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 CWP44/12

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 empirical processes themselves in the sup-norm. We prove an abstract approximation theorem that is applicable to a wide variety of problems, primarily in statistics. Especially, 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 processes 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 considerably weak regularity conditions, especially concerning the bandwidth and the number of series functions, in those examples.

1. Introduction

1.1. **Overview.** 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 $X_1, \dots, X_n$ 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 $\mathbb{G}_n f = n^{-1/2} \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_1)])$ indexed by $\mathcal{F}_n$. For a moment, we implicitly assume that each $\mathcal{F}_n$ is "nice" enough and leave the measurability question. 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 $\sup_{f \in \mathcal{F}} B_n f$, where each B_n is a centered

Date: December 28, 2012.

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

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

K.Kato is supported by the Grant-in-Aid for Young Scientists (B) (22730179), the Japan Society for the Promotion of Science. V. Chernozhukov and D. Chetverikov are supported by a National Science Foundation grant.

Gaussian process indexed by $\mathcal{F}_n$ with covariance function $\mathbb{E}[B_n(f)B_n(g)] = \mathbb{E}[f(X_1)g(X_1)]$ for $f, g \in \mathcal{F}_n$. We look for conditions under which there is a sequence of such random variables $\widetilde{Z}_n$ with

$$|Z_n - \widetilde{Z}_n| = O_{\mathbb{P}}(r_n), \tag{1.1}$$

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

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 [22]. The (tractable) distributional approximation of the supremum of the empirical process is of particular importance in 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, for example, 2, 25, 35, 20, 19]. 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. [8] 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 empirical processes *themselves* by a sequence of Gaussian processes in the sup-norm. This problem is stronger than (1.1). Indeed, (1.1) is implied if there is a sequence of versions of B_n (which we denote by the same symbol B_n) such that

$$\|\mathbb{G}_n - B_n\|_{\mathcal{F}_n} := \sup_{f \in \mathcal{F}_n} |(\mathbb{G}_n - B_n)f| = O_{\mathbb{P}}(r_n). \tag{1.2}$$

There is a large literature on the problem (1.2). Notably, Komlós et al. [24] (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 [29] and [5] for refinements of KMT's result. [30], [23] and [35] developed extensions of the KMT construction to more general classes of functions.

The KMT construction is a powerful tool in addressing the problem (1.2), but when applied to general empirical processes, it requires side conditions on classes of functions and distributions. For examples, Rio [35] required that $\mathcal{F}_n$ are classes of functions of uniformly bounded variations on $S = [0,1]^d$, and P has a continuous and positive Lebesgue density on $[0,1]^d$. Such side conditions are essential to the KMT construction since it depends crucially on the Haar approximation of indicator functions and binomial coupling inequalities of Tusnády. [13], [1] and [36] considered the problem of Gaussian approximation of general empirical processes with different approaches and thereby without such side conditions. [13] used a finite approximation of a (possibly uncountably) infinite class of functions and apply a coupling inequality of [42] to the discretized empirical process (more precisely, Dudley and Philipp used a version of Yurinskii's inequality proved by [11]). [1] and [36], on the other hand, used a coupling inequality of [43] instead of Yurinskii's and some recent empirical process techniques such as Talagrand's [39] concentration inequality, which leads to refinements of Dudley and Philipp's results in some cases. However, the "global" rates that [11], [1] and [36] established do not lead to tight conditions for the Gaussian approximation to, say, the supremum deviation of kernel type statistics.

We develop here a new direct approach to the problem (1.1), without taking the route of approximating empirical processes themselves in the sup-norm and with different technical tools than those used in the aforementioned papers. We prove an abstract approximation theorem (Theorem 2.1) that leads to results of type (1.1) 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), in which Stein's method for normal approximation, originally due to [37, 38], plays an important role; 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 [3], complemented with a new "local" maximal inequality for the expectation of suprema of empirical processes that extends the work of [41] (Theorem 5.2). We study applications of this approximation theorem to local empirical processes and series estimation in nonparametric regression. We find that our new technique is able to prove the Gaussian approximation for the supremum type statistics under considerably 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 (that is, there is 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 in which the "complexity" of $\mathcal{F}_n$ increases with n and even the $\mathbb{G}_n$ process does not have a limit process (in a suitable sense) in the asymptotic situation. Second, the theorem only requires finite moments of the envelope function, which should be contrasted with [23, 35, 1, 36] where the class of functions must be uniformly bounded. Hence the theorem is applicable to an even wide class of problems to which the previous results in those works are not applicable. Third, the bound in Theorem 2.1 is able to exploit the "local" property 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 the detail about these features).

In this paper, we substantially rely on modern empirical process theory. For general references on empirical process theory, we refer to [27], [40] and [12]. Especially, [12], Section 9.5, gives excellent historical remarks on the

Gaussian approximation of empirical processes. For a textbook treatment of Yurinskii's and KMT's couplings, we refer to [34], Chapter 10.

1.2. **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 empirical processes and series estimation in nonparametric regression. Sections 4 and 5 are devoted to developing some technical tools needed to prove Theorem 2.1. In Section 4, we prove a new coupling inequality for maxima of sums of random vectors, and in Section 5, we prove some inequalities for empirical processes. We put some proofs in Appendix.

1.3. **Notation.** We shall obey the following notation. Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote an underlying probability space. We assume that the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is "rich enough" in the sense that there is 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 $\|\xi\|_q = (\mathbb{E}[|\xi|^q])^{1/q}$, $1 \leq q < \infty$. 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 $\mathcal{L}^p(Q), p \geq 1$ 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 $t_\varepsilon \in T_\varepsilon$ with $d(t, t_\varepsilon) \leq \varepsilon$. The *$\varepsilon$-covering number* $N(T, d, \varepsilon)$ of T is the infimum of the cardinality of ε-nets of T, that is, $N(T, d, \varepsilon) := \inf\{\text{Card}(T_\varepsilon) : \ T_\varepsilon \text{ is an } \varepsilon\text{-net of } T\}$.

The standard Euclidean norm is denoted by $|\cdot|$. 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.

For a subset A of a semimetric space (U, ρ), let A^δ denotes the δ-enlargement of A, that is, $A^\delta = \{x \in S : \rho(x, A) \leq \delta\}$ where $\rho(x, A) = \inf_{x \in A} \rho(x, y)$.

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

2. Abstract approximation theorem

We begin with reviewing the setup. Let $X_1, \dots, X_n$ 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 $n \geq 3$. Suppose that there is a class $\mathcal{F}$ of measurable functions $S \to \mathbb{R}$, to which a a measurable *envelope* F is attached, that is, F is a non-negative measurable function $S \to \mathbb{R}$ such that

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

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 distributions (note: $\sup_Q$ can be replaced by the supremum over all finite discrete distributions). 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*, that is, 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) $J(1, \mathcal{F}, F) < \infty$.

Assumption (A1) is made to avoid a measurability complication. See Section 2.3.1 of [40] for further discussion. This assumption ensures that, for example, $\sup_{f \in \mathcal{F}} \mathbb{G}_n f = \sup_{f \in \mathcal{G}} \mathbb{G}_n f$, and hence the former supremum is a measurable map from Ω to $\mathbb{R}$. Assumptions (A2) and (A3) in particular ensure that $\mathcal{F}$ is P-pre-Gaussian:

Lemma 2.1. *Under assumptions (A2) and (A3), the class $\mathcal{F}$ is P-pre-Gaussian, that is, there is a tight Gaussian random element G_P in $\ell^\infty(\mathcal{F})$ with mean zero and covariance function $\mathbb{E}[G_P(f)G_P(g)] = P(f - Pf)(g - Pg)$ for $f, g \in \mathcal{F}$.*

Proof. This is a standard fact. For the sake of completeness, we shall verify this lemma in Appendix. □

Here is the main theorem. For the notational convenience, let us write

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

Also, write $M = \max_{1 \leq i \leq n} F(X_i)$ and $\mathcal{F} \cdot \mathcal{F} = \{fg : f \in \mathcal{F}, g \in \mathcal{F}\}$.

