Probability Theory and Related Fields (2020) 176:1097–1163
https://doi.org/10.1007/s00440-019-00936-y

Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications

Xiaohui Chen[1] · **Kengo Kato**[2]

Received: 26 April 2018 / Revised: 13 February 2019 / Published online: 31 July 2019

Abstract
This paper is concerned with finite sample approximations to the supremum of a non-degenerate U-process of a general order indexed by a function class. We are primarily interested in situations where the function class as well as the underlying distribution change with the sample size, and the U-process itself is not weakly convergent as a process. Such situations arise in a variety of modern statistical problems. We first consider Gaussian approximations, namely, approximate the U-process supremum by the supremum of a Gaussian process, and derive coupling and Kolmogorov distance bounds. Such Gaussian approximations are, however, not often directly applicable in statistical problems since the covariance function of the approximating Gaussian process is unknown. This motivates us to study bootstrap-type approximations to the U-process supremum. We propose a novel jackknife multiplier bootstrap (JMB) tailored to the U-process, and derive coupling and Kolmogorov distance bounds for the proposed JMB method. All these results are non-asymptotic, and established under fairly general conditions on function classes and underlying distributions. Key technical tools in the proofs are new local maximal inequalities for U-processes, which may be useful in other problems. We also discuss applications of the general approximation results to testing for qualitative features of nonparametric functions based on generalized local U-processes.

X. Chen is supported by NSF DMS-1404891, NSF CAREER Award DMS-1752614, and UIUC Research Board Awards (RB17092, RB18099).

✉ Xiaohui Chen
xhchen@illinois.edu

Kengo Kato
kk976@cornell.edu

[1] Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, IL 61820, USA

[2] Department of Statistical Science, Cornell University, 1194 Comstock Hall, Ithaca, NY 14853, USA

Keywords Gaussian approximation · Jackknife multiplier bootstrap · Coupling · U-process · Local maximal inequality

Mathematics Subject Classification 60F17 · 62E17 · 62F40 · 62G10

1 Introduction

This paper is concerned with finite sample approximations to the supremum of a U-process of a general order indexed by a function class. We begin with describing our setting. Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in a measurable space $(S, \mathcal{S})$ with common distribution P. For a given integer $r \geqslant 2$, let $\mathcal{H}$ be a class of jointly measurable functions (kernels) $h : S^r \to \mathbb{R}$ equipped with a measurable envelope H (i.e., H is a nonnegative function on S^r such that $H \geqslant \sup_{h\in\mathcal{H}} |h|$). Consider the associated U-process

$$U_n(h) := U_n^{(r)}(h) := \frac{1}{|I_{n,r}|} \sum_{(i_1,\ldots,i_r)\in I_{n,r}} h(X_{i_1}, \ldots, X_{i_r}), \ h \in \mathcal{H}, \tag{1}$$

where $I_{n,r} = \{(i_1, \ldots, i_r) : 1 \leqslant i_1, \ldots, i_r \leqslant n, i_j \neq i_k \text{ for } j \neq k\}$ and $|I_{n,r}| = n!/(n-r)!$ denotes the cardinality of $I_{n,r}$. Without loss of generality, we may assume that each $h \in \mathcal{H}$ is symmetric, i.e., $h(x_1, \ldots, x_r) = h(x_{i_1}, \ldots, x_{i_r})$ for every permutation $i_1, \ldots, i_r$ of $1, \ldots, r$, and the envelope H is symmetric as well. Consider the normalized U-process

$$\mathbb{U}_n(h) = \sqrt{n}\{U_n(h) - \mathbb{E}[U_n(h)]\}, \quad h \in \mathcal{H}. \tag{2}$$

The main focus of this paper is to derive finite sample approximation results for the supremum of the normalized U-process, namely, $Z_n := \sup_{h\in\mathcal{H}} \mathbb{U}_n(h)/r$, in the case where the U-process is *non-degenerate*, i.e., $\mathrm{Var}(\mathbb{E}[h(X_1, \ldots, X_r) \mid X_1]) > 0$ for all $h \in \mathcal{H}$. The function class $\mathcal{H}$ is allowed to depend on n, i.e., $\mathcal{H} = \mathcal{H}_n$, and we are primarily interested in situations where the normalized U-process $\mathbb{U}_n$ is not weakly convergent as a process (beyond finite dimensional convergence). For example, there are situations where $\mathcal{H}_n$ depends on n but $\mathcal{H}_n$ is further indexed by a parameter set Θ independent of n. In such cases, one can think of $\mathbb{U}_n$ as a U-process indexed by Θ and can consider weak convergence of the U-process in the space of bounded functions on Θ, i.e., $\ell^\infty(\Theta)$. However, even in such cases, there are a variety of statistical problems where the U-process is not weakly convergent in $\ell^\infty(\Theta)$, even after a proper normalization. The present paper covers such "difficult" (and in fact yet more general) problems.

U-processes are powerful tools for a broad range of statistical applications such as testing for qualitative features of functions in nonparametric statistics [1,25,38], cross-validation for density estimation [43], and establishing limiting distributions of M-estimators (see, e.g., [4,18,50,51]). There are two perspectives on U-processes: (1)

they are *infinite-dimensional* versions of U-statistics (with one kernel); (2) they are stochastic processes that are *nonlinear* generalizations of empirical processes. Both views are useful in that: (1) statistically, it is of greater interest to consider a rich class of statistics rather than a single statistic; (2) mathematically, we can borrow the insights from empirical process theory to derive limit or approximation theorems for U-processes. Importantly, however, (1) extending U-statistics to U-processes requires substantial efforts and different techniques; and (2) generalization from empirical processes to U-processes is highly nontrivial especially when U-processes are not weakly convergent as processes. In classical settings where indexing function classes are fixed (i.e., independent of n), it is known that Uniform Central Limit Theorems (UCLTs) in the Hoffmann-Jørgensen sense hold for U-processes under metric (or bracketing) entropy conditions, where U-processes are weakly convergent in spaces of bounded functions [4,8,18,44] (these references also cover degenerate U-processes where limiting processes are Gaussian chaoses rather than Gaussian processes). Under such classical settings, [5,56] study limit theorems for bootstrapping U-processes; see also [3,6,9,19,32–34,55] as references on bootstraps for U-statistics. Giné and Mason [27] introduce a notion of the local U-process motivated by a density estimator of a function of several variables proposed by [24] and establish a version of UCLTs for local U-processes. More recently, [11] studies Gaussian and bootstrap approximations for high-dimensional (order-two) U-statistics, which can be viewed as U-processes indexed by *finite* function classes $\mathcal{H}_n$ with increasing cardinality in n. To the best of our knowledge, however, no existing work covers the case where the indexing function class $\mathcal{H} = \mathcal{H}_n$ (1) may change with n; (2) may have infinite cardinality for each n; and (3) need not verify UCLTs. This is indeed the situation for many of nonparametric specification testing problems [1,25,38]; see examples in Sect. 4 for details.

In this paper, we develop a general non-asymptotic theory for directly approximating the supremum $Z_n = \sup_{h\in\mathcal{H}} \mathbb{U}_n(h)/r$ without referring a weak limit of the underlying U-process $\{\mathbb{U}_n(h) : h \in \mathcal{H}\}$. Specifically, we first establish a general Gaussian coupling result to approximate Z_n by the supremum of a Gaussian process W_P in Sect. 2. Our Gaussian approximation result builds upon recent development in modern empirical process theory [13–15] and high-dimensional U-statistics [11]. As a significant departure from the existing literature [4,14,15,27], our Gaussian approximation for U-processes has a multi-resolution nature, which is neither parallel with the theory of U-processes with fixed function classes nor that of empirical processes. In particular, unlike U-processes with fixed function classes, the higher-order degenerate components are not necessarily negligible compared with the Hájek (empirical) process (in the sense of the Hoeffding projections [31]) and they may impact error bounds of the Gaussian approximation.

However, the covariance function of the Gaussian process W_P depends on the underlying distribution P which is unknown and hence the Gaussian approximation developed in Sect. 2 is not directly applicable to statistical problems such as computing critical values of a test statistic defined by the supremum of a U-process. On the other hand, the (Gaussian) multiplier bootstrap developed in [13,15] for empirical processes is not directly applicable to U-processes since the Hájek process also depends on P and hence is unknown. Our second main contribution is to develop a fully data-dependent procedure for approximating the distribution of Z_n. Specifically, we propose a novel

jackknife multiplier bootstrap (JMB) tailored to U-processes in Sect. 3. The key insight of the JMB is to replace the (unobserved) Hájek process by its jackknife estimate (cf. [10]). We establish finite sample validity of the JMB (i.e., conditional multiplier CLT) with explicit error bounds. As a distinguished feature, our error bounds involve a delicate interplay among *all* levels of the Hoeffding projections. In particular, the key innovations are a collection of new powerful local maximal inequalities for level-dependent degenerate components associated with the U-process (see Sect. 5). To the best of our knowledge, there has been no theoretical guarantee on bootstrap consistency for U-processes whose function classes change with n and which do not converge weakly as processes. Our finite sample bootstrap validity results with explicit error bounds fill this important gap in literature, although we only focus on the supremum functional.

It should be emphasized that our approximation problem is different from the problem of approximating the *whole* U-process $\{\mathbb{U}_n(h) : h \in \mathcal{H}\}$. In testing monotonicity of nonparametric regression functions, [25] consider a test statistic defined by the supremum of a bounded U-process of order-two and derive a Gaussian approximation result for the normalized U-process. Their idea is a two-step approximation procedure: first approximate the U-process by its Hájek process and then apply Rio's coupling result [47], which is a Komlós–Major–Tusnády (KMT) [36] type strong approximation for empirical processes indexed by Vapnik-Červonenkis (VC) type classes of functions. See also [35,41] for extensions of the KMT construction to other function classes. It is worth noting that the two-step approximation of U-processes based on KMT type approximations in general requires more restrictive conditions on the function class and the underlying distribution in statistical applications. Our regularity conditions on the function class and the underlying distribution for the Gaussian and bootstrap approximations are easy to verify and are less restrictive than those required for KMT type approximations since we directly approximate the supremum of a U-process rather than the whole U-process; in fact, our approximation results can cover examples of statistical applications for which KMT type approximations are not applicable or difficult to apply; see Sect. 4 for details. In particular, both Gaussian and bootstrap approximation results of the present paper allow classes of functions with *unbounded* envelopes provided suitable moment conditions are satisfied.

To illustrate the general approximation results for suprema of U-processes, we consider the problem of testing qualitative features of the conditional distribution and regression functions in nonparametric statistics [1,25,38]. In Sect. 4, we propose a unified test statistic for specifications (such as monotonicity, linearity, convexity, concavity, etc.) of nonparametric functions based on the *generalized local U-process* (the name is inspired by [27]). Instead of attempting to establish a Gumbel type limiting distribution for the extreme-value test statistic (which is known to have slow rates of convergence; see [30,46]), we apply the JMB to approximate the finite sample distribution of the proposed test statistic. Notably, the JMB is valid for a larger spectrum of bandwidths, allows for an unbounded envelope, and the size error of the JMB is decreasing polynomially fast in n, which should be contrasted with the fact that tests based on Gumbel approximations have size errors of order $1/\log n$. It is worth noting that [38], who develop a test for the stochastic monotonicity based on the supremum of a (second-order) U-process and derive a Gumbel limiting distribution

for their test statistic under the null, state a conjecture that a bootstrap resampling method would yield the test whose size error is decreasing polynomially fast in n [38, p. 594]. The results of the present paper formally solve this conjecture for a different version of bootstrap, namely, the JMB, in a more general setting. In addition, our general theory can be used to develop a version of the JMB test that is uniformly valid in compact bandwidth sets. Such "uniform-in-bandwidth" type results allow one to consider tests with data-dependent bandwidth selection procedures, which are not covered in [1,25,38].

