\overset{(**)}{\le} \log n \cdot \mathbb{E}\sup_{C\in\mathscr{C}_d}|(\mathbb{P}_n - P_0)(C)| + \mathcal{O}_{\mathbf{P}}(n^{-1/2}).$$

Here, $\mathscr{C}_d$ is the set of all convex bodies in $\mathbb{R}^d$. The inequality $(*)$ follows as for any s-concave density p, $\{(\log p(x))_+ \ge t\} = \{\log p(x) \vee 0 \ge t\} = \{p(x) \vee 1 \ge e^t\} = \{\varphi(x) \wedge 1 \le e^{st}\}$ and $\{(\log p(x))_- \le t\} = \{-(\log p(x) \wedge 0) \le t\} = \{p(x) \wedge 1 \ge e^{-t}\} = \{\varphi(x) \vee 1 \le e^{-ts}\}$ are convex sets, and $-C_1 \log n \le \log c_n \le \log \widetilde{p}_n \le \log(M)$ for n large. The inequality $(**)$ follows from (5.11).

Hence we only need to bound the entropy $\mathcal{N}_I(\varepsilon, \mathscr{C}_d, P_0)$. To this end, for a multiindex $\ell = (\ell_1, \ldots, \ell_d) \in \mathbb{Z}^d_{\ge 0}$, let $I_\ell \equiv \prod_{k=1}^d [2^{\ell_k} - 1, 2^{\ell_k+1} - 1]$. Then $|I_\ell| \asymp 2^{\sum_k \ell_k}$. Let $\{(A_{\ell,j}, B_{\ell,j}) : 1 \le j \le N_\ell\}$ be an ε_ℓ-bracket for $\{C|_{I_\ell} : C \in \mathscr{C}\}$ under the Lebesgue measure. By [13], Theorem 8.25, we have $\log N_\ell \lesssim_d (|I_\ell|^{-1}\varepsilon_\ell)^{(1-d)/2}$. Let $\varepsilon_\ell = a_\ell \cdot \varepsilon$, where $a_\ell \equiv |I_\ell|^{(1+\delta)}$ for some $\delta > 0$ such that $1 < 1 + \delta < (-sd)^{-1} \in (1, \infty)$. Then $\{(\sum_\ell \mathbf{1}_{A_{\ell,j_\ell}}\mathbf{1}_{I_\ell}, \sum_\ell \mathbf{1}_{B_{\ell,j_\ell}}\mathbf{1}_{I_\ell}) : 1 \le j_\ell \le N_\ell, \ell \in \mathbb{Z}^d_{\ge 0}\}$ forms a bracket for $\mathscr{C} \cap \mathbb{R}^d_{\ge 0}$ with P_0-size

$$\Big|P_0\Big(\sum_\ell \mathbf{1}_{B_{\ell,j_\ell}}\mathbf{1}_{I_\ell} - \sum_\ell \mathbf{1}_{A_{\ell,j_\ell}}\mathbf{1}_{I_\ell}\Big)\Big| \le \sum_\ell \varepsilon_\ell \sup_{x\in I_\ell} p_0(x) \lesssim \varepsilon \sum_\ell a_\ell |I_\ell|^{-(-sd)^{-1}} \lesssim \varepsilon.$$

The logarithm of the number of the brackets can be bounded by

$$C_d \sum_\ell |I_\ell|^{(d-1)/2}\varepsilon_\ell^{(1-d)/2} \lesssim \varepsilon^{(1-d)/2}\sum_\ell \big(a_\ell |I_\ell|^{-1}\big)^{(1-d)/2} \lesssim \varepsilon^{(1-d)/2}.$$

Other quadrants can be handled similarly. This means that $\log \mathcal{N}_I(\varepsilon, \mathscr{C}, P_0) \lesssim \varepsilon^{(1-d)/2}$, and hence Theorem 2.3 applies to (5.12).

(*Case* 2: $d = 3$). The situation for $d = 3$ is similar; but with an additional $\log n$ term in the estimate for the empirical process $\mathbb{E}\sup_{C\in\mathscr{C}_3}|\mathbb{G}_n(C)| \lesssim \log n$ (cf. Remark 2.4) and, therefore, the rate in squared Hellinger comes with an additional $\log n$.

(*Case* 3: $d = 2$). We employ an idea in [11] in $d = 1$, which first calculates the L_2 entropy of the class of bounded s-concave functions on $[0, 1]$ by discretization of the range of the underlying convex functions until a prescribed L_2 error is reached at the level of s-concave densities, and then use the integrability of $\mathcal{P}_{s,M}$ to extend the brackets from $[0, 1]$ to $\mathbb{R}$. Our arguments below substantially simplify those presented in [11]. A similar idea is exploited in [23] in $d = 2, 3$ in the context of log-concave densities, with further technicalities due to the unknown shapes of domain at the truncated levels in dimensions 2 and 3 (for $d = 1$ they are simply intervals).