Theorem 2.1. *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 $\kappa^3 \geq \mathbb{E}[\sup_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^n |f(X_i)|^3]$. Then for every $\varepsilon \in (0, 1]$ and*

$\gamma \in (0,1)$, *there is a random variable* $\widetilde{Z} \stackrel{d}{=} \sup_{f\in\mathcal{F}} G_P f$ *such that*

$$\begin{aligned}&\mathbb{P}\left\{|Z-\widetilde{Z}| > K(q)\Delta_n(\varepsilon,\gamma)\right\}\\ &\qquad\le \gamma\left[1+\frac{1}{4}P[(F/\kappa)^3 1(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\varepsilon)^{-1/3})]\right] + \frac{C\log n}{n},\end{aligned}$$

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

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

Proof. See Section 6. □

Remark 2.1. 1. Theorem 2.1 is a non-asymptotic result. In applications $\mathcal{F}$ and F (and even S) may change with n, that is, $\mathcal{F}=\mathcal{F}_n$ and $F=F_n$. In that case, assumption (A3) is interpreted as $J(1,\mathcal{F}_n,F_n)<\infty$ for each n, but it allows for the case in which $J(1,\mathcal{F}_n,F_n)$ diverges as $n\to\infty$.

2. 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.

3. By [12], Theorem 3.1.1, one can extend 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.

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.*

Henceforth we only deal with $\sup_{f\in\mathcal{F}}\mathbb{G}_n f$.

When applying Theorem 2.1, one has to derive suitable bounds on

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n |f(X_i)|^3\right] \text{ and } \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}].$$

We can simply bound these terms by

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n |f(X_i)|^3\right] \le \|F\|_{P,3}^3,\ \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}] \lesssim J(1,\mathcal{F},F)\|F\|_{P,4}^2.$$

The latter estimate is deduced from Theorem 2.14.1 of [40] together with the fact that

$$\sup_Q N(\mathcal{F}\cdot\mathcal{F}, e_Q, 2\varepsilon\|F^2\|_{Q,2}) \le \sup_Q N^2(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}). \tag{2.1}$$

See Lemma A.1 in Appendix. These simple bounds are, however, too crude when PF^3 and PF^4 are significantly larger than the "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. The following lemma will be useful for handling such cases.

Lemma 2.2. *Suppose that assumptions (A1), (A2) with $q = 4$, and (A3) are satisfied. 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*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} |f(X_i)|^3\right] - \sup_{f\in\mathcal{F}} P|f|^3 \lesssim \frac{\|M\|_3^3 J^2(\delta_3^{3/2}, \mathcal{F}, F)}{n\delta_3^3},$$

and

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}] \lesssim J(\delta_4^2, \mathcal{F}, F)\|F\|_{P,4}^2 + \frac{\|M\|_4^2 J^2(\delta_4^2, \mathcal{F}, F)}{\sqrt{n}\delta_4^4}.$$

Proof. See Appendix. □

Armed with Lemma 2.2, Theorem 2.1 is relatively easy to apply for concrete examples. Before going to the applications, we discuss the key features of Theorem 2.1. First, Theorem 2.1 does not require the uniform boundedness of $\mathcal{F}$, and requires only finite moments of the envelope function. At first, this sounds not surprising. However, most papers working on the Gaussian approximation of empirical processes in the sup-norm, such as [23, 35, 1, 36], required that the class of functions at hand is uniformly bounded. The uniform boundedness of the class of functions is a strong assumption in, say, statistical applications, and the generality of Theorem 2.1 in this direction will turn out to be useful. The cost of this generality 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" (inspection of the proof shows that the $n^{-1}\log n$ term can be replaced by n^{-2}, at the cost of increasing the constant $K(q)$). However, this should be considered as a consequence of the generality of allowing for non-uniformly bounded classes of functions, and should not be considered as a genuine drawback of the theorem. The second feature of Theorem 2.1 is that it is able to exploit the "local" property 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 $\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}] \approx \sup_{f\in\mathcal{F}} \sqrt{Pf^4}$ (up to a log term). In some applications, for example, the supremum deviation of kernel type statistics, the class $\mathcal{F} = \mathcal{F}_n$ changes with n and $\sup_{f\in\mathcal{F}_n} P|f|^k$ with $k = 3, 4$ decrease to 0 (but the envelope F_n is such that $PF_n^k = O(1)$ for $k = 3, 4$). The bound in Theorem 2.1 can effectively exploit this information and lead to tight conditions on, say, the bandwidth, for the Gaussian approximation. This feature will turn out to be clear from the proofs for the applications in the following section.

3. Applications

This section studies applications of Theorem 2.1 to local empirical processes and series estimation in nonparametric regression. All the proofs in this section are gathered in Appendix. 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.

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 $U \times \mathbb{R}^d$, where $(U,\mathcal{U})$ is an arbitrary measurable space. Suppose that there is a class $\mathcal{G}$ of measurable functions $U \to \mathbb{R}$. 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 $\mathbb{R}^d$. We do not ask $k(\cdot)$ to be non-negative nor the integral of $k(\cdot)$ over $\mathbb{R}^d$ to be 1. Let h_n be a sequence of positive constants such that $h_n \to 0$ as $n \to \infty$, and let I be an arbitrary Borel subset of $\mathbb{R}^d$. Consider the kernel-type statistics

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

When the kernel function $k(\cdot)$ is such that $\int_{\mathbb{R}^d} k(t)dt = 1$, 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. For example, when $g \equiv 1$, $S_n(x, g)$ will be a consistent estimator of $p(x)$; when $U = \mathbb{R}$ 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 $U = \mathbb{R}$ 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:

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

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

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

Limit theorems on W_n are developed in [2], [25], [10], [35], [15], and [28], among others.

[15] 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 [15] 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 certain Gaussian processes. For each $n \geq 1$, let B_n be a centered Gaussian process indexed by $I \times \mathcal{G}$ with covariance function

$$\begin{aligned} &\mathbb{E}[B_n(x,g)B_n(\check{x},\check{g})] \\ &= h_n^{-d} c_g(x) c_{\check{g}}(\check{x}) \operatorname{Cov}[g(Y_1)k(h_n^{-1}(X_1 - x)), \check{g}(Y_1)k(h_n^{-1}(X_1 - \check{x}))]. \end{aligned} \tag{3.1}$$

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

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

Definition 3.1. 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 $a > 0$ and $v > 0$ such that $\sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}) \leq (a/\varepsilon)^v$ for all $0 < \varepsilon \leq 1$.

We make the following assumptions.

- (B1) $\mathcal{G}$ is a pointwise measurable class of functions $U \to \mathbb{R}$ 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) $C_{I\times\mathcal{G}} := \sup_{(x,g)\in I\times\mathcal{G}} |c_g(x)| < \infty$. Furthermore, for every $(x_m, g_m) \in I\times\mathcal{G}$ with $x_m \to x \in I$ and $g_m \to g \in \mathcal{G}$ pointwise, $c_{g_m}(x_m) \to c_g(x)$.

