Chernozhukov, Victor; Chetverikov, Denis; Kato, Kengo

Working Paper

Gaussian approximation of suprema of empirical processes

cemmap working paper, No. CWP75/13

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Chetverikov, Denis; Kato, Kengo (2013) : Gaussian approximation of suprema of empirical processes, cemmap working paper, No. CWP75/13, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2013.7513

This Version is available at:
https://hdl.handle.net/10419/97391

cemmap
centre for microdata methods and practice

Gaussian approximation of suprema of empirical processes

Victor Chernozhukov
Denis Chetverikov
Kengo Kato

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP75/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

GAUSSIAN APPROXIMATION OF SUPREMA OF EMPIRICAL PROCESSES

VICTOR CHERNOZHUKOV, DENIS CHETVERIKOV, AND KENGO KATO

ABSTRACT. We develop a new direct approach to approximating suprema of general empirical processes by a sequence of suprema of Gaussian processes, without taking the route of approximating whole empirical processes in the supremum norm. We prove an abstract approximation theorem that is applicable to a wide variety of problems, primarily in statistics. In particular, the bound in the main approximation theorem is non-asymptotic and the theorem does not require uniform boundedness of the class of functions. The proof of the approximation theorem builds on a new coupling inequality for maxima of sums of random vectors, the proof of which depends on an effective use of Stein's method for normal approximation, and some new empirical process techniques. We study applications of this approximation theorem to local empirical processes and series estimation in nonparametric regression where the classes of functions change with the sample size and are not Donsker-type. Importantly, our new technique is able to prove the Gaussian approximation for the supremum type statistics under weak regularity conditions, especially concerning the bandwidth and the number of series functions, in those examples.

1. INTRODUCTION

This paper is concerned with the problem of approximating suprema of empirical processes by a sequence of suprema of Gaussian processes. To formulate the problem, let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$ with common distribution P. Suppose that there is a sequence $\mathcal{F}_n$ of classes of measurable functions $S \to \mathbb{R}$, and consider the empirical process indexed by $\mathcal{F}_n$:

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_1)]), \ f \in \mathcal{F}_n.$$

Date: First version: December 31, 2012. This version: September 25, 2013.

2000 *Mathematics Subject Classification.* 60F17, 62E17, 62G20.

Key words and phrases. coupling, empirical process, Gaussian approximation, kernel estimation, local empirical process, series estimation, supremum.

V. Chernozhukov and D. Chetverikov are supported by a National Science Foundation grant. K.Kato is supported by the Grant-in-Aid for Young Scientists (B) (25780152), the Japan Society for the Promotion of Science.

For a moment, we implicitly assume that each $\mathcal{F}_n$ is "nice" enough and postpone the measurability issue. This paper tackles the problem of approximating $Z_n = \sup_{f\in\mathcal{F}_n} \mathbb{G}_n f$ by a sequence of random variables $\widetilde{Z}_n$ equal in distribution to $\sup_{f\in\mathcal{F}_n} B_n f$, where each B_n is a centered Gaussian process indexed by $\mathcal{F}_n$ with covariance function $\mathbb{E}[B_n(f)B_n(g)] = \mathrm{Cov}(f(X_1), g(X_1))$ for all $f, g \in \mathcal{F}_n$. We look for conditions under which there exists a sequence of such random variables $\widetilde{Z}_n$ with

$$|Z_n - \widetilde{Z}_n| = O_{\mathbb{P}}(r_n), \tag{1}$$

where $r_n \to 0$ as $n \to \infty$ is a sequence of constants[1].

The study of asymptotic and non-asymptotic behaviors of the supremum of the empirical process is one of the central issues in probability theory, and dates back to the classical work of [31]. The (tractable) distributional approximation of the supremum of the empirical process is of particular importance in mathematical statistics. A leading example is uniform inference in nonparametric estimation, such as construction of uniform confidence bands and specification testing in nonparametric density and regression estimation where critical values are given by quantiles of supremum type statistics [see, e.g., 3, 34, 46, 27, 26, 12]. Another interesting example appears in econometrics where there is an interest in estimating a parameter that is given as the extremum of an unknown function such as a conditional mean function. [13] proposed a precision-corrected estimate for such a parameter. In construction of their estimate, approximation of quantiles of a supremum type statistic is needed, to which the Gaussian approximation of the supremum type statistics plays a crucial role.

A related but different problem is that of approximating *whole* empirical processes by a sequence of Gaussian processes in the supremum norm. This problem is stronger than (1). Indeed, (1) is implied if there exists a sequence of versions of B_n (which we denote by the same symbol B_n) such that

$$\|\mathbb{G}_n - B_n\|_{\mathcal{F}_n} := \sup_{f\in\mathcal{F}_n} |(\mathbb{G}_n - B_n)f| = O_{\mathbb{P}}(r_n). \tag{2}$$

There is a large literature on the latter problem (2). Notably, Komlós et al. [33] (henceforth, abbreviated as KMT) proved that $\|\mathbb{G}_n - B_n\|_{\mathcal{F}} = O_{a.s.}(n^{-1/2}\log n)$ for $S = [0,1], P =$ uniform distribution on $[0,1]$, and $\mathcal{F} = \{1_{[0,t]} : t \in [0,1]\}$. See [38] and [7] for refinements of KMT's result. [39], [32] and [46] developed extensions of the KMT construction to more general classes of functions.

The KMT construction is a powerful tool in addressing the problem (2), but when applied to general empirical processes, it typically requires strong conditions on classes of functions and distributions. For example, Rio [46] required that $\mathcal{F}_n$ are uniformly bounded classes of functions having uniformly bounded variations on $S = [0,1]^d$, and P has a continuous and positive

[1]These results have immediate statistical implications; see Remark 2.5 ahead.

Lebesgue density on $[0,1]^d$. Such conditions are essential to the KMT construction since it depends crucially on the Haar approximation and binomial coupling inequalities of Tusnády. Note that [32] directly made an assumption on the accuracy of the Haar approximation of the class of functions, but still required similar side conditions to [46] in concrete applications; see Section 11 in [32]. [19], [2] and [47] considered the problem of Gaussian approximation of general empirical processes with different approaches and thereby without such side conditions. [19] used a finite approximation of a (possibly uncountably) infinite class of functions and apply a coupling inequality of [54] to the discretized empirical process (more precisely, [19] used a version of Yurinskii's inequality proved by [17]). [2] and [47], on the other hand, used a coupling inequality of [55] instead of Yurinskii's and some recent empirical process techniques such as Talagrand's [50] concentration inequality, which leads to refinements of Dudley and Philipp's results in some cases. However, the rates that [17], [2] and [47] established do not lead to tight conditions for the Gaussian approximations in non-Donsker cases, with important examples being the suprema of empirical processes arising in nonparametric statistics, namely the the suprema of local and series empirical processes (see Section 3 for detailed treatment).

We develop here a new direct approach to the problem (1), without taking the route of approximating whole empirical processes in the supremum norm and with different technical tools than those used in the aforementioned papers (especially the approach taken does not rely on the Haar expansion and hence differs from the KMT type approximation). We prove an abstract approximation theorem (Theorem 2.1) that leads to results of type (1) in several situations. The proof of the approximation theorem builds on a number of technical tools that are of interest in their own rights: notably, 1) a new coupling inequality for maxima of sums of random vectors (Theorem 4.1), where Stein's method for normal approximation (building here on [9] and originally due to [48, 49]) plays an important role (see also [45] and [40]); 2) a deviation inequality for suprema of empirical processes that only requires finite moments of envelope functions (Theorem 5.1), due essentially to the recent work of [5], complemented with a new "local" maximal inequality for the expectation of suprema of empirical processes that extends the work of [53] (Theorem 5.2). We study applications of this approximation theorem to local and series empirical processes arising in nonparametric regression, and demonstrate that our new technique is able to provide the Gaussian approximation for the supremum type statistics under weak regularity conditions, especially concerning the bandwidth and the number of series functions, in those examples.

It is instructive to briefly summarize here the key features of the main approximation theorem. First, the theorem establishes a non-asymptotic bound between Z_n and its Gaussian analogue $\widetilde{Z}_n$. The theorem requires that each $\mathcal{F}_n$ is pre-Gaussian (i.e., assuming the existence of a version of B_n that is a tight Gaussian random element in $\ell^\infty(\mathcal{F}_n)$; see below for the

notation), but allows for the case where the "complexity" of $\mathcal{F}_n$ increases with n, which places the function classes outside any fixed Donsker class; moreover, neither the process $\mathbb{G}_n$ nor the supremum statistic Z_n need to be weakly convergent as $n \to \infty$ (even after suitable normalization). Second, the bound in Theorem 2.1 is able to exploit the "local" properties of the class of functions, thereby, when applied to, say, the supremum deviation of kernel type statistics, it leads to tight conditions on the bandwidth for the Gaussian approximation (see the discussion after Theorem 2.1 for details about these features). Note that our bound does not rely on "smoothness" of $\mathcal{F}_n$ (in contrast to [46] where the bound on the Gaussian approximation for empirical processes depends on the total variation norm of functions), which is helpful in deriving good conditions on the number of series functions for the Gaussian approximation of the supremum deviation of projection type statistics handled in Section 3.2 since, e.g., the total variation norm is typically large or difficult to control well for such type of statistics. Lastly, the theorem only requires finite moments of the envelope function, which should be contrasted with [32, 46, 2, 47] where the classes of functions studied are assumed to be uniformly bounded. Hence the theorem is readily applicable to a wide class of statistical problems to which the previous results are not, at least immediately.

We note that, to the best of our knowledge, [43] is the only previous work that considered the problem of directly approximating the distribution of the supremum of the empirical process by that of the corresponding Gaussian process. However, they only cover the case where the class of functions is independent of n and Donsker as the constant C in their master Theorem 2 is dependent on $\mathcal{F}$ (and how C depends on $\mathcal{F}$ is not specified), and their condition (1.4) essentially excludes the case where the "complexity" of $\mathcal{F}$ grows with n, which means that their results are not applicable to the statistical problems handled in this paper (see Remark 2.5 or Lemma 6.1 ahead). Moreover, their approach is significantly different from ours.

In this paper, we substantially rely on modern empirical process theory. For general references on empirical process theory, we refer to [36], [52], [18] and also [4]. Section 9.5 of [18] has excellent historical remarks on the Gaussian approximation of empirical processes. For textbook treatments of Yurinskii's and KMT's couplings, we refer to [15] and Chapter 10 in [44].

1.1. **Organization.** The rest of the paper is organized as follows. In Section 2, we present the main approximation theorem (Theorem 2.1). We give a proof of Theorem 2.1 in Section 6. In Section 3, we study applications of Theorem 2.1 to local and series empirical processes arising in nonparametric regression. Sections 4 and 5 are devoted to developing some technical tools needed to prove Theorem 2.1 and its supporting Lemma 2.2. In Section 4, we prove a new coupling inequality for maxima of sums of random vectors, and in Section 5, we prepare some inequalities for empirical processes. We put some additional technical proofs in Appendix.

1.2. **Notation.** We shall obey the following notation. Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote the underlying probability space. We assume that the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is rich enough, in the sense that there exists a uniform random variable on $(0,1)$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$ independent of the sample at hand. For a real-valued random variable ξ, let $\|\xi\|_q = (\mathbb{E}[|\xi|^q])^{1/q},\ 1 \le q < \infty$. For two random variables ξ and η, we write

$$\xi \stackrel{d}{=} \eta$$

if they have the same distribution.

For any probability measure Q on a measurable space $(S, \mathcal{S})$, we use the notation $Qf := \int f dQ$. Let $\mathcal{L}^p(Q), p \ge 1$ denote the space of all measurable functions $f : S \to \mathbb{R}$ such that $\|f\|_{Q,p} := (Q|f|^p)^{1/p} < \infty$. We also use the notation $\|f\|_\infty := \sup_{x \in S} |f(x)|$. Denote by e_Q the $\mathcal{L}^2(Q)$-semimetric:

$$e_Q(f,g) = \|f - g\|_{Q,2},\ f, g \in \mathcal{L}^2(Q).$$

For an arbitrary set T, let $\ell^\infty(T)$ denote the space of all bounded functions $T \to \mathbb{R}$, equipped with the uniform norm $\|f\|_T := \sup_{t \in T} |f(t)|$. We endow $\ell^\infty(T)$ with the Borel σ-field induced from the norm topology. A *random element* in $\ell^\infty(T)$ refers to a Borel measurable map from Ω to $\ell^\infty(T)$. For $\varepsilon > 0$, an *ε-net* of a semimetric space (T, d) is a subset T_ε of T such that for every $t \in T$ there exists a point $t_\varepsilon \in T_\varepsilon$ with $d(t, t_\varepsilon) < \varepsilon$. The *$\varepsilon$-covering number* $N(T, d, \varepsilon)$ of T is the infimum of the cardinality of ε-nets of T, i.e., $N(T, d, \varepsilon) := \inf\{\mathrm{Card}(T_\varepsilon) : T_\varepsilon \text{ is an } \varepsilon\text{-net of } T\}$ (formally define $N(T, d, 0) := \lim_{\varepsilon \downarrow 0} N(T, d, \varepsilon)$, where the right limit, possibly being infinite, exists as the map $\varepsilon \mapsto N(T, d, \varepsilon)$ is non-increasing). For a subset A of a semimetric space (T, d), let A^δ denotes the δ-enlargement of A, i.e., $A^\delta = \{x \in T : d(x, A) \le \delta\}$ where $d(x, A) = \inf_{y \in A} d(x, y)$.

The standard Euclidean norm is denoted by $|\cdot|$. The transpose of a vector x is denoted by x^T. For a smooth function $f : \mathbb{R}^p \to \mathbb{R}$, we use the notation $\partial_j f(x) = \partial f(x)/\partial x_j$, $\partial_j \partial_k f(x) = \partial^2 f(x)/\partial x_j \partial x_k$, and so on.

We write $a \lesssim b$ if there exists a universal constant $C > 0$ such that $a \le Cb$. For a given parameter q, we write $a \lesssim_q b$ if there exists a constant $C(q) > 0$ depending only on q such that $a \le C(q)b$. For $a, b \in \mathbb{R}$, $a \vee b = \max\{a, b\}$, $a_+ = a \vee 0$. Unless otherwise stated, $c, C > 0$ denote universal constants of which the values may change from line to line.

Lastly, for a sequence $\{z_i\}_{i=1}^n$, we write $\mathbb{E}_n[z_i] = n^{-1} \sum_{i=1}^n z_i$, i.e., $\mathbb{E}_n$ abbreviates the symbol $n^{-1} \sum_{i=1}^n$. For example, $\mathbb{E}_n[f(X_i)] = n^{-1} \sum_{i=1}^n f(X_i)$.

2. Abstract approximation theorem

We begin with reviewing the setup. Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$ with common distribution P. In all what follows, we assume $n \ge 3$. Let $\mathcal{F}$ be a class of measurable functions $S \to \mathbb{R}$. We assume that the class $\mathcal{F}$ is P-centered, i.e.,

$$Pf = 0,\ \forall f \in \mathcal{F}.$$

This does not lose generality since otherwise we may replace $\mathcal{F}$ by $\{f - Pf : f \in \mathcal{F}\}$. Denote by F a measurable *envelope* of $\mathcal{F}$, i.e., F is a non-negative measurable function $S \to \mathbb{R}$ such that

$$F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|, \ \forall x \in S.$$

In this section the sample size n is fixed, and hence the possible dependence of $\mathcal{F}$ and F (and other quantities) on n is dropped.

We make the following assumptions.

(A1) The class $\mathcal{F}$ is *pointwise measurable*, i.e., it contains a countable subset $\mathcal{G}$ such that for every $f \in \mathcal{F}$ there exists a sequence $g_m \in \mathcal{G}$ with $g_m(x) \to f(x)$ for every $x \in S$.

(A2) For some $q \geq 2$, $F \in \mathcal{L}^q(P)$.

(A3) The class $\mathcal{F}$ is *P-pre-Gaussian*, i.e, there exists a tight Gaussian random element G_P in $\ell^\infty(\mathcal{F})$ with mean zero and covariance function

$$\mathbb{E}[G_P(f)G_P(g)] = P(fg) = \mathbb{E}[f(X_1)g(X_1)], \ \forall f, g \in \mathcal{F}.$$

Assumption (A1) is made to avoid measurability complications. See Section 2.3.1 of [52] for further discussion. This assumption ensures that, e.g., $\sup_{f \in \mathcal{F}} \mathbb{G}_n f = \sup_{f \in \mathcal{G}} \mathbb{G}_n f$, and hence the former supremum is a measurable map from Ω to $\mathbb{R}$. Note that by Example 1.5.10 in [52], assumption (A3) implies that $\mathcal{F}$ is totally bounded for e_P, and G_P has sample paths a.s. uniformly e_P-continuous.

To state the main result, we prepare some notation. For $\varepsilon > 0$, define $\mathcal{F}_\varepsilon = \{f - g : f, g \in \mathcal{F}, e_P(f, g) < \varepsilon \|F\|_{P,2}\}$. Note that by Theorem 3.1.1 in [18], under assumption (A3), one can extend G_P to the linear hull of $\mathcal{F}$ in such a way that G_P has linear sample paths. With this in mind, let

$$\phi_n(\varepsilon) = \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon}] \vee \mathbb{E}[\|G_P\|_{\mathcal{F}_\varepsilon}].$$

For the notational convenience, let us write

$$H_n(\varepsilon) = \log(N(\mathcal{F}, e_P, \varepsilon \|F\|_{P,2}) \vee n). \tag{3}$$

Note that since $\mathcal{F}$ is totally bounded for e_P (because of assumption (A3)), $H_n(\varepsilon)$ is finite for every $0 < \varepsilon \leq 1$. Moreover, write $M = \max_{1 \leq i \leq n} F(X_i)$ and $\mathcal{F} \cdot \mathcal{F} = \{fg : f \in \mathcal{F}, g \in \mathcal{F}\}$. The following is the main theorem of this paper. The proof of the following theorem will be given in Section 6.