1.1 Organization

The rest of the paper is organized as follows. In Sect. 2, we derive non-asymptotic Gaussian approximation error bounds for the U-process supremum in the non-degenerate case. In Sect. 3, we develop and study a jackknife multiplier bootstrap (with Gaussian weights) tailored to the U-process to further approximate the distribution of the U-process supremum in a data-dependent manner. In Sect. 4, we discuss applications of the general results developed in Sects. 2 and 3 to testing for qualitative features of nonparametric functions based on generalized local U-processes. In Sect. 5, we prove new *multi-resolution* and *local* maximal inequalities for degenerate U-processes with respect to the degeneracy levels of their kernel. These inequalities are key technical tools in the proofs for the results in the previous sections. In Sect. 6, we present the proofs for Sects. 2, 3. Appendix contains additional proofs, discussions, and auxiliary technical results.

1.2 Notation

For a nonempty set T, let $\ell^\infty(T)$ denote the Banach space of bounded real-valued functions $f : T \to \mathbb{R}$ equipped with the sup norm $\|f\|_T := \sup_{t\in T} |f(t)|$. For a pseudometric space (T, d), let $N(T, d, \varepsilon)$ denote the ε-covering number for (T, d), i.e., the minimum number of closed d-balls with radius at most ε that cover T. See [53, Section 2.1] or [29, Section 2.3] for details. For a probability space $(T, \mathcal{T}, Q)$ and a measurable function $f : T \to \mathbb{R}$, we use the notation $Qf := \int f dQ$ whenever the integral is defined. For $q \in [1, \infty]$, let $\|\cdot\|_{Q,q}$ denote the $L^q(Q)$-seminorm, i.e., $\|f\|_{Q,q} := (Q|f|^q)^{1/q} := (\int |f|^q dQ)^{1/q}$ for finite q while $\|f\|_{Q,\infty}$ denotes the essential supremum of $|f|$ with respect to Q. For a measurable space $(S, \mathcal{S})$ and a positive integer r, $S^r = S \times \cdots \times S$ (r times) denotes the product space equipped with the product σ-field $\mathcal{S}^r$. For a generic random variable Y (not necessarily real-valued), let $\mathcal{L}(Y)$ denote the law (distribution) of Y. For $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Let $\lfloor a \rfloor$ denote the integer part of $a \in \mathbb{R}$. "Constants" refer to finite, positive, and non-random numbers.

2 Gaussian approximation for suprema of *U*-processes

In this section, we derive non-asymptotic Gaussian approximation error bounds for the U-process supremum in the non-degenerate case, which is essential for establishing the bootstrap validity in Sect. 3. The goal is to approximate the supremum of the normalized U-process, $\sup_{h\in\mathcal{H}} \mathbb{U}_n(h)/r$, by the supremum of a suitable Gaussian process, and derive bounds on such approximations.

We first recall the setting. Let $X_1, \ldots, X_n$ be i.i.d. random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in a measurable space $(S, \mathcal{S})$ with common distribution P. For a technical reason, we assume that S is a separable metric space and $\mathcal{S}$ is its Borel σ-field. For a given integer $r \geqslant 2$, let $\mathcal{H}$ be a class of symmetric measurable functions $h : S^r \to \mathbb{R}$ equipped with a symmetric measurable envelope H. Recall the U-process $\{U_n(h) : h \in \mathcal{H}\}$ defined in (1) and its normalized version $\{\mathbb{U}_n(h) : h \in \mathcal{H}\}$ defined in (2). In applications, the function class $\mathcal{H}$ may depend on n, i.e., $\mathcal{H} = \mathcal{H}_n$. However, in Sects. 2 and 3, we will derive non-asymptotic results that are valid for each sample size n, and therefore suppress the possible dependence of $\mathcal{H} = \mathcal{H}_n$ on n for the notational convenience.

We will use the following notation. For a symmetric measurable function $h : S^r \to \mathbb{R}$ and $k = 1, \ldots, r$, let $P^{r-k}h$ denote the function on S^k defined by

$$P^{r-k}h(x_1, \ldots, x_k) = \mathbb{E}[h(x_1, \ldots, x_k, X_{k+1}, \ldots, X_r)] = \int \cdots \int h(x_1, \ldots, x_k, x_{k+1}, \ldots, x_r) dP(x_{k+1}) \cdots dP(x_r),$$

whenever the latter integral exists and is finite for every $(x_1, \ldots, x_k) \in S^k$ ($P^0 h = h$). Provided that $P^{r-k}h$ is well-defined, $P^{r-k}h$ is symmetric and measurable.

In this paper, we focus on the case where the function class $\mathcal{H}$ is *VC (Vapnik-Červonenkis) type*, whose formal definition is stated as follows.

Definition 2.1 (*VC type class*) A function class $\mathcal{H}$ on S^r with envelope H is said to be *VC type* with characteristics (A, v) if $\sup_Q N(\mathcal{H}, \|\cdot\|_{Q,2}, \varepsilon \|H\|_{Q,2}) \leqslant (A/\varepsilon)^v$ for all $0 < \varepsilon \leqslant 1$, where $\sup_Q$ is taken over all finitely discrete distributions on S^r.

We make the following assumptions on the function class $\mathcal{H}$ and the distribution P.

(PM) The function class $\mathcal{H}$ is *pointwise measurable*, i.e., there exists a countable subset $\mathcal{H}' \subset \mathcal{H}$ such that for every $h \in \mathcal{H}$, there exists a sequence $h_k \in \mathcal{H}'$ with $h_k \to h$ pointwise.

(VC) The function class $\mathcal{H}$ is VC type with characteristics $A \geqslant (e^{2(r-1)}/16) \vee e$ and $v \geqslant 1$ for envelope H. The envelope H satisfies that $H \in L^q(P^r)$ for some $q \in [4, \infty]$ and $P^{r-k}H$ is everywhere finite for every $k = 1, \ldots, r$.

(MT) Let $\mathcal{G} := P^{r-1}\mathcal{H} := \{P^{r-1}h : h \in \mathcal{H}\}$ and $G := P^{r-1}H$. There exist (finite) constants

$$b_{\mathfrak{h}} \geqslant b_{\mathfrak{g}} \vee \sigma_{\mathfrak{h}} \geqslant b_{\mathfrak{g}} \wedge \sigma_{\mathfrak{h}} \geqslant \overline{\sigma}_{\mathfrak{g}} > 0$$

such that the following hold:

$$\|G\|_{P,q} \leqslant b_{\mathfrak{g}}, \qquad \sup_{g\in\mathcal{G}} \|g\|_{P,\ell}^{\ell} \leqslant \overline{\sigma}_{\mathfrak{g}}^2 b_{\mathfrak{g}}^{\ell-2},\ \ell = 2,3,4,$$
$$\|P^{r-2}H\|_{P^2,q} \leqslant b_{\mathfrak{h}}, \text{ and } \sup_{h\in\mathcal{H}} \|P^{r-2}h\|_{P^2,\ell}^{\ell} \leqslant \sigma_{\mathfrak{h}}^2 b_{\mathfrak{h}}^{\ell-2},\ \ell = 2,4,$$

where q appears in Condition (VC).

Some comments on the conditions are in order. Conditions (PM), (VC), and (MT) are inspired by Conditions (A)–(C) in [15]. Condition (PM) is made to avoid measurability difficulties. Our definition of "pointwise measurability" is borrowed from Example 2.3.4 in [53]; [29, p. 262] calls a pointwise measurable function class a function class satisfying the *pointwise countable approximation property*. Condition (PM) ensures that, e.g., $\sup_{h\in\mathcal{H}} \mathbb{U}_n(h) = \sup_{h\in\mathcal{H}'} \mathbb{U}_n(h)$, so that $\sup_{h\in\mathcal{H}} \mathbb{U}_n(h)$ is a (proper) random variable. See [53, Section 2.2] for details.

Condition (VC) ensures that $\mathcal{G}$ is VC type as well with characteristics $4\sqrt{A}$ and $2v$ for envelope $G = P^{r-1}H$; see Lemma 5.4 ahead. Since $G \in L^2(P)$ by Condition (VC), it is seen from Dudley's criterion on sample continuity of Gaussian processes (see, e.g., [29, Theorem 2.3.7]) that the function class $\mathcal{G}$ is *P-pre-Gaussian*, i.e., there exists a tight Gaussian random variable W_P in $\ell^\infty(\mathcal{G})$ with mean zero and covariance function

$$\mathbb{E}[W_P(g)W_P(g')] = \mathrm{Cov}(g(X_1), g'(X_1)),\ g, g' \in \mathcal{G}.$$

Recall that a Gaussian process $W = \{W(g) : g \in \mathcal{G}\}$ is a tight Gaussian random variable in $\ell^\infty(\mathcal{G})$ if and only if $\mathcal{G}$ is totally bounded for the intrinsic pseudometric $d_W(g, g') = (\mathbb{E}[(W(g) - W(g'))^2])^{1/2}$, $g, g' \in \mathcal{G}$, and W has sample paths almost surely uniformly d_W-continuous [53, Section 1.5]. In applications, $\mathcal{G}$ may depend on n and so the Gaussian process W_P (and its distribution) may depend on n as well, although such dependences are suppressed in Sects. 2 and 3. The VC type assumption made in Condition (VC) covers many statistical applications. However, it is worth noting that in principle, we can derive corresponding results for Gaussian and bootstrap approximations under more general complexity assumptions on the function class beyond the VC type, as our local maximal inequalities for the U-process in Theorem 5.1 ahead, which are key technical results in the proofs of the Gaussian and bootstrap approximation results, can cover more general function classes than VC type classes; but the resulting bounds would be more complicated and may not be clear enough. For the clarity of exposition, we focus on VC type function classes and present a Gaussian coupling bound for general function classes in "Appendix E".

Condition (MT) imposes suitable moment bounds on the kernel and its Hájek projection. Specifically, this moment condition contains interpolated parameters which control the lower moments (i.e., L^2, L^3, and L^4 sizes) and the envelopes of $\mathcal{H}$ and $\mathcal{G}$.

Under these conditions on the function class $\mathcal{H}$ and the distribution P, we will first construct a random variable, defined on the same probability space as $X_1, \ldots, X_n$, which is equal in distribution to $\sup_{g\in\mathcal{G}} W_P(g)$ and "close" to Z_n with high-probability. To ensure such constructions, a common assumption is that the probability space is

rich enough. For the sake of clarity, we will assume in Sects. 2 and 3 that the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is such that

$$(\Omega, \mathcal{A}, \mathbb{P}) = (S^n, \mathcal{S}^n, P^n) \times (\Xi, \mathcal{C}, R) \times ((0,1), \mathcal{B}(0,1), U(0,1)), \tag{3}$$

where $X_1, \ldots, X_n$ are the coordinate projections of $(S^n, \mathcal{S}^n, P^n)$, multiplier random variables $\xi_1, \ldots, \xi_n$ to be introduced in Sect. 3 depend only on the "second" coordinate $(\Xi, \mathcal{C}, R)$, and $U(0,1)$ denotes the uniform distribution (Lebesgue measure) on $(0,1)$ ($\mathcal{B}(0,1)$ denotes the Borel σ-field on $(0,1)$). The augmentation of the last coordinate is reserved to generate a $U(0,1)$ random variable independent of $X_1, \ldots, X_n$ and $\xi_1, \ldots, \xi_n$, which is needed when applying the Strassen–Dudley theorem and its conditional version in the proofs of Proposition 2.1 and Theorem 3.1; see "Appendix B" for the Strassen–Dudley theorem and its conditional version. We will also assume that the Gaussian process W_P is defined on the same probability space (e.g. one can generate W_P by the previous $U(0,1)$ random variable), but of course $\sup_{g\in\mathcal{G}} W_P(g)$ is not what we want since there is no guarantee that $\sup_{g\in\mathcal{G}} W_P(g)$ is close to Z_n.

Now, we are ready to state the first result of this paper. Recall the notation given in Condition (MT) and define

$$K_n = v \log(A \vee n) \quad \text{and} \quad \chi_n = \sum_{k=3}^{r} n^{-(k-1)/2} \| P^{r-k} H \|_{P^k,2} K_n^{k/2}$$

with the convention that $\sum_{k=3}^{r} = 0$ if $r = 2$. The following proposition derives Gaussian coupling bounds for $Z_n = \sup_{h\in\mathcal{H}} \mathbb{U}_n(h)/r$.