We first assume that $\mathcal{G}$ is uniformly bounded, which will be relaxed later. [33], Lemma 22, gives simple sufficient conditions under which $\mathcal{K}$ is VC type.

Proposition 3.1. *Suppose that assumptions (B1)-(B5) are satisfied. Then for every $n \geq 1$, there is a tight Gaussian random element B_n in $\ell^\infty(I \times \mathcal{G})$ with mean zero and covariance function (3.1), and there is a sequence $\widetilde{W}_n$ of random variables such that $\widetilde{W}_n \stackrel{d}{=} \sup_{(x,g)\in I\times\mathcal{G}} B_n(x,g)$ 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 $U \to \mathbb{R}$ with measurable envelope G such that $\mathbb{E}[G^q(Y_1)] < \infty$ for some $q \geq 4$ and $\sup_{x\in\mathbb{R}^d} \mathbb{E}[G^4(Y_1) \mid X_1 = x] < \infty$. Furthermore, $\mathcal{G}$ is VC type with envelope G.

Then we have the following proposition.

Proposition 3.2. *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. It is instructive to compare Propositions 3.1 and 3.2 with the implications of Theorem 1.1 of Rio [35].

1. Apparently, Rio's Theorem 1.1 is not applicable to the case in which G is not bounded. Hence Proposition 3.2 is not implied by that theorem. Indeed, we do not are of any previous result that leads to the conclusion of Theorem 3.2, at least in this generality.

2. In the special case of kernel density estimation (that is, $g \equiv 1$), 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 [35]). Hence Rio's rates are better than ours when $d = 2, 3$, but worse when $d \geq 4$ (aside from the difference between "in probability" and almost sure bounds).

3. On the other hand, consider, as a second example, kernel regression estimation (that is, $U = \mathbb{R}$ and $g(y) = y$). In order to formally apply Rio's Theorem 1.1 to this example, we need to assume that the *joint* distribution of (Y_1, X_1) is supported on $[0,1]^{d+1}$ (this condition can be weakened in such a way that the support of (Y_1, X_1) is a bounded rectangular in $\mathbb{R}^{d+1}$), and admits a continuous and positive Lebesgue density on $[0,1]^{d+1}$. Subject to such side conditions, Rio's Theorem 1.1 leads to $|W_n - \widetilde{W}_n| = O_{a.s.}\{(n^{d/(d+1)}h_n^d)^{-1/(2d)}\sqrt{\log n} + (nh_n^d)^{-1/2}\log n\}$. See, for example, [8], Theorem 8. Hence, aside from the difference between "in probability" and almost sure bounds, as long as $h_n = O(n^{-a})$ for some $a > 0$, our rates are always better when $d \geq 2$. When $d = 1$, ignoring logs, our rate is better as long as $nh_n^4 \to 0$ (and vice versa).

Remark 3.2. Importantly, both Propositions 3.1 and 3.2 require only mild restrictions on the bandwidth h_n. Ignoring the log terms, Proposition 3.1 requires $nh_n^d \to \infty$, and Proposition 3.2 requires $n^{1-2/q}h_n^d \to \infty$. Interestingly, they essentially coincide with the conditions on the bandwidth used in establishing *exact* rates of uniform strong consistency of kernel type estimators in [16, 17].

3.2. **Series estimation.** This section considers an application of Theorem 2.1 to series estimation in nonparametric regression. Consider a canonical nonparametric regression model

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

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 $m(x) = \mathbb{E}[Y_1 \mid X_1 = x]$.

We consider series estimation of $m(x)$. Suppose that for each $K \geq 1$, there are K basis functions $\psi_{K,1}, \dots, \psi_{K,K}$ defined on $[0,1]^d$. Let $\psi^K(x) = (\psi_{K,1}(x), \dots, \psi_{K,K}(x))^T$. Examples of such basis functions are Fourier series, splines, Cohen-Daubechies-Vial (CDV) wavelet bases [9], Hermite polynomials and so on. Let K_n be a sequence of positive constant such that $K_n \to \infty$ as $n \to \infty$. The idea of series estimation is to approximate $m(x)$ by $\sum_{j=1}^{K_n} \theta_{K_n,j}\psi_{K,j}(x)$ and to estimate the vector $\theta^{K_n} = (\theta_{K_n,1}, \dots, \theta_{K_n,K_n})^T$ 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, $\widehat{m}(x)$ admits an asymptotic linear form

$$\widehat{m}(x) - m(x) \approx \psi^{K_n}(x)^T (\mathbb{E}[\psi^{K_n}(X_1)\psi^{K_n}(X_1)])^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} \eta_i \psi^{K_n}(X_i) \right].$$

See, for example, [32]. [To make this approximation precise, we have to use undersmoothing to make the effect of the bias negligible relative to the right side. However, we skip the discussion on control of the bias since it is out of the scope of this paper.] Redefining $\psi^{K_n}(x)$ by

$$\psi^{K_n}(x) \leftarrow (\mathbb{E}[\psi^{K_n}(X_1)\psi^{K_n}(X_1)])^{-1/2} \psi^{K_n}(x),$$

we have the following formal approximation:

$$\frac{\sqrt{n}(\widehat{m}(x) - m(x))}{\sigma |\psi^{K_n}(x)|} \approx \frac{1}{\sigma} \cdot \frac{\psi^{K_n}(x)^T}{|\psi^{K_n}(x)|} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \psi^{K_n}(X_i) \right] =: \frac{1}{\sigma} S_n(x).$$

Therefore, for the purpose of making uniform inference on $m(x)$ over a Borel subset I of $[0,1]^d$, it is desirable to have a (tractable) distributional approximation of the following quantity:

$$W_n = \sup_{x \in I} S_n(x).$$

We address this problem in what follows. Let B_n be a centered Gaussian process indexed by I with covariance function

$$\mathbb{E}[B_n(x)B_n(x')] = \sigma^2 \alpha_{K_n}(x)^T \alpha_{K_n}(x'), \tag{3.2}$$

where $\alpha_K(x) = \psi^K(x)/|\psi^K(x)|$. Intuitively, it is expected that under suitable regularity conditions, there is a sequence $\widetilde{W}_n$ of random variables such that $\widetilde{W}_n \stackrel{d}{=} \sup_{x \in I} B_n(x)$ and as $n \to \infty$, $|W_n - \widetilde{W}_n| \stackrel{\mathbb{P}}{\to} 0$. We shall argue the speed of this approximation.

We make the following assumptions.

(C1) For each $K \geq 1$, $\mathbb{E}[\psi^K(X_1)\psi^K(X_1)] = I_K$, $b_K := \sup_{x\in[0,1]^d} |\psi^K(x)| \vee 1 < \infty$, and the map $x \mapsto \psi^K(x)/|\psi^K(x)| =: \alpha_K(x)$ is Lipschitz continuous with Lipschitz constant $\leq L_K(\geq 1)$, that is,

$$|\alpha_K(x) - \alpha_K(x')| \leq L_K|x - x'|, \ \forall x, x' \in [0,1]^d.$$

(C2) $\log b_{K_n} = O(\log n)$ and $\log L_{K_n} = O(\log n)$ as $n \to \infty$.