Theorem 2.1 (Gaussian approximation to suprema of empirical processes)**.** *Suppose that assumptions (A1), (A2) with $q \geq 3$, and (A3) are satisfied. Let $Z = \sup_{f \in \mathcal{F}} \mathbb{G}_n f$. Let $\kappa > 0$ be any positive constant such that $\kappa^3 \geq \mathbb{E}[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}]$. Then for every $\varepsilon \in (0, 1]$ and $\gamma \in (0, 1)$, there exists a random variable $\widetilde{Z} \stackrel{d}{=} \sup_{f \in \mathcal{F}} G_P f$ such that*

$$\mathbb{P}\left\{|Z - \widetilde{Z}| > K(q)\Delta_n(\varepsilon, \gamma)\right\} \leq \gamma \{1 + \delta_n(\varepsilon, \gamma)\} + \frac{C \log n}{n},$$

where $K(q) > 0$ is a constant that depends only on q, and

$$\Delta_n(\varepsilon,\gamma) := \phi_n(\varepsilon) + \gamma^{-1/q}\varepsilon\|F\|_{P,2} + n^{-1/2}\gamma^{-1/q}\|M\|_q + n^{-1/2}\gamma^{-2/q}\|M\|_2 + n^{-1/4}\gamma^{-1/2}(\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}])^{1/2}H_n^{1/2}(\varepsilon) + n^{-1/6}\gamma^{-1/3}\kappa H_n^{2/3}(\varepsilon).$$

$$\delta_n(\varepsilon,\gamma) := \frac{1}{4}P\{(F/\kappa)^3 1(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\varepsilon)^{-1/3})\}.$$

Remark 2.1. The factor $1/4$ on the right side has no special meaning. It can be replaced by a smaller positive constant, but at the cost of increasing the constant $K(q)$. We do not pursue the generality in this direction. ∎

Recall that we have extended G_P to the linear hull of $\mathcal{F}$ in such a way that G_P has linear sample paths. Hence

$$\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}\cup(-\mathcal{F})} \mathbb{G}_n f,\ \|G_P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}\cup(-\mathcal{F})} G_P f,$$

where $-\mathcal{F} := \{-f : f \in \mathcal{F}\}$, from which one can readily deduce the following corollary. Henceforth we only deal with $\sup_{f\in\mathcal{F}} \mathbb{G}_n f$.

Corollary 2.1. *The conclusion of Theorem 2.1 continues to hold with Z replaced by $Z = \|\mathbb{G}_n\|_{\mathcal{F}}$, $\widetilde{Z}$ replaced by $\widetilde{Z} \stackrel{d}{=} \|G_P\|_{\mathcal{F}}$, and with different constants.*

Theorem 2.1 is useful only if there are suitable bounds on the following triple of terms, appearing in its statement:

$$\phi_n(\varepsilon),\ \mathbb{E}[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}] \text{ and } \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}]. \tag{4}$$

To this end, the entropy method or the more general generic chaining method [51] are useful. We will derive bounds on these terms using the entropy method since typically it leads to readily computable bounds. However, we leave the option of bounding the terms in (4) by other means , e.g., generic chaining methods (in some applications the latter is known to give sharper bounds than the entropy approach).

Consider the (uniform) entropy integral

$$J(\delta) = J(\delta,\mathcal{F},F) = \int_0^\delta \sup_Q \sqrt{1+\log N(\mathcal{F},e_Q,\varepsilon\|F\|_{Q,2})}d\varepsilon,$$

where the supremum is taken over all finitely discrete probability measures on $(S,\mathcal{S})$. We assume the integral is finite:

(A4) $J(1,\mathcal{F},F) < \infty$.

Remark 2.2. In applications $\mathcal{F}$ and F (and even S) may change with n, i.e., $\mathcal{F} = \mathcal{F}_n$ and $F = F_n$. In that case, assumption (A4) is interpreted as $J(1,\mathcal{F}_n,F_n) < \infty$ for each n, but let us keep in mind that it does allow for the case where $J(1,\mathcal{F}_n,F_n) \to \infty$ as $n \to \infty$. ∎

We first note the following (standard) fact.

Lemma 2.1. *Assumptions (A2) and (A4) imply assumption (A3).*

For the sake of completeness, we verify this lemma in Appendix. The following lemma provides bounds on the quantities in (4). The proof of the following lemma is given in Appendix.

Lemma 2.2 (Entropy-based bounds on the triple (4))**.** *Suppose that assumptions (A1), (A2) and (A4) are satisfied. Then for $\varepsilon \in (0,1]$,*

$$\phi_n(\varepsilon) \lesssim J(\varepsilon)\|F\|_{P,2} + n^{-1/2}\varepsilon^{-2}J^2(\varepsilon)\|M\|_2.$$

Moreover, suppose that assumption (A2) is satisfied with $q \geq 4$, and for $k = 3, 4$, let $\delta_k \in (0,1]$ be any positive constant such that $\delta_k \geq \sup_{f\in\mathcal{F}} \|f\|_{P,k}/\|F\|_{P,k}$. Then

$$\begin{aligned}
&\mathbb{E}[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}] - \sup_{f\in\mathcal{F}} P|f|^3 \\
&\qquad \lesssim n^{-1/2}\|M\|_3^{3/2}\left[J(\delta_3^{3/2},\mathcal{F},F)\|F\|_{P,3}^{3/2} + \frac{\|M\|_3^{3/2}J^2(\delta_3^{3/2},\mathcal{F},F)}{\sqrt{n}\delta_3^3}\right], \\
&\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}] \lesssim J(\delta_4^2,\mathcal{F},F)\|F\|_{P,4}^2 + \frac{\|M\|_4^2J^2(\delta_4^2,\mathcal{F},F)}{\sqrt{n}\delta_4^4}.
\end{aligned}$$

Remark 2.3 (The necessity of the above bounds)**.** The bounds provided above are carefully derived to give sharp results in applications of Section 3. Some readers may wonder if the following simpler bounds would suffice instead:

$$\mathbb{E}[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}] \leq \|F\|_{P,3}^3,\ \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}] \lesssim J(1,\mathcal{F},F)\|F\|_{P,4}^2.$$

The latter estimate is deduced from Theorem 2.14.1 of [52] together with the fact that

$$\sup_Q N(\mathcal{F}\cdot\mathcal{F}, e_Q, 2\varepsilon\|F^2\|_{Q,2}) \leq \sup_Q N^2(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}), \tag{5}$$

which is deduced from Lemma A.5 in Appendix. These simple bounds, however, become too non-sharp when PF^3 and PF^4 are significantly larger than the supremum of "weak" moments $\sup_{f\in\mathcal{F}} P|f|^3$ and $\sup_{f\in\mathcal{F}} Pf^4$, respectively, which is the case for all the examples studied in Section 3. ∎

Remark 2.4 (**Key features of Theorem 2.1**)**.** Before going to the applications, we discuss the key features of Theorem 2.1. First, Theorem 2.1 does not require uniform boundedness of $\mathcal{F}$, and requires only finite moments of the envelope function. This should be contrasted with the fact that many papers working on the Gaussian approximation of empirical processes in the supremum norm, such as [32, 46, 2, 47], required that classes of functions are uniformly bounded. There are, however, many statistical applications where uniform boundedness of the class of functions is too restrictive, and the generality of Theorem 2.1 in this direction will turn out to be useful. One drawback is that γ, which in applications we take as $\gamma = \gamma_n \to 0$, is typically at most $O(n^{-1/6})$, and hence Theorem 2.1 gives only "in probability bounds" rather than "almost sure bounds". The second feature of

Theorem 2.1 is that it is able to exploit the "local" properties of the class of functions $\mathcal{F}$. By Lemma 2.2, typically, we may take $\kappa^3 \approx \sup_{f\in\mathcal{F}} P|f|^3$ and $\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}.\mathcal{F}}] \approx \sup_{f\in\mathcal{F}} \sqrt{Pf^4}$ (up to logarithmic in n factors). In some applications, e.g., nonparametric kernel and series problems considered in the next section, the class $\mathcal{F} = \mathcal{F}_n$ changes with n and $\sup_{f\in\mathcal{F}_n} \|f\|_{P,k}/\|F_n\|_{P,k}$ with $k = 3, 4$ decrease to 0 where F_n is an envelope function of $\mathcal{F}_n$. The bound in Theorem 2.1 (with help of Lemma 2.2) effectively exploits this information and leads to tight conditions on, say, the bandwidth and the number of series functions, for the Gaussian approximation. This feature will be clear from the proofs for the applications in the following section. ■

Remark 2.5 (Gaussian approximation in Kolmogorov distance)**.** Theorem 2.1 combined with Lemma 2.2 can be used to show that the result (1) holds for some sequence of constants $r_n \to 0$ (subject to some conditions; possible rates of r_n are problem-specific). In statistical applications, however, one is typically interested in the result of the form (here we follow the notation used in Section 1)

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(Z_n \le t) - \mathbb{P}(\widetilde{Z}_n \le t)| = o(1), \ n \to \infty. \tag{6}$$

That is, the approximation of the distribution of Z_n by that of $\widetilde{Z}_n$ in the Kolmogorov distance (recall that the Kolmogorov distance between the distributions of two random variables ξ_1 and ξ_2 is defined by $\sup_{t\in\mathbb{R}} |\mathbb{P}(\xi_1 \le t) - \mathbb{P}(\xi_2 \le t)|$). To derive (6) from (1), we invoke the following lemma.

Lemma 2.3 (Gaussian approximation in Kolmogorov distance: non-asymptotic result)**.** *Consider the setting described in the beginning of this section. Suppose that assumptions (A1)-(A3) are satisfied, and that there exist constants $\underline{\sigma}, \bar{\sigma} > 0$ such that $\underline{\sigma}^2 \le Pf^2 \le \bar{\sigma}^2$ for all $f \in \mathcal{F}$. Moreover, suppose that there exist constants $r_1, r_2 > 0$ and a random variable $\widetilde{Z} \overset{d}{=} \sup_{f\in\mathcal{F}} G_P f$ such that $\mathbb{P}\{|Z - \widetilde{Z}| > r_1\} \le r_2$. Then*

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(Z \le t) - \mathbb{P}(\widetilde{Z} \le t)| \le C_\sigma r_1 \left\{\mathbb{E}[\widetilde{Z}] + \sqrt{1 \vee \log(\underline{\sigma}/r_1)}\right\} + r_2,$$

where C_σ is a constant depending only on $\underline{\sigma}$ and $\bar{\sigma}$.

It is now not difficult to give conditions to deduce (6) from (1). Formally, we state the following lemma.

Lemma 2.4 (Gaussian approximation in Kolmogorov distance: asymptotic result)**.** *Suppose that there exists a sequence of (P-centered) classes $\mathcal{F}_n$ of measurable functions $S \to \mathbb{R}$ satisfying assumptions (A1)-(A3) with $\mathcal{F} = \mathcal{F}_n$ for each n, and that there exist constants $\underline{\sigma}, \bar{\sigma} > 0$ (independent of n) such that $\underline{\sigma}^2 \le Pf^2 \le \bar{\sigma}^2$ for all $f \in \mathcal{F}_n$. Let $Z_n = \sup_{f\in\mathcal{F}_n} \mathbb{G}_n f$, and denote by B_n a tight Gaussian random element in $\ell^\infty(\mathcal{F}_n)$ with mean zero and covariance function $\mathbb{E}[B_n(f)B_n(g)] = P(fg)$ for all $f, g \in \mathcal{F}_n$. Moreover, suppose that there exist a sequence of random variables $\widetilde{Z}_n \overset{d}{=} \sup_{f\in\mathcal{F}_n} B_n f$*

and a sequence of constants $r_n \to 0$ *such that* $|Z_n - \widetilde{Z}_n| = O_{\mathbb{P}}(r_n)$ *and* $r_n \mathbb{E}[\widetilde{Z}_n] = o(1)$ *as* $n \to \infty$. *Then as* $n \to \infty$, $\sup_{t \in \mathbb{R}} |\mathbb{P}(Z_n \leq t) - \mathbb{P}(\widetilde{Z}_n \leq t)| = o(1)$.

Note here that we allow the case where $\mathbb{E}[\widetilde{Z}_n] \to \infty$. In the examples handled in the following section, typically, we have $\mathbb{E}[\widetilde{Z}_n] = O(\sqrt{\log n})$.

■

3. Applications

This section studies applications of Theorem 2.1 to local and series empirical processes arising in nonparametric regression. In both examples, the classes of functions change with the sample size n and the corresponding $\mathbb{G}_n$ processes do not have tight limits. Hence regularity conditions for the Gaussian approximation for the suprema will be of interest. All the proofs in this section are gathered in Appendix.

3.1. **Local empirical processes.** This section applies Theorem 2.1 to the supremum deviation of kernel type statistics. Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be i.i.d. random variables taking values in the product space $\mathcal{Y} \times \mathbb{R}^d$, where $(\mathcal{Y}, \mathcal{A}_{\mathcal{Y}})$ is an arbitrary measurable space. Suppose that there is a class $\mathcal{G}$ of measurable functions $\mathcal{Y} \to \mathbb{R}$. Let $k(\cdot)$ be a kernel function on $\mathbb{R}^d$. By "kernel function", we simply mean that $k(\cdot)$ is integrable with respect to the Lebesgue measure on $\mathbb{R}^d$ and its integral on $\mathbb{R}^d$ is normalized to be 1, but do not assume that $k(\cdot)$ is non-negative, i.e., higher order kernels are allowed. Let h_n be a sequence of positive constants such that $h_n \to 0$ as $n \to \infty$, and let $\mathcal{I}$ be an arbitrary Borel subset of $\mathbb{R}^d$. Consider the kernel-type statistics

$$S_n(x, g) = \frac{1}{nh_n^d} \sum_{i=1}^n g(Y_i) k(h_n^{-1}(X_i - x)), \ (x, g) \in \mathcal{I} \times \mathcal{G}.$$

Typically, under suitable regularity conditions, $S_n(x, g)$ will be a consistent estimator of $\mathbb{E}[g(Y_1) \mid X_1 = x]p(x)$, where $p(\cdot)$ denotes a Lebesgue density of the distribution of X_1 (assuming its existence). For example, when $g \equiv 1$, $S_n(x, g)$ will be a consistent estimator of $p(x)$; when $\mathcal{Y} = \mathbb{R}$ and $g(y) = y$, $S_n(x, g)$ will be a consistent estimator of $\mathbb{E}[Y_1 \mid X_1 = x]p(x)$; and when $\mathcal{Y} = \mathbb{R}$ and $g(\cdot) = 1(\cdot \leq y), y \in \mathbb{R}$, $S_n(x, g)$ will be a consistent estimator of $\mathbb{P}(Y_1 \leq y \mid X_1 = x)p(x)$. In statistical applications, it is often of interest to approximate the distribution of the following quantity:

$$W_n = \sup_{(x,g) \in \mathcal{I} \times \mathcal{G}} c_n(x, g) \sqrt{nh_n^d} (S_n(x, g) - \mathbb{E}[S_n(x, g)]),$$

where $c_n(x, g)$ is a suitable normalizing constant. A typical choice of $c_n(x, g)$ would be such that

$$\mathrm{Var}\left(\sqrt{nh_n^d} S_n(x, g) \right) = c_n(x, g)^{-2} + o(1).$$

Limit theorems for W_n are developed in [3], [34], [16], [46], [21], and [37], among others.

[21] called the process $g \mapsto \sqrt{nh_n^d}(S_n(x,g) - \mathbb{E}[S_n(x,g)])$ a "local" empirical process at x (the original definition of the local empirical process in [21] is slightly more general in that h_n is replaced by a sequence of bimeasurable functions). With a slight abuse of terminology, we also call the process $(x,g) \mapsto \sqrt{nh_n^d}(S_n(x,g) - \mathbb{E}[S_n(x,g)])$ a local empirical process.

We consider the problem of approximating W_n by a sequence of suprema of Gaussian processes. For each $n \geq 1$, let B_n be a centered Gaussian process indexed by $\mathcal{I} \times \mathcal{G}$ with covariance function

$$\begin{aligned}&\mathbb{E}[B_n(x,g)B_n(\check{x},\check{g})] \\ &\quad = h_n^{-d} c_n(x,g)c_n(\check{x},\check{g}) \operatorname{Cov}[g(Y_1)k(h_n^{-1}(X_1 - x)), \check{g}(Y_1)k(h_n^{-1}(X_1 - \check{x}))].\end{aligned} \tag{7}$$

Intuitively, it is expected that under suitable regularity conditions, there is a sequence $\widetilde{W}_n$ of random variables such that $\widetilde{W}_n \stackrel{d}{=} \sup_{(x,g)\in\mathcal{I}\times\mathcal{G}} B_n(x,g)$ and as $n \to \infty$, $|W_n - \widetilde{W}_n| \stackrel{\mathbb{P}}{\to} 0$. We shall argue the validity of this approximation with explicit rates.

Before stating the assumptions, we recall the notion of VC type class.

Definition 3.1 (VC type class)**.** Let $\mathcal{F}$ be a class of measurable functions on a measurable space $(S, \mathcal{S})$, to which a measurable envelope F is attached. We say that $\mathcal{F}$ is *VC type* with envelope F if there are constants $A, v > 0$ such that $\sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}) \leq (A/\varepsilon)^v$ for all $0 < \varepsilon \leq 1$, where the supremum is taken over all finitely discrete probability measures on $(S, \mathcal{S})$.

We make the following assumptions.

(B1) $\mathcal{G}$ is a pointwise measurable class of functions $\mathcal{Y} \to \mathbb{R}$ uniformly bounded by a constant $b > 0$, and is VC type with envelope $\equiv b$.

(B2) $k(\cdot)$ is a bounded and continuous kernel function on $\mathbb{R}^d$, and such that the class of functions $\mathcal{K} = \{t \mapsto k(ht + x) : h > 0, x \in \mathbb{R}^d\}$ is VC type with envelope $\equiv \|k\|_\infty$.

(B3) The distribution of X_1 has a bounded Lebesgue density $p(\cdot)$ on $\mathbb{R}^d$.

(B4) $h_n \to 0$ and $\log(1/h_n) = O(\log n)$ as $n \to \infty$.

(B5) $C_{\mathcal{I}\times\mathcal{G}} := \sup_{n\geq 1} \sup_{(x,g)\in\mathcal{I}\times\mathcal{G}} |c_n(x,g)| < \infty$. Moreover, for every fixed $n \geq 1$ and for every $(x_m, g_m) \in \mathcal{I} \times \mathcal{G}$ with $x_m \to x \in \mathcal{I}$ and $g_m \to g \in \mathcal{G}$ pointwise, $c_n(x_m, g_m) \to c_n(x, g)$.

We note that [42], Lemma 22, gives simple sufficient conditions under which $\mathcal{K}$ is VC type.

We first assume that $\mathcal{G}$ is uniformly bounded, which will be relaxed later.