Proposition 2.1 (Gaussian coupling bounds) *Let $Z_n = \sup_{h\in\mathcal{H}} \mathbb{U}_n(h)/r$. Suppose that Conditions (PM), (VC), and (MT) hold, and that $K_n^3 \leqslant n$. Then, for every $n \geqslant r+1$ and $\gamma \in (0,1)$, one can construct a random variable $\widetilde{Z}_{n,\gamma}$ such that $\mathcal{L}(\widetilde{Z}_{n,\gamma}) = \mathcal{L}(\sup_{g\in\mathcal{G}} W_P(g))$ and*

$$\mathbb{P}(|Z_n - \widetilde{Z}_{n,\gamma}| > C\varpi_n) \leqslant C'(\gamma + n^{-1}),$$

where C, C' are constants depending only on r, and

$$\varpi_n := \varpi_n(\gamma) := \frac{(\overline{\sigma}_{\mathfrak{g}}^2 b_{\mathfrak{g}} K_n^2)^{1/3}}{\gamma^{1/3} n^{1/6}} + \frac{1}{\gamma}\left(\frac{b_{\mathfrak{g}} K_n}{n^{1/2-1/q}} + \frac{\sigma_{\mathfrak{h}} K_n}{n^{1/2}} + \frac{b_{\mathfrak{h}} K_n^2}{n^{1-1/q}} + \chi_n\right). \tag{4}$$

In the case of $q = \infty$, "$1/q$" is interpreted as 0.

In statistical applications, bounds on the Kolmogorov distance are often more useful than coupling bounds. For two real-valued random variables V, Y, let $\rho(V, Y)$ denote the Kolmogorov distance between the distributions of V and Y, i.e., $\rho(V, Y) := \sup_{t\in\mathbb{R}} |\mathbb{P}(V \leqslant t) - \mathbb{P}(Y \leqslant t)|$. To derive a Kolomogorov distance bound, we will assume that there exists a constant $\underline{\sigma}_{\mathfrak{g}} > 0$ such that

$$\inf_{g\in\mathcal{G}} \|g - Pg\|_{P,2} \geqslant \underline{\sigma}_{\mathfrak{g}}. \tag{5}$$

Condition (5) implies that the U-process is non-degenerate. For the notational convenience, let $\widetilde{Z} = \sup_{g\in\mathcal{G}} W_P(g)$.

Corollary 2.2 (Bounds on the Kolmogorov distance between Z_n and $\sup_{g\in\mathcal{G}} W_P(g)$) *Assume that all the conditions in Proposition 2.1 and (5) hold. Then, there exists a constant C depending only on $r, \overline{\sigma}_{\mathfrak{g}}$ and $\underline{\sigma}_{\mathfrak{g}}$ such that*

$$\rho(Z_n, \widetilde{Z}) \leqslant C \left\{ \left(\frac{b_{\mathfrak{g}}^2 K_n^7}{n} \right)^{1/8} + \left(\frac{b_{\mathfrak{g}}^2 K_n^3}{n^{1-2/q}} \right)^{1/4} + \left(\frac{\sigma_{\mathfrak{h}}^2 K_n^3}{n} \right)^{1/4} + \left(\frac{b_{\mathfrak{h}} K_n^{5/2}}{n^{1-1/q}} \right)^{1/2} + \chi_n^{1/2} K_n^{1/4} \right\}.$$

In particular, if the function class $\mathcal{H}$ and the distribution P are independent of n, then $\rho(Z_n, \widetilde{Z}) = O(\{(\log n)^7/n\}^{1/8})$.

Condition (5) is used to apply the "anti-concentration" inequality for the Gaussian supremum (see Lemma A.1), which is a key technical ingredient of the proof of Corollary 2.2. The dependence of the constant C on the variance parameters $\underline{\sigma}_{\mathfrak{g}}$ and $\overline{\sigma}_{\mathfrak{g}}$ is not a serious restriction in statistical applications. In statistical applications, the function class $\mathcal{H}$ is often normalized in such a way that each function $g \in \mathcal{G}$ has (approximately) unit variance. In such cases, we may take $\underline{\sigma}_{\mathfrak{g}} = \overline{\sigma}_{\mathfrak{g}} = 1$ or $(\underline{\sigma}_{\mathfrak{g}}, \overline{\sigma}_{\mathfrak{g}})$ as positive constants independent of n; see Sect. 4 for details.

Remark 2.1 (*Comparisons with Gaussian approximations to suprema of empirical processes*) Our Gaussian coupling (Proposition 2.1) and approximation (Corollary 2.2) results are level-dependent on the Hoeffding projections of the U-process $\mathbb{U}_n$ (cf. (17) and (18) for formal definitions of the Hoeffding projections and decomposition). Specifically, we observe that: (1) $\underline{\sigma}_{\mathfrak{g}}, \overline{\sigma}_{\mathfrak{g}}, b_{\mathfrak{g}}$ quantify the contribution from the Hájek (empirical) process associated with $\mathbb{U}_n$; (2) $\sigma_{\mathfrak{h}}, b_{\mathfrak{h}}$ are related to the second-order degenerate component associated with $\mathbb{U}_n$; (3) χ_n contains the effect from all higher order projection terms of $\mathbb{U}_n$. For statistical applications in Sect. 4 where the function class $\mathcal{H} = \mathcal{H}_n$ changes with n, the second and higher order projections terms are not necessarily negligible and we have to take into account the contributions of all higher order projection terms. Hence, the Gaussian approximation for the U-process supremum of a general order is not parallel with the approximation results for the empirical process supremum [14,15].

3 Bootstrap approximation for suprema of *U*-processes

The Gaussian approximation results derived in the previous section are often not directly applicable in statistical applications such as computing critical values of a test statistic defined by the supremum of a U-process. This is because the covariance function of the approximating Gaussian process $W_P(g)$, $g \in \mathcal{G}$, is often unknown. In this section, we study a Gaussian multiplier bootstrap, tailored to the U-process, to

further approximate the distribution of the random variable $Z_n = \sup_{h\in\mathcal{H}} \mathbb{U}_n(h)/r$ in a data-dependent manner. The Gaussian approximation results will be used as building blocks for establishing validity of the Gaussian multiplier bootstrap.

We begin with noting that, in contrast to the empirical process case studied in [13,15], devising (Gaussian) multiplier bootstraps for the U-process is not straightforward. From the Gaussian approximation results, the distribution of Z_n is well approximated by the Gaussian supremum $\sup_{g\in\mathcal{G}} W_P(g)$. Hence, one might be tempted to approximate the distribution of $\sup_{g\in\mathcal{G}} W_P(g)$ by the conditional distribution of the supremum of the the multiplier process

$$\mathcal{G} \ni g \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \{g(X_i) - \overline{g}\}, \tag{6}$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. $N(0, 1)$ random variables independent of the data $X_1^n := \{X_1, \ldots, X_n\}$ and $\overline{g} = n^{-1} \sum_{i=1}^n g(X_i)$. However, a major problem of this approach is that, in statistical applications, functions in $\mathcal{G}$ are unknown to us since functions in $\mathcal{G}$ are of the form $P^{r-1}h$ for some $h \in \mathcal{H}$ and depend on the (unknown) underlying distribution P. Therefore, we must devise a multiplier bootstrap properly tailored to the U-process.

Motivated by this fundamental challenge, we propose and study the following version of Gaussian multiplier bootstrap. Let $\xi_1, \ldots, \xi_n$ be i.i.d. $N(0, 1)$ random variables independent of the data X_1^n [these multiplier variables will be assumed to depend only on the "second" coordinate in the probability space construction (3)]. We introduce the following multiplier process:

$$\mathbb{U}_n^\sharp(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \left[\frac{1}{|I_{n-1,r-1}|} \sum_{(i,i_2,\ldots,i_r)\in I_{n,r}} h(X_i, X_{i_2}, \ldots, X_{i_r}) - U_n(h) \right], \quad h \in \mathcal{H}, \tag{7}$$

where $\sum_{(i,i_2,\ldots,i_r)}$ is taken with respect to $(i_2, \ldots, i_r)$ while keeping i fixed. The process $\{\mathbb{U}_n^\sharp(h) : h \in \mathcal{H}\}$ is a centered Gaussian process conditionally on the data X_1^n and can be regarded as a version of the (infeasible) multiplier process (6) with each $g(X_i)$ replaced by a jackknife estimate. In fact, the multiplier process (6) can be alternatively represented as

$$\mathcal{H} \ni h \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \{(P^{r-1}h)(X_i) - \overline{P^{r-1}h}\}, \tag{8}$$

where $\overline{P^{r-1}h} = n^{-1} \sum_{i=1}^n P^{r-1}h(X_i)$. For $x \in S$, denote by δ_x the Dirac measure at x and denote by $\delta_x h$ the function on S^{r-1} defined by $(\delta_x h)(x_2, \ldots, x_r) = h(x, x_2, \ldots, x_r)$ for $(x_2, \ldots, x_r) \in S^{r-1}$. For each $i = 1, \ldots, n$ and a function f on S^{r-1}, let $U_{n-1,-i}^{(r-1)}(f)$ denote the U-statistic with kernel f for the sample without the i-th observation, i.e.,

$$U_{n-1,-i}^{(r-1)}(f) = \frac{1}{|I_{n-1,r-1}|} \sum_{(i,i_2,\ldots,i_r)\in I_{n,r}} f(X_{i_2},\ldots,X_{i_r}).$$

Then the proposed multiplier process (7) can be alternatively written as

$$\mathbb{U}_n^\sharp(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \left[U_{n-1,-i}^{(r-1)}(\delta_{X_i} h) - U_n(h) \right],$$

that is, our multiplier process (7) replaces each $(P^{r-1}h)(X_i)$ in the infeasible multiplier process (8) by its jackknife estimate $U_{n-1,-i}^{(r-1)}(\delta_{X_i}h)$.

In practice, we approximate the distribution of Z_n by the conditional distribution of the supremum of the multiplier process $Z_n^\sharp := \sup_{h\in\mathcal{H}} \mathbb{U}_n^\sharp(h)$ given X_1^n, which can be further approximated by Monte Carlo simulations on the multiplier variables.

To the best of our knowledge, our multiplier bootstrap method for U-processes is new in the literature, at least in this generality; see Remark 3.1 for comparisons with other bootstraps for U-processes. We call the resulting bootstrap method the *jackknife multiplier bootstrap* (JMB) for U-processes.