For many commonly used basis functions such as Fourier series, splines and CDV wavelet bases, $b_K = O(\sqrt{K})$ as $K \to \infty$. See [32]. The Lipschitz continuity of $\alpha_K(x)$ is implied if $\inf_{x\in[0,1]^d} |\psi^K(x)| > 0$ and $\psi^K(x)$ is Lipschitz continuous. Condition (C2) states mild growth restrictions on K_n and L_{K_n} and is usually satisfied.

Proposition 3.3. *Suppose that assumptions (C1) and (C2) are satisfied. Furthermore, suppose either (i) η_1 is bounded, or (ii) $\mathbb{E}[|\eta_1|^q] < \infty$ for some $q \geq 4$ and $\sup_{x\in[0,1]^d} \mathbb{E}[\eta_1^4 \mid X_1 = x] < \infty$. Then for every $n \geq 1$, there is a tight Gaussian random element B_n in $\ell^\infty(I)$ with mean zero and covariance function (3.2), and there is a sequence $\widetilde{W}_n$ of random variables such that $\widetilde{W}_n \stackrel{d}{=} \sup_{x\in I} B_n(x)$ and as $n \to \infty$,*

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

Proposition 3.3 is a new result. Suppose that $b_K = O(\sqrt{K})$ as $K \to \infty$. Ignoring the log terms, Proposition 3.3 requires $K_n/n \to 0$ in case (i) and $K_n/n^{1-2/q} \to 0$ in case (ii). These requirements 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 21]. Another approach to deduce a result similar to Proposition 3.3 is to apply Yurinskii's coupling to random vectors $\eta_i\psi^{K_n}(X_i)$, $i = 1, \dots, n$, which, however, requires a rather stringent restriction on K_n, namely $K_n^5/n \to 0$ (up to logs), for ensuring $|W_n - \widetilde{W}_n| \stackrel{\mathbb{P}}{\to} 0$. See, for example, [8], Theorem 7.

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 as follows.

Theorem 4.1. *Let $X_1, \dots, X_n$ be independent random vectors in $\mathbb{R}^p$ with mean zero and finite absolute third moments, that is, $\mathbb{E}[X_{ij}] = 0$ 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 $Y_1, \dots, Y_n$ be independent random vectors in $\mathbb{R}^p$ such that

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

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

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

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. *Consider the same setup as in Theorem 4.1. Then for every $\delta > 0$, there is a random variable $\widetilde{Z} \stackrel{d}{=} \max_{1\le j\le p}\sum_{i=1}^n Y_{ij}$ such that*

$$\mathbb{P}(|Z-\widetilde{Z}| > 16\delta) \lesssim \delta^{-2}(B_1+\delta^{-1}(B_2+B_4)\log(p\vee n))\log(p\vee n)+\frac{\log n}{n}, \quad (4.1)$$

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. □

Remark 4.1. Inspection of the proof shows that the $n^{-1}\log n$ term on the right side can be replaced by n^{-a} for any $a > 0$, with changing the constant "16" on the left side.

Theorem 4.1 is a coupling inequality similar in nature to Yurinskii's [42]. Before proving Theorem 4.1, let us first recall Yurinskii's coupling inequality.

Theorem 4.2 ([42]; see also [26])**.** *Consider the same setup as in Theorem 4.1. Let $S_n = \sum_{i=1}^n X_i$. Then for every $\delta > 0$, there is 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 [34], 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 $\widetilde{Z} = \max_{1 \le j \le p} (T_n)_j$,

$$\mathbb{P}(|Z - \widetilde{Z}| > 3\delta) \lesssim B_0 \left(1 + \frac{|\log(1/B_0)|}{p}\right). \tag{4.2}$$

Unfortunately, when p is large, the right side is typically too crude. 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 (4.1) and (4.2), consider the situation in which p is indexed by n and $p = p_n \to \infty$ as $n \to \infty$. Moreover, consider the simple case in which $X_{ij} = x_{ij}/\sqrt{n}$ and $|x_{ij}| \le b$ (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 Pisier's inequality conditional on $X_1, \dots, X_n$, one has

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

On the other hand,

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

Therefore, the former (4.1) allows p_n to be of an exponential order (p_n can be as large as $\log p_n = o(n^{1/4})$; hence, for example, p_n can be of order e^{n^α} for $0 < \alpha < 1/4$), while the latter (4.2) restricts p_n to be $p_n = o(n^{1/5})$.

Remark 4.2. 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 is translated into 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 log: $\log N(\mathcal{F}, e_P, \varepsilon\|F\|_{P,2})$.

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. *Let μ and ν be Borel probability measures on $\mathbb{R}$. Let $\varepsilon > 0$ and $\delta > 0$ be two positive constants. Suppose that $\mu(A) \leq \nu(A^\delta) + \varepsilon$ for every Borel subset A of $\mathbb{R}$. Let V be a random variable with distribution μ. Then there is a random variable W with distribution ν such that $\mathbb{P}(|V - W| > \delta) \leq \varepsilon$.*

Proof. By Strassen's theorem [see 34, 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 $F(w \mid v)$, that is, $F^{-1}(\tau \mid v) = \inf\{w : F(w \mid v) \geq \tau\}$. Generate a uniform random variable θ on $(0, 1)$ independent of V and take $W(\omega) = F^{-1}(\theta(\omega) \mid V(\omega))$. Then it is routine to verify that $(V, W) \stackrel{d}{=} (V^*, W^*)$. □

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

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

Without loss of generality, we may assume that $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ are independent. Let

$$S_n = \sum_{i=1}^n X_i, \ T_n = \sum_{i=1}^n Y_i.$$

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, as in [6], 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 $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$,

$$\max_{1 \leq j \leq p} x_j \leq F_\beta(x) \leq \max_{1 \leq j \leq p} x_j + \beta^{-1} \log p. \tag{4.3}$$

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 is a smooth function $g : \mathbb{R} \to \mathbb{R}$ such that $\|g'\|_\infty \le \delta^{-1}, \|g''\|_\infty \le C\beta\delta^{-1}, \|g'''\|_\infty \le C\beta^2\delta^{-1}$, and*

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

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 [34], 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 $\le \delta^{-1}$. 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

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

The rest of the proof is the same as [34], Lemma 10.18. Hence we omit the detail. □

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

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

Step 3: The next step uses Stein's method to compare

$$\mathbb{E}[g \circ F_\beta(S_n)] \text{ and } \mathbb{E}[g \circ F_\beta(T_n)].$$

The following argument is inspired by [7], Theorem 7. We first make some complimentary computations.

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

$$\sum_{j,k=1}^p |\partial_j\partial_k(g \circ F_\beta)(x)| \le \|g''\|_\infty + 2\|g'\|_\infty\beta, \tag{4.4}$$

$$\sum_{j,k,l=1}^p |\partial_j\partial_k\partial_l(g \circ F_\beta)(x)| \le \|g'''\|_\infty + 6\|g''\|_\infty\beta + 6\|g'\|_\infty\beta^2. \tag{4.5}$$

Furthermore, letting $D(\beta^{-1}) = \{y \in \mathbb{R}^p : |y_j| \le \beta^{-1}, 1 \le \forall j \le p\}$,

$$\sum_{j,k,l=1}^p \sup_{y \in D(\beta^{-1})} |\partial_j\partial_k\partial_l(g \circ F_\beta)(x+y)| \le C(\|g'''\|_\infty + \|g''\|_\infty\beta + \|g'\|_\infty\beta^2). \tag{4.6}$$

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

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

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 (4.4) and (4.5) follow from these relations and the following computation.