Proposition 3.1 (Gaussian approximation to suprema of local empirical processes: bounded case)**.** *Suppose that assumptions (B1)-(B5) are satisfied. Then for every $n \geq 1$, there is a tight Gaussian random element B_n in $\ell^\infty(\mathcal{I} \times \mathcal{G})$ with mean zero and covariance function (7), and there is a*

sequence $\widetilde{W}_n$ *of random variables such that* $\widetilde{W}_n \stackrel{d}{=} \sup_{(x,g)\in\mathcal{I}\times\mathcal{G}} B_n(x,g)$ *and as* $n \to \infty$,

$$|W_n - \widetilde{W}_n| = O_{\mathbb{P}}\{(nh_n^d)^{-1/6}\log n + (nh_n^d)^{-1/4}\log^{5/4} n + (nh_n^d)^{-1/2}\log^{3/2} n\}.$$

Even when $\mathcal{G}$ is not uniformly bounded, a version of Proposition 3.1 continues to hold provided that suitable restrictions on the moments of the envelope of $\mathcal{G}$ are assumed. Instead of assumption (B1), we make the following assumption.

(B1)′ $\mathcal{G}$ is a pointwise measurable class of functions $\mathcal{Y} \to \mathbb{R}$ with measurable envelope G such that $\mathbb{E}[G^q(Y_1)] < \infty$ for some $q \geq 4$ and $\sup_{x\in\mathbb{R}^d} \mathbb{E}[G^4(Y_1) \mid X_1 = x] < \infty$. Moreover, $\mathcal{G}$ is VC type with envelope G.

Then we have the following proposition.

Proposition 3.2 (Gaussian approximation to suprema of local empirical processes: unbounded case)**.** *Suppose that assumptions (B1)′ and (B2)-(B5) are satisfied. Then the conclusion of Proposition 3.1 continues to hold, except for that the speed of approximation is*

$$O_{\mathbb{P}}\{(nh_n^d)^{-1/6}\log n + (nh_n^d)^{-1/4}\log^{5/4} n + (n^{1-2/q}h_n^d)^{-1/2}\log^{3/2} n\}.$$

Remark 3.1 (Discussion and comparison to other results)**.** It is instructive to compare Propositions 3.1 and 3.2 with the implications of Theorem 1.1 of Rio [46], which is a very sharp result on the Gaussian approximation (in the supremum norm) of general empirical processes indexed by uniformly bounded VC type classes of functions having locally uniformly bounded variation.

1. Rio's [46] Theorem 1.1 is not applicable to the case where the envelope function G is not bounded. Hence Proposition 3.2 is not covered by [46]. Indeed, we are not aware of any previous result that leads to the conclusion of Proposition 3.2, at least in this generality. For example, [34] considered the Gaussian approximation of W_n in the case where $\mathcal{Y} = \mathbb{R}$ and $g(y) = y$, but also assumed that the support of Y_1 is bounded. [21] proved in their Theorem 1.1 a weak convergence result for local empirical processes, which, combined with the Skorohod representation and Lemma 4.1 ahead, implies a Gaussian approximation result for W_n even when $\mathcal{G}$ is not uniformly bounded (but without explicit rates); however, their Theorem 1.1 (and also Theorem 1.2) is tied with the single value of x , i.e. x is fixed, since both theorems assume that the "localized" probability measure, localized at a given x, converges (in a suitable sense) to a fixed probability measure (see assumption (F.ii) in [21]). The same comment applies to [22]. In contrast, our results apply to the case where the supremum is taken over an uncountable set of values of x, which is relevant to statistical applications such as construction of uniform confidence bands.

2. In the special case of kernel density estimation (i.e., $g \equiv 1$), Rio's Theorem 1.1 implies (subject to some regularity conditions) that $|W_n -$

$\widetilde{W}_n| = O_{a.s.}\{(nh_n^d)^{-1/(2d)}\sqrt{\log n} + (nh_n^d)^{-1/2}\log n\}$ for $d \geq 2$ (the $d = 1$ case is formally excluded from [46]). Hence Rio's error rates are better than ours when $d = 2, 3$, but ours are better when $d \geq 4$ (aside from the difference between "in probability" and almost sure bounds).

3. Consider, as a second example, kernel regression estimation (i.e., $\mathcal{Y} = \mathbb{R}$ and $g(y) = y$). In order to formally apply Rio's Theorem 1.1 to this example, we need to assume that, e.g., (Y_1, X_1) is generated in such a way that $(Y_1, X_1) = (h(U, X_1), X_1)$ where the joint distribution (U, X_1) has support $[0,1]^{d+1}$ with continuous and positive Lebesgue density on $[0,1]^{d+1}$, and h is a function $[0,1]^{d+1} \to \mathbb{R}$ which is bounded and of bounded variation.[2] Subject to such side conditions, Rio's Theorem 1.1 leads to the following error rate: $|W_n - \widetilde{W}_n| = O_{a.s.}\{(n^{d/(d+1)}h_n^d)^{-1/(2d)}\sqrt{\log n} + (nh_n^d)^{-1/2}\log n\}$. See, e.g., [13], Theorem 8. In contrast, Propositions 3.1 and 3.2 do not require such side conditions. Moreover, aside from the difference between "in probability" and almost sure bounds, as long as $h_n = O(n^{-a})$ for some $a > 0$, our error rates are always better when $d \geq 2$. When $d = 1$, our rate is better as long as $nh_n^4/\log^c n \to 0$ (and vice versa) where $c > 0$ is some constant. ∎

Remark 3.2 (Converting coupling to convergence in Kolmogorov distance)**.** By Remark 2.5, we can convert the results in Propositions 3.1 and 3.2 into convergence of the Kolmogorov distance between the distributions of W_n and its Gaussian analogue $\widetilde{W}_n$. In fact, under either the assumptions of Proposition 3.1 or 3.2, by Dudley's inequality for Gaussian processes [52, Corollary 2.2.8], it is not difficult to deduce that $\mathbb{E}[\widetilde{W}_n] = O(\sqrt{\log n})$. Hence if moreover there exists a constant $\underline{\sigma} > 0$ (independent of n) such that $\mathrm{Var}(c_n(x,g)\sqrt{nh_n^d}S_n(x,g)) \geq \underline{\sigma}^2$ for all $(x,g) \in \mathcal{I} \times \mathcal{G}$ (giving primitive regularity conditions for this assumption is a standard task)[3], we have

$$|W_n - \widetilde{W}_n| = o_{\mathbb{P}}(\log^{-1/2} n) \Rightarrow \sup_{t\in\mathbb{R}} |\mathbb{P}(W_n \leq t) - \mathbb{P}(\widetilde{W}_n \leq t)| = o(1).$$

Note that $|W_n - \widetilde{W}_n| = o_{\mathbb{P}}(\log^{-1/2} n)$ (i) if $nh_n^d/\log^c n \to \infty$ under the assumptions of Proposition 3.1, and (ii) if $n^{(1-2/q)}h_n^d/\log^c n \to \infty$ under the assumptions of Proposition 3.2, where $c > 0$ is some constant. These conditions on the bandwidth h_n are mild, and interestingly they essentially coincide with the conditions on the bandwidth used in establishing exact rates of uniform strong consistency of kernel type estimators in [23, 24]. ∎

[2]For example, let $F^{-1}_{Y_1|X_1}(\cdot \mid x)$ denote the quantile function of the conditional distribution of Y_1 given $X_1 = x$ and take U uniformly distributed on $(0,1)$ independent of X_1. Then $(Y_1, X_1) \overset{d}{=} (F^{-1}_{Y_1|X_1}(U \mid X_1), X_1)$, but for the above condition to be met, we need to assume that $F^{-1}_{Y_1|X_1}(u \mid x)$ is (bounded and) of bounded variation as a function of u *and* x, which is not a typical assumption in estimation of the conditional mean.

[3]Under either the assumptions of Proposition 3.1 or 3.2, $\mathrm{Var}(c_n(x,g)\sqrt{nh_n^d}S_n(x,g))$ is bounded from above uniformly in $(x,g) \in \mathcal{I} \times \mathcal{G}$.

3.2. **Series empirical processes.** Here we consider the following problem. Let $(\eta_1, X_1), \ldots, (\eta_n, X_n)$ be i.i.d. random variables taking values in the product space $\mathcal{E} \times \mathbb{R}^d$, where $(\mathcal{E}, \mathcal{A}_\mathcal{E})$ is an arbitrary measurable space. Suppose that the support of X_1 is normalized to be $[0,1]^d$, and for each $K \geq 1$, there are K basis functions $\psi_{K,1}, \ldots, \psi_{K,K}$ defined on $[0,1]^d$. Let $\psi^K(x) = (\psi_{K,1}(x), \ldots, \psi_{K,K}(x))^T$. Examples of such basis functions are Fourier series, splines, Cohen-Daubechies-Vial (CDV) wavelet bases [14], Hermite polynomials and so on. Let K_n be a sequence of positive constants such that $K_n \to \infty$ as $n \to \infty$. Let $\mathcal{G}$ be a class of measurable functions $\mathcal{E} \to \mathbb{R}$ such that $\mathbb{E}[g^2(\eta_1)] < \infty$ and $\mathbb{E}[g(\eta_1) \mid X_1] = 0$ a.s. for all $g \in \mathcal{G}$, and let $\mathcal{I}$ be an arbitrary Borel measurable subset of $[0,1]^d$. Suppose that there are sequences of $K_n \times K_n$ matrices $A_{1n}(g)$ and $A_{2n}(g)$ indexed by $g \in \mathcal{G}$. We assume that $s_{\min}(A_{2n}(g)) > 0$ for all $g \in \mathcal{G}$. In what follows, we let $s_{\min}(A)$ and $s_{\max}(A)$ denote the minimum and maximum singular values of a matrix A, respectively. Consider the following empirical process:

$$S_n(x,g) = \frac{\psi^{K_n}(x)^T A_{1n}(g)^T}{|A_{2n}(g)\psi^{K_n}(x)|} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n g(\eta_i)\psi^{K_n}(X_i) \right], \ x \in \mathcal{I}, g \in \mathcal{G},$$

which we shall call the "series empirical process" (we shall formally follow the convention $0/0 = 0$). The problem here is the Gaussian approximation of the supremum of this series empirical process:

$$W_n := \sup_{(x,g) \in \mathcal{I} \times \mathcal{G}} S_n(x,g).$$

We address this problem in what follows. The study of distributional approximation of this statistic is motivated by the following statistical problems.

Example 3.1 (Forms of $S_n(x,g)$ arising in nonparametric mean regression)**.** Here we explain which forms of $S_n(x,g)$ arise in the nonparametric series or sieve mean regression. Consider a (generally heteroscedastic) nonparametric regression model

$$Y_i = m(X_i) + \eta_i, \ \mathbb{E}[\eta_i \mid X_i] = 0, \ \mathbb{E}[\eta_i^2 \mid X_i = x] = \sigma^2(x), \ 1 \leq i \leq n,$$

where Y_i is a scalar response variable, X_i is a d-vector of covariates of which the support $= [0,1]^d$, and η_i is a scalar unobservable error term. We assume that the data $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d. The parameter of interest is the conditional mean function $m(x) = \mathbb{E}[Y_1 \mid X_1 = x]$.

Consider series estimation of $m(x)$. The idea of series estimation is to approximate $m(x)$ by $\sum_{j=1}^{K_n} \theta_{K_n,j}\psi_{K_n,j}(x)$ with $K_n \to \infty$ as $n \to \infty$ and to estimate the vector $\theta^{K_n} = (\theta_{K_n,1}, \ldots, \theta_{K_n,K_n})^T$ by the least squares method:

$$\widehat{\theta}^{K_n} = \arg \min_{\theta^{K_n} \in \mathbb{R}^{K_n}} \sum_{i=1}^n \left(Y_i - \psi^{K_n}(X_i)^T \theta^{K_n} \right)^2.$$

The resulting estimate of $m(x)$ is given by $\widehat{m}(x) = \psi^{K_n}(x)^T \widehat{\theta}^{K_n}$.

The asymptotic properties of the series estimate have been thoroughly investigated in the literature. Importantly, under suitable regularity conditions, the rescaled and recentered estimator admits an asymptotic linear form:

$$\tilde{S}_n(x) = \frac{\sqrt{n}(\widehat{m}(x) - m(x))}{|A_{2n}\psi^{K_n}(x)|} \approx \frac{\psi^{K_n}(x)^T A_{1n}}{|A_{2n}\psi^{K_n}(x)|} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \psi^{K_n}(X_i) \right] =: S_n(x),$$

where $A_{1n} = (\mathbb{E}[\psi^{K_n}(X_1)\psi^{K_n}(X_1)^T])^{-1}$ and

$$A_{2n} = (\mathbb{E}[\sigma^2(X_1)\psi^{K_n}(X_1)\psi^{K_n}(X_1)^T])^{1/2} A_{1n}.$$

See, e.g., [41]. Here $\tilde{S}_n(x) \approx S_n(x)$ means that $\tilde{S}_n(x) = S_n(x) + o_{\mathbb{P}}(\log^{-1/2} n)$ uniformly in $x \in \mathcal{I}$ (the remainder term could be faster, but $o_{\mathbb{P}}(\log^{-1/2} n)$ is fast enough to make the remainder term negligible in approximating (in the Kolmogorov distance) the distribution of $\sup_{x\in\mathcal{I}} \tilde{S}_n(x)$ by that of the Gaussian analogue of $\sup_{x\in\mathcal{I}} S_n(x)$ as the expectation of the latter is typically $O(\sqrt{\log n})$; see Remark 2.5 and Lemma 6.1). Hence, for the purpose of making uniform inference on $m(x)$ over a Borel subset $\mathcal{I}$ of $[0,1]^d$, it is desirable to have a (tractable) distributional approximation of the quantity $W_n = \sup_{x\in\mathcal{I}} S_n(x)$.

Note that obtaining this approximation requires the use of undersmoothing to make the effect of the approximation bias negligible relative to the variance, i.e., we have to take $K_n \to \infty$ faster than those leading to the optimal rate of convergence in the supremum norm. However, we skip the discussion of regularity conditions for obtaining this approximation, because this is outside the scope of the paper. Rather the focus of this paper is on studying the supremum of the empirical process S_n and not on how we obtain the process S_n per se. ∎

Example 3.2 (Forms of $S_n(x, g)$ arising in nonparametric quantile regression)**.** Here we explain which forms of $S_n(x, g)$ arise in the nonparametric series or sieve quantile regression. Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be i.i.d. random variables taking values in $\mathbb{R} \times \mathbb{R}^d$ where the support of $X_1 = [0,1]^d$. Suppose that the parameter of interest is the conditional quantile function:

$$Q(\tau, x) = \inf\{y : F_{Y|X}(y \mid x) \geq \tau\},\ x \in [0,1]^d, \tau \in (0,1),$$

where $F_{Y|X}(y \mid x) = \mathbb{P}(Y_1 \leq y \mid X_1 = x)$ is the conditional distribution function. Consider series estimation of $Q(\tau, x)$. A standard way is to solve the following minimization problem:

$$\widehat{\theta}^{K_n}(\tau) = \arg \min_{\theta^{K_n} \in \mathbb{R}^{K_n}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \psi^{K_n}(X_i)^T \theta^{K_n} \right),$$

where $\rho_\tau(y) = \{\tau - 1(y \leq 0)\}y$ is called the check function [30], and where $K_n \to \infty$ as $n \to \infty$. A series estimate of $Q(\tau, x)$ is obtained by $\widehat{Q}(\tau, x) = \psi^{K_n}(x)^T \widehat{\theta}^{K_n}(\tau)$. Let $\mathcal{T}$ be an arbitrary closed interval in $(0,1)$. Suppose that the conditional distribution function $F_{Y|X}(y \mid x)$ has a Lebesgue density

$f_{Y|X}(y \mid x)$. Then, subject to some regularity conditions, the rescaled and recentered estimator admits an asymptotically linear form:

$$\begin{aligned}\tilde{S}_n(x,\tau) &= \frac{\sqrt{n}(\widehat{Q}(\tau,x) - Q(\tau,x))}{\sqrt{\tau(1-\tau)}|A_{2n}(\tau)\psi^{K_n}(x)|} \\ &\approx \frac{\psi^{K_n}(x)^T A_{1n}(\tau)}{\sqrt{\tau(1-\tau)}|A_{2n}(\tau)\psi^{K_n}(x)|}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \{\tau - 1(Y_i \le Q(\tau,X_i))\}\psi^{K_n}(X_i)\right] \\ &=: S_n(x,\tau),\end{aligned}$$

where $A_{1n}(\tau) = J_n(\tau)^{-1}, J_n(\tau) = \mathbb{E}[f_{Y|X}(Q(\tau,X_1) \mid X_1)\psi^{K_n}(X_1)\psi^{K_n}(X_1)^T]$, $A_{2n}(\tau) = (\mathbb{E}[\psi^{K_n}(X_1)\psi^{K_n}(X_1)^T])^{1/2}J_n(\tau)^{-1}$ (note that $\tau(1-\tau)$ comes from the conditional variance of $1(Y_i \le Q(\tau,X_i))$ given X_i). Here too $\tilde{S}_n(x,\tau) \approx S_n(x,\tau)$ means that $\tilde{S}_n(x,\tau) = S_n(x,\tau) + o_{\mathbb{P}}(\log^{-1/2} n)$ uniformly in $(x,\tau) \in \mathcal{I}\times\mathcal{T}$; see [28] and Belloni et al. [1, Theorem 2]. Note that

$$Y_i \le Q(\tau,X_i) \Leftrightarrow \eta_i \le \tau, \text{ with } \eta_i = F_{Y|X}(Y_i \mid X_i),$$

and η_i are uniform random variables on $(0,1)$, independent of $X_1,\dots,X_n$. So letting $g_\tau(\eta) = \tau - 1(\eta \le \tau)$, we have the expression

$$S_n(x,\tau) = \frac{\psi^{K_n}(x)^T A_{1n}(\tau)}{\sqrt{\tau(1-\tau)}|A_{2n}(\tau)\psi^{K_n}(x)|}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n g_\tau(\eta_i)\psi^{K_n}(X_i)\right].$$

For the purpose of making uniform inference on $Q(\tau,x)$ over $(\tau,x) \in \mathcal{T}\times\mathcal{I}$, it is desirable to have a (tractable) distributional approximation of the quantity $W_n = \sup_{(x,\tau)\in\mathcal{I}\times\mathcal{T}} S_n(x,\tau)$. ∎

The preceding examples explain and motivate various forms of S_n arising in mathematical statistics. We now go back to the analysis of the supremum W_n of S_n. Let B_n be a centered Gaussian process indexed by $\mathcal{I}\times\mathcal{G}$ with covariance function

$$\begin{aligned}&\mathbb{E}[B_n(x,g)B_n(\check{x},\check{g})] \\ &\quad = \alpha_n(x,g)^T\mathbb{E}[g(\eta_1)\check{g}(\eta_1)\psi^{K_n}(X_1)\psi^{K_n}(X_1)^T]\alpha_n(\check{x},\check{g}),\end{aligned} \tag{8}$$

where $\alpha_n(x,g) = A_{1n}(g)\psi^{K_n}(x)/|A_{2n}(g)\psi^{K_n}(x)|$. Intuitively, it is expected that under suitable regularity conditions, there is a sequence $\widetilde{W}_n$ of random variables such that $\widetilde{W}_n \stackrel{d}{=} \sup_{(x,g)\in\mathcal{I}\times\mathcal{G}} B_n(x,g)$ and as $n\to\infty$, $|W_n - \widetilde{W}_n| \stackrel{\mathbb{P}}{\to} 0$. We shall establish the validity of this approximation with explicit rates.