Now, we turn to proving validity of the proposed JMB. We will first construct couplings $Z_n^\sharp$ and $\widetilde{Z}_n^\sharp := \widetilde{Z}_{n,\gamma}^\sharp$ (a real-valued random variable that may depend on the coupling error $\gamma \in (0,1)$) such that: 1) $\mathcal{L}(\widetilde{Z}_n^\sharp \mid X_1^n) = \mathcal{L}(\widetilde{Z})$, where $\mathcal{L}(\cdot \mid X_1^n)$ denotes the conditional law given X_1^n (i.e., $\widetilde{Z}_n^\sharp$ is independent of X_1^n and has the same distribution as $\widetilde{Z} = \sup_{g\in\mathcal{G}} W_P(g)$); and at the same time 2) $Z_n^\sharp$ and $\widetilde{Z}_n^\sharp$ are "close" to each other. Construction of such couplings leads to validity of the JMB. To see this, suppose that $Z_n^\sharp$ and $\widetilde{Z}_n^\sharp$ are close to each other, namely, $\mathbb{P}(|Z_n^\sharp - \widetilde{Z}_n^\sharp| > r_1) \leqslant r_2$ for some small $r_1, r_2 > 0$. To ease the notation, denote by $\mathbb{P}_{|X_1^n}$ and $\mathbb{E}_{|X_1^n}$ the conditional probability and expectation given X_1^n, respectively (i.e., the notation $\mathbb{P}_{|X_1^n}$ corresponds to taking probability with respect to the "latter two" coordinates in (3) while fixing X_1^n). Then,

$$\mathbb{P}\left\{ \mathbb{P}_{|X_1^n}(|Z_n^\sharp - \widetilde{Z}_n^\sharp| > r_1) > r_2^{1/2} \right\} \leqslant r_2^{1/2}$$

by Markov's inequality, so that, on the event $\{\mathbb{P}_{|X_1^n}(|Z_n^\sharp - \widetilde{Z}_n^\sharp| > r_1) \leqslant r_2^{1/2}\}$ whose probability is at least $1 - r_2^{1/2}$, for every $t \in \mathbb{R}$,

$$\mathbb{P}_{|X_1^n}(Z_n^\sharp \leqslant t) \leqslant \mathbb{P}_{|X_1^n}(\widetilde{Z}_n^\sharp \leqslant t + r_1) + r_2^{1/2} = \mathbb{P}(\widetilde{Z} \leqslant t + r_1) + r_2^{1/2},$$

and likewise $\mathbb{P}_{|X_1^n}(Z_n^\sharp \leqslant t) \geqslant \mathbb{P}(\widetilde{Z} \leqslant t - r_1) - r_2^{1/2}$. Hence, on that event,

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}_{|X_1^n}(Z_n^\sharp \leqslant t) - \mathbb{P}(\widetilde{Z} \leqslant t) \right| \leqslant \sup_{t\in\mathbb{R}} \mathbb{P}(|\widetilde{Z} - t| \leqslant r_1) + r_2^{1/2}.$$

The first term on the right hand side can be bounded by using the anti-concentration inequality for the supremum of a Gaussian process (cf. [14, Lemma A.1] which is

stated in Lemma A.1 in "Appendix A"), and combining the Gaussian approximation results, we obtain a bound on the Kolmogorov distance between $\mathcal{L}(Z_n^\sharp \mid X_1^n)$ and $\mathcal{L}(Z_n)$ on an event with probability close to one, which leads to validity of the JMB.

The following theorem is the main result of this paper and derives bounds on such couplings. To state the next theorem, we need the additional notation. For a symmetric measurable function f on S^2, define $f^{\odot 2} = f_P^{\odot 2}$ by

$$f^{\odot 2}(x_1, x_2) := \int f(x_1, x) f(x, x_2) dP(x).$$

Let $\nu_{\mathfrak{h}} := \|(P^{r-2}H)^{\odot 2}\|_{P^2,q/2}^{1/2}$.

Theorem 3.1 (Bootstrap coupling bounds) *Let $Z_n^\sharp = \sup_{h\in\mathcal{H}} \mathbb{U}_n^\sharp(h)$. Suppose that Conditions (PM), (VC), and (MT) hold. Furthermore, suppose that*

$$\begin{gathered}\sigma_{\mathfrak{h}} K_n^{1/2} \leqslant \overline{\sigma}_{\mathfrak{g}} n^{1/2},\ \nu_{\mathfrak{h}} K_n \leqslant \overline{\sigma}_{\mathfrak{g}} n^{3/4-1/q},\ (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/2} K_n^{3/4} \leqslant \overline{\sigma}_{\mathfrak{g}} n^{3/4}, \\ b_{\mathfrak{h}} K_n^{3/2} \leqslant \overline{\sigma}_{\mathfrak{g}} n^{1-1/q},\ \text{and } \chi_n \leqslant \overline{\sigma}_{\mathfrak{g}}.\end{gathered} \tag{9}$$

Then, for every $n \geqslant r+1$ and $\gamma \in (0,1)$, one can construct a random variable $\widetilde{Z}_{n,\gamma}^\sharp$ such that $\mathcal{L}(\widetilde{Z}_{n,\gamma}^\sharp \mid X_1^n) = \mathcal{L}(\sup_{g\in\mathcal{G}} W_P(g))$ and

$$\mathbb{P}(|Z_n^\sharp - \widetilde{Z}_{n,\gamma}^\sharp| > C\varpi_n^\sharp) \leqslant C'(\gamma + n^{-1}),$$

where C, C' are constants depending only on r, and

$$\begin{aligned}\varpi_n^\sharp &:= \varpi_n^\sharp(\gamma) \\ &:= \frac{1}{\gamma^{3/2}} \left\{ \frac{\{(b_{\mathfrak{g}} \vee \sigma_{\mathfrak{h}})\overline{\sigma}_{\mathfrak{g}} K_n^{3/2}\}^{1/2}}{n^{1/4}} + \frac{b_{\mathfrak{g}} K_n}{n^{1/2-1/q}} + \frac{(\overline{\sigma}_{\mathfrak{g}} \nu_{\mathfrak{h}})^{1/2} K_n}{n^{3/8-1/(2q)}} \right. \\ &\quad \left. + \frac{\overline{\sigma}_{\mathfrak{g}}^{1/2} (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/4} K_n^{7/8}}{n^{3/8}} + \frac{(\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{h}})^{1/2} K_n^{5/4}}{n^{1/2-1/(2q)}} + \overline{\sigma}_{\mathfrak{g}}^{1/2} \chi_n^{1/2} K_n^{1/2} \right\}.\end{aligned} \tag{10}$$

In the case of $q = \infty$, "$1/q$" is interpreted as 0.

We note that $\nu_{\mathfrak{h}}^q \leqslant \|P^{r-2}H\|_{P^2,q}^q \leqslant b_{\mathfrak{h}}^q$, but in our applications $\nu_{\mathfrak{h}} \ll b_{\mathfrak{h}}$ and this is why we introduced such a seemingly complicated definition for $\nu_{\mathfrak{h}}$. To see that $\nu_{\mathfrak{h}} \leqslant b_{\mathfrak{h}}$, observe that by the Cauchy–Schwarz and Jensen inequalities,

$$\begin{aligned}\nu_{\mathfrak{h}}^q &= \iint \left\{ \int (P^{r-2}H)(x_1, x)(P^{r-2}H)(x, x_2) dP(x) \right\}^{q/2} dP(x_1) dP(x_2) \\ &\leqslant \left\{ \iint (P^{r-2}H)^{q/2}(x_1, x_2) dP(x_1) dP(x_2) \right\}^2 \\ &\leqslant \iint (P^{r-2}H)^q(x_1, x_2) dP(x_1) dP(x_2) \leqslant b_{\mathfrak{h}}^q.\end{aligned}$$

Condition (9) is not restrictive. In applications, the function class $\mathcal{H}$ is often normalized in such a way that $\overline{\sigma}_{\mathfrak{g}}$ is of constant order, and under this normalization, Condition (9) is a merely necessary condition for the coupling bound (10) to tend to zero.

The proof of Theorem 3.1 is lengthy and involved. A delicate part of the proof is to sharply bound the sup-norm distance between the conditional covariance function of the multiplier process $\mathbb{U}_n^\sharp$ and the covariance function of W_P, which boils down to bounding the term

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \{U_{n-1,-i}^{(r-1)}(\delta_{X_i} h) - P^{r-1} h(X_i)\}^2 \right\|_{\mathcal{H}}.$$

To this end, we make use of the following observation: for a P^{r-1}-integrable function f on S^{r-1}, $U_{n-1,-i}^{(r-1)}(f)$ is a U-statistic of order $(r-1)$, and denote by $S_{n-1,-i}(f)$ its first Hoeffding projection term. Conditionally on X_i, $U_{n-1,-i}^{(r-1)}(\delta_{X_i} h) - P^{r-1} h(X_i) - S_{n-1,-i}(\delta_{X_i} h)$ is a degenerate U-process, and we will bound the expectation of the squared supremum of this term conditionally on X_i using "simpler" maximal inequalities (Corollary 5.6 ahead). On the other hand, the term $n^{-1} \sum_{i=1}^{n} \{S_{n-1,-i}(\delta_{X_i} h)\}^2$ is decomposed into

$$n^{-1}(\text{non-degenerate } U\text{-statistic of order 2}) + (\text{degenerate } U\text{-statistic of order 3}),$$

where the order of degeneracy of the latter term is 1, and we will apply "sharper" local maximal inequalities (Corollary 5.5 ahead) to bound the suprema of both terms. Such a delicate combination of different maximal inequalities turns out to be crucial to yield sharper regularity conditions for validity of the JMB in our applications. In particular, if we bound the sup-norm distance between the conditional covariance function of $\mathbb{U}_n^\sharp$ and the covariance function of W_P in a cruder way, then this will lead to more restrictive conditions on bandwidths in our applications, especially for the "uniform-in-bandwidth" results [cf. Condition (T5′) in Theorem 4.4].

The following corollary derives a "high-probability" bound for the Kolmogorov distance between $\mathcal{L}(Z_n^\sharp \mid X_1^n)$ and $\mathcal{L}(\widetilde{Z})$ (here a high-probability bound refers to a bound holding with probability at least $1 - Cn^{-c}$ for some constants C, c).

Corollary 3.2 (Validity of the JMB) *Suppose that Conditions (PM), (VC), (MT), and (5) hold. Let*

$$\eta_n := \frac{\{(b_{\mathfrak{g}} \vee \sigma_{\mathfrak{h}}) K_n^{5/2}\}^{1/2}}{n^{1/4}} + \frac{b_{\mathfrak{g}} K_n^{3/2}}{n^{1/2-1/q}} + \frac{v_{\mathfrak{h}}^{1/2} K_n^{3/2}}{n^{3/8-1/(2q)}} + \frac{(\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/4} K_n^{11/8}}{n^{3/8}} + \frac{b_{\mathfrak{h}}^{1/2} K_n^{7/4}}{n^{1/2-1/(2q)}} + \chi_n^{1/2} K_n$$

with the convention that $1/q = 0$ when $q = \infty$. Then, there exist constants C, C' depending only on r, $\overline{\sigma}_{\mathfrak{g}}$, and $\underline{\sigma}_{\mathfrak{g}}$ such that, with probability at least $1 - C\eta_n^{1/4}$,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}_{|X_1^n}(Z_n^\sharp \leqslant t) - \mathbb{P}(\widetilde{Z} \leqslant t)\right| \leqslant C'\eta_n^{1/4}.$$

If the function class $\mathcal{H}$ and the distribution P are independent of n, then $\eta_n^{1/4}$ is of order $n^{-1/16}$, which is polynomially decreasing in n but appears to be non-sharp. Sharper bounds could be derived by improving on $\gamma^{-3/2}$ in front of the $n^{-1/4}$ term in (10). The proof of Theorem 3.1 consists of constructing a "high-probability" event on which, e.g., the sup-norm distance between the conditional covariance function of $\mathbb{U}_n^\sharp$ and the covariance function of W_P is small. To construct such a high-probability event, the current proof repeatedly relies on Markov's inequality, which could be replaced by more sophisticated deviation inequalities. However, this is at the cost of more technical difficulties and more restrictive moment conditions. In addition, we derive a conditional UCLT for the JMB in "Appendix D" when $\mathcal{H}$ is fixed and P does not depend on n.