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

For the last inequality (4.6), it is standard to see that when $|y_j| \leq \beta^{-1}, 1 \leq \forall j \leq p$,

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

from which the desired inequality follows. □

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

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

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

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

Hence

$$S_n' \stackrel{d}{=} S_n.$$

Also, with $X_1^n = \{X_1, \dots, X_n\}$,

$$\begin{aligned}\mathbb{E}[S_n' - S_n \mid X_1^n] &= \mathbb{E}[X_I' - X_I \mid X_1^n] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i' - X_i \mid X_1^n] = -n^{-1} S_n, \end{aligned} \tag{4.7}$$

and

$$\begin{aligned}\mathbb{E}[(S_n' - S_n)(S_n' - S_n)^T \mid X_1^n] &= \mathbb{E}[(X_I' - X_I)(X_I' - X_I)^T \mid X_1^n] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i' - X_i)(X_i' - X_i)^T \mid X_1^n] \\ &= \frac{1}{n}\sum_{i=1}^n (\mathbb{E}[X_i X_i^T] + X_i X_i^T) \\ &= \frac{2}{n}\sum_{i=1}^n \mathbb{E}[X_i X_i^T] + \frac{1}{n}\sum_{i=1}^n (X_i X_i^T - \mathbb{E}[X_i X_i^T]) \\ &= \frac{2}{n}\sum_{i=1}^n \mathbb{E}[X_i X_i^T] + n^{-1} V, \end{aligned} \tag{4.8}$$

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

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

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 [31] 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

$$\begin{aligned}\mathbb{E}[f(S_n)] - \mathbb{E}[f(T_n)] &= \mathbb{E}\left[\sum_{j=1}^p \sum_{i=1}^n X_{ij} \partial_j h(S_n)\right] \\ &\quad - \mathbb{E}\left[\sum_{j,k=1}^p \sum_{i=1}^n \mathbb{E}[X_{ij} X_{ik}] \partial_j \partial_k h(S_n)\right]. \end{aligned} \tag{4.9}$$

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

$$\begin{aligned} 0 &= n\mathbb{E}[h(S_n') - h(S_n)] \quad (S_n' \stackrel{d}{=} S_n) \\ &= n\mathbb{E}[(S_n' - S_n)^T \nabla h(S_n) + 2^{-1}(S_n' - S_n)^T (\operatorname{Hess} h(S_n))(S_n' - S_n) + R] \\ &= n\mathbb{E}\Big[\mathbb{E}[(S_n' - S_n)^T \mid X_1^n] \nabla h(S_n) \\ &\qquad + 2^{-1} \operatorname{Tr}\Big((\operatorname{Hess} h(S_n))\mathbb{E}[(S_n' - S_n)(S_n' - S_n)^T \mid X_1^n]\Big) + R\Big] \\ &= \mathbb{E}\left[-\sum_{j=1}^p \sum_{i=1}^n X_{ij}\partial_j h(S_n) + \sum_{j,k=1}^p \sum_{i=1}^n \mathbb{E}[X_{ij}X_{ik}]\partial_j\partial_k h(S_n)\right] \\ &\qquad + \mathbb{E}\left[\frac{1}{2}\sum_{j,k=1}^p V_{jk}\partial_j\partial_k h(S_n) + nR\right] \quad \text{(by (4.7) and (4.8))} \\ &= -\mathbb{E}[f(S_n)] + \mathbb{E}[f(T_n)] + \mathbb{E}\left[\frac{1}{2}\sum_{j,k=1}^p V_{jk}\partial_j\partial_k h(S_n) + nR\right], \quad \text{(by (4.9))} \end{aligned}$$

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

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

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

$$\begin{aligned} |\mathbb{E}[nR]| &= \left|\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^n \sum_{j,k,l=1}^p \Delta_{ij}\Delta_{ik}\Delta_{il}(1-\theta)^2 \partial_j\partial_k\partial_l h(S_n + \theta\Delta_i)\right]\right| \\ &\qquad (\theta \sim U(0,1) \text{ independent of all the other variables}) \\ &\le \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^n \sum_{j,k,l=1}^p |\Delta_{ij}\Delta_{ik}\Delta_{il}| \cdot |\partial_j\partial_k\partial_l h(S_n + \theta\Delta_i)|\right]. \end{aligned} \tag{4.10}$$

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

$$(4.10) = \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^n \chi_i *\right] + \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^n \chi_i^c *\right]$$
$$=: \frac{1}{2}\left[(A)+(B)\right].$$

Observe that

$$(A) \le \mathbb{E}\left[\sum_{j,k,l=1}^p \max_{1\le i\le n}(\chi_i\cdot|\partial_j\partial_k\partial_l h(S_n+\theta\Delta_i)|)\times \max_{1\le j,k,l\le p}\sum_{i=1}^n |\Delta_{ij}\Delta_{ik}\Delta_{il}|\right]$$
$$\le C\beta^2\delta^{-1}\mathbb{E}\left[\max_{1\le j,k,l\le p}\sum_{i=1}^n|\Delta_{ij}\Delta_{ik}\Delta_{il}|\right] \quad \text{(by (4.6))}$$
$$\le C\beta^2\delta^{-1}\mathbb{E}\left[\max_{1\le j\le p}\sum_{i=1}^n|\Delta_{ij}|^3\right]$$
$$\le C\beta^2\delta^{-1}\mathbb{E}\left[\max_{1\le j\le p}\sum_{i=1}^n|X_{ij}|^3\right] = C\beta^2\delta^{-1}B_2,$$

and

$$(B) \le C\beta^2\delta^{-1}\sum_{i=1}^n\mathbb{E}\left[\chi_i^c\max_{1\le j\le p}|\Delta_{ij}|^3\right] \quad \text{(by (4.5))}$$
$$\le C\beta^2\delta^{-1}\sum_{i=1}^n\mathbb{E}\left[\chi_i^c\max_{1\le j\le p}|X_{ij}|^3\right]. \quad \text{(by symmetry)}$$

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

$$\mathbb{E}\left[\chi_i^c\max_{1\le j\le p}|X_{ij}|^3\right] \le \mathbb{E}\left[\max_{1\le j\le p}|X_{ij}|^3\cdot 1\left(\max_{1\le j\le p}|X_{ij}|>\beta^{-1}/2\right)\right]$$
$$+\mathbb{E}\left[\max_{1\le j\le p}|X_{ij}|^3\right]\cdot\mathbb{P}\left(\max_{1\le j\le p}|X_{ij}|>\beta^{-1}/2\right). \quad (4.11)$$

We recall the following lemma.