We make the following assumptions.

(C1) $\mathcal{G}$ is a pointwise measurable VC type class of functions $\mathcal{E}\to\mathbb{R}$ with measurable envelope G such that $\mathbb{E}[g^2(\eta_1)] < \infty$ and $\mathbb{E}[g(\eta_1) \mid X_1] = 0$ a.s. for all $g \in \mathcal{G}$.

(C2) There exist some constants $c_1, C_1 > 0$ such that $s_{\max}(A_{2n}(g)) \le C_1$ and $s_{\min}(A_{2n}(g)) \ge c_1$ for all $g \in \mathcal{G}$ and $n \ge 1$.

(C3) $b_n := \sup_{x\in[0,1]^d} |\psi^{K_n}(x)| \vee 1 < \infty$ and there exists a constant $C_2 > 0$ such that $s_{\max}(\mathbb{E}[\psi^{K_n}(X_1)\psi^{K_n}(X_1)']) \leq C_2$ for all $n \geq 1$. The map $(x, g) \mapsto A_{1n}(g)\psi^{K_n}(x)/|A_{2n}(g)\psi^{K_n}(x)| =: \alpha_n(x, g)$ is Lipschitz continuous with Lipschitz constant $\leq L_n(\geq 1)$ in the following sense:

$$|\alpha_n(x,g) - \alpha_n(\check{x},\check{g})| \leq L_n\{|x-\check{x}| + (\mathbb{E}[(g(\eta_1)-\check{g}(\eta_1))^2])^{1/2}\}, \quad \forall x,\check{x}\in[0,1]^d, \forall g,\check{g}\in\mathcal{G}. \tag{9}$$

Here b_n and L_n are allowed to diverge as $n \to \infty$.

(C4) $\log b_n = O(\log n)$ and $\log L_n = O(\log n)$ as $n \to \infty$.

For many commonly used basis functions such as Fourier series, splines and CDV wavelet bases, $b_n = O(\sqrt{K_n})$ as $n \to \infty$. See [41]. The Lipschitz condition (9) is satisfied if $\inf_{x\in[0,1]^d} |\psi^{K_n}(x)| \geq c_2 > 0$, $|\psi^{K_n}(x) - \psi^{K_n}(\check{x})| \leq L_{1n}|x-\check{x}|$, and $\|A_{1n}(g) - A_{1n}(\check{g})\|_{\mathrm{op}} \vee \|A_{2n}(g) - A_{2n}(\check{g})\|_{\mathrm{op}} \leq L_{2n}(\mathbb{E}[(g(\eta_1) - \check{g}(\eta_1))^2])^{1/2}$, where $c_2 > 0$ is a fixed constant and L_{1n}, L_{2n} are sequences of constants possibly divergent as $n \to \infty$ ($\|A\|_{\mathrm{op}}$ denotes the operator norm of a matrix A). Then (9) is satisfied with $L_n = O(L_{1n} \vee L_{2n})$. Assumption (C4) states mild growth restrictions on K_n and L_n and is usually satisfied.

Proposition 3.3 (Gaussian approximation to suprema of series empirical processes)**.** *Suppose that assumptions (C1)-(C4) are satisfied. Moreover, suppose either (i) G is bounded (i.e., $\|G\|_\infty < \infty$), or (ii) $\mathbb{E}[G^q(\eta_1)] < \infty$ for some $q \geq 4$ and $\sup_{x\in[0,1]^d} \mathbb{E}[G^4(\eta_1) \mid X_1 = x] < \infty$. Then for every $n \geq 1$, there is a tight Gaussian random element B_n in $\ell^\infty(\mathcal{I}\times\mathcal{G})$ with mean zero and covariance function (8), and there exists a sequence $\widetilde{W}_n$ of random variables such that $\widetilde{W}_n \overset{d}{=} \sup_{(x,g)\in\mathcal{I}\times\mathcal{G}} B_n(x,g)$ and as $n \to \infty$,*

$$|W_n - \widetilde{W}_n| = \begin{cases} O_{\mathbb{P}}\{n^{-1/6}b_n^{1/3}\log n + n^{-1/4}b_n^{1/2}\log^{5/4} n + n^{-1/2}b_n\log^{3/2} n\}, & \text{(i)}, \\ O_{\mathbb{P}}\{n^{-1/6}b_n^{1/3}\log n + n^{-1/4}b_n^{1/2}\log^{5/4} n + n^{-1/2+1/q}b_n\log^{3/2} n\}, & \text{(ii)}. \end{cases}$$

Remark 3.3 (Discussion and comparisons with other approximations)**.** Proposition 3.3 is a new result, and its principal attractive feature is the weak requirement on the number of series functions (K_n). Another approach to deduce a result similar to Proposition 3.3 is to apply Yurinskii's coupling (see Theorem 4.2 ahead) to random vectors $g(\eta_i)\psi^{K_n}(X_i)$, which, however, requires a rather stringent restriction on K_n, namely $K_n^5/n \to 0$, for ensuring $|W_n - \widetilde{W}_n| \xrightarrow{\mathbb{P}} 0$ even in the simplest case where $\mathcal{E} = \mathbb{R}$ and $g(\eta) = \eta$. See, e.g., [13], Theorem 7. Moreover, the use of Rio's [46] Theorem 1.1 here is not effective since the total variation bound is large or difficult to control well in this example, which results in restrictive conditions on K_n (also Rio's [46] Theorem 1.1 does not cover case (ii) where G may not be bounded). ■

Remark 3.4 (Converting coupling to convergence in Kolmogorov distance)**.** As before, we can convert the results in Proposition 3.3 into convergence of

the Kolmogorov distance between the distributions of W_n and its Gaussian analogue $\widetilde{W}_n$. By Dudley's inequality for Gaussian processes [52, Corollary 2.2.8], it is not difficult to deduce that $\mathbb{E}[\widetilde{W}_n] = O(\sqrt{\log n})$ under the assumptions of Proposition 3.3. Hence if moreover there exists a constant $\underline{\sigma} > 0$ (independent of n) such that $\mathrm{Var}(S_n(x,g)) \geq \underline{\sigma}^2$ for all $(x,g) \in \mathcal{I} \times \mathcal{G}$, by Lemma 2.4, we have

$$|W_n - \widetilde{W}_n| = o_{\mathbb{P}}(\log^{-1/2} n) \Rightarrow \sup_{t \in \mathbb{R}} |\mathbb{P}(W_n \leq t) - \mathbb{P}(\widetilde{W}_n \leq t)| = o(1).$$

Note that $|W_n - \widetilde{W}_n| = o_{\mathbb{P}}(\log^{-1/2} n)$ if $K_n(\log n)^c/n \to 0$ in case (i) and $K_n(\log n)^c/n^{1-2/q} \to 0$ in case (ii), where $c > 0$ is some constant. These requirements on K_n are mild, in view of the fact that at least $K_n/n \to 0$ is needed for consistency (in the L^2-norm) of the series estimator [see 29]. ∎

4. A coupling inequality for maxima of sums of random vectors

The main ingredient in the proof of Theorem 2.1 is a new coupling inequality for maxima of sums of random vectors, which is stated below. A related Gaussian approximation inequality was obtained in [10] with a different technique, but the current Theorem 4.1 is more convenient for the purpose of this paper.

Theorem 4.1 (A coupling inequality for maxima of sums of random vectors)**.** *Let $X_1, \ldots, X_n$ be independent random vectors in $\mathbb{R}^p$ with mean zero and finite absolute third moments, i.e., $\mathbb{E}[X_{ij}] = 0$ and $\mathbb{E}[|X_{ij}|^3] < \infty$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$. Consider the statistic*

$$Z = \max_{1 \leq j \leq p} \sum_{i=1}^n X_{ij}.$$

Let $Y_1, \ldots, Y_n$ be independent random vectors in $\mathbb{R}^p$ such that

$$Y_i \sim N(0, \mathbb{E}[X_i X_i^T]), 1 \leq i \leq n.$$

Then for every $\beta > 0$ and $\delta > 1/\beta$, there exists a random variable $\widetilde{Z} \stackrel{d}{=} \max_{1 \leq j \leq p} \sum_{i=1}^n Y_{ij}$ such that

$$\mathbb{P}(|Z - \widetilde{Z}| > 2\beta^{-1} \log p + 3\delta) \leq \frac{\varepsilon + C\beta\delta^{-1}\{B_1 + \beta(B_2 + B_3)\}}{1 - \varepsilon},$$

where $\varepsilon = \varepsilon_{\beta,\delta}$ is given by

$$\varepsilon = \sqrt{e^{-\alpha}(1+\alpha)} < 1, \ \alpha = \beta^2\delta^2 - 1 > 0,$$

and

$$B_1 = \mathbb{E}\left[\max_{1\le j,k\le p} |\sum_{i=1}^n (X_{ij}X_{ik} - \mathbb{E}[X_{ij}X_{ik}])|\right],$$
$$B_2 = \mathbb{E}\left[\max_{1\le j\le p} \sum_{i=1}^n |X_{ij}|^3\right],$$
$$B_3 = \sum_{i=1}^n \mathbb{E}\left[\max_{1\le j\le p} |X_{ij}|^3 \cdot 1\left(\max_{1\le j\le p} |X_{ij}| > \beta^{-1}/2\right)\right].$$

The following corollary is useful for many applications. Recall $n \ge 3$.

Corollary 4.1 (An applied coupling inequality for maxima of sums of random vectors)**.** *Consider the same setup as in Theorem 4.1. Then for every $\delta > 0$, there exists a random variable $\widetilde{Z} \stackrel{d}{=} \max_{1\le j\le p} \sum_{i=1}^n Y_{ij}$ such that*

$$\mathbb{P}(|Z - \widetilde{Z}| > 16\delta) \lesssim \delta^{-2}\{B_1 + \delta^{-1}(B_2 + B_4)\log(p\vee n)\}\log(p\vee n) + \frac{\log n}{n}, \quad (10)$$

where B_1 and B_2 are as in Theorem 4.1, and

$$B_4 = \sum_{i=1}^n \mathbb{E}\left[\max_{1\le j\le p} |X_{ij}|^3 \cdot 1\left(\max_{1\le j\le p} |X_{ij}| > \delta/\log(p\vee n)\right)\right].$$

Proof of Corollary 4.1. In Theorem 4.1, take $\beta = 2\delta^{-1}\log(p\vee n)$. Then $\alpha = \beta^2\delta^2 - 1 = 4\log^2(p\vee n) - 1 \ge 2\log(p\vee n)$ (recall $n \ge 3 > e$), so that $\varepsilon \le 2\log(p\vee n)/(p\vee n) \le 2n^{-1}\log n$. This completes the proof. ∎

Theorem 4.1 is a coupling inequality similar in nature to Yurinskii's [54] coupling for sums of random vectors (as opposed to the maxima of such vectors as in the current theorem). Before proving Theorem 4.1, let us first recall Yurinskii's coupling inequality.

Theorem 4.2 (Yurinskii's coupling for sums of random vectors; [54]; see also [35])**.** *Consider the same setup as in Theorem 4.1. Let $S_n = \sum_{i=1}^n X_i$. Then for every $\delta > 0$, there exists a random vector $T_n \stackrel{d}{=} \sum_{i=1}^n Y_i$ such that*

$$\mathbb{P}(|S_n - T_n| > 3\delta) \lesssim B_0\left(1 + \frac{|\log(1/B_0)|}{p}\right),$$

where $B_0 = p\delta^{-3}\sum_{i=1}^n \mathbb{E}[|X_i|^3]$.

For the proof, see [44], Section 10.4. Because of the general fact that $\max_{1\le j\le n}|x_j| \le |x|$ for $x \in \mathbb{R}^p$, one has

$$|\max_{1\le j\le p}(S_n)_j - \max_{1\le j\le n}(T_n)_j| \le \max_{1\le j\le p}|(S_n - T_n)_j| \le |S_n - T_n|.$$

Hence if we take $\widetilde{Z} = \max_{1\le j\le p}(T_n)_j$,

$$\mathbb{P}(|Z - \widetilde{Z}| > 3\delta) \lesssim B_0\left(1 + \frac{|\log(1/B_0)|}{p}\right). \quad (11)$$

Unfortunately, when p is large, the right side needs not be small. This is because B_0 is proportional to $\sum_{i=1}^n \mathbb{E}[|X_i|^3]$ and this quantity may be larger than what we want.

To better understand the difference between (10) and (11), consider the situation where p is indexed by n and $p = p_n \to \infty$ as $n \to \infty$. Moreover, consider the simple case where $X_{ij} = x_{ij}/\sqrt{n}$ and $|x_{ij}| \leq b$ (x_{ij} are random; b is a fixed constant). Then

$$B_1 = O(n^{-1/2} \log^{1/2} p_n), \ B_2 + B_4 = O(n^{-1/2}).$$

The former estimate is deduced from the fact that, using the symmetrization and the maximal inequality for Rademacher averages conditional on $X_1, \ldots, X_n$ [use 52, Lemmas 2.2.2 and 2.2.7], one has

$$B_1 \lesssim \sqrt{\log(1+p)} \mathbb{E}\left[\max_{1 \leq j \leq p} (\sum_{i=1}^n X_{ij}^4)^{1/2} \right].$$

On the other hand,

$$p_n \sum_{i=1}^n |X_i|^3 = O(n^{-1/2} p_n^{5/2}).$$

Therefore, to make $|Z - \widetilde{Z}| \xrightarrow{\mathbb{P}} 0$, the former (10) allows p_n to be of an exponential order (p_n can be as large as $\log p_n = o(n^{1/4})$; hence, e.g., p_n can be of order e^{n^α} for $0 < \alpha < 1/4$), while the latter (11) restricts p_n to be $p_n = o(n^{1/5})$. Note that, under the exponential moment condition, instead of Yurinskii's coupling, we can use Zaitsev's coupling inequality [55, Theorem 1.1] but it still requires $p_n = o(n^{1/5})$ to deduce that $|Z - \widetilde{Z}| \xrightarrow{\mathbb{P}} 0$ (although by using Zaitsev's coupling, we indeed have an exponential type inequality for $|Z - \widetilde{Z}|$).

Remark 4.1 (Connection to Theorem 2.1)**.** The importance of Theorem 4.1 in the context of the proof of Theorem 2.1 is described as follows. In the proof of Theorem 2.1, we make a finite approximation of $\mathcal{F}$ by a minimal $\varepsilon \|F\|_{P,2}$-net of $(\mathcal{F}, e_P)$ and apply Theorem 4.1 to the "discretized" empirical process; hence in this application, $p = N(\mathcal{F}, e_P, \varepsilon \|F\|_{P,2})$. The fact that Theorem 4.1 allows for "large" p means that a "finer" discretization is possible, and as a result, the bound in Theorem 2.1 depends on the covering number $N(\mathcal{F}, e_P, \varepsilon \|F\|_{P,2})$ only through its logarithm: $\log N(\mathcal{F}, e_P, \varepsilon \|F\|_{P,2})$. ■

We will use a version of Strassen's theorem to prove Theorem 4.1. We state it for the reader's convenience.

Lemma 4.1 (An implication of Strassen's theorem)**.** *Let μ and ν be Borel probability measures on $\mathbb{R}$, and let V be a random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution μ. Suppose that the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ admits a uniform random variable on $(0,1)$ independent of V. Let $\varepsilon > 0$ and $\delta > 0$ be two positive constants. Then there exists*

a random variable W, defined on $(\Omega, \mathcal{A}, \mathbb{P})$, with distribution ν such that $\mathbb{P}(|V - W| > \delta) \leq \varepsilon$ if and only if $\mu(A) \leq \nu(A^\delta) + \varepsilon$ for every Borel subset A of $\mathbb{R}$.

Proof. The "only if" part is trivial, and hence we prove the "if" part. By Strassen's theorem [see 44, Section 10.3], there are random variables V^* and W^* with distributions μ and ν such that $\mathbb{P}(|V^* - W^*| > \delta) \leq \varepsilon$. V^* may be different from V. Let $F(w \mid v)$ be a regular conditional distribution function of W^* given $V^* = v$. Denote by $F^{-1}(\tau \mid v)$ the quantile function of $F(w \mid v)$, i.e., $F^{-1}(\tau \mid v) = \inf\{w : F(w \mid v) \geq \tau\}$. Generate a uniform random variable U on $(0,1)$ independent of V and take $W(\omega) = F^{-1}(U(\omega) \mid V(\omega))$. Then it is routine to verify that $(V, W) \stackrel{d}{=} (V^*, W^*)$. ∎

Proof of Theorem 4.1. For the notational convenience, write $e_\beta = \beta^{-1} \log p$. Construct $Y_1, \ldots, Y_n$ independent of $X_1, \ldots, X_n$. By Lemma 4.1, the conclusion follows if we can prove that for every Borel subset A of $\mathbb{R}$,

$$\mathbb{P}(Z \in A) \leq \mathbb{P}(\widetilde{Z}^* \in A^{2e_\beta + 3\delta}) + \frac{\varepsilon + C\beta\delta^{-1}\{B_1 + \beta(B_2 + B_3)\}}{1 - \varepsilon},$$

where $\widetilde{Z}^* := \max_{1 \leq j \leq p} \sum_{i=1}^n Y_{ij}$. Let $S_n = \sum_{i=1}^n X_i$ and $T_n = \sum_{i=1}^n Y_i$. Fix any Borel subset A of $\mathbb{R}$. We divide the proof into several steps.