Let $\widetilde{\mathcal{P}}_s(I, B) \equiv \{p \text{ is s-concave on } I \subset \mathbb{R}^2 : 0 \le p(x) \le B, \forall x \in I\}$. We write $\widetilde{\mathcal{P}}_s = \widetilde{\mathcal{P}}_s([0, 1]^2, 1)$ for simplicity. We claim that for $s > -1$,

$$\log \mathcal{N}_{[\cdot]}(\varepsilon, \widetilde{\mathcal{P}}_s, L_2) \lesssim_s \varepsilon^{-1}\log(1/\varepsilon). \tag{5.13}$$

Fix $\varepsilon > 0$. Let $y_k \equiv 2^k$, $1 \le k \le k_0$, where k_0 is the smallest integer such that $y_{k_0}^{1/s} \le \varepsilon$, that is, $k_0 = \lceil \log_2((1/\varepsilon)^{-s})\rceil$. Let $\{(A_j, B_j) : A_j \supset B_j\}_{j=1}^{N_1}$ be an ε_0^2-bracket for all convex sets in $[0, 1]^2$ under the Lebesgue measure. By [13], Theorem 8.25, Corollary 8.26, we have $\log N_1 \lesssim \varepsilon_0^{-1}$. By [23], Proposition 4, supplement, for each $j = 1, \ldots, N_1$, and $k = 1, \ldots, k_0$, we may find a lower $\varepsilon_{j,k}$-bracket $\{\underline{f}_{j,k,m} : 1 \le m \le \underline{N}_{j,k}\}$ in L_2 (resp., upper $\varepsilon_{j,k}$-bracket $\{\bar{f}_{j,k,m} : 1 \le m \le \bar{N}_{j,k}\}$ in L_2) for nonnegative convex functions defined on B_j with an upper bound 2^k, such that $\log(\bar{N}_{j,k} \vee \underline{N}_{j,k}) \lesssim (2^k/\varepsilon_{j,k})\log(2^k/\varepsilon_{j,k})$.

For any $p \in \widetilde{\mathcal{P}}_s$, let $\varphi = p^s$ be the underlying convex function. Let $C_k \equiv \{\varphi \le y_k\}$. Let (A_{j_k}, B_{j_k}), $A_{j_k} \supset B_{j_k}$ be a bracket for $A_{j_k} \supset C_k \supset B_{j_k}$, and let $\underline{f}_{j_k,k,m}$ (resp. $\bar{f}_{j_k,k,m}$) be a lower (resp., upper) bracket for $\varphi|_{B_{j_k}}$.

Let $A_{j_0} \equiv \varnothing$ and $y_0 \equiv 1$. Consider an upper bracket for p of form

$$\sum_{k=1}^{k_0} \big[(\underline{f}_{j_k,k,m} \vee y_{k-1}) \mathbf{1}_{B_{j_k} \setminus A_{j_{k-1}}}\big]^{1/s} + \sum_{k=1}^{k_0} \big[(y_{k-1})^{1/s} \wedge 1\big] \mathbf{1}_{A_{j_k} \setminus B_{j_k}} + \varepsilon \mathbf{1}_{[0,1]^2 \setminus A_{j_{k_0}}},$$

and a lower bracket of p of form $\sum_{k=1}^{k_0} [(\bar{f}_{j_k,k,m} \wedge y_k) \mathbf{1}_{B_{j_k} \setminus A_{j_{k-1}}}]^{1/s}$. For the choice $\varepsilon_0 \equiv \varepsilon$ and $\varepsilon_{j,k} \equiv \varepsilon \cdot 2^{2k}$, this bracket has squared L_2 size bounded, up to a constant depending only on s, by

$$\sum_{k=1}^{k_0} \varepsilon_{j_k,k}^2 \cdot (2^k)^{2(1/s-1)} + \varepsilon_0^2 \cdot \sum_{k=1}^{k_0} \big[(y_{k-1})^{1/s} \wedge 1\big] + \varepsilon^2 \lesssim \varepsilon^2.$$