Lemma 4.4. *Let φ and ψ be two functions defined on an interval $\mathcal{I}$ in $\mathbb{R}$. 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. Denote by μ the distribution of ξ. Then

$$\begin{aligned}\operatorname{Cov}(\varphi(\xi), \psi(\xi)) &= \int_{\mathcal{I}} \varphi \cdot \psi d\mu - \int_{\mathcal{I}} \varphi d\mu \cdot \int_{\mathcal{I}} \psi d\mu \\ &= \frac{1}{2} \int_{\mathcal{I}} \int_{\mathcal{I}} (\varphi(t) - \varphi(s))(\psi(t) - \psi(s)) d\mu(s) d\mu(t).\end{aligned}$$

The conclusion of the lemma is deduced directly from this expression. □

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 (4.11) is not larger than the first term. Hence

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

Therefore,

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

Step 4: Combining Steps 1-3, one has

$$\begin{aligned}\mathbb{P}(Z \in A) &\leq (1-\varepsilon)^{-1}\mathbb{E}[g \circ F_\beta(T_n)] + \frac{C\beta\delta^{-1}(B_1 + \beta(B_2 + B_3))}{1-\varepsilon} \\ &\leq \mathbb{P}(F_\beta(T_n) \in A^{e_\beta + 3\delta}) + \frac{\varepsilon + C\beta\delta^{-1}(B_1 + \beta(B_2 + B_3))}{1-\varepsilon} \\ &\qquad \text{(by construction of } g) \\ &\leq \mathbb{P}(\widetilde{Z} \in A^{2e_\beta + 3\delta}) + \frac{\varepsilon + C\beta\delta^{-1}(B_1 + \beta(B_2 + B_3))}{1-\varepsilon}. \quad \text{(Lemma 4.3)}\end{aligned}$$

This completes the proof. □

5. Inequalities for empirical processes

In this section, we shall prove some inequalities for empirical processes that will be used in the proof of Theorem 2.1. These inequalities are of interest in their own rights. Consider the same setup as in Section 2, that is, let $X_1, \dots, X_n$ 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. 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

$$\sup_{f \in \mathcal{F}} Pf^2 \leq \sigma^2 \leq \|F\|_{P,2}^2.$$

Let

$$M = \max_{1 \leq i \leq n} F(X_i).$$

Theorem 5.1. *Suppose that $F \in \mathcal{L}^q(P)$ for some $q \geq 2$. Then for every $t \geq 1$, with probability $> 1 - t^{-q/2}$,*

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

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 [39] 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 [14] (see their Theorem 3.1). For the purpose of this paper, however, Theorem 5.1 is more suitable.

Proof of Theorem 5.1. The theorem essentially follows from [3], 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,

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

Hence, using the simple inequality $2\sqrt{ab} \leq \beta a + \beta^{-1}b, \forall \beta > 0$, 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 $t \geq 1$, with probability $> 1 - t^{-q}$,

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

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

Theorem 5.1 will be complemented with the following moment inequality for suprema of empirical processes, which is an extension of [41], Theorem 2.1, to possibly unbounded classes of functions.

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

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

We give a full proof of Theorem 5.2 for the sake of completeness. We first prove the following preliminary lemma.

Lemma 5.1. *Write $J(\delta)$ for $J(\delta, \mathcal{F}, F)$. Then (i) the map $\delta \mapsto J(\delta)$ is concave; (ii) $J(c\delta) \leq cJ(\delta)$, $\forall c \geq 1$; (iii) the map $\delta \mapsto J(\delta)/\delta$ is non-increasing; (iv) the map $\mathbb{R}_+ \times (0, \infty) \ni (x, y) \mapsto J(\sqrt{x/y})\sqrt{y}$ 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

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

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, for example, [4], Section 3.2.

Lemma 5.2. *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 5.3. *Let D_1 be a convex subset of $\mathbb{R}^n$, and let $g_i : D_1 \to \mathbb{R}, 1 \leq i \leq k$ be concave functions. Let D_2 denote the convex hull of the set $\{(g_1(x), \ldots, g_k(x)) : x \in D\}$. Let $h : D_2 \to \mathbb{R}$ be concave and nondecreasing in each coordinate. Then $f(x) = h(g_1(x), \ldots, g_k(x)), D_1 \to \mathbb{R}$, 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 5.4. *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 [27], Theorem 4.12, and the discussion following the theorem. □

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

Theorem 5.3. *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, for example, [27], Theorem 6.20. □

We are now in position to prove Theorem 5.2.

Proof of Theorem 5.2. 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 $\varepsilon_1, \dots, \varepsilon_n$ independent of $X_1, \dots, X_n$, 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

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

Hence by Lemma 5.1 (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 5.4) and the Cauchy-Schwarz inequality,

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

Here by the Hoffmann-Jørgensen inequality (Theorem 5.3),

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

so that,

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

where $\Delta^2 := \max\{\sigma^2, n^{-1}\|M\|_2^2\}/\|F\|_{P,2}^2 \ge \delta^2$ and $D := \|M\|_2/(\sqrt{n}\|F\|_{P,2}^2)$. Therefore, using Lemma 5.1 (ii), we have

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

We consider the following two cases:

(i) $\sqrt{DZ} \le \Delta$. In this case, $J(\Delta \vee \sqrt{DZ}) \le J(\Delta)$, so that $Z \le C\|F\|_{P,2}J(\Delta)$. Since the map $\delta \mapsto J(\delta)/\delta$ is non-increasing (Lemma 5.1 (iii)),

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

Since $J(\delta)/\delta \ge J(1) \ge 1$, 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) $\sqrt{DZ} \ge \Delta$. In this case, $J(\Delta \vee \sqrt{DZ}) \le J(\sqrt{DZ})$, and since the map $\delta \mapsto J(\delta)/\delta$ is non-increasing (Lemma 5.1 (iii)),

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

Therefore,

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

that is

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

This completes the proof. □

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 [18], Proposition 2.1.

Corollary 5.1. *Consider the same setup as in Theorem 5.2. Suppose that there exist constants $a \ge e$ and $v \ge 1$ such that*

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

Then

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

Proof. We observe that

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

An integration by parts gives

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

by which we have

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

Since $a/\delta \ge a \ge e$, we have

$$J(\delta) \le 2\sqrt{2v}\delta\sqrt{\log(a/\delta)}.$$

Applying Theorem 5.2, we obtain the desired conclusion. □

6. 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 Lemma 2.1, independent of $X_1, \dots, X_n$. 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 12, Theorem 3.1.1]. Let $\{f_1, \dots, f_N\}$ 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 is a function $f_j, 1 \le j \le N$ such that $e_P(f, f_j) \le \varepsilon\|F\|_{P,2}$. Define

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

and $\mathcal{F}_\varepsilon = \{f - g : f, g \in \mathcal{F}, e_P(f,g) \le \varepsilon\|F\|_{P,2}\}$. It is standard to see that

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

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 $\delta > 0$,

$$\begin{aligned}\mathbb{P}(Z^\varepsilon \in A) &\le \mathbb{P}(\widetilde{Z}^\varepsilon \in A^{16\delta}) \\ &\quad + C\left[\delta^{-2}(B_1 + \delta^{-1}(B_2 + B_4)H_n(\varepsilon))H_n(\varepsilon) + \frac{\log n}{n}\right],\end{aligned}$$

where

$$B_1 = n^{-1}\mathbb{E}\left[\max_{1\le j,k\le N} |\sum_{i=1}^n (f_j(X_i)f_k(X_i) - P(f_j f_k))|\right],$$

$$B_2 = n^{-3/2}\mathbb{E}\left[\max_{1\le j\le N}\sum_{i=1}^n |f_j(X_i)|^3\right],$$

$$B_4 = n^{-1/2}\mathbb{E}\left[\max_{1\le j\le N}|f_j(X_1)|^3 \cdot 1\left(\max_{1\le j\le N}|f_j(X_1)| > \delta\sqrt{n}H_n(\varepsilon)^{-1}\right)\right].$$

Clearly

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

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

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

that is,

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

one has

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

Note that

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

so that

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

Therefore,

$$\begin{aligned}\mathbb{P}(Z^\varepsilon \in A) &\le \mathbb{P}(\widetilde{Z}^\varepsilon \in A^{16\delta}) + \frac{\gamma}{2} \\ &\quad + \frac{\gamma}{4}P[(F/\kappa)^3 1(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\varepsilon)^{-1/3})] + \frac{C\log n}{n} \\ &=: \mathbb{P}(\widetilde{Z}^\varepsilon \in A^{16\delta}) + \frac{\gamma}{2} + \text{error}. \end{aligned} \tag{6.1}$$