Remark 3.1 (*Connections to other bootstraps*) There are several versions of bootstraps for *non-degenerate* U-processes. The most celebrated one is the *empirical bootstrap*

$$\mathbb{U}_n^*(h) = \frac{\sqrt{n}}{r|I_{n,r}|}\sum_{(i_1,\ldots,i_r)\in I_{n,r}}\left\{h(X_{i_1}^*,\ldots,X_{i_r}^*) - V_n(h)\right\},\ h\in\mathcal{H},$$

where $X_1^*,\ldots,X_n^*$ are i.i.d. draws from the empirical distribution $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ and $V_n(h) = n^{-r}\sum_{i_1,\ldots,i_r=1}^n h(X_{i_1},\ldots,X_{i_r})$ is the V-statistic associated with kernel h (cf. [5,6,11]). A slightly different bootstrap procedure

$$\mathbb{U}_n^\natural(h) = n^{-r+1/2}\sum_{1\leqslant i_1,\ldots,i_r\leqslant n}\left\{h(X_{i_1}^*,X_{i_2},\ldots,X_{i_r}) - h(X_{i_1},X_{i_2},\ldots,X_{i_r}))\right\},$$
$$h\in\mathcal{H},$$

is proposed in [3]; see Remark 2.7 therein. If $\mathcal{H} = \{h\}$ is a singleton and the associated U-statistic $U_n(h)$ is non-degenerate, then $\mathbb{U}_n^\natural(h)$ and $\mathbb{U}_n^*(h)$ are asymptotically equivalent in the sense that they have the same weak limit that is given by the centered Gaussian random variable $W_P(P^{r-1}h)$; see Theorem 2.4 and Corollary 2.6 in [3]. Since the bootstrap $\mathbb{U}_n^\natural(h)$ can be viewed as the empirical bootstrap applied to a V-statistic estimate of the Hájek projection, i.e., $\mathbb{U}_n^\natural(h) = n^{-1/2}\sum_{i=1}^n(\delta_{X_i^*} - P_n)P_n^{r-1}h$, our JMB is connected to (but still different from) $\mathbb{U}_n^\natural(h)$ in the sense that we apply the multiplier bootstrap to a jackknife U-statistic estimate of the Hajek projection. Another example is the *Bayesian bootstrap* (with Dirichlet weights)

$$\mathbb{U}_n^\flat(h) = \frac{\sqrt{n}}{r|I_{n,r}|}\sum_{(i_1,\ldots,i_r)\in I_{n,r}}(w_{i_1}\cdots w_{i_r} - 1)h(X_{i_1},\ldots,X_{i_r}),\ h\in\mathcal{H},$$

where $w_i = \eta_i/(n^{-1}\sum_{j=1}^n\eta_j)$ for $i = 1,\ldots,n$ and $\eta_1,\ldots,\eta_n$ are i.i.d. exponential random variables with mean one (i.e., $(w_1,\ldots,w_n)$ follows a scaled Dirichlet distri-

bution) independent of $X_1^n = \{X_1, \ldots, X_n\}$ [39,40,48,56]. If $\mathcal{H}$ is a fixed VC type function class and the distribution P is independent of n (hence the distribution of the approximating Gaussian process W_P is independent of n), then the conditional distributions (given X_1^n) of the empirical bootstrap process $\{\mathbb{U}_n^*(h) : h \in \mathcal{H}\}$ and the Bayesian bootstrap process $\{\mathbb{U}_n^\flat(h) : h \in \mathcal{H}\}$ (with Dirichlet weights) are known to have the same weak limit as the U-process $\{r^{-1}\mathbb{U}_n(h) : h \in \mathcal{H}\}$, where the weak limit is the Gaussian process $W_P \circ P^{r-1}$ in the non-degenerate case [5,56]. The proposed multiplier process in (7) is also connected to the empirical and Baysian bootstraps (or more general randomly reweighted bootstraps) in the sense that the latter two bootstraps also implicitly construct an empirical process whose conditional covariance function is close to that of W_P under the supremum norm (cf. [11]). Recall that the conditional covariance function of $\mathbb{U}_n^\sharp$ can be viewed as a jackknife estimate of the covariance function of W_P. For the special case where $r = 2$ and $\mathcal{H} = \mathcal{H}_n$ is such that $|\mathcal{H}_n| < \infty$ and $|\mathcal{H}_n|$ is allowed to increase with n, [11] shows that the Gaussian multiplier, empirical and randomly reweighted bootstraps ($\mathbb{U}_n^\flat(h)$ with i.i.d. Gaussian weights $w_i \sim N(1, 1)$) all achieve similar error bounds. In the U-process setting, it would be possible to establish finite sample validity for the empirical and more general randomly reweighted bootstraps, but this is at the price of a much more involved technical analysis which we do not pursue in the present paper.

4 Applications: testing for qualitative features based on generalized local *U*-processes

In this section, we discuss applications of the general results in the previous sections to *generalized local U-processes*, which are motivated from testing for qualitative features of functions in nonparametric statistics (see below for concrete statistical problems).

Let $m \geqslant 1, r \geqslant 2$ be fixed integers and let $\mathcal{V}$ be a separable metric space. Suppose that $n \geqslant r + 1$, and let $D_i = (X_i, V_i), i = 1, \ldots, n$ be i.i.d. random variables taking values in $\mathbb{R}^m \times \mathcal{V}$ with joint distribution P defined on the product σ-field on $\mathbb{R}^m \times \mathcal{V}$ (we equip $\mathbb{R}^m$ and $\mathcal{V}$ with the Borel σ-fields). The variable V_i may include some components of X_i. Let Φ be a class of *symmetric* measurable functions $\varphi : \mathcal{V}^r \to \mathbb{R}$, and let $L : \mathbb{R}^m \to \mathbb{R}$ be a (fixed) "kernel function", i.e., an integrable function on $\mathbb{R}^m$ (with respect to the Lebesgue measure) such that $\int_{\mathbb{R}^m} L(x)dx = 1$. For $b > 0$ ("bandwidth"), we use the notation $L_b(\cdot) = b^{-m}L(\cdot/b)$. For a given sequence of bandwidths $b_n \to 0$, let

$$h_{n,\vartheta}(d_1, \ldots, d_r) := \varphi(v_1, \ldots, v_r) \prod_{k=1}^{r} L_{b_n}(x - x_k), \ \vartheta = (x, \varphi) \in \Theta := \mathcal{X} \times \Phi,$$

where $\mathcal{X} \subset \mathbb{R}^m$ is a (nonempty) compact subset. Consider the U-process

$$U_n(h_{n,\vartheta}) := U_n^{(r)}(h_{n,\vartheta}) := \frac{1}{|I_{n,r}|} \sum_{(i_1,\ldots,i_r)\in I_{n,r}} h_{n,\vartheta}(D_{i_1}, \ldots, D_{i_r}),$$

which we call, following [27], the *generalized local U-process*. The indexing function class is $\{h_{n,\vartheta} : \vartheta \in \Theta\}$ which depends on the sample size n. The U-process $U_n(h_{n,\vartheta})$ can be seen as a process indexed by Θ, but in general is not weakly convergent in the space $\ell^\infty(\Theta)$, even after a suitable normalization (an exception is the case where $\mathcal{X}$ and Φ are finite sets, and in that case, under regularity conditions, the vector $\{\sqrt{nb_n^m}(U_n(h_{n,\vartheta}) - P^r h_{n,\vartheta})\}_{\vartheta\in\Theta}$ converges weakly to a multivariate normal distribution). In addition, we will allow the set Θ to depend on n.

We are interested in approximating the distribution of the normalized version of this process

$$S_n = \sup_{\vartheta\in\Theta} \frac{\sqrt{nb_n^m}\{U_n(h_{n,\vartheta}) - P^r h_{n,\vartheta}\}}{rc_n(\vartheta)},$$

where $c_n(\vartheta) > 0$ is a suitable normalizing constant. The goal of this section is to characterize conditions under which the JMB developed in the previous section is consistent for approximating the distribution of S_n (more generally we will allow the normalizing constant $c_n(\vartheta)$ to be data-dependent). There are a number of statistical applications where we are interested in approximating distributions of such statistics. We provide a couple of examples. All the test statistics discussed in Examples in 4.1 and 4.2 are covered by our general framework. In Examples 4.1 and 4.2, $\alpha \in (0, 1)$ is a nominal level.

Example 4.1 (*Testing stochastic monotonicity*) Let X, Y be real-valued random variables and denote by $F_{Y|X}(y \mid x)$ the conditional distribution function of Y given X. Consider the problem of testing the stochastic monotonicity

$$H_0 : F_{Y|X}(y \mid x) \leqslant F_{Y|X}(y \mid x') \ \forall y \in \mathbb{R} \text{ whenever } x \geqslant x'.$$

Testing for the stochastic monotonicity is an important topic in a variety of applied fields such as economics [7,23,52]. For this problem, [38] consider a test for H_0 based on a local Kendall's tau statistic, inspired by [25]. Let $(X_i, Y_i), i = 1, \ldots, n$ be i.i.d. copies of (X, Y). Lee et al. [38] consider the U-process

$$U_n(x, y) = \frac{1}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} \{1(Y_i \leqslant y) - 1(Y_j \leqslant y)\} \\ \operatorname{sign}(X_i - X_j) L_{b_n}(x - X_i) L_{b_n}(x - X_j),$$

where $b_n \to 0$ is a sequence of bandwidths and $\operatorname{sign}(x) = 1(x > 0) - 1(x < 0)$ is the sign function. They propose to reject the null hypothesis if $S_n = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} U_n(x, y)/c_n(x)$ is large, where $\mathcal{X}, \mathcal{Y}$ are subsets of the supports of X, Y, respectively and $c_n(x) > 0$ is a suitable normalizing constant. Lee et al. [38] argue that as far as the size control is concerned, it is enough to choose, as a critical value, the $(1 - \alpha)$-quantile of S_n when X, Y are independent, under which $U_n(x, y)$ is centered. Under independence between X and Y, and under regularity conditions, they derive a Gumbel limiting distribution for a properly scaled version of S_n using

techniques from [45], but do not consider bootstrap approximations to S_n. It should be noted that [38] consider a slightly more general setup than that described above in the sense that they allow X_i not to be directly observed but assume that estimated X_i are available, and also cover the case where X is multidimensional.

Example 4.2 (*Testing curvature and monotonicity of nonparametric regression*) Consider the nonparametric regression model $Y = f(X) + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$, where Y is a scalar outcome variable, X is an m-dimensional vector of regressors, ε is an error term, and f is the conditional mean function $f(x) = \mathbb{E}[Y \mid X = x]$. We observe i.i.d. copies $V_i = (X_i, Y_i)$, $i = 1, \ldots, n$ of $V = (X, Y)$. We are interested in testing for qualitative features (e.g., curvature, monotonicity) of the regression function f.

Abrevaya and Jiang [1] consider a simplex statistic to test linearity, concavity, convexity of f under the assumption that the conditional distribution of ε given X is symmetric. To define their test statistics, for $x_1, \ldots, x_{m+1} \in \mathbb{R}^m$, let $\Delta^\circ(x_1, \ldots, x_{m+1}) = \{\sum_{i=1}^{m+1} a_i x_i : 0 < a_j < 1, j = 1, \ldots, m+1, \sum_{i=1}^{m+1} a_i = 1\}$ denote the interior of the simplex spanned by $x_1, \ldots, x_{m+1}$, and define $\mathcal{D} = \bigcup_{j=1}^{m+2} \mathcal{D}_j$, where

$$\mathcal{D}_j = \left\{(x_1, \ldots, x_{m+2}) \in \mathbb{R}^{m\times(m+2)} : \begin{array}{l} x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{m+2} \text{ are affinely independent} \\ \text{and } x_j \in \Delta^\circ(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{m+2}) \end{array}\right\}.$$

The sets $\mathcal{D}_1, \ldots, \mathcal{D}_{m+2}$ are disjoint. For given $v_i = (x_i, y_i) \in \mathbb{R}^m \times \mathbb{R}$, $i = 1, \ldots, m+2$, if $(x_1, \ldots, x_{m+2}) \in \mathcal{D}$ then there exist a unique index $j = 1, \ldots, m+2$ and a unique vector $(a_i)_{1\leqslant i\leqslant m+2, i\neq j}$ such that $0 < a_i < 1$ for all $i \neq j$, $\sum_{i\neq j} a_i = 1$, and $x_j = \sum_{i\neq j} a_i x_i$; then, define $w(v_1, \ldots, v_{m+2}) = \sum_{i\neq j} a_i y_i - y_j$. The index j and vector $(a_i)_{1\leqslant i\leqslant m+2, i\neq j}$ are functions of x_i's. The set $\mathcal{D}$ is symmetric (i.e., its indicator function is symmetric) and $w(v_1, \ldots, v_{m+2})$ is symmetric in its arguments.