Step 1: We approximate the non-smooth map $x \mapsto 1_A(\max_{1 \leq j \leq p} x_j)$ by a smooth function. The first step is to approximate the map $x \mapsto \max_{1 \leq j \leq p} x_j$ by a smooth function. Consider the function $F_\beta : \mathbb{R}^p \to \mathbb{R}$ defined by

$$F_\beta(x) = \beta^{-1} \log \left(\sum_{j=1}^p e^{\beta x_j} \right),$$

which gives a smooth approximation of $\max_{1 \leq j \leq p} x_j$. Indeed, an elementary calculation gives the following inequality: for every $x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$,

$$\max_{1 \leq j \leq p} x_j \leq F_\beta(x) \leq \max_{1 \leq j \leq p} x_j + \beta^{-1} \log p. \tag{12}$$

See [8]. Hence we have

$$\mathbb{P}(Z \in A) \leq \mathbb{P}(F_\beta(S_n) \in A^{e_\beta}) = \mathbb{E}[1_{A^{e_\beta}}(F_\beta(S_n))].$$

Step 2: The next step is to approximate the indicator function $t \mapsto 1_A(t)$ by a smooth function. This step is rather standard.

Lemma 4.2. *Let $\beta > 0$ and $\delta > 1/\beta$. For every Borel subset A of $\mathbb{R}$, there exists a smooth function $g : \mathbb{R} \to \mathbb{R}$ such that $\|g'\|_\infty \leq \delta^{-1}, \|g''\|_\infty \leq C\beta\delta^{-1}, \|g'''\|_\infty \leq C\beta^2\delta^{-1}$, and*

$$(1 - \varepsilon)1_A(t) \leq g(t) \leq \varepsilon + (1 - \varepsilon)1_{A^{3\delta}}(t), \ \forall t \in \mathbb{R},$$

where $\varepsilon = \varepsilon_{\beta,\delta}$ is given by

$$\varepsilon = \sqrt{e^{-\alpha}(1 + \alpha)} < 1, \ \alpha = \beta^2\delta^2 - 1 > 0.$$

Proof of Lemma 4.2. The proof is due to [44], Lemma 10.18 (p. 248). Let $\rho(\cdot,\cdot)$ denote the Euclidean distance on $\mathbb{R}$. Then consider the function $h(t) = (1-\rho(t, A^\delta)/\delta)_+$. Note that h is Lipschitz continuous with Lipschitz constant $\leq \delta^{-1}$. Construct a smooth approximation of $h(t)$ by

$$g(t) = \frac{\beta}{\sqrt{2\pi}} \int_{\mathbb{R}} h(s) e^{-\frac{1}{2}\beta^2(s-t)^2} ds = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h(t+\beta^{-1}s) e^{-\frac{1}{2}s^2} ds.$$

Then the map $t \mapsto g(t)$ is infinitely differentiable, and

$$\|g'\|_\infty \leq \delta^{-1},\ \|g''\|_\infty \leq C\beta\delta^{-1},\ \|g'''\|_\infty \leq C\beta^2\delta^{-1}.$$

The rest of the proof is the same as [44], Lemma 10.18 and omitted. ∎

Apply Lemma 4.2 to $A = A^{e_\beta}$ to construct a suitable function g. Then

$$\mathbb{E}[1_{A^{e_\beta}}(F_\beta(S_n))] \leq (1-\varepsilon)^{-1}\mathbb{E}[g \circ F_\beta(S_n)].$$

Step 3: The next step uses Stein's method to compare $\mathbb{E}[g \circ F_\beta(S_n)]$ and $\mathbb{E}[g \circ F_\beta(T_n)]$. The following argument is inspired by [9], Theorem 7. We first make some complimentary computations.

Lemma 4.3. *Let $\beta > 0$. For every $g \in C^3(\mathbb{R})$,*

$$\sum_{j,k=1}^p |\partial_j \partial_k (g \circ F_\beta)(x)| \leq \|g''\|_\infty + 2\|g'\|_\infty \beta, \tag{13}$$

$$\sum_{j,k,l=1}^p |\partial_j \partial_k \partial_l (g \circ F_\beta)(x)| \leq \|g'''\|_\infty + 6\|g''\|_\infty \beta + 6\|g'\|_\infty \beta^2. \tag{14}$$

Moreover, let $U_{jkl}(x) := \sup\{|\partial_j\partial_k\partial_l(g\circ F_\beta)(x+y)| : y \in \mathbb{R}^p, |y_j| \leq \beta^{-1}, 1 \leq \forall j \leq p\}$. Then

$$\sum_{j,k,l=1}^p U_{jkl}(x) \leq C(\|g'''\|_\infty + \|g''\|_\infty \beta + \|g'\|_\infty \beta^2). \tag{15}$$

Proof of Lemma 4.3. Let $\delta_{jk} = 1(j=k)$. A direct calculation gives

$$\partial_j F_\beta(x) = \pi_j(z),\ \partial_j\partial_k F_\beta(x) = \beta w_{jk}(x),\ \partial_j\partial_k\partial_l F_\beta(x) = \beta^2 q_{jkl}(x),$$

where

$$\pi_j(x) = e^{\beta x_j}/\textstyle\sum_{k=1}^p e^{\beta x_k},\ w_{jk}(x) = (\pi_j\delta_{jk} - \pi_j\pi_k)(x),$$
$$q_{jkl}(x) = (\pi_j\delta_{jl}\delta_{jk} - \pi_j\pi_l\delta_{jk} - \pi_j\pi_k(\delta_{jl}+\delta_{kl}) + 2\pi_j\pi_k\pi_l)(x).$$

By these expressions, we have

$$\pi_j(x) \geq 0,\ \sum_{j=1}^p \pi_j(x) = 1,\ \sum_{j,k=1}^p |w_{jk}(x)| \leq 2,\ \sum_{j,k,l=1}^p |q_{jkl}(x)| \leq 6.$$

Inequalities (13) and (14) follow from these relations and the following computation.

$$\begin{aligned}
&\partial_j(g\circ F_\beta)(x) = (g'\circ F_\beta)(x)\pi_j(x),\\
&\partial_j\partial_k(g\circ F_\beta)(x) = (g''\circ F_\beta)(x)\pi_j(x)\pi_k(x) + (g'\circ F_\beta)(x)\beta w_{jk}(x),\\
&\partial_j\partial_k\partial_l(g\circ F_\beta)(x) = (g'''\circ F_\beta)(x)\pi_j(x)\pi_k(x)\pi_l(x)\\
&\qquad\qquad + (g''\circ F_\beta)(x)\beta(w_{jk}(x)\pi_l(x) + w_{jl}(x)\pi_k(x) + w_{kl}(x)\pi_j(x))\\
&\qquad\qquad + (g'\circ F_\beta)(x)\beta^2 q_{jkl}(x).
\end{aligned}$$

For the last inequality (15), it is standard to see that whenever $|y_j| \le \beta^{-1}, 1 \le \forall j \le p$,

$$\pi_j(x+y) \le e^2\pi_j(x),$$

from which the desired inequality follows. ∎

For $i = 1, \dots, n$, let X_i' be an independent copy of X_i. Let I be a uniform random variable on $\{1, \dots, n\}$ independent of all the other variables. Define

$$S_n' = S_n - X_I + X_I'.$$

For $\lambda \in \mathbb{R}^p$,

$$\begin{aligned}
&\mathbb{E}[e^{\sqrt{-1}\lambda^T S_n'}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[e^{\sqrt{-1}\lambda^T(S_n - X_i + X_i')}]\\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[e^{\sqrt{-1}\lambda^T(S_n - X_i)}]\mathbb{E}[e^{\sqrt{-1}\lambda^T X_i'}] = \frac{1}{n}\sum_{i=1}^n \prod_{j\ne i}\mathbb{E}[e^{\sqrt{-1}\lambda^T X_j}]\mathbb{E}[e^{\sqrt{-1}\lambda^T X_i}]\\
&= \prod_{i=1}^n \mathbb{E}[e^{\sqrt{-1}\lambda^T X_i}] = \mathbb{E}[e^{\sqrt{-1}\lambda^T S_n}].
\end{aligned}$$

Hence $S_n' \stackrel{d}{=} S_n$. Also with $X_1^n = \{X_1, \dots, X_n\}$,

$$\begin{aligned}
\mathbb{E}[S_n' - S_n \mid X_1^n] &= \mathbb{E}[X_I' - X_I \mid X_1^n]\\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i' - X_i \mid X_1^n] = -n^{-1}S_n, \qquad (16)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[(S_n' - S_n)(S_n' - S_n)^T \mid X_1^n] = \mathbb{E}[(X_I' - X_I)(X_I' - X_I)^T \mid X_1^n]\\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i' - X_i)(X_i' - X_i)^T \mid X_1^n] = \frac{1}{n}\sum_{i=1}^n (\mathbb{E}[X_i X_i^T] + X_i X_i^T)\\
&= \frac{2}{n}\sum_{i=1}^n \mathbb{E}[X_i X_i^T] + \frac{1}{n}\sum_{i=1}^n (X_i X_i^T - \mathbb{E}[X_i X_i^T])\\
&= \frac{2}{n}\sum_{i=1}^n \mathbb{E}[X_i X_i^T] + n^{-1}V, \qquad (17)
\end{aligned}$$

where V is the $p \times p$ matrix defined by

$$V = (V_{jk})_{1 \le j,k \le p} = \frac{1}{n} \sum_{i=1}^{n} (X_i X_i^T - \mathbb{E}[X_i X_i^T]).$$

For the notational convenience, write $f = g \circ F_\beta$. Consider

$$h(x) = \int_0^1 \frac{1}{2t} \mathbb{E}[f(\sqrt{t}x + \sqrt{1-t}T_n) - f(T_n)]dt.$$

Then Lemma 1 of [40] implies

$$\sum_{j=1}^{p} x_j \partial_j h(x) - \sum_{j,k=1}^{p} \sum_{i=1}^{n} \mathbb{E}[X_{ij}X_{ik}] \partial_j \partial_k h(x) = f(x) - \mathbb{E}[f(T_n)],$$

and especially

$$\begin{aligned} \mathbb{E}[f(S_n)] - \mathbb{E}[f(T_n)] = \mathbb{E}\left[\sum_{j=1}^{p} \sum_{i=1}^{n} X_{ij} \partial_j h(S_n) \right] \\ - \mathbb{E}\left[\sum_{j,k=1}^{p} \sum_{i=1}^{n} \mathbb{E}[X_{ij}X_{ik}] \partial_j \partial_k h(S_n) \right]. \end{aligned} \tag{18}$$

Denote by $\nabla h(x)$ and $\operatorname{Hess} h(x)$ the gradient vector and the Hessian matrix of $h(x)$, respectively. Let

$$\begin{aligned} R = h(S_n') - h(S_n) - (S_n' - S_n)^T \nabla h(S_n) \\ - 2^{-1} (S_n' - S_n)^T (\operatorname{Hess} h(S_n))(S_n' - S_n). \end{aligned}$$

Then one has

$$\begin{aligned} 0 &= n\mathbb{E}[h(S_n') - h(S_n)] \quad (\text{as } S_n' \stackrel{d}{=} S_n) \\ &= n\mathbb{E}[(S_n' - S_n)^T \nabla h(S_n) + 2^{-1}(S_n' - S_n)^T (\operatorname{Hess} h(S_n))(S_n' - S_n) + R] \\ &= n\mathbb{E}\Big[\mathbb{E}[(S_n' - S_n)^T \mid X_1^n] \nabla h(S_n) \\ &\qquad + 2^{-1} \operatorname{Tr}\Big((\operatorname{Hess} h(S_n)) \mathbb{E}[(S_n' - S_n)(S_n' - S_n)^T \mid X_1^n] \Big) + R \Big] \\ &= \mathbb{E}\left[-\sum_{j=1}^{p} \sum_{i=1}^{n} X_{ij} \partial_j h(S_n) + \sum_{j,k=1}^{p} \sum_{i=1}^{n} \mathbb{E}[X_{ij}X_{ik}] \partial_j \partial_k h(S_n) \right] \\ &\quad + \mathbb{E}\left[\frac{1}{2} \sum_{j,k=1}^{p} V_{jk} \partial_j \partial_k h(S_n) + nR \right] \quad (\text{by (16) and (17)}) \\ &= -\mathbb{E}[f(S_n)] + \mathbb{E}[f(T_n)] + \mathbb{E}\left[\frac{1}{2} \sum_{j,k=1}^{p} V_{jk} \partial_j \partial_k h(S_n) + nR \right], \quad (\text{by (18)}) \end{aligned}$$

that is,

$$\mathbb{E}[f(S_n)] - \mathbb{E}[f(T_n)] = \mathbb{E}\left[\frac{1}{2}\sum_{j,k=1}^{p} V_{jk}\partial_j\partial_k h(S_n) + nR\right].$$

Using Lemma 4.3, one has

$$|\sum_{j,k=1}^{p} V_{jk}\partial_j\partial_k h(S_n)| \le \max_{1\le j,k\le p}|V_{jk}| \sum_{j,k=1}^{p}|\partial_j\partial_k h(S_n)| \le C\beta\delta^{-1}\max_{1\le j,k\le p}|V_{jk}|,$$

and with $\Delta_i := (\Delta_{i1},\dots,\Delta_{ip})^T := X_i' - X_i$,

$$\begin{aligned}|\mathbb{E}[nR]| &= \left|\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j,k,l=1}^{p}\Delta_{ij}\Delta_{ik}\Delta_{il}(1-\theta)^2\partial_j\partial_k\partial_l h(S_n+\theta\Delta_i)\right]\right| \\ &\qquad (\theta\sim U(0,1)\text{ independent of all the other variables}) \\ &\le \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j,k,l=1}^{p}|\Delta_{ij}\Delta_{ik}\Delta_{il}|\cdot|\partial_j\partial_k\partial_l h(S_n+\theta\Delta_i)|\right]. \qquad (19)\end{aligned}$$

Let $\chi_i = 1(\max_{1\le j\le p}|\Delta_{ij}| \le \beta^{-1})$ and $\chi_i^c := 1-\chi_i$. Then

$$(19) = \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}\chi_i *\right] + \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}\chi_i^c *\right] =: \frac{1}{2}\left[(\mathrm{A})+(\mathrm{B})\right].$$

Observe that

$$\begin{aligned}(\mathrm{A}) &\le \mathbb{E}\left[\sum_{j,k,l=1}^{p}\max_{1\le i\le n}(\chi_i\cdot|\partial_j\partial_k\partial_l h(S_n+\theta\Delta_i)|)\times\max_{1\le j,k,l\le p}\sum_{i=1}^{n}|\Delta_{ij}\Delta_{ik}\Delta_{il}|\right] \\ &\le C\beta^2\delta^{-1}\mathbb{E}\left[\max_{1\le j,k,l\le p}\sum_{i=1}^{n}|\Delta_{ij}\Delta_{ik}\Delta_{il}|\right] \quad (\text{by (15)}) \\ &\le C\beta^2\delta^{-1}\mathbb{E}\left[\max_{1\le j\le p}\sum_{i=1}^{n}|\Delta_{ij}|^3\right] \le C\beta^2\delta^{-1}\mathbb{E}\left[\max_{1\le j\le p}\sum_{i=1}^{n}|X_{ij}|^3\right] = C\beta^2\delta^{-1}B_2,\end{aligned}$$

and

$$\begin{aligned}(\mathrm{B}) &\le C\beta^2\delta^{-1}\sum_{i=1}^{n}\mathbb{E}\left[\chi_i^c\max_{1\le j\le p}|\Delta_{ij}|^3\right] \quad (\text{by (14)}) \\ &\le C\beta^2\delta^{-1}\sum_{i=1}^{n}\mathbb{E}\left[\chi_i^c\max_{1\le j\le p}|X_{ij}|^3\right]. \quad (\text{by symmetry})\end{aligned}$$

Because

$$\chi_i^c \le 1\left(\max_{1\le j\le p}|X_{ij}| > \beta^{-1}/2\right) + 1\left(\max_{1\le j\le p}|X_{ij}'| > \beta^{-1}/2\right),$$

we have

$$\mathbb{E}\left[\chi_i^c \max_{1\le j\le p}|X_{ij}|^3\right] \le \mathbb{E}\left[\max_{1\le j\le p}|X_{ij}|^3 \cdot 1\left(\max_{1\le j\le p}|X_{ij}| > \beta^{-1}/2\right)\right] + \mathbb{E}\left[\max_{1\le j\le p}|X_{ij}|^3\right] \cdot \mathbb{P}\left(\max_{1\le j\le p}|X_{ij}| > \beta^{-1}/2\right). \quad (20)$$

We here recall Chebyshev's association inequalities stated in the following lemma.

Lemma 4.4 (Chebyshev's association inequalities)**.** *Let φ and ψ be functions defined on an interval $\mathcal{I}$ in $\mathbb{R}$, and let ξ be a random variable such that $\mathbb{P}(\xi \in \mathcal{I}) = 1$. Suppose that $\mathbb{E}[|\varphi(\xi)|] < \infty, \mathbb{E}[|\psi(\xi)|] < \infty$ and $\mathbb{E}[|\varphi(\xi)\psi(\xi)|] < \infty$. Then $\mathrm{Cov}(\varphi(\xi), \psi(\xi)) \ge 0$ if φ and ψ are monotone in the same direction, and $\mathrm{Cov}(\varphi(\xi), \psi(\xi)) \le 0$ if φ and ψ are monotone in the opposite direction.*

Proof of Lemma 4.4. See, e.g., Theorem 2.14 in [4]. ∎

Since the maps $t \mapsto t^3$ and $t \mapsto 1(t > \beta^{-1}/2)$ are non-decreasing on $[0, \infty)$, the second term on the right side of (20) is not larger than the first term. Hence

$$\text{(B)} \le C\beta^2\delta^{-1}\sum_{i=1}^n \mathbb{E}\left[\max_{1\le j\le p}|X_{ij}|^3 \cdot 1\left(\max_{1\le j\le p}|X_{ij}| > \beta^{-1}/2\right)\right] = C\beta^2\delta^{-1}B_3.$$

Therefore,

$$|\mathbb{E}[f(S_n)] - \mathbb{E}[f(T_n)]| \le C\beta\delta^{-1}\{B_1 + \beta(B_2 + B_3)\}.$$

Step 4: Combining Steps 1-3, one has

$$\begin{aligned}\mathbb{P}(Z \in A) &\le (1-\varepsilon)^{-1}\mathbb{E}[g \circ F_\beta(T_n)] + \frac{C\beta\delta^{-1}\{B_1 + \beta(B_2 + B_3)\}}{1-\varepsilon} \\ &\le \mathbb{P}(F_\beta(T_n) \in A^{e_\beta + 3\delta}) + \frac{\varepsilon + C\beta\delta^{-1}\{B_1 + \beta(B_2 + B_3)\}}{1-\varepsilon} \\ &\qquad\qquad \text{(by construction of } g\text{)} \\ &\le \mathbb{P}(\widetilde{Z}^* \in A^{2e_\beta + 3\delta}) + \frac{\varepsilon + C\beta\delta^{-1}\{B_1 + \beta(B_2 + B_3)\}}{1-\varepsilon}. \text{ (Lemma 12)}\end{aligned}$$

This completes the proof. ∎

5. Inequalities for empirical processes

In this section, we shall provide some inequalities for empirical processes that will be used in the proofs of Theorem 2.1 and Lemma 2.2. These inequalities are of interest in their own rights. Consider the same setup as in Section 2, i.e., let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$ with common distribution P. Let $\mathcal{F}$ be a pointwise measurable class of functions $S \to \mathbb{R}$, to which a measurable envelope F is attached. In this section, however, *we do not assume that $\mathcal{F}$ is*

P-centered. Consider the empirical process $\mathbb{G}_n f = n^{-1/2}\sum_{i=1}^n (f(X_i) - Pf)$. Let $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}} Pf^2 \le \sigma^2 \le \|F\|_{P,2}^2$. Let $M = \max_{1\le i\le n} F(X_i)$.