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

$$\begin{aligned}\|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon} \lesssim_q \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon}] + (\varepsilon\|F\|_{P,2} + n^{-1/2}\|M\|_q)\gamma^{-1/q} \\ + n^{-1/2}\|M\|_2\gamma^{-2/q}.\end{aligned}$$

Here by Theorem 5.2,

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

Hence

$$\begin{aligned} a := (1-\gamma/4)\text{-quantile of } \|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon} \lesssim_q\ & J(\varepsilon)\|F\|_{P,2} + n^{-1/2}\varepsilon^{-2}J^2(\varepsilon)\|M\|_2 \\ &+ (\varepsilon\|F\|_{P,2} + n^{-1/2}\|M\|_q)\gamma^{-1/q} \\ &+ n^{-1/2}\|M\|_2\gamma^{-2/q}. \end{aligned} \tag{6.2}$$

On the other hand, by the Borell-Sudakov-Tsirel'son inequality [40, Proposition A.1], with probability $> 1-\gamma/4$,

$$\|G_P\|_{\mathcal{F}_\varepsilon} \le \mathbb{E}[\|G_P\|_{\mathcal{F}_\varepsilon}] + \varepsilon\|F\|_{P,2}\sqrt{2\log(4/\gamma)}.$$

We can bound the expectation $\mathbb{E}[\|G_P\|_{\mathcal{F}_\varepsilon}]$ by Dudley's inequality [40, Corollary 2.2.8]:

$$\mathbb{E}[\|G_P\|_{\mathcal{F}_\varepsilon}] \lesssim J(\varepsilon)\|F\|_{P,2}.$$

Hence

$$\begin{aligned} b &:= (1-\gamma/4)\text{-quantile of } \|G_P\|_{\mathcal{F}_\varepsilon} \\ &\lesssim J(\varepsilon)\|F\|_{P,2} + \varepsilon\|F\|_{P,2}\sqrt{\log(1/\gamma)}. \end{aligned} \tag{6.3}$$

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

$$\begin{aligned} \mathbb{P}(Z \in A) &\le \mathbb{P}(Z^\varepsilon \in A^a) + \frac{\gamma}{4} \quad \text{(by (6.2))} \\ &\le \mathbb{P}(\widetilde{Z}^\varepsilon \in A^{a+16\delta}) + \frac{3}{4}\gamma + \text{error} \quad \text{(by (6.1))} \\ &\le \mathbb{P}(\widetilde{Z} \in A^{a+b+16\delta}) + \gamma + \text{error}. \quad \text{(by (6.3))} \end{aligned}$$

The conclusion follows from Lemma 4.1. □

Appendix A. Additional proofs

A.1. **Proof of Lemma 2.1.** Let G_P be a centered Gaussian process indexed by $\mathcal{F}$ with covariance function $\mathbb{E}[G_P(f)G_P(g)] = P(f-Pf)(g-Pg)$. Recall that $\mathcal{F}$ is P-pre-Gaussian if and only if $(\mathcal{F}, \rho_P)$ is totally bounded and G_P has a version that has sample paths almost surely uniformly ρ_P-continuous, where $\rho_P(f,g) := \sqrt{\mathrm{Var}_P(f-g)}$ [40, Example 1.3.10]. Dudley's criterion for sample continuity of Gaussian processes states that when

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

there is a version of G_P that has sample paths uniformly ρ_P-continuous [40, p.100-101]. The lemma readily follows from these observations and the simple fact $\rho_P \le e_P$. □

A.2. **Proof of Lemma 2.2.** Before proving Lemma 2.2, we shall prepare the following lemmas.

Lemma A.1. *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*

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

Proof. Implicit in [40], Section 2.10.3. Hence we omit the detail. □

Lemma A.2. *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*

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

Proof. Let us write $\mathcal{F}/F = \{f/F : f \in \mathcal{F}\}$ with the convention $0/0 = 0$. We first point out that

$$\sup_Q N(\mathcal{F}/F, e_Q, \varepsilon) \leq \sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}). \tag{A.1}$$

Indeed, for a given distribution Q, let $\{f_1, \dots, f_N\}$ be an $\varepsilon\|F\|_{Q,2}$-net of $\mathcal{F}$. Define the new distribution Q' by $dQ' = (F^2/\|F\|_{Q,2}^2)dQ$. Then for every $f \in \mathcal{F}$, there is a function $f_j, 1 \leq j \leq N$ with $\|f - f_j\|_{Q,2} \leq \varepsilon\|F\|_{Q,2}$, so that

$$\|f/F - f_j/F\|_{Q',2} \leq \|f - f_j\|_{Q,2}/\|F\|_{Q,2} \leq \varepsilon.$$

Returning to the original problem, for a given distribution Q, define the new distribution Q' by $dQ' = (F^{2q}/\|F^q\|_{Q,2}^2)dQ$. For every $f, g \in \mathcal{F}$,

$$\begin{aligned}||f|^q - |g|^q| &\leq q\max\{|f|^{q-1}, |g|^{q-1}\}|f - g| \\ &\leq qF^{q-1}|f - g| = qF^q|f/F - g/F|.\end{aligned}$$

Hence

$$\begin{aligned}\||f|^q - |g|^q\|_{Q,2} &\leq q\|F^q|f/F - g/F|\|_{Q,2} \\ &\leq q\|F^q\|_{Q,2}\|f/F - g/F\|_{Q',2}.\end{aligned}$$

This implies that

$$N(\mathcal{F}(q), e_Q, q\varepsilon\|F^q\|_{Q,2}) \leq N(\mathcal{F}/F, e_{Q'}, \varepsilon).$$

Combining (A.1) leads to the desired conclusion. □

Proof of Lemma 2.2. The second inequality is deduced from Theorem 5.2 together with the covering number estimate (2.1). Hence we shall prove the first inequality. We first observe that

$$\frac{1}{n}\sum_{i=1}^n |f(X_i)|^3 \leq P|f|^3 + \left|\frac{1}{n}\sum_{i=1}^n (|f(X_i)|^3 - P|f|^3)\right|,$$

by which we have

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

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

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

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

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

Furthermore, 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.2, 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

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

A further simplification is possible. By the simple inequality $2ab \leq a^2+b^2$, the right side is

$$\lesssim n^{-1}\|M\|_3^3 + \frac{\|M\|_3^3 J^2(\delta_3^{3/2},\mathcal{F},F)}{n\delta_3^3}.$$

By Lemma 5.1 (iii), the map $\delta \mapsto J(\delta,\mathcal{F},F)$ is non-increasing, so that $J^2(\delta_3^{3/2},\mathcal{F},F)/\delta_3^3 \geq J^2(1,\mathcal{F},F) \geq 1$. Hence the first term above is not larger than the second term. This completes the proof. □

A.3. **Proofs of Propositions 3.1-3.3.**

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

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

Consider the class of functions $\mathcal{F}_n = \{f_{x,g,h_n} : (x,g) \in I \times \mathcal{G}\}$. 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 $|f_{x,g,h}(y,t)| \leq C_{I\times\mathcal{G}} b \|k\|_\infty \equiv F$. It is not difficult to see that $\mathcal{F}_n$ is pointwise measurable for every $n \geq 1$. Using Lemma A.1, we can prove that there are constants $a > 0$ and $v > 0$ such that