Under this notation, [1] consider the following *localized simplex statistic*

$$U_n(x) = \frac{1}{|I_{n,m+2}|} \sum_{(i_1,\ldots,i_{m+2})\in I_{n,m+2}} \varphi(V_{i_1}, \ldots, V_{i_{m+2}}) \prod_{k=1}^{m+2} L_{b_n}(x - X_{i_k}), \quad (11)$$

where $\varphi(v_1, \ldots, v_{m+2}) = 1\{(x_1, \ldots, x_{m+2}) \in \mathcal{D}\}\text{sign}(w(v_1, \ldots, v_{m+2}))$, which is a U-process of order $(m+2)$. To test concavity and convexity of f, [1] propose to reject the hypotheses if $\overline{S}_n = \sup_{x\in\mathcal{X}} U_n(x)/c_n(x)$ and $\underline{S}_n = \inf_{x\in\mathcal{X}} U_n(x)/c_n(x)$ are large and small, respectively, where $\mathcal{X}$ is a subset of the support of X and $c_n(x) > 0$ is a suitable normalizing constant. The infimum statistic $\underline{S}_n$ can be written as the supremum of a U-process by replacing φ with $-\varphi$, so we will focus on $\overline{S}_n$. Precisely speaking, they consider to take discrete deign points $x_1, \ldots, x_G$ with $G = G_n \to \infty$, and take the supremum or infimum on the discrete grids $\{x_1, \ldots, x_G\}$. Abrevaya and Jiang [1] argue that as far as the size control is concerned, it is enough to choose, as a critical value, the $(1-\alpha)$-quantile of $\overline{S}_n$ when f is linear, under which $U_n(x)$ is centered due to the symmetry assumption on the distribution of ε conditionally on X. Under linearity of f, [1, Theorem 6] claims to derive a Gumbel limiting distribution for a properly scaled version of $\overline{S}_n$, but the authors think that their proof needs a further justification.

The proof of Theorem 6 in [1] proves that, in their notation, the *marginal* distributions of $\widetilde{U}_{n,h}(x_g^*)$ converge to $N(0,1)$ uniformly in $g = 1, \ldots, G$ (see their equation (A.1)), and the covariances between $\widetilde{U}_{n,h}(x_g^*)$ and $\widetilde{U}_{n,h}(x_{g'}^*)$ for $g \neq g'$ are approaching zero faster than the variances, but what they need to show is that the *joint* distribution of $(\widetilde{U}_{n,h}(x_1^*), \ldots, \widetilde{U}_{n,h}(x_G^*))$ is approximated by $N(0, I_G)$ in a suitable sense, which is lacking in their proof. An alternative proof strategy is to apply Rio's coupling [47] to the Hájek process associated to U_n, but it seems non-trivial to apply Rio's coupling since it is non-trivial to verify that the function φ is of bounded variation.

On the other hand, [25] study testing monotonicity of f when $m = 1$ and ε is independent of X. Specifically, they consider testing whether f is increasing, and propose to reject the hypothesis if $S_n = \sup_{x \in \mathcal{X}} \check{U}_n(x)/c_n(x)$ is large, where $\mathcal{X}$ is a subset of the support of X,

$$\check{U}_n(x) = \frac{1}{n(n-1)} \sum_{1 \leqslant i \neq j \leqslant n} \operatorname{sign}(Y_j - Y_i) \operatorname{sign}(X_i - X_j) L_{b_n}(x - X_i) L_{b_n}(x - X_j), \tag{12}$$

and $c_n(x) > 0$ is a suitable normalizing constant. Ghosal et al. [25] argue that as far as the size control is concerned, it is enough to choose, as a critical value, the $(1-\alpha)$-quantile of S_n when $f \equiv 0$, under which $U_n(x)$ is centered. Under $f \equiv 0$ and under regularity conditions, [25] derive a Gumbel limiting distribution for a properly scaled version of S_n but do not study bootstrap approximations to S_n.

In Appendix F, we discuss some alternative tests in the literature for concavity/convexity and monotonicity of regression functions.

Now, we go back to the general case. In applications, a typical choice of the normalizing constant $c_n(\vartheta)$ is $c_n(\vartheta) = b_n^{m/2} \sqrt{\operatorname{Var}_P(P^{r-1} h_{n,\vartheta})}$ where $\operatorname{Var}_P(\cdot)$ denotes the variance under P, so that each $b_n^{m/2} c_n(\vartheta)^{-1} P^{r-1} h_{n,\vartheta}$ is normalized to have unit variance, but other choices (such as $c_n(\vartheta) \equiv 1$) are also possible. The choice $c_n(\vartheta) = b_n^{m/2} \sqrt{\operatorname{Var}_P(P^{r-1} h_{n,\vartheta})}$ depends on the unknown distribution P and needs to be estimated in practice. Suppose in general (i.e., $c_n(\vartheta)$ need not to be $b_n^{m/2} \sqrt{\operatorname{Var}_P(P^{r-1} h_{n,\vartheta})}$) that there is an estimator $\widehat{c}_n(\vartheta) = \widehat{c}_n(\vartheta; D_1^n) > 0$ for $c_n(\vartheta)$ for each $\vartheta \in \Theta$, and instead of original S_n, consider

$$\widehat{S}_n := \sup_{\vartheta \in \Theta} \frac{\sqrt{n b_n^m} \{U_n(h_{n,\vartheta}) - P^r h_{n,\vartheta}\}}{r \widehat{c}_n(\vartheta)}.$$

We consider to approximate the distribution of $\widehat{S}_n$ by the conditional distribution of the JMB analogue of $\widehat{S}_n$: $\widehat{S}_n^\sharp := \sup_{\vartheta \in \Theta} b_n^{m/2} \mathbb{U}_n^\sharp(h_{n,\vartheta})/\widehat{c}_n(\vartheta)$, where

$$\mathbb{U}_n^\sharp(h_{n,\vartheta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \left[U_{n-1,-i}^{(r-1)}(\delta_{D_i} h_{n,\vartheta}) - U_n(h_{n,\vartheta}) \right], \ \vartheta \in \Theta,$$

and $\xi_1, \dots, \xi_n$ are i.i.d. $N(0,1)$ random variables independent of $D_1^n = \{D_i\}_{i=1}^n$. Recall that for a function f on $(\mathbb{R}^m \times \mathcal{V})^{r-1}$, $U_{n-1,-i}^{(r-1)}(f)$ denotes the U-statistic with kernel f for the sample without the i-th observation, i.e., $U_{n-1,-i}^{(r-1)}(f) = |I_{n-1,r-1}|^{-1} \sum_{(i,i_2,\dots,i_r)\in I_{n,r}} f(D_{i_2}, \dots, D_{i_r})$.

Let ζ, c_1, c_2, and C_1 be given positive constants such that $C_1 > 1$ and $c_2 \in (0,1)$, and let $q \in [4, \infty]$. Denote by $\mathcal{X}^\zeta$ the ζ-enlargement of $\mathcal{X}$, i.e., $\mathcal{X}^\zeta := \{x \in \mathbb{R}^m : \inf_{x' \in \mathcal{X}} |x - x'| \leqslant \zeta\}$ where $|\cdot|$ denotes the Euclidean norm. Let $\mathrm{Cov}_P(\cdot,\cdot)$ and $\mathrm{Var}_P(\cdot)$ denote the covariance and variance under P, respectively. For the notational convenience, for arbitrary r variables $d_1, \dots, d_r$, we use the notation $d_{k:\ell} = (d_k, d_{k+1}, \dots, d_\ell)$ for $1 \leqslant k \leqslant \ell \leqslant r$. We make the following assumptions.

(T1) Let $\mathcal{X}$ be a non-empty compact subset of $\mathbb{R}^m$ such that its diameter is bounded by C_1.

(T2) The random vector X has a Lebesgue density $p(\cdot)$ such that $\|p\|_{\mathcal{X}^\zeta} \leqslant C_1$.

(T3) Let $L : \mathbb{R}^m \to \mathbb{R}$ be a continuous kernel function supported in $[-1,1]^m$ such that the function class $\mathfrak{L} := \{x \mapsto L(ax+b) : a \in \mathbb{R}, b \in \mathbb{R}^m\}$ is VC type for envelope $\|L\|_{\mathbb{R}^m} = \sup_{x\in\mathbb{R}^m} |L(x)|$.

(T4) Let Φ be a pointwise measurable class of symmetric functions $\mathcal{V}^r \to \mathbb{R}$ that is VC type with characteristics (A, v) for a finite and symmetric envelope $\overline{\varphi} \in L^q(P^r)$ such that $\log A \leqslant C_1 \log n$ and $v \leqslant C_1$. In addition, the envelope $\overline{\varphi}$ satisfies that $(\mathbb{E}[\overline{\varphi}^q(V_{1:r}) \mid X_{1:r} = x_{1:r}])^{1/q} \leqslant C_1$ for all $x_{1:r} \in \mathcal{X}^\zeta \times \cdots \times \mathcal{X}^\zeta$ if q is finite, and $\|\overline{\varphi}\|_{P^r,\infty} \leqslant C_1$ if $q = \infty$

(T5) $nb_n^{3mq/[2(q-1)]} \geqslant C_1 n^{c_2}$ with the convention that $q/(q-1) = 1$ when $q = \infty$, and $2m(r-1)b_n \leqslant \zeta/2$.

(T6) $b_n^{m/2}\sqrt{\mathrm{Var}_P(P^{r-1}h_{n,\vartheta})} \geqslant c_1$ for all n and $\vartheta \in \Theta$.

(T7) $c_1 \leqslant c_n(\vartheta) \leqslant C_1$ for all n and $\vartheta \in \Theta$. For each fixed n, if $x_k \to x$ in $\mathcal{X}$ and $\varphi_k \to \varphi$ pointwise in Φ, then $c_n(x_k, \varphi_k) \to c_n(x, \varphi)$.

(T8) With probability at least $1 - C_1 n^{-c_2}$, $\sup_{\vartheta\in\Theta} \left|\frac{\widehat{c}_n(\vartheta)}{c_n(\vartheta)} - 1\right| \leqslant C_1 n^{-c_2}$.

Some comments on the conditions are in order. Condition (T1) allows the set $\mathcal{X}$ to depend on n, i.e., $\mathcal{X} = \mathcal{X}_n$, but its diameter is bounded (by C_1). For example, $\mathcal{X}$ can be discrete grids whose cardinality increases with n but its diameter must be bounded (an implicit assumption here is that the dimension m is fixed; in fact the constants appearing in the following results depend on the dimension m, so that m should be considered as fixed). Condition (T2) is a mild restriction on the density of X. It is worth mentioning that V may take values in a generic measurable space, and even if V takes values in a Euclidean space, V need not be absolutely continuous with respect to the Lebesgue measure (we will often omit the qualification "with respect to the Lebesgue measure"). In Examples 4.1 and 4.2, the variable V consists of the pair of regressor vector and outcome variable, i.e., $V = (X, Y)$ with Y being real-valued, and our conditions allow the distribution of Y to be generic. In contrast, [25,38] assume that the *joint* distribution of X and Y have a continuous density (or at least they require the distribution function of Y to be continuous) and thereby ruling out the case where the distribution of Y has a discrete component. This is essentially because they rely on Rio's coupling [47] when deriving limiting null distributions of their test statistics. Rio's coupling is a powerful KMT [36] type strong approximation result for general

empirical processes, but requires the underlying distribution to be defined on a hyper-cube and to have a density bounded away from zero on the hyper-cube. In contrast, our analysis is conditional on X and we only require some moment conditions and VC type conditions on the function class. Thus our JMB does not require Y to have a density for its validity and thereby having a wider applicability in this respect.

Condition (T3) is a standard regularity condition on kernel functions L. Sufficient conditions under which $\mathfrak{L}$ is VC type are found in [28,29,43]. Condition (T4) allows the envelope $\overline{\varphi}$ to be unbounded. Condition (T4) allows the function class Φ to depend on n, as long as the VC characteristics A and v satisfy that $\log A \leqslant C_1 \log n$ and $v \leqslant C_1$. For example, Φ can be a discrete set whose cardinality is bounded by Cn^c for some constants $c, C > 0$. Condition (T5) relaxes bandwidth requirements in [25,38] where $m = 1$ and $q = \infty$. For example, [25] assume $nb_n^2/(\log n)^4 \to \infty$ and $b_n \log n \to 0$ for size control. For the problem of testing for regression/stochastic monotonicity of univariate functions, our test statistic is of order $r = 2$. If we choose a bounded kernel (such as the sign kernel), then we only need $n^{-2/3+c} \lesssim b_n \lesssim 1$ for some small constant $c > 0$. Further, our general theory allows us to develop a version of the JMB that is uniformly valid in compact bandwidth sets, which can be used to develop versions of tests that are valid with data-dependent bandwidths in Examples 4.1 and 4.2; see Sect. 4.1 ahead for details.