Theorem 5.1 (A useful deviation inequality for suprema of empirical processes)**.** *Suppose that $F \in \mathcal{L}^q(P)$ for some $q \ge 2$. Then for every $t \ge 1$, with probability $> 1 - t^{-q/2}$,*

$$\|\mathbb{G}_n\|_{\mathcal{F}} \le (1+\alpha)\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + K(q)\Big[(\sigma + n^{-1/2}\|M\|_q)\sqrt{t} + \alpha^{-1}n^{-1/2}\|M\|_2 t\Big],\ \forall \alpha > 0,$$

where $K(q) > 0$ is a constant depending only on q.

Remark 5.1. Theorem 5.1 gives a deviation inequality for suprema of empirical processes that only requires finite moments of envelope functions. Talagrand's [50] inequality gives an exponential type deviation inequality for the supremum but requires uniform boundedness of $\mathcal{F}$, which is violated in our applications. Another known deviation inequality similar in nature to Theorem 5.1 is a Fuk-Nagaev type inequality proved in [20] (see their Theorem 3.1). For the purpose of this paper, however, Theorem 5.1 is more convenient.

Proof of Theorem 5.1. The theorem essentially follows from [5], Theorem 12, which states that

$$\|(\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}])_+\|_q \lesssim \sqrt{q}(\Sigma + \sigma) + qn^{-1/2}(\|M\|_q + \sigma),$$

where $\Sigma^2 = \mathbb{E}[\|n^{-1}\sum_{i=1}^n (f(X_i) - Pf)^2\|_{\mathcal{F}}]$. By Lemma 7 of the same paper,

$$\Sigma^2 \le \sigma^2 + 64n^{-1/2}\|M\|_2\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + 32n^{-1}\|M\|_2^2.$$

Hence, using the simple inequality $2\sqrt{ab} \le \beta a + \beta^{-1}b, \forall \beta > 0$, one has

$$\|(\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}])_+\|_q \lesssim \sqrt{q}\beta\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + \sqrt{q}(1+\beta^{-1})n^{-1/2}\|M\|_2 + \sqrt{q}\sigma + qn^{-1/2}(\|M\|_q + \sigma).$$

Therefore, by Markov's inequality, for every $t \ge 1$, with probability $> 1 - t^{-q}$,

$$\begin{aligned}\|\mathbb{G}_n\|_{\mathcal{F}} &\le \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + (\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}])_+ \\ &\le (1 + C\sqrt{q}\beta t)\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] \\ &\quad + C\sqrt{q}(1+\beta^{-1})n^{-1/2}\|M\|_2 t \\ &\quad + C\sqrt{q}\sigma t + Cqn^{-1/2}(\|M\|_q + \sigma)t,\ \forall \beta > 0.\end{aligned}$$

The final conclusion follows from taking $\beta = C^{-1}q^{-1/2}t^{-1}\alpha$. ∎

The proof of Lemma 2.2 relies on the following moment inequality for suprema of empirical processes, which is an extension of [53], Theorem 2.1, to possibly unbounded classes of functions (Theorem 3.1 of [53] derives a moment inequality applicable to the case where the envelope F has $q > 4$

moments, but the form of the inequality in Theorem 5.2 is more convenient in our applications; note that Theorem 5.2 only requires $F \in \mathcal{L}^2(P)$, as opposed to $F \in \mathcal{L}^q(P)$ with $q > 4$ in Theorem 3.1 of [53], and Theorem 5.2 is not covered by [53]). Recall the uniform entropy integral $J(\delta, \mathcal{F}, F)$.

Theorem 5.2 (A useful maximal inequality)**.** *Suppose that $F \in \mathcal{L}^2(P)$. Let $\delta = \sigma/\|F\|_{P,2}$. Then*

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim J(\delta, \mathcal{F}, F)\|F\|_{P,2} + \frac{\|M\|_2 J^2(\delta, \mathcal{F}, F)}{\delta^2\sqrt{n}}.$$

In Appendix, we give a full proof of Theorem 5.2 for the sake of completeness, although the the proof is essentially similar to the proof of Theorem 2.1 in [53].

The bound in Theorem 5.2 will be explicit as soon as a suitable bound on the covering number is available. For example, the following corollary is an extension of [25], Proposition 2.1. For its proof, see Appendix A.3.

Corollary 5.1 (Maximal inequality specialized to VC type classes)**.** *Consider the same setup as in Theorem 5.2. Suppose that there exist constants $A \geq e$ and $v \geq 1$ such that*

$$\sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}) \leq (A/\varepsilon)^v, \ 0 < \forall \varepsilon \leq 1.$$

Then

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim \sqrt{v\sigma^2 \log\left(\frac{A\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_2}{\sqrt{n}} \log\left(\frac{A\|F\|_{P,2}}{\sigma}\right).$$

6. Proofs of Theorem 2.1, Lemmas 2.3 and 2.4

6.1. **Proof of Theorem 2.1.** We make use of Lemma 4.1 to prove the theorem. Construct a tight Gaussian random element G_P in $\ell^\infty(\mathcal{F})$ given in assumption (A3), independent of $X_1, \ldots, X_n$. We note that one can extend G_P to the linear hull of $\mathcal{F}$ in such a way that G_P has linear sample paths [see 18, Theorem 3.1.1]. Let $\{f_1, \ldots, f_N\}$ be a minimal $\varepsilon\|F\|_{P,2}$-net of $(\mathcal{F}, e_P)$ with $N = N(\mathcal{F}, e_P, \varepsilon\|F\|_{P,2})$. Then for every $f \in \mathcal{F}$, there exists a function $f_j, 1 \leq j \leq N$ such that $e_P(f, f_j) < \varepsilon\|F\|_{P,2}$. Recall $\mathcal{F}_\varepsilon = \{f - g : f, g \in \mathcal{F}, e_P(f, g) < \varepsilon\|F\|_{P,2}\}$ and define

$$Z^\varepsilon = \max_{1\leq j\leq N} \mathbb{G}_n f_j, \ \widetilde{Z}^* = \sup_{f\in\mathcal{F}} G_P f, \ \widetilde{Z}^{*\varepsilon} = \max_{1\leq j\leq N} G_P f_j.$$

Observe that

$$|Z - Z^\varepsilon| \leq \|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon}, \ |\widetilde{Z}^{*\varepsilon} - \widetilde{Z}^*| \leq \|G_P\|_{\mathcal{F}_\varepsilon}.$$

We shall apply Corollary 4.1 to Z^ε. Recall that $\log(N \vee n) = H_n(\varepsilon)$. Then for every Borel subset A of $\mathbb{R}$ and $\delta > 0$,

$$\mathbb{P}(Z^\varepsilon \in A) - \mathbb{P}(\widetilde{Z}^{*\varepsilon} \in A^{16\delta}) \lesssim \delta^{-2}\{B_1 + \delta^{-1}(B_2 + B_4)H_n(\varepsilon)\}H_n(\varepsilon) + n^{-1}\log n,$$

where

$$B_1 = n^{-1}\mathbb{E}\left[\max_{1\le j,k\le N}|\sum_{i=1}^n (f_j(X_i)f_k(X_i) - P(f_jf_k))|\right],$$
$$B_2 = n^{-3/2}\mathbb{E}\left[\max_{1\le j\le N}\sum_{i=1}^n |f_j(X_i)|^3\right],$$
$$B_4 = n^{-1/2}\mathbb{E}\left[\max_{1\le j\le N}|f_j(X_1)|^3\cdot 1\left(\max_{1\le j\le N}|f_j(X_1)| > \delta\sqrt{n}H_n(\varepsilon)^{-1}\right)\right].$$

Clearly $B_1 \le n^{-1/2}\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}.\mathcal{F}}], B_2 \le n^{-1/2}\kappa^3$, and

$$B_4 \le n^{-1/2}P[F^3 1(F > \delta\sqrt{n}H_n(\varepsilon)^{-1})].$$

Hence choosing $\delta > 0$ in such a way that

$$C\delta^{-2}n^{-1/2}\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}.\mathcal{F}}]H_n(\varepsilon) \le \frac{\gamma}{4},\ C\delta^{-3}n^{-1/2}\kappa^3 H_n^2(\varepsilon) \le \frac{\gamma}{4},$$

that is,

$$\delta \ge C\max\left\{\gamma^{-1/2}n^{-1/4}(\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}.\mathcal{F}}])^{1/2}H_n^{1/2}(\varepsilon), \gamma^{-1/3}n^{-1/6}\kappa H_n^{2/3}(\varepsilon)\right\},$$

we have

$$\mathbb{P}(Z^\varepsilon \in A) \le \mathbb{P}(\widetilde{Z}^{*\varepsilon} \in A^{16\delta}) + \frac{\gamma}{2} + \frac{\gamma}{4}\kappa^{-3}P[F^3 1(F > \delta\sqrt{n}H_n(\varepsilon)^{-1})] + \frac{C\log n}{n}.$$

Note that $\delta \ge c\gamma^{-1/3}n^{-1/6}\kappa H_n^{2/3}(\varepsilon)$, so that

$$P[F^3 1(F > \delta\sqrt{n}H_n(\varepsilon)^{-1})] \le P[F^3 1(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\varepsilon)^{-1/3})].$$

Hence

$$\begin{aligned}\mathbb{P}(Z^\varepsilon \in A) &\le \mathbb{P}(\widetilde{Z}^{*\varepsilon} \in A^{16\delta}) + \frac{\gamma}{2} \\ &\quad + \frac{\gamma}{4}P[(F/\kappa)^3 1(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\varepsilon)^{-1/3})] + \frac{C\log n}{n} \\ &=: \mathbb{P}(\widetilde{Z}^{*\varepsilon} \in A^{16\delta}) + \frac{\gamma}{2} + \text{error}. \end{aligned} \tag{21}$$

By Theorem 5.1, with probability $> 1 - \gamma/4$,

$$\begin{aligned}\|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon} \le K(q)\big\{\phi_n(\varepsilon) + (\varepsilon\|F\|_{P,2} + n^{-1/2}\|M\|_q)\gamma^{-1/q} \\ + n^{-1/2}\|M\|_2\gamma^{-2/q}\big\} =: a,\end{aligned} \tag{22}$$

where $K(q)$ is a constant that depends only on q. Moreover, by the Borell-Sudakov-Tsirel'son inequality [52, Proposition A.1], with probability $> 1 - \gamma/4$, we have

$$\|G_P\|_{\mathcal{F}_\varepsilon} \le \phi_n(\varepsilon) + \varepsilon\|F\|_{P,2}\sqrt{2\log(4/\gamma)} =: b. \tag{23}$$

Therefore, for every Borel subset A of $\mathbb{R}$,

$$\begin{aligned}\mathbb{P}(Z \in A) &\leq \mathbb{P}(Z^\varepsilon \in A^a) + \frac{\gamma}{4} \quad \text{(by (22))} \\ &\leq \mathbb{P}(\widetilde{Z}^{*\varepsilon} \in A^{a+16\delta}) + \frac{3}{4}\gamma + \text{error} \quad \text{(by (21))} \\ &\leq \mathbb{P}(\widetilde{Z}^* \in A^{a+b+16\delta}) + \gamma + \text{error}. \quad \text{(by (23))}\end{aligned}$$

The conclusion follows from Lemma 4.1. ∎

6.2. **Proof of Lemma 2.3.** The proof of Lemma 2.3 depends on the following lemma on *anti-concentration* of suprema of Gaussian processes.

Lemma 6.1 (An anti-concentration inequality)**.** *Let $(S, \mathcal{S}, P)$ be a probability space, and let $\mathcal{F} \subset \mathcal{L}^2(P)$ be a P-pre-Gaussian class of functions. Denote by G_P a tight Gaussian random element in $\ell^\infty(\mathcal{F})$ with mean zero and covariance function $\mathbb{E}[G_P(f)G_P(g)] = \mathrm{Cov}_P(f,g)$ for all $f, g \in \mathcal{F}$ where $\mathrm{Cov}_P(\cdot,\cdot)$ denotes the covariance under P. Suppose that there exist constants $\underline{\sigma}, \bar{\sigma} > 0$ such that $\underline{\sigma}^2 \leq \mathrm{Var}_P(f) \leq \bar{\sigma}^2$ for all $f \in \mathcal{F}$. Then for every $\epsilon > 0$,*

$$\sup_{x \in \mathbb{R}} \mathbb{P}\left\{\left|\sup_{f\in\mathcal{F}} G_P f - x\right| \leq \epsilon\right\} \leq C_\sigma \epsilon \left\{\mathbb{E}\left[\sup_{f\in\mathcal{F}} G_P f\right] + \sqrt{1 \vee \log(\underline{\sigma}/\epsilon)}\right\},$$

where C_σ is a constant depending only on $\underline{\sigma}$ and $\bar{\sigma}$.

Proof of Lemma 6.1. The proof of this lemma is the same as that of Theorem 2.1 in [12] with the exception that we now apply Theorem 3, part (ii) instead of Theorem 3, part (i) from [11]. ∎

Going back to the proof of Lemma 2.3, for every $t \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{P}(Z \leq t) &= \mathbb{P}(\{Z \leq t\} \cap \{|Z - \widetilde{Z}| \leq r_1\}) + \mathbb{P}(\{Z \leq t\} \cap \{|Z - \widetilde{Z}| > r_1\}) \\ &\leq \mathbb{P}(\widetilde{Z} \leq t + r_1) + r_2 \\ &\leq \mathbb{P}(\widetilde{Z} \leq t) + C_\sigma r_1 \{\mathbb{E}[\widetilde{Z}] + \sqrt{1 \vee \log(\underline{\sigma}/r_1)}\} + r_2,\end{aligned}$$

where we have used Lemma 6.1 to deduce the last inequality. A similar argument leads to the reverse inequality. This completes the proof. ∎

6.3. **Proof of Lemma 2.4.** Take $\beta_n \to \infty$ sufficiently slowly such that $\beta_n r_n (1 \vee \mathbb{E}[\widetilde{Z}_n]) = o(1)$. Then since $\mathbb{P}(|Z_n - \widetilde{Z}_n| > \beta_n r_n) = o(1)$, by Lemma 2.3, we have

$$\sup_{t\in\mathbb{R}} |\mathbb{P}(Z_n \leq t) - \mathbb{P}(\widetilde{Z}_n \leq t)| = O\{r_n(\mathbb{E}[\widetilde{Z}_n] + |\log(\beta_n r_n)|)\} + o(1) = o(1).$$

This completes the proof. ∎

References

[1] Belloni, A., Chernozhukov, V. and Fernández-Val, I. (2011). Conditional quantile processes based on series or many regressors. arXiv:1105.6154.

[2] Berthet, P. and Mason, D.M. (2006). Revisiting two strong approximation results of Dudley and Philipp. In: *High Dimensional Probability*, IMS Lecture Notes-Monograph Series, Vol. 51, pp.155-172.

[3] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071-1095.

[4] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

[5] Boucheron, S., Bousquet, O., Lugosi, G. and Massart, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514-560.

[6] Boyd, S. and Vanderberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[7] Bretagnolle, J. and Massart, P. (1989). Hungarian construction from the non asymptotic viewpoint. *Ann. Probab.* **17** 239-256.

[8] Chatterjee, S. (2005). An error bound in the Sudakov-Fernique inequality. arXiv:math/0510424.

[9] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.* **4** 257-283.

[10] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, to appear.

[11] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. arXiv:1301.4807v3.

[12] Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Anti-concentration and adaptive honest confidence bands. arXiv:1303:7152.

[13] Chernozhukov, V., Lee, S., and Rosen, A. (2013). Intersection bounds: estimation and inference. arXiv:0907.3503v4. To appear in *Econometrica*.

[14] Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54-81.

[15] Csörgo, M. and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics*. Wiley.

[16] Deheuvels, P. and Mason, D.M. (1994). Functional laws of the iterated logarithm for local empirical processes indexed by sets. *Ann. Probab.* **22** 1619-1661.

[17] Dehling, H. (1983). Limit theorems for sums of weakly dependent Banach space valued random variables. *Z. Warhsch. Verw. Gabiete* **63**

393-432.

[18] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

[19] Dudley, R.M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Warhsch. Verw. Gabiete* **62** 509-552.

[20] Einmahl, U. and Li, D. (2008). Characterization of LIL behavior in Banach space. *Trans. Amer. Math. Soc.* **360** 6677-6693.

[21] Einmahl, U. and Mason, D.M. (1997). Gaussian approximation of local empirical processes indexed by functions. *Probab. Theory Related Fields* **107** 283-311.

[22] Einmahl, U. and Mason, D.M. (1998). Strong approximations to the local empirical process. In: *High Dimensional Probability* (eds. E. Eberlein, M. Hahn and M. Talagrand) pp. 75-92.

[23] Einmahl, U. and Mason, D.M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13** 1-37.

[24] Einmahl, U. and Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380-1403.

[25] Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist.* **37** 503-522.

[26] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122-1170.

[27] Ghosal, S., Sen, A. and van der Vaart, A.W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28** 1054-1082.

[28] He, X. and Shao, Q.-M. (2000). On parameters on increasing dimensions. *J. Multivariate Anal.* **73** 125-135.

[29] Huang, J.Z. (2003). Asymptotics for polynomial spline regression under weak conditions. *Statist. Probab. Lett.* **65** 207-216.

[30] Koenker, R. and Bassett G.W. (1978). Regression quantiles. *Econometrica* **46** 33-50.

[31] Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Atti. Giorn.* **4** 83-91.