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

Hence for every $n \geq 1$, $\mathcal{F}_n$ is pre-Gaussian and there is 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 (A.2), $J(\delta, \mathcal{F}_n, F) = O(\delta\sqrt{\log 1/\delta})$ as $\delta \to 0$ uniformly in n. Furthermore,

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

and

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

Thus, by Lemma 2.2,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_n} \frac{1}{n}\sum_{i=1}^n |f(Y_i,X_i)|^3\right] = O(h_n^d + n^{-1}\log n)$$

and

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_n\cdot\mathcal{F}_n}] = O(h_n^{d/2}\sqrt{\log n} + n^{-1/2}\log n).$$

Choosing $\kappa = \kappa_n = \text{const.} \times (h_n^{d/3} + n^{-1/3}\log^{1/3} n), \varepsilon = \varepsilon_n = n^{-1/6}\kappa_n$ and $\gamma = \gamma_n = (\log n)^{-1}$, we have, after an elementary calculation,

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

Furthermore, 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 is 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. Indeed, let

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

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 $\widetilde{W}_n \stackrel{d}{=} \sup_{(x,g) \in I \times \mathcal{G}} B_n(x,g)$ 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 $F(y,x) = C_{I \times \mathcal{G}} \|k\|_\infty G(y)$ as an envelope of $\mathcal{F}_n$. A version of inequality (A.2) continues to hold with $C_{I \times \mathcal{G}} b \|k\|_\infty$ 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

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

and

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

Thus, using Lemma 2.2, we have

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n |f(Y_i, X_i)|^3\right] = O(h_n^d + n^{-1+3/q} \log n)$$

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 $\kappa = \kappa_n = \text{const.} \times (h_n^{d/3} + n^{-1/3+1/q} \log^{1/3} n), \varepsilon = \varepsilon_n = n^{-1/6} \kappa_n$ 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 have to check that

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

Indeed, 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). For given $x \in I$ and $K \geq 1$, define

$$f_{x,K}(\eta, t) = \eta \alpha_K(x)^T \psi^K(t), \ (\eta, t) \in \mathbb{R} \times [0,1]^d.$$

Consider the class of functions $\mathcal{F}_n = \{f_{x,K_n} : x \in I\}$. 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 $|f_{x,K_n}(\eta,t)| \leq b_{K_n}|\eta| =: F_n(\eta,t)$. Second, since

$$|\eta \alpha_{K_n}(x)^T \psi^{K_n}(t) - \eta \alpha_{K_n}(x')^T \psi^{K_n}(t)| \leq L_{K_n} b_{K_n} |\eta||x - x'|, \ \forall x, x' \in [0,1]^d,$$

there is a constant $a > 0$ such that

$$\sup_Q N(\mathcal{F}_n, e_Q, \varepsilon \|F_n\|_{Q,2}) \leq (a L_{K_n}/\varepsilon)^d, \ 0 < \forall \varepsilon \leq 1, \ \forall n \geq 1. \tag{A.3}$$

Hence, for every $n \geq 1$, there is 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 (A.3), 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 $D = \sup_{x \in [0,1]^d} \mathbb{E}[|\eta_1|^4 \mid X_1 = x]$. Then for each $K \geq 1$,

$$\begin{aligned}
\mathbb{E}[|\eta_1 \alpha_K(x)^T \psi^K(X_1)|^3] &\leq \mathbb{E}[\mathbb{E}[|\eta_1|^3 \mid X_1] |\alpha_K(x)^T \psi^K(X_1)|^3] \\
&\leq (1 + D) b_K \mathbb{E}[|\alpha_K(x)^T \psi^K(X_1)|^2] \\
&\leq (1 + D) b_K \alpha_K(x)^T \mathbb{E}[\psi^K(X_1) \psi^K(X_1)^T] \alpha_K(x) \\
&= (1 + D) b_K,
\end{aligned}$$

and

$$\mathbb{E}[|\eta_1 \alpha_K(x)^T \psi^K(X_1)|^4] \leq D b_K^2.$$

Thus, using Lemma 2.2, we have

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n |f(\eta_i, X_i)|^3\right] = O(b_{K_n} + n^{-1+3/q} b_{K_n}^3 \log n)$$

and

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_n \cdot \mathcal{F}_n}] = O(b_{K_n} \sqrt{\log n} + n^{-1/2+2/q} b_{K_n}^2 \log n).$$

Choosing $\kappa = \kappa_n = \text{const.} \times (b_{K_n}^{1/3} + n^{-1/3+1/q} b_{K_n} \log^{1/3} n), \varepsilon = \varepsilon_n = n^{-1/2}$ and $\gamma = \gamma_n = (\log n)^{-1}$, we have

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

We have to check that

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

Indeed, the left side is bounded by

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

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

References

[1] 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.

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

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

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

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

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

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

[8] Chernozhukov, V., Lee, S., and Rosen, A. (2012). Intersection bounds: estimation and inference. arXiv:0907.3503v4. To appear in *Econometrica.*

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

[10] 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.

[11] Dehling, H. (1983). Limit theorems for sums of weakly dependent Banach space valued random variables. *Z. Warhsch. Verw. Gabiete* **63** 393-432.

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

[13] 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.

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

[15] Einmahl, U. and Mason, D.M. (1997). Gaussian approximation of local empirical processes indexed by functions. *Probab. Theory Related Fields* **107** 283-311.
[16] 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.
[17] Einmahl, U. and Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380-1403.
[18] 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.
[19] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122-1170.
[20] Ghosal, S., Sen, A. and van der Vaart, A.W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28** 1054-1082.
[21] Huang, J.Z. (2003). Asymptotics for polynomial spline regression under weak conditions. *Statist. Probab. Lett.* **65** 207-216.
[22] Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Atti. Giorn.* **4** 83-91.
[23] 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.
[24] 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.
[25] Konakov, V.D. and Piterberg, V.I. (1984). On the convergence rate of maximal deviations distributions for kernel regression estimates. *J. Multivariate Anal.* **15** 279-294.
[26] 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.
[27] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces.* Springer.
[28] Mason, D.M. (2004). A uniform functional law of the logarithm for the local empirical process. *Ann. Probab.* **32** 1391-1418.
[29] 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.
[30] Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT construction. *Ann. Probab.* **17** 266-291.
[31] 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.
[32] Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147-168.

[33] Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *Ann. Statist.* **15** 780-799.
[34] Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability.* Cambridge University Press.
[35] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21-45.
[36] Settati, A. (2009). Gaussian approximation of the empirical process under random entropy conditions. *Stochastic Process. Appl.* **119** 1541-1560.
[37] 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.
[38] Stein, C. (1986). *Approximate Computation of Expectations.* IMS Lecture Notes-Monograph Series, Vol.7.
[39] Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505-563.
[40] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
[41] van der Vaart, A.W. and Wellner, J.A. (2011). A local maximal inequality under uniform entropy. *Electronic J. Statist.* **5** 192-203.
[42] Yurinskii, V.V. (1977). On the error of the Gaussian approximation for convolutions. *Theory of Probability and Its Applications* **2** 236-247.
[43] 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.

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

(D. Chetverikov) Department of Economics, MIT, 50 Memorial Drive, Cambridge, MA 02142, USA.
E-mail address: dchetver@mit.edu

(K. Kato) Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan.
E-mail address: kkato@hiroshima-u.ac.jp