Condition (T6) is a high-level condition and implies the U-process to be non-degenerate. Let $\varphi_{[r-1]}(v_1, x_{2:r}) := \mathbb{E}[\varphi(v_1, V_{2:r}) \mid X_{2:r} = x_{2:r}] \prod_{j=2}^r p(x_j)$, and observe that

$$(P^{r-1}h_{n,\vartheta})(x_1, v_1) = L_{b_n}(x - x_1) \int \varphi_{[r-1]}(v_1, x - b_n x_{2:r}) \prod_{j=2}^r L(x_j) dx_{2:r}$$

for $\vartheta = (x, \varphi)$, where $x - b_n x_{2:r} = (x - b_n x_2, \ldots, x - b_n x_r)$. From this expression, in applications, it is not difficult to find primitive regularity conditions that guarantee Condition (T6). To keep the presentation concise, however, we assume Condition (T6).

Condition (T7) is concerned with the normalizing constant $c_n(\vartheta)$. For the special case where $c_n(\vartheta) = b_n^{m/2}\sqrt{\mathrm{Var}_P(P^{r-1}h_{n,\vartheta})}$, Condition (T7) is implied by Conditions (T4) and (T6). Condition (T8) is also a high-level condition, which together with (T7) implies that there is a uniformly consistent estimate $\widehat{c}_n(\vartheta)$ of $c_n(\vartheta)$ in Θ with polynomial error rates. Construction of $\widehat{c}_n(\vartheta)$ is quite flexible: for $c_n(\vartheta) = b_n^{m/2}\sqrt{\mathrm{Var}_P(P^{r-1}h_{n,\vartheta})}$, one natural example is the jackknife estimate

$$\widehat{c}_n(\vartheta) = \sqrt{\frac{b_n^m}{n} \sum_{i=1}^n \left\{ U_{n-1,-i}^{(r-1)}(\delta_{D_i} h_{n,\vartheta}) - U_n(h_{n,\vartheta}) \right\}^2}, \ \vartheta \in \Theta. \tag{13}$$

The following lemma verifies that the jackknife estimate (13) obeys Condition (T8) for $c_n(\vartheta) = b_n^{m/2}\sqrt{\mathrm{Var}_P(P^{r-1}h_{n,\vartheta})}$. However, it should be noted that other estimates for this normalizing constant are possible depending on applications of interest; see [1,25,38].

Lemma 4.1 (Estimation error of the normalizing constant) *Suppose that Conditions (T1)–(T7) hold. Let* $c_n(\vartheta) = b_n^{m/2}\sqrt{\mathrm{Var}_P(P^{r-1}h_{n,\vartheta})}$, $\vartheta \in \Theta$ *and* $\widehat{c}_n(\vartheta)$ *be defined in* (13). *Then there exist constants* c, C *depending only on* $r, m, \zeta, c_1, c_2, C_1, L$ *such that*

$$\mathbb{P}\left\{\sup_{\vartheta\in\Theta}\left|\frac{\widehat{c}_n(\vartheta)}{c_n(\vartheta)} - 1\right| > Cn^{-c}\right\} \leqslant Cn^{-c}.$$

Now, we are ready to state finite sample validity of the JMB for approximating the distribution of the supremum of the generalized local U-process.

Theorem 4.2 (JMB validity for the supremum of a generalized local U-process) *Suppose that Conditions (T1)–(T8) hold. Then there exist constants* c, C *depending only on* $r, m, \zeta, c_1, c_2, C_1, L$ *such that the following holds: for every n, there exists a tight Gaussian random variable* $W_{P,n}(\vartheta)$, $\vartheta \in \Theta$ *in* $\ell^\infty(\Theta)$ *with mean zero and covariance function*

$$\mathbb{E}[W_{P,n}(\vartheta)W_{P,n}(\vartheta')] = b_n^m \mathrm{Cov}_P(P^{r-1}h_{n,\vartheta},\, P^{r-1}h_{n,\vartheta'})/\{c_n(\vartheta)c_n(\vartheta')\} \quad (14)$$

for $\vartheta, \vartheta' \in \Theta$, *and it follows that*

$$\begin{aligned} &\sup_{t\in\mathbb{R}}\left|\mathbb{P}(\widehat{S}_n \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t)\right| \leqslant Cn^{-c}, \\ &\mathbb{P}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{P}_{|D_1^n}(\widehat{S}_n^\sharp \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t)\right| > Cn^{-c}\right\} \leqslant Cn^{-c}, \end{aligned} \quad (15)$$

where $\widetilde{S}_n := \sup_{\vartheta\in\Theta} W_{P,n}(\vartheta)$.

Theorem 4.2 leads to the following corollary, which is another form of validity of the JMB. For $\alpha \in (0,1)$, let $q_{\widehat{S}_n^\sharp}(\alpha) = q_{\widehat{S}_n^\sharp}(\alpha; D_1^n)$ denote the conditional α-quantile of $\widehat{S}_n^\sharp$ given D_1^n, i.e., $q_{\widehat{S}_n^\sharp}(\alpha) = \inf\left\{t \in \mathbb{R} : \mathbb{P}_{|D_1^n}(\widehat{S}_n^\sharp \leqslant t) \geqslant \alpha\right\}$.

Corollary 4.3 (Size validity of the JMB test) *Suppose that Conditions (T1)–(T8) hold. Then there exist constants* c, C *depending only on* $r, m, \zeta, c_1, c_2, C_1, L$ *such that*

$$\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left\{\widehat{S}_n \leqslant q_{\widehat{S}_n^\sharp}(\alpha)\right\} - \alpha\right| \leqslant Cn^{-c}.$$

4.1 Uniformly valid JMB test in bandwidth

A version of Theorem 4.2 continues to hold even if we additionally take the supremum over a set of possible bandwidths. For a given bandwidth $b \in (0,1)$, let

$$h_{\vartheta,b}(d_1,\ldots,d_r) = \varphi(v_1,\ldots,v_r)\prod_{k=1}^{r} L_b(x - x_k),$$

and for a given candidate set of bandwidths $\mathcal{B}_n \subset [\underline{b}_n, \overline{b}_n]$ with $0 < \underline{b}_n \leqslant \overline{b}_n < 1$, consider

$$S_n := \sup_{(\vartheta,b)\in\Theta\times\mathcal{B}_n} \frac{\sqrt{nb^m}\{U_n(h_{\vartheta,b}) - P^r h_{\vartheta,b}\}}{rc(\vartheta, b)} \quad \text{and}$$
$$\widehat{S}_n := \sup_{(\vartheta,b)\in\Theta\times\mathcal{B}_n} \frac{\sqrt{nb^m}\{U_n(h_{\vartheta,b}) - P^r h_{\vartheta,b}\}}{r\widehat{c}(\vartheta, b)},$$

where $c_n(\vartheta, b) > 0$ is a suitable normalizing constant and $\widehat{c}(\vartheta, b) > 0$ is an estimate of $c(\vartheta, b)$. Following a similar argument used in the proof of Theorem 4.2, we are able to derive a version of the JMB test that is also valid uniformly in bandwidth, which opens new possibilities to develop tests that are valid with data-dependent bandwidths in Examples 4.1 and 4.2. For related discussions, we refer the readers to Remark 3.2 in [38] for testing stochastic monotonicity and [22] for kernel type estimators.

Consider the JMB analogue of $\widehat{S}_n$:

$$\widehat{S}_n^\sharp = \sup_{(\vartheta,b)\in\Theta\times\mathcal{B}_n} \frac{b^{m/2}}{\widehat{c}_n(\vartheta, b)\sqrt{n}} \sum_{i=1}^n \xi_i \left[U_{n-1,-i}^{(r-1)}(\delta_{D_i} h_{\vartheta,b}) - U_n(h_{\vartheta,b})\right].$$

Let $\kappa_n = \overline{b}_n/\underline{b}_n$ denote the ratio of the largest and smallest possible values in the bandwidth set $\mathcal{B}_n$, which intuitively quantifies the size of $\mathcal{B}_n$. To ease the notation and to facilitate comparisons, we only consider $q = \infty$. We make the following assumptions instead of Conditions (T5)–(T8).

(T5′) $n\underline{b}_n^{3m/2} \geqslant C_1 n^{c_2}\kappa_n^{m(r-2)}$, $\kappa_n \leqslant C_1\underline{b}_n^{-1/(2r)}$, and $2m(r-1)\overline{b}_n \leqslant \zeta/2$.
(T6′) $b^{m/2}\sqrt{\mathrm{Var}_P(P^{r-1}h_{\vartheta,b})} \geqslant c_1$ for all n and $(\vartheta, b) \in \Theta \times \mathcal{B}_n$.
(T7′) $c_1 \leqslant c_n(\vartheta, b) \leqslant C_1$ for all n and $(\vartheta, b) \in \Theta \times \mathcal{B}_n$. For each fixed n, if $x_k \to x$ in $\mathcal{X}$, $\varphi_k \to \varphi$ pointwise in Φ, and $b_k \to b$ in $\mathcal{B}_n$, then $c_n(x_k, \varphi_k, b_k) \to c_n(x, \varphi, b)$.
(T8′) With probability at least $1 - C_1 n^{-c_2}$, $\sup_{(\vartheta,b)\in\Theta\times\mathcal{B}_n} \left|\frac{\widehat{c}_n(\vartheta,b)}{c_n(\vartheta,b)} - 1\right| \leqslant C_1 n^{-c_2}$.

Theorem 4.4 (Bootstrap validity for the supremum of a generalized local U-process: uniform-in-bandwidth result) *Suppose that Conditions (T1)–(T4) with $q = \infty$, and Conditions (T5′)–(T8′) hold. Then there exist constants c, C depending only on $r, m, \zeta, c_1, c_2, C_1, L$ such that the following holds: for every n, there exists a tight Gaussian random variable $W_{P,n}(\vartheta, b)$, $(\vartheta, b) \in \Theta \times \mathcal{B}_n$ in $\ell^\infty(\Theta \times \mathcal{B}_n)$ with mean zero and covariance function*

$$\mathbb{E}[W_{P,n}(\vartheta, b)W_{P,n}(\vartheta', b')] = b^{m/2}(b')^{m/2}\mathrm{Cov}_P(P^{r-1}h_{\vartheta,b}, P^{r-1}h_{\vartheta',b'})/\{c_n(\vartheta, b)c_n(\vartheta', b')\}$$

for $(\vartheta, b), (\vartheta', b') \in \Theta \times \mathcal{B}_n$, and the result (15) continues to hold with $\widetilde{S}_n := \sup_{(\vartheta,b)\in\Theta\times\mathcal{B}_n} W_{P,n}(\vartheta, b)$.

Table 1 Empirical rejection probability of the JMB test for regression monotonicity at the nominal sizes 0.05 and 0.10 with Gaussian and Rademacher error distributions

Nominal size	Sample size	Gaussian	Rademacher
$\alpha = 0.05$	$n = 100$	0.0374	0.0372
	$n = 200$	0.0362	0.0408
	$n = 500$	0.0412	0.0430
$\alpha = 0.10$	$n = 100$	0.0846	0.0796
	$n = 200$	0.0860	0.0872
	$n = 500$	0.0886	0.0844

If $\underline{b}_n = \overline{b}_n = b_n$ (i.e., $\mathcal{B}_n = \{b_n\}$ is a singleton set), then Conditions (T5′)–(T8′) reduce to (T5)–(T8) and Theorem 4.4 covers Theorem 4.2 with $q = \infty$ as a special case. Condition (T5′) states that the size of the bandwidth set $\mathcal{B}_n$ cannot be too large. Conditions (T6′)–(T8′) are completely parallel with Conditions (T6)–(T8). Such "uniform-in-bandwidth" type results are not covered in [1,25,38].