[32] Koltchinskii, V.I. (1994). Komlós-Major-Tusnády approximation for the general empirical process and Haar expansions of classes of functions. *J. Theoret. Probab.* **7** 73-118.

[33] Komlós, J., Major, P., and Tusnády, G. (1975). An approximation for partial sums of independent rv's and the sample df I. *Z. Warhsch. Verw. Gabiete* **32** 111-131.

[34] Konakov, V.D. and Piterbarg, V.I. (1984). On the convergence rate of maximal deviations distributions for kernel regression estimates. *J. Multivariate Anal.* **15** 279-294.

[35] Le Cam, L. (1988). On the Prokhorov distance between the empirical process and the associated Gaussian bridge. Technical Report No. 170,

Department of Statistics, University of California, Berkeley.
[36] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer.
[37] Mason, D.M. (2004). A uniform functional law of the logarithm for the local empirical process. *Ann. Probab.* **32** 1391-1418.
[38] Mason, D.M. and van Zwet, W.R. (1987). A refinement of the KMT inequality for the uniform empirical process. *Ann. Probab.* **15** 871-884.
[39] Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT construction. *Ann. Probab.* **17** 266-291.
[40] Meckes, E. (2009). On Stein's method for multivariate normal approximation. In: *High Dimensional Probability V: The Luminy Volume*, IMS Collections, Vol.5, pp.159-178.
[41] Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147-168.
[42] Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *Ann. Statist.* **15** 780-799.
[43] Norvaiša, R. and Paulauskas, V. (1991). Rate of convergence in the Central Limit Theorem for empirical processes. *J.Theoret. Probab.* **4** 511-534.
[44] Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.
[45] Reinert, G. and Röllin, A. (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.* **37** 2150-2173.
[46] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21-45.
[47] Settati, A. (2009). Gaussian approximation of the empirical process under random entropy conditions. *Stochastic Process. Appl.* **119** 1541-1560.
[48] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: *Proc. of the Sixth Berkeley Symp. on Math. Statist. and Probab.*, Vol. II: Probability theory. pp.583-602.
[49] Stein, C. (1986). *Approximate Computation of Expectations*. IMS Lecture Notes-Monograph Series, Vol.7.
[50] Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505-563.
[51] Talagrand, M. (2005). *The Generic Chaining*. Springer.
[52] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
[53] van der Vaart, A.W. and Wellner, J.A. (2011). A local maximal inequality under uniform entropy. *Electronic J. Statist.* **5** 192-203.
[54] Yurinskii, V.V. (1977). On the error of the Gaussian approximation for convolutions. *Theory of Probability and Its Applications* **2** 236-247.

[55] Zaitsev, Y. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of S.N. Bernstein's inequality conditions. *Probab. Theory Related Fields* **74** 535-566.

Appendix A. Additional proofs

A.1. **Proof of Lemma 2.1.** We first note that by approximation [see 52, Problem 2.5.1], assumption (A4) implies that

$$\int_0^1 \sqrt{\log N(\mathcal{F}, e_P, \varepsilon \|F\|_{P,2})} d\varepsilon < \infty.$$

Let G_P be a centered Gaussian process indexed by $\mathcal{F}$ with covariance function $\mathbb{E}[G_P(f)G_P(g)] = P(fg)$. Recall that $\mathcal{F}$ is P-centered, and by Example 1.3.10 in [52], $\mathcal{F}$ is P-pre-Gaussian if and only if $(\mathcal{F}, e_P)$ is totally bounded and G_P has a version that has sample paths almost surely uniformly e_P-continuous. Dudley's criterion for sample continuity of Gaussian processes states that when

$$\int_0^\infty \sqrt{\log N(\mathcal{F}, e_P, \varepsilon)} d\varepsilon < \infty, \tag{24}$$

there exists a version of G_P that has sample paths uniformly e_P-continuous [52, p.100-101] (note that (24) implies that $N(\mathcal{F}, e_P, \varepsilon)$ is finite for every $\varepsilon > 0$, i.e., $\mathcal{F}$ is totally bounded for e_P). The lemma readily follows from these observations. ∎

A.2. **Proof of Theorem 5.2.** We first prove the following technical lemma.

Lemma A.1. *Write $J(\delta)$ for $J(\delta, \mathcal{F}, F)$ and suppose that $J(1)$ is finite (and hence $J(\delta)$ is finite for all δ). Then (i) the map $\delta \mapsto J(\delta)$ is concave; (ii) $J(c\delta) \le cJ(\delta)$, $\forall c \ge 1$; (iii) the map $\delta \mapsto J(\delta)/\delta$ is non-increasing; (iv) the map $[0,\infty) \times (0,\infty) \ni (x,y) \mapsto J(\sqrt{x/y})\sqrt{y}$ is concave.*

Proof. Let $\lambda(\varepsilon) = \sup_Q \sqrt{1 + \log N(\mathcal{F}, e_Q, \varepsilon \|F\|_{Q,2})}$. Part (i) follows from the fact that the map $\varepsilon \mapsto \lambda(\varepsilon)$ is non-increasing. Part (ii) follows from the inequality

$$\int_0^{c\delta} \lambda(\varepsilon) d\varepsilon = c \int_0^{\delta} \lambda(c\varepsilon) d\varepsilon \le c \int_0^{\delta} \lambda(\varepsilon) d\varepsilon.$$

Part (iii) follows from the identity

$$\frac{J(\delta)}{\delta} = \int_0^1 \lambda(\delta\varepsilon) d\varepsilon.$$

The proof of part (iv) uses some facts in convex analysis. Proofs of the following lemmas can be found in, e.g., [6], Section 3.2.

Lemma A.2. *Let D be a convex subset of $\mathbb{R}^n$, and let $f : D \to \mathbb{R}$ be a concave function. Then the* perspective *$(x,t) \mapsto tf(x/t)$, $\{(x,t) \in \mathbb{R}^{n+1} : x/t \in D, t > 0\} \to \mathbb{R}$, is also concave.*

Lemma A.3. *Let D_1 be a convex subset of $\mathbb{R}^n$, and let $g_i : D_1 \to \mathbb{R}, 1 \le i \le k$ be concave functions. Let D_2 denote the convex hull of the set $\{(g_1(x), \dots, g_k(x)) : x \in D_1\}$. Let $h : D_2 \to \mathbb{R}$ be concave and nondecreasing in each coordinate. Then $f(x) = h(g_1(x), \dots, g_k(x)), D_1 \to \mathbb{R}$, is concave.*

Let $h(s,t) = J(s/t)t$, $g_1(x,y) = \sqrt{x}$ and $g_2(x,y) = \sqrt{y}$. Then h is concave and nondecreasing in each coordinate, and $g_i, i = 1,2$ are concave. Hence $J(\sqrt{x/y})\sqrt{y} = h(g_1(x,y), g(x,y))$ is concave. ∎

We will use a version of the contraction principle for Rademacher averages. Recall that a Rademacher random variable is a random variables taking ± 1 with equal probability.

Lemma A.4 (A contraction principle, [36]). *Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables independent of $X_1, \ldots, X_n$. Then*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i f^2(X_i)\right\|_{\mathcal{F}}\right] \leq 4\mathbb{E}\left[M\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right].$$

Proof. See [36], Theorem 4.12, and the discussion following the theorem. ∎

We will also use the following form of the Hoffmann-Jørgensen inequality.

Theorem A.1 (A Hoffmann-Jørgensen-type inequality, [36]). *Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables independent of $X_1, \ldots, X_n$. Then for every $1 < q < \infty$,*

$$\left(\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}^q\right]\right)^{1/q} \lesssim_q \mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right] + \|M\|_q.$$

Proof. See, e.g., [36], Theorem 6.20. ∎

We are now in position to prove Theorem 5.2.

Proof of Theorem 5.2. We may assume that $J(1)$ is finite since otherwise $J(\delta)$ is infinite and there is nothing to prove. Moreover, without loss of generality, we may assume that F is everywhere positive. Let P_n denote the empirical distribution that assigns probability n^{-1} to each X_i. Let $\sigma_n^2 = \sup_{f\in\mathcal{F}} n^{-1}\sum_{i=1}^n f^2(X_i)$. For i.i.d. Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_n$ independent of $X_1, \ldots, X_n$, the symmetrization inequality gives

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] \leq 2\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right].$$

Here the standard entropy integral inequality gives

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{F}} \mid X_1, \ldots, X_n\right] &\leq C\int_0^{\sigma_n} \sqrt{1 + \log N(\mathcal{F}, e_{P_n}, \varepsilon)}d\varepsilon \\
&\leq C\|F\|_{P_n,2}\int_0^{\sigma_n/\|F\|_{P_n,2}} \sqrt{1 + \log N(\mathcal{F}, e_{P_n}, \varepsilon\|F\|_{P_n,2})}d\varepsilon \\
&\leq C\|F\|_{P_n,2} J(\sigma_n/\|F\|_{P_n,2}).
\end{aligned}$$

Hence by Lemma A.1 (iv) and Jensen's inequality,

$$Z := \mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}}\right] \leq C\|F\|_{P,2}J(\sqrt{\mathbb{E}[\sigma_n^2]}/\|F\|_{P,2}).$$

By the symmetrization inequality, the contraction principle (Lemma A.4) and the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}[\sigma_n^2] &\leq \sigma^2 + \mathbb{E}\left[\left\|\mathbb{E}_n[(f^2(X_i) - Pf^2)]\right\|_{\mathcal{F}}\right] \leq \sigma^2 + 2\mathbb{E}\left[\left\|\mathbb{E}_n[\varepsilon_i f^2(X_i)]\right\|_{\mathcal{F}}\right] \\ &\leq \sigma^2 + 8\mathbb{E}\left[M\left\|\mathbb{E}_n[\varepsilon_i f(X_i)]\right\|_{\mathcal{F}}\right] \leq \sigma^2 + 8\|M\|_2\left(\mathbb{E}\left[\left\|\mathbb{E}_n[\varepsilon_i f(X_i)]\right\|_{\mathcal{F}}^2\right]\right)^{1/2}.\end{aligned}$$

Here by the Hoffmann-Jørgensen inequality (Theorem A.1),

$$\left(\mathbb{E}\left[\left\|\mathbb{E}_n[\varepsilon_i f(X_i)]\right\|_{\mathcal{F}}^2\right]\right)^{1/2} \lesssim \mathbb{E}\left[\left\|\mathbb{E}_n[\varepsilon_i f(X_i)]\right\|_{\mathcal{F}}\right] + n^{-1}\|M\|_2,$$

so that,

$$\sqrt{\mathbb{E}[\sigma_n^2]} \leq C\|F\|_{P,2}(\Delta \vee \sqrt{DZ}),$$

where $\Delta^2 := \max\{\sigma^2, n^{-1}\|M\|_2^2\}/\|F\|_{P,2}^2 \geq \delta^2$ and $D := \|M\|_2/(\sqrt{n}\|F\|_{P,2}^2)$. Therefore, using Lemma A.1 (ii), we have

$$Z \leq C\|F\|_{P,2}J(\Delta \vee \sqrt{DZ})$$

We consider the following two cases:

(i) $\sqrt{DZ} \leq \Delta$. In this case, $J(\Delta \vee \sqrt{DZ}) \leq J(\Delta)$, so that $Z \leq C\|F\|_{P,2}J(\Delta)$. Since the map $\delta \mapsto J(\delta)/\delta$ is non-increasing (Lemma A.1 (iii)),

$$J(\Delta) = \Delta\frac{J(\Delta)}{\Delta} \leq \Delta\frac{J(\delta)}{\delta} = \max\left\{J(\delta), \frac{\|M\|_2 J(\delta)}{\sqrt{n}\delta\|F\|_{P,2}}\right\}.$$

Since $J(\delta)/\delta \geq J(1) \geq 1$, the last expression is bounded by

$$\max\left\{J(\delta), \frac{\|M\|_2 J^2(\delta)}{\sqrt{n}\delta^2\|F\|_{P,2}}\right\}.$$

(ii) $\sqrt{DZ} \geq \Delta$. In this case, $J(\Delta \vee \sqrt{DZ}) \leq J(\sqrt{DZ})$, and since the map $\delta \mapsto J(\delta)/\delta$ is non-increasing (Lemma A.1 (iii)),

$$J(\sqrt{DZ}) = \sqrt{DZ}\frac{J(\sqrt{DZ})}{\sqrt{DZ}} \leq \sqrt{DZ}\frac{J(\Delta)}{\Delta} \leq \sqrt{DZ}\frac{J(\delta)}{\delta}.$$

Therefore,

$$Z \leq C\|F\|_{P,2}\sqrt{DZ}\frac{J(\delta)}{\delta},$$

that is

$$Z \leq C\|F\|_{P,2}^2 D\frac{J^2(\delta)}{\delta^2} = \frac{C\|M\|_2 J^2(\delta)}{\sqrt{n}\delta^2}.$$

This completes the proof. ∎

A.3. **Proof of Corollary 5.1.** Observe that

$$J(\delta) \leq \int_0^\delta \sqrt{1 + v \log(A/\varepsilon)} d\varepsilon \leq A\sqrt{v} \int_{A/\delta}^\infty \frac{\sqrt{1+\log \varepsilon}}{\varepsilon^2} d\varepsilon.$$

An integration by parts gives

$$\begin{aligned}\int_c^\infty \frac{\sqrt{1+\log \varepsilon}}{\varepsilon^2} d\varepsilon &= \left[-\frac{\sqrt{1+\log \varepsilon}}{\varepsilon}\right]_c^\infty + \frac{1}{2}\int_c^\infty \frac{1}{\varepsilon^2\sqrt{1+\log \varepsilon}} d\varepsilon \\ &\leq \frac{\sqrt{1+\log c}}{c} + \frac{1}{2}\int_c^\infty \frac{\sqrt{1+\log \varepsilon}}{\varepsilon^2} d\varepsilon, \text{ if } c \geq e.\end{aligned}$$

by which we have

$$\int_c^\infty \frac{\sqrt{1+\log \varepsilon}}{\varepsilon^2} d\varepsilon \leq \frac{2\sqrt{1+\log c}}{c} \leq \frac{2\sqrt{2}\sqrt{\log c}}{c}, \text{ if } c \geq e,$$

Since $A/\delta \geq A \geq e$, we have

$$J(\delta) \leq 2\sqrt{2v}\delta\sqrt{\log(A/\delta)}.$$

Applying Theorem 5.2, we obtain the desired conclusion. ∎

A.4. **Proof of Lemma 2.2.** Before proving Lemma 2.2, we shall recall the following lemmas. The proofs of these lemmas are implicit in [52], Section 2.10.3, and hence omitted.

Lemma A.5. *Let $\mathcal{F}$ and $\mathcal{G}$ be classes of measurable functions $S \to \mathbb{R}$, to which measurable envelopes F and G are attached, respectively. Denote by $\mathcal{F} \cdot \mathcal{G}$ the pointwise product of $\mathcal{F}$ and $\mathcal{G}$. Then for every $0 < \varepsilon \leq 1$,*

$$\sup_Q N(\mathcal{F}\cdot\mathcal{G}, e_Q, 2\varepsilon\|FG\|_{Q,2}) \leq \sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}) \sup_Q N(\mathcal{G}, e_Q, \varepsilon\|G\|_{Q,2}),$$

where the suprema are taken over all finitely discrete probability measures Q on $(S, \mathcal{S})$.

Lemma A.6. *Let $\mathcal{F}$ be a class of measurable functions $S \to \mathbb{R}$, to which a measurable envelope F is attached. For every $q \geq 1$, let $\mathcal{F}(q) = \{|f|^q : f \in \mathcal{F}\}$. Then*

$$\sup_Q N(\mathcal{F}(q), e_Q, q\varepsilon\|F^q\|_{Q,2}) \leq \sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}), \ 0 < \forall\varepsilon \leq 1,$$

where the suprema are taken over all finitely discrete probability measures Q on $(S, \mathcal{S})$.

Proof of Lemma 2.2. For the first inequality, noting that $J(\delta, \mathcal{F}_\varepsilon, 2F) \lesssim J(\delta, \mathcal{F}, F) = J(\delta)$, by Theorem 5.2, we have

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon}] \lesssim J(\varepsilon)\|F\|_{P,2} + n^{-1/2}\varepsilon^{-2}J^2(\varepsilon)\|M\|_2.$$

Moreover, by Dudley's inequality [52, Corollary 2.2.8], $\mathbb{E}[\|G_P\|_{\mathcal{F}_\varepsilon}] \lesssim J(\varepsilon)\|F\|_{P,2}$. Note that by approximation [see 52, Problem 2.5.1], we have

$$\int_0^\delta \sqrt{1 + \log N(\mathcal{F}, e_P, \tau\|F\|_{P,2})} d\tau \lesssim J(\delta).$$

Hence the first inequality is proved.

The third inequality is deduced from Theorem 5.2 together with the covering number estimate (5). Hence we shall prove the second inequality. We first observe that

$$\mathbb{E}_n[|f(X_i)|^3] = P|f|^3 + n^{-1/2}\mathbb{G}_n(|f|^3),$$

by which we have

$$\mathbb{E}\left[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}\right] \le \sup_{f\in\mathcal{F}} P|f|^3 + n^{-1/2}\mathbb{E}[\|\mathbb{G}_n(|f|^3)\|_{\mathcal{F}}].$$

Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. Rademacher random variables independent of $X_1, \dots, X_n$. By the symmetrization inequality,

$$\mathbb{E}[\|\mathbb{G}_n(|f|^3)\|_{\mathcal{F}}] \le 2\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i |f(X_i)|^3\right\|_{\mathcal{F}}\right].$$

By the contraction principle together with the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i |f(X_i)|^3\right\|_{\mathcal{F}}\right] &\lesssim \mathbb{E}\left[M^{3/2}\left\|\sum_{i=1}^n \varepsilon_i |f(X_i)|^{3/2}\right\|_{\mathcal{F}}\right] \\ &\le \|M\|_3^{3/2}\left(\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i |f(X_i)|^{3/2}\right\|_{\mathcal{F}}^2\right]\right)^{1/2}.\end{aligned}$$

Moreover, by the Hoffmann-Jørgensen inequality,

$$\left(\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i |f(X_i)|^{3/2}\right\|_{\mathcal{F}}^2\right]\right)^{1/2} \lesssim \mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i |f(X_i)|^{3/2}\right\|_{\mathcal{F}}\right] + \|M\|_3^{3/2}.$$

By Theorem 5.2 together with Lemma A.6, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i |f(X_i)|^{3/2}\right\|_{\mathcal{F}}\right] &\lesssim J(\delta_3^{3/2}, \mathcal{F}, F)\|F^{3/2}\|_{P,2} \\ &\quad + \frac{\|M^{3/2}\|_2 J^2(\delta_3^{3/2}, \mathcal{F}, F)}{\sqrt{n}\delta_3^3},\end{aligned}$$

by which we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}\right] - \sup_{f\in\mathcal{F}} P|f|^3 &\lesssim n^{-1}\|M\|_3^3 \\ &\quad + n^{-1/2}\|M\|_3^{3/2}\left[J(\delta_3^{3/2}, \mathcal{F}, F)\|F\|_{P,3}^{3/2} + \frac{\|M\|_3^{3/2} J^2(\delta_3^{3/2}, \mathcal{F}, F)}{\sqrt{n}\delta_3^3}\right].\end{aligned}$$

A further simplification is possible. By Lemma A.1 (iii), the map $\delta \mapsto J(\delta, \mathcal{F}, F)/\delta$ is non-increasing, so that $J^2(\delta_3^{3/2}, \mathcal{F}, F)/\delta_3^3 \ge J^2(1, \mathcal{F}, F) \ge 1$. Hence the first term on the right side is not larger than

$$\|M\|_3^3 J^2(\delta_3^{3/2}, \mathcal{F}, F)/(n\delta_3^3).$$

This completes the proof. ∎

A.5. **Proofs of Propositions 3.1-3.3.** We will freely use the following simple lemmas. The proofs are straightforward and hence omitted.