The logarithm of the total number of brackets can be bounded by

$$\log\left[\prod_{k=1}^{k_0} N_1^2 \bar{N}_{j_k,k} \underline{N}_{j_k,k}\right] \lesssim \sum_{k=1}^{k_0} \left(\varepsilon_0^{-1} + \frac{2^k}{\varepsilon_{j_k,k}} \log\left(\frac{2^k}{\varepsilon_{j_k,k}}\right)\right) \lesssim \varepsilon^{-1} \log(1/\varepsilon),$$

proving the claim (5.13). Let I_ℓ be the same as in the proof for $d \ge 4$. By rescaling, it follows that

$$\log \mathcal{N}_{[\cdot]}\big(\varepsilon, \widetilde{\mathcal{P}}_s(I_\ell, B), L_2\big) \lesssim \frac{(B^2|I_\ell|)^{1/2}}{\varepsilon} \log\left(\frac{(B^2|I_\ell|)^{1/2}}{\varepsilon}\right).$$

By (5.7), on I_ℓ, $\sup_{x \in I_\ell} \sup_{p \in \mathcal{P}_{s,M}} p(x) \le |I_\ell|^{1/2s}$. Let $b_\ell = |I_\ell|^{-\delta'}$ for some $\delta' \in (0, (-1/s - 1)/2)$, and $\{\underline{f}_{j,\ell}, \bar{f}_{j,\ell} : 1 \le j \le N_\ell\}$ be a $b_\ell \varepsilon$-bracket for $\mathcal{P}_{s,M}|_{I_\ell}$ under L_2. A global bracket for $\mathcal{P}_{s,M}$ can be obtained by assembling these local brackets for all(= four) quadrants, with squared L_2-size at most $\varepsilon^2 \sum_\ell b_\ell^2 \lesssim \varepsilon^2$, and the logarithm of the number of brackets is

$$\sum_\ell \log N_\ell \lesssim \sum_\ell \frac{|I_\ell|^{(1/s+1)/2}}{b_\ell \varepsilon} \log\left(\frac{|I_\ell|^{(1/s+1)/2}}{b_\ell \varepsilon}\right) \lesssim \varepsilon^{-1} \log(1/\varepsilon).$$

Hence for $s > -1/2$, $\log \mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{P}_{s,M}, h) = \log \mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{P}_{2s,M}, L_2) \lesssim \varepsilon^{-1} \log(1/\varepsilon)$. The rest of the proof is a standard computation of the size of the localized empirical process via Hellinger bracketing numbers (cf. [47], Theorem 3.4.4), so we omit the details. □

Acknowledgments. The major part of this work (materials presented before Section 3.3 and their proofs) is based on Chapter 4 of the author's University of Washington Ph.D. thesis in 2018. The author would like to thank Jon Wellner and Cun-Hui Zhang for helpful discussion and encouragements. He would also like to thank four referees, an Associate Editor and the Editor for their very helpful comments and suggestions that significantly improved the quality of the paper.

Funding. The author's research is supported in part by NSF Grant DMS-1916221.

SUPPLEMENTARY MATERIAL

Supplement: Additional proofs and results. (DOI: 10.1214/21-AOS2049SUPP; .pdf). In the supplement, we provide additional proofs and results.

REFERENCES

[1] BARAUD, Y. (2016). Bounding the expectation of the supremum of an empirical process over a (weak) VC-major class. *Electron. J. Stat.* **10** 1709–1728. MR3522658 https://doi.org/10.1214/15-EJS1055

[2] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028 https://doi.org/10.1007/s004400050210

[3] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129 https://doi.org/10.1007/BF00532480

[4] BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719 https://doi.org/10.1007/BF01199316

[5] BRUNEL, V.-E. (2013). Adaptive estimation of convex polytopes and convex sets from noisy data. *Electron. J. Stat.* **7** 1301–1327. MR3063609 https://doi.org/10.1214/13-EJS804

[6] CARPENTER, T., DIAKONIKOLAS, I., SIDIROPOULOS, A. and STEWART, A. (2018). Near-optimal sample complexity bounds for maximum likelihood estimation of multivariate log-concave densities. In *Conference on Learning Theory* 1–29.

[7] CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982 https://doi.org/10.1214/14-AOS1254

[8] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072–1100. MR3706788 https://doi.org/10.3150/16-BEJ865

[9] DAGAN, Y. and KUR, G. (2019). The log-concave maximum likelihood estimator is optimal in high dimensions. Preprint. Available at arXiv:1903.05315.