Lemma A.7. *Let ξ be a real-valued random variable such that $\mathbb{E}[|\xi|^q] < \infty$ for some $q \geq 1$. Then for every $1 \leq r \leq q$ and $\tau > 0$,*

$$\mathbb{E}[|\xi|^r 1(|\xi| > \tau)] \leq \frac{\mathbb{E}[|\xi|^q]}{\tau^{q-r}}.$$

Lemma A.8. *Let $\xi_1, \ldots, \xi_m$ be arbitrary real-valued random variables such that $\max_{1\leq i\leq m} \mathbb{E}[|\xi_i|^q] < \infty$ for some $q \geq 1$. Then for every $1 \leq r \leq q$*

$$\| \max_{1\leq i\leq m} |\xi_i| \|_r \leq \| \max_{1\leq i\leq m} |\xi_i| \|_q \leq m^{1/q} \max_{1\leq i\leq m} \|\xi_i\|_q.$$

In particulrar, $\mathbb{E}[\max_{1\leq i\leq m} |\xi_i|^r] \leq m^{r/q} \max_{1\leq i\leq m} \|\xi_i\|_q^r$.

Proof of Proposition 3.1. For given $x \in \mathcal{I}, g \in \mathcal{G}$ and $h > 0$, define

$$f_{x,g,h}(y,t) = c_n(x,g)g(y)k(h^{-1}(t-x)),\ (y,t) \in \mathcal{Y} \times \mathbb{R}^d.$$

Consider the class of functions $\mathcal{F}_n = \{f_{x,g,h_n} - \mathbb{E}[f_{x,g,h_n}(Y_1, X_1)] : (x,g) \in \mathcal{I} \times \mathcal{G}\}$. We shall apply Theorem 2.1 to $\mathcal{F}_n$. Let $Z_n = \sup_{f\in\mathcal{F}_n} \mathbb{G}_n f$. We first note that $|f_{x,g,h}(y,t)| \leq C_{\mathcal{I}\times\mathcal{G}} b \|k\|_\infty$ so that $|f_{x,g,h}(y,t) - \mathbb{E}[f_{x,g,h}(Y_1,X_1)]| \leq 2C_{\mathcal{I}\times\mathcal{G}} b\|k\|_\infty \equiv F$. It is not difficult to see that $\mathcal{F}_n$ is pointwise measurable. Using Lemma A.5, we can prove that there are constants $A, v > 0$ such that

$$\sup_Q N(\mathcal{F}_n, e_Q, 2\varepsilon C_{\mathcal{I}\times\mathcal{G}} b\|k\|_\infty) \leq (A/\varepsilon)^v,\ 0 < \forall \varepsilon \leq 1,\ \forall n \geq 1. \tag{25}$$

Hence for every $n \geq 1$, $\mathcal{F}_n$ is pre-Gaussian and there exists a tight Gaussian random element G_n in $\ell^\infty(\mathcal{F}_n)$ with mean zero and covariance function

$$\mathbb{E}[G_n(f)G_n(\check{f})] = \mathrm{Cov}(f(Y_1,X_1), \check{f}(Y_1,X_1)),\ f, \check{f} \in \mathcal{F}_n.$$

To apply Theorem 2.1, we make some complimentary calculations. By (25), $J(\delta, \mathcal{F}_n, F) = O(\delta\sqrt{\log 1/\delta})$ as $\delta \to 0$ uniformly in n. Moreover,

$$\begin{aligned}
&\mathbb{E}[|f_{x,g,h_n}(Y_1,X_1)|^3] \\
&\quad = |c_n(x,g)|^3 \int_{\mathbb{R}^d} \mathbb{E}[|g(Y_1)|^3 \mid X_1 = t]|k(h_n^{-1}(t-x))|^3 p(t)dt \\
&\quad = |c_n(x,g)|^3 h_n^d \int_{\mathbb{R}^d} \mathbb{E}[|g(Y_1)|^3 \mid X_1 = x + h_n t]|k(t)|^3 p(x+h_n t)dt \\
&\quad \leq C_{\mathcal{I}\times\mathcal{G}}^3 b^3 \|p\|_\infty h_n^d \int_{\mathbb{R}^d} |k(t)|^3 dt,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[|f_{x,g,h_n}(Y_1,X_1)|^4] \\
&\quad = |c_n(x,g)|^4 h_n^d \int_{\mathbb{R}^d} \mathbb{E}[|g(Y_1)|^4 \mid X_1 = x + h_n t]|k(t)|^4 p(x+h_n t)dt \\
&\quad \le C_{\mathcal{I}\times\mathcal{G}}^4 b^4 \|p\|_\infty h_n^d \int_{\mathbb{R}^d} |k(t)|^4 dt.
\end{aligned}$$

Thus, by using Lemma 2.2, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbb{E}_n[|f(Y_i,X_i)|^3]\|_{\mathcal{F}_n}\right] = O(h_n^d + n^{-1}\log n), \text{ and} \\
&\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_n \cdot \mathcal{F}_n}] = O(h_n^{d/2}\sqrt{\log n} + n^{-1/2}\log n).
\end{aligned}$$

Choosing $\kappa = \kappa_n = C_1(h_n^{d/3} + n^{-1/3}\log^{1/3} n)$ with a sufficiently large constant C_1, and $\varepsilon = \varepsilon_n = n^{-1/6}\kappa_n$ and $\gamma = \gamma_n = (\log n)^{-1}$, we have, after an elementary calculation,

$$\Delta_n(\varepsilon_n, \gamma_n) = O(n^{-1/6} h_n^{d/3} \log n + n^{-1/4} h_n^{d/4} \log^{5/4} n + n^{-1/2}\log^{3/2} n).$$

Moreover, as $\kappa_n \gamma_n^{-1/3} n^{1/3} H_n(\varepsilon_n)^{-1/3} \to \infty$, for large n,

$$1(F > c\kappa_n \gamma_n^{-1/3} n^{1/3} H_n(\varepsilon_n)^{-1/3}) = 0.$$

Therefore, by Theorem 2.1, there exists a sequence $\widetilde{Z}_n$ of random variables such that $\widetilde{Z}_n \stackrel{d}{=} \sup_{f\in\mathcal{F}_n} G_n f$ and as $n \to \infty$,

$$|Z_n - \widetilde{Z}_n| = O_{\mathbb{P}}(n^{-1/6} h_n^{d/3} \log n + n^{-1/4} h_n^{d/4} \log^{5/4} n + n^{-1/2}\log^{3/2} n).$$

This implies the conclusion of the theorem. In fact, let

$$B_n(x,g) = h_n^{-d/2} G_n(f_{x,g,h_n}),\ (x,g) \in \mathcal{I}\times\mathcal{G},$$

and $\widetilde{W}_n = h_n^{-d/2}\widetilde{Z}_n$. Then B_n is the desired Gaussian process, and as $W_n = h_n^{-d/2} Z_n$, we have $\widetilde{W}_n \stackrel{d}{=} \sup_{(x,g)\in\mathcal{I}\times\mathcal{G}} B_n(x,g)$ and

$$\begin{aligned}
|W_n - \widetilde{W}_n| &= h_n^{-d/2}|Z_n - \widetilde{Z}_n| \\
&= O_{\mathbb{P}}\{(nh_n^d)^{-1/6}\log n + (nh_n^d)^{-1/4}\log^{5/4} n + n^{-1/2}h_n^{-d/2}\log^{3/2} n\}.
\end{aligned}$$

This completes the proof. ∎

Proof of Proposition 3.2. We shall follow the notation used in the proof of Proposition 3.1. Take $F(y,x) = C_{\mathcal{I}\times\mathcal{G}}\|k\|_\infty(G(y) + \mathbb{E}[G(Y_1)])$ as an envelope of $\mathcal{F}_n$. A version of inequality (25) continues to hold with $2C_{\mathcal{I}\times\mathcal{G}} b\|k\|_\infty$ replaced by $\|F\|_{Q,2}$. Let $D = \sup_{x\in\mathbb{R}^d}\mathbb{E}[G^4(Y_1) \mid X_1 = x]$. Then we have

$$\mathbb{E}[|f_{x,g,h_n}(Y_1,X_1)|^3] \le (1+D)C_{\mathcal{I}\times\mathcal{G}}^3 \|p\|_\infty h_n^d \int_{\mathbb{R}^d} |k(t)|^3 dt,$$

and

$$\mathbb{E}[|f_{x,g,h_n}(Y_1,X_1)|^4] \le DC_{\mathcal{I}\times\mathcal{G}}^4 \|p\|_\infty h_n^d \int_{\mathbb{R}^d} |k(t)|^4 dt.$$

Thus, using Lemma 2.2, we have

$$\mathbb{E}\left[\|\mathbb{E}_n[|f(Y_i,X_i)|^3]\|_{\mathcal{F}_n}\right] = O(h_n^d + n^{-1+3/q}\log n), \text{ and}$$
$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_n\cdot\mathcal{F}_n}] = O(h_n^{d/2}\sqrt{\log n} + n^{-1/2+2/q}\log n).$$

Choosing $\kappa = \kappa_n = C_1(h_n^{d/3} + n^{-1/3+1/q}\log^{1/3} n)$ with a sufficiently large constant C_1, and $\varepsilon = \varepsilon_n = n^{-1/6}\kappa_n$ and $\gamma = \gamma_n = (\log n)^{-1}$, we have, after an elementary calculation,

$$\Delta_n(\varepsilon_n,\gamma_n) = O(n^{-1/6}h_n^{d/3}\log n + n^{-1/4}h_n^{d/4}\log^{5/4} n + n^{-1/2+1/q}\log^{3/2} n).$$

We wish to check that

$$\mathbb{E}[(F/\kappa_n)^3 1(F/\kappa_n > c\gamma_n^{-1/3}n^{1/3}H_n(\varepsilon_n)^{-1/3})] = o(1).$$

In fact, the left side is bounded by

$$\kappa_n^{-3}(c\gamma_n^{-1/3}\kappa_n n^{1/3}H_n(\varepsilon_n)^{-1/3})^{3-q}\mathbb{E}[F^q] = O(n^{1-q/3}\kappa_n^{-q}) = o(1).$$

The rest of the proof is the same as in the previous one. ∎

Proof of Proposition 3.3. We only deal with case (ii). The proof for case (i) is similar. Observe first that by condition (C2),

$$|\alpha_n(x,g)| \le \frac{C_1|\psi^{K_n}(x)|}{c_1|\psi^{K_n}(x)|} \le C_3,$$

where $C_3 = C_1/c_1$. For given $n \ge 1, x \in \mathcal{I}$ and $g \in \mathcal{G}$, define

$$f_{n,x,g}(\eta,t) = g(\eta)\alpha_n(x,g)^T\psi^{K_n}(t),\ (\eta,t) \in \mathcal{E}\times[0,1]^d.$$

Consider the class of functions $\mathcal{F}_n = \{f_{n,x,g} : (x,g) \in \mathcal{I}\times\mathcal{G}\}$. We shall apply Theorem 2.1 to $\mathcal{F}_n$. Note that $W_n = \sup_{f\in\mathcal{F}_n}\mathbb{G}_n f$. First, we have $|f_{n,x,g}(\eta,t)| \le C_3 b_n|G(\eta)| =: F_n(\eta,t)$. Second, observe that $\mathcal{F}_n = \mathcal{H}_1\cdot\mathcal{H}_{2n}$, where $\mathcal{H}_1 = \{(\eta,t)\mapsto g(\eta) : g\in\mathcal{G}\}$ and $\mathcal{H}_{2n} = \{(\eta,t)\mapsto \alpha_n(x,g)^T\psi^{K_n}(t) : (x,g)\in\mathcal{I}\times\mathcal{G}\}$. By condition (C3),

$$|\alpha_n(x,g)^T\psi^{K_n}(t) - \alpha_n(\check{x},\check{g})^T\psi^{K_n}(t)| \le L_n b_n\{|x-\check{x}| + (\mathbb{E}[(g(\eta_1)-\check{g}(\eta_1))^2)^{1/2}\},$$

so that, using the fact that $\mathcal{G}$ is VC type, we deduce that there are constants $A, v > 0$ such that

$$\sup_Q N(\mathcal{H}_{2n}, e_Q, \varepsilon C_3 b_n) \le (AL_n/\varepsilon)^v,\ 0<\forall\varepsilon\le 1,\ \forall n \ge 1.$$

Using again the fact that $\mathcal{G}$ is VC type and Lemma A.5, we deduce that there are constants $A', v' > 0$ such that

$$\sup_Q N(\mathcal{F}_n, e_Q, \varepsilon\|F_n\|_{Q,2}) \le (A'L_n/\varepsilon)^{v'},\ 0<\forall\varepsilon\le 1,\ \forall n\ge 1. \tag{26}$$

Hence for every $n \ge 1$, there exists a tight Gaussian random element G_n in $\ell^\infty(\mathcal{F}_n)$ with mean zero and covariance function

$$\mathbb{E}[G_n(f)G_n(\check{f})] = \mathrm{Cov}(f(\eta_1,X_1),\check{f}(\eta_1,X_1)),\ f,\check{f}\in\mathcal{F}_n.$$

To apply Theorem 2.1, we make some complimentary calculations. Note that, by (26), for every $\delta_n \downarrow 0$ with $\log(1/\delta_n) = O(\log n)$, $J(\delta_n, \mathcal{F}_n, F_n) = O(\delta_n\sqrt{\log n})$. Let $D = \sup_{x\in[0,1]^d} \mathbb{E}[G^4(\eta_1) \mid X_1 = x]$. Then for $n \geq 1$,

$$\begin{aligned}
&\mathbb{E}[|g(\eta_1)\alpha_n(x,g)^T\psi^{K_n}(X_1)|^3] \\
&\quad \leq \mathbb{E}[\mathbb{E}[G^3(\eta_1) \mid X_1]|\alpha_n(x,g)^T\psi^{K_n}(X_1)|^3] \\
&\quad \leq C_3(1+D)b_n\mathbb{E}[|\alpha_n(x,g)^T\psi^{K_n}(X_1)|^2] \\
&\quad = C_3(1+D)b_n\alpha_n(x,g)^T\mathbb{E}[\psi^{K_n}(X_1)\psi^{K_n}(X_1)^T]\alpha_n(x,g) \\
&\quad \leq C_3^3C_2(1+D)b_n,
\end{aligned}$$

and

$$\mathbb{E}[|g(\eta_1)\alpha_n(x,g)^T\psi^{K_n}(X_1)|^4] \leq C_2^4C_2Db_n^2.$$

Thus, using Lemma 2.2, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbb{E}_n[|f(\eta_i,X_i)|^3]\|_{\mathcal{F}_n}\right] = O(b_n + n^{-1+3/q}b_n^3\log n), \text{ and} \\
&\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_n\cdot\mathcal{F}_n}] = O(b_n\sqrt{\log n} + n^{-1/2+2/q}b_n^2\log n).
\end{aligned}$$

Choosing $\kappa = \kappa_n = C_4(b_n^{1/3} + n^{-1/3+1/q}b_n\log^{1/3} n)$ with a sufficiently large constant C_4, and $\varepsilon = \varepsilon_n = n^{-1/2}$ and $\gamma = \gamma_n = (\log n)^{-1}$, we have

$$\Delta_n(\varepsilon_n,\gamma_n) = O(n^{-1/6}b_n^{1/3}\log n + n^{-1/4}b_n^{1/2}\log^{5/4} n + n^{-1/2+1/q}b_n\log^{3/2} n).$$

We wish to check that

$$\mathbb{E}[(F_n/\kappa_n)^3 1(F_n/\kappa_n > c\gamma_n^{-1/3}n^{1/3}H_n(\varepsilon_n)^{-1/3})] = o(1).$$

In fact, the left side is bounded by

$$\kappa_n^{-3}(c\kappa_n\gamma_n^{-1/3}n^{1/3}H_n(\varepsilon_n)^{-1/3})^{3-q}\mathbb{E}[F_n^q] = O(n^{1-q/3}\kappa_n^{-q}b_n^q) = o(1).$$

Finally, let $B_n(x,g) = G_n(f_{n,x,g}), (x,g) \in \mathcal{I}\times\mathcal{G}$. Then B_n is the desired Gaussian process, and by Theorem 2.1, there exists a sequence $\widetilde{W}_n$ of random variables such that $\widetilde{W}_n \stackrel{d}{=} \sup_{f\in\mathcal{F}_n} G_nf = \sup_{(x,g)\in\mathcal{I}\times\mathcal{G}} B_n(x,g)$ and as $n \to \infty$, $|W_n - \widetilde{W}_n| = O_{\mathbb{P}}(\Delta_n(\varepsilon_n,\gamma_n))$. This completes the proof. ∎

(V. Chernozhukov) Department of Economics and Operations Research Center, MIT, 50 Memorial Drive, Cambridge, MA 02142, USA.
E-mail address: vchern@mit.edu

(D. Chetverikov) Department of Economics, UCLA, Bunche Hall, 8283, 315 Portola Plaza, Los Angeles, CA 90095, USA.
E-mail address: chetverikov@econ.ucla.edu

(K. Kato) Graduate School of Economics, University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033, Japan.
E-mail address: kkato@e.u-tokyo.ac.jp