[10] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

[11] DOSS, C. R. and WELLNER, J. A. (2016). Global rates of convergence of the MLEs of log-concave and s-concave densities. *Ann. Statist.* **44** 954–981. MR3485950 https://doi.org/10.1214/15-AOS1394

[12] DUDLEY, R. M. (1982). Empirical and Poisson processes on classes of sets or functions too large for central limit theorems. *Z. Wahrsch. Verw. Gebiete* **61** 355–368. MR0679680 https://doi.org/10.1007/BF00539835

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

[14] GAO, F. and WELLNER, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *J. Multivariate Anal.* **98** 1751–1764. MR2392431 https://doi.org/10.1016/j.jmva.2006.09.003

[15] GINÉ, E. and KOLTCHINSKII, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143–1216. MR2243881 https://doi.org/10.1214/009117906000000070

[16] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

[17] GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York. MR0599175

[18] GUNTUBOYINA, A. (2012). Optimal rates of convergence for convex set estimation from support functions. *Ann. Statist.* **40** 385–411. MR3014311 https://doi.org/10.1214/11-AOS959

[19] HAN, Q. (2021). Supplement to "Set structured global empirical risk minimizers are rate optimal in general dimensions." https://doi.org/10.1214/21-AOS2049SUPP.

[20] HAN, Q., WANG, T., CHATTERJEE, S. and SAMWORTH, R. J. (2019). Isotonic regression in general dimensions. *Ann. Statist.* **47** 2440–2471. MR3988762 https://doi.org/10.1214/18-AOS1753

[21] HAN, Q. and WELLNER, J. A. (2016). Approximation and estimation of s-concave densities via Rényi divergences. *Ann. Statist.* **44** 1332–1359. MR3485962 https://doi.org/10.1214/15-AOS1408

[22] HAN, Q. and WELLNER, J. A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *Ann. Statist.* **47** 2286–2319. MR3953452 https://doi.org/10.1214/18-AOS1748

[23] KIM, A. K. H. and SAMWORTH, R. J. (2016). Global rates of convergence in log-concave density estimation. *Ann. Statist.* **44** 2756–2779. MR3576560 https://doi.org/10.1214/16-AOS1480

[24] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442 https://doi.org/10.1214/009053606000001019

[25] KOROSTELËV, A. P. and TSYBAKOV, A. B. (1992). Asymptotically minimax image reconstruction problems. In *Topics in Nonparametric Estimation*. *Adv. Soviet Math.* **12** 45–86. Amer. Math. Soc., Providence, RI. MR1191691 https://doi.org/10.1137/1136015

[26] KOROSTELËV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. *Lecture Notes in Statistics* **82**. Springer, New York. MR1226450 https://doi.org/10.1007/978-1-4612-2712-0

[27] KUR, G., GAO, F., GUNTUBOYINA, A. and SEN, B. (2020). Convex regression in multidimensions: Suboptimality of least squares estimators. Preprint. Available at arXiv:2006.02044.
[28] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381
[29] LECUÉ, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* **35** 1698–1721. MR2351102 https://doi.org/10.1214/009053607000000055
[30] MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524. MR1332579 https://doi.org/10.1214/aos/1176324533
[31] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240
[32] MASSART, P. and NÉDÉLEC, É. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366. MR2291502 https://doi.org/10.1214/009053606000000786
[33] OSSIANDER, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15** 897–919. MR0893905
[34] POLLARD, D. (2002). Maximal inequalities via bracketing with adaptive truncation *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 1039–1052. MR1955351 https://doi.org/10.1016/S0246-0203(02)01138-X
[35] ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR0961262
[36] SAUMARD, A. (2012). Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Stat.* **6** 579–655. MR2988421 https://doi.org/10.1214/12-EJS679
[37] SEREGIN, A. and WELLNER, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Ann. Statist.* **38** 3751–3781. MR2766867 https://doi.org/10.1214/10-AOS840
[38] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131
[39] VAN DE GEER, S. (1987). A new approach to least-squares estimation, with applications. *Ann. Statist.* **15** 587–602. MR0888427 https://doi.org/10.1214/aos/1176350362
[40] VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343 https://doi.org/10.1214/aos/1176347632
[41] VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44. MR1212164 https://doi.org/10.1214/aos/1176349013
[42] VAN DE GEER, S. (1995). The method of sieves and minimum contrast estimators. *Math. Methods Statist.* **4** 20–38. MR1324688
[43] VAN DE GEER, S. and WAINWRIGHT, M. J. (2017). On concentration for (regularized) empirical risk minimization. *Sankhya A* **79** 159–200. MR3707417 https://doi.org/10.1007/s13171-017-0111-9
[44] VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge Univ. Press, Cambridge. MR1739079
[45] VAN DER VAART, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.* **24** 862–878. MR1394993 https://doi.org/10.1214/aos/1032894470
[46] VAN DER VAART, A. (1996). New Donsker classes. *Ann. Probab.* **24** 2128–2140. MR1415244 https://doi.org/10.1214/aop/1041903221
[47] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2
[48] WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362. MR1332570 https://doi.org/10.1214/aos/1176324524
[49] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500 https://doi.org/10.1214/aos/1017939142