4.2 A simulation study on testing for monotonicity of regression

We provide a numerical example to verify the size validity of the JMB test for monotonicity of regression in Example 4.2. We generate i.i.d. univariate covariates $X_1, \ldots, X_n$ from the uniform distribution on $[0, 1]$ and consider the zero regression function $f \equiv 0$ (which implies that the covariate X and the response Y are stochastically independent). As argued in [25], $f \equiv 0$ is the hardest case in terms of size control under the null hypothesis $H_0 : f$ is increasing on $[0, 1]$. We consider two error distributions: (i) Gaussian distribution $\varepsilon_i \sim N(0, 0.1^2)$; (ii) (scaled) Rademacher distribution $\mathbb{P}(\varepsilon_i = \pm 0.1) = 1/2$. For both error distributions, the (unnormalized) U-process $\check{U}_n(x)$ defined in (12) has mean zero (i.e., $\mathbb{E}[\check{U}_n(x)] = 0$ for all $x \in [0, 1]$). The Rademacher distribution is not covered in [25]. We use the Epanechnikov kernel $L(x) = 0.75(1 - x^2)$ for $x \in [-1, 1]$ and $L(x) = 0$ otherwise, together with bandwidth parameter $b_n = n^{-1/5}$. We consider three sample sizes $n = 100, 200, 500$. For each setup, we generate 2000 bootstrap samples. We consider test of the form

$$\sup_{x \in [0.05, 0.95]} \frac{\check{U}_n(x)}{\widehat{c}_n(x)} > q \Rightarrow \text{reject } H_0,$$

where $\widehat{c}_n(x)$ is given in (13) and the critical value q is calibrated by the JMB. In particular, for any nominal size $\alpha \in (0, 1)$, the value of $q := q(\alpha)$ is chosen as the $(1-\alpha)$-th conditional quantile of the JBM. Empirical rejection probability of the JMB test is obtained by averaging over 5000 simulations. We observe that the empirical rejection probability is close to the nominal size of the JMB test. Table 1 shows the proportion of rejections at the nominal sizes $\alpha = 0.05, 0.10$, and Fig. 1 shows the JMB approximation of the proportion of rejections uniformly in $\alpha \in (0, 1)$.

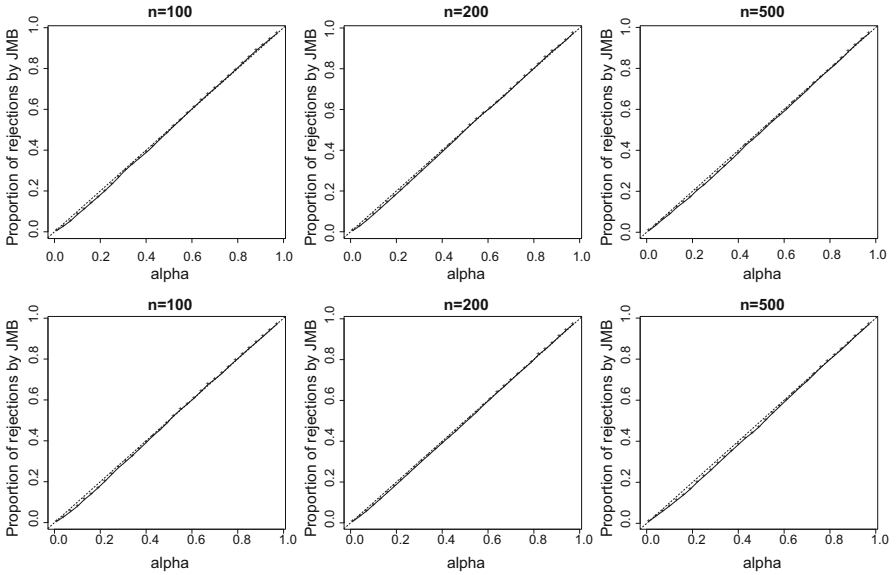

Fig. 1 JMB approximation of sizes of the regression monotonicity test. Top row: Gaussian errors. Bottom row: Rademacher errors

5 Local maximal inequalities for *U*-processes

In this section, we prove *local maximal inequalities* for U-processes, which are of independent interest and can be useful for other applications. These multi-resolution local maximal inequalities are key technical tools in proving the results stated in the previous sections.

We first review some basic terminologies and facts about U-processes. For a textbook treatment on U-processes, we refer to [18]. Let $r \geqslant 1$ be a fixed integer and let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$ with common distribution P.

Definition 5.1 (*Kernel degeneracy; Definition 3.5.1 in* [18]) A symmetric measurable function $f : S^r \to \mathbb{R}$ with $P^r f = 0$ is said to be *degenerate of order k* with respect to P if $P^{r-k} f(x_1, \ldots, x_k) = 0$ for all $x_1, \ldots, x_k \in S$. In particular, f is said to be *completely degenerate* if f is degenerate of order $r - 1$, and f is said to be *non-degenerate* if f is not degenerate of any positive order.

Let $\mathcal{F}$ be a class of symmetric measurable functions $f : S^r \to \mathbb{R}$. We assume that there is a symmetric measurable envelope F for $\mathcal{F}$ such that $P^r F^2 < \infty$. Furthermore, we assume that each $P^{r-k} F$ is everywhere finite. Consider the associated U-process

$$U_n^{(r)}(f) = \frac{1}{|I_{n,r}|} \sum_{(i_1,\ldots,i_r) \in I_{n,r}} f(X_{i_1}, \ldots, X_{i_r}), \quad f \in \mathcal{F}. \tag{16}$$

For each $k = 1, \ldots, r$, the *Hoeffding projection* (with respect to P) is defined by

$$(\pi_k f)(x_1, \ldots, x_k) := (\delta_{x_1} - P) \cdots (\delta_{x_k} - P) P^{r-k} f. \tag{17}$$

The Hoeffding projection $\pi_k f$ is a completely degenerate kernel of k variables. Then, the *Hoeffding decomposition* of $U_n^{(r)}(f)$ is given by

$$U_n^{(r)}(f) - P^r f = \sum_{k=1}^{r} \binom{r}{k} U_n^{(k)}(\pi_k f). \tag{18}$$

In what follows, let σ_k be any positive constant such that $\sup_{f \in \mathcal{F}} \|P^{r-k} f\|_{P^k,2} \leqslant \sigma_k \leqslant \|P^{r-k} F\|_{P^k,2}$ whenever $\|P F^{r-k}\|_{P^k,2} > 0$ (take $\sigma_k = 0$ when $\|P^{r-k} F\|_{P^k,2} = 0$), and let

$$M_k = \max_{1 \leqslant i \leqslant \lfloor n/k \rfloor} (P^{r-k} F)(X_{(i-1)k+1}^{ik}),$$

where $X_{(i-1)k+1}^{ik} = (X_{(i-1)k+1}, \ldots, X_{ik})$.

We will assume certain uniform covering number conditions for the function class $\mathcal{F}$. For $k = 1, \ldots, r$, define the uniform entropy integral

$$\begin{aligned} J_k(\delta) &:= J_k(\delta, \mathcal{F}, F) \\ &:= \int_0^\delta \sup_Q \left[1 + \log N(P^{r-k}\mathcal{F}, \|\cdot\|_{Q,2}, \tau \|P^{r-k} F\|_{Q,2}) \right]^{k/2} d\tau, \end{aligned} \tag{19}$$

where $P^{r-k}\mathcal{F} = \{P^{r-k} f : f \in \mathcal{F}\}$ and $\sup_Q$ is taken over all finitely discrete distributions on S^k. We note that $P^{r-k} F$ is an envelope for $P^{r-k}\mathcal{F}$. To avoid measurablity difficulties, we will assume that $\mathcal{F}$ is pointwise measurable. If $\mathcal{F}$ is pointwise measurable and $P^r F < \infty$ (which we have assumed) then $\pi_k \mathcal{F} := \{\pi_k f : f \in \mathcal{F}\}$ and $P^{r-k}\mathcal{F}$ for $k = 1, \ldots, r$ are all pointwise measurable by the dominated convergence theorem.

Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables such that $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$. A real-valued Rademacher chaos variable of order k, X, is a polynomial of order k in the Rademacher random variables ε_i with real coefficients, i.e.,

$$X = a + \sum_{i=1}^{n} a_i \varepsilon_i + \sum_{(i_1,i_2) \in I_{n,2}} a_{i_1 i_2} \varepsilon_{i_1} \varepsilon_{i_2} + \cdots + \sum_{(i_1,\ldots,i_k) \in I_{n,k}} a_{i_1 \ldots i_k} \varepsilon_{i_1} \cdots \varepsilon_{i_k},$$

where $a, a_i, a_{i_1 i_2}, \ldots, a_{i_1 \ldots i_k} \in \mathbb{R}$. If only the monomials of degree k in the variables ε_i in X are not zero, then X is a homogeneous Rademacher chaos of order k; see Section 3.2 in [18].

Definition 5.2 (*Rademacher chaos process of order k*; *page 220 in* [18]) A stochastic process $X(t), t \in T$ is said to be a *Rademacher chaos process of order k* if for

all $s, t \in T$, the joint law of $(X(s), X(t))$ coincides with the joint law of two (not necessarily homogeneous) Rademacher chaos variables of order k.

In the remainder of this section, the notation $\lesssim$ signifies that the left hand side is bounded by the right hand side up to a constant that depends only on r. Recall that $\|\cdot\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\cdot|$.

Theorem 5.1 (Local maximal inequalities for U-processes) *Suppose that $\mathcal{F}$ is pointwise measurable and that $J_k(1) < \infty$ for $k = 1, \ldots, r$. Let $\delta_k = \sigma_k / \|P^{r-k}F\|_{P^k,2}$ for $k = 1, \ldots, r$. Then*

$$n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \lesssim J_k(\delta_k)\|P^{r-k}F\|_{P^k,2} + \frac{J_k^2(\delta_k)\|M_k\|_{\mathbb{P},2}}{\delta_k^2\sqrt{n}} \tag{20}$$

for every $k = 1, \ldots, r$. If $\|P^{r-k}F\|_{P^k,2} = 0$, then the right hand side is interpreted as 0.

The proof of Theorem 5.1 relies on the following lemma on the uniform entropy integrals.

Lemma 5.2 (Properties of the maps $\delta \mapsto J_k(\delta)$) *Assume that $J_k(1) < \infty$ for $k = 1, \ldots, r$. Then, the following properties hold for every $k = 1, \ldots, r$. (i) The map $\delta \mapsto J_k(\delta)$ is non-decreasing and concave. (ii) For $c \geqslant 1$, $J_k(c\delta) \leqslant cJ_k(\delta)$. (iii) The map $\delta \mapsto J_k(\delta)/\delta$ is non-increasing. (iv) The map $(x, y) \mapsto J_k(\sqrt{x/y})\sqrt{y}$ is jointly concave in $(x, y) \in [0, \infty) \times (0, \infty)$.*

Proof of Lemma 5.2 The proof is almost identical to [14, Lemma A.2] and hence omitted. □

Proof of Theorem 5.1 Pick any $k = 1, \ldots, r$. It suffices to prove (20) when $\|P^{r-k}F\|_{P^k,2} > 0$ since otherwise there is nothing to prove (recall that we have assumed that $P^r F^2 < \infty$, which ensures that $\|P^{r-k}F\|_{P^k,2} < \infty$). Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables independent of X_1^n. In addition, let $\{X_i^j\}$ and $\{\varepsilon_i^j\}$ be independent copies of $\{X_i\}$ and $\{\varepsilon_i\}$. From the randomization theorem for U-processes [18, Theorem 3.5.3] and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] &\lesssim \mathbb{E}\left[\left\|\frac{1}{|I_{n,k}|}\sum_{(i_1,\ldots,i_k)\in I_{n,k}} \varepsilon_{i_1}^1 \cdots \varepsilon_{i_k}^k (\pi_k f)(X_{i_1}^1, \ldots, X_{i_k}^k)\right\|_{\mathcal{F}}\right] \\
&\lesssim \mathbb{E}\left[\left\|\frac{1}{|I_{n,k}|}\sum_{(i_1,\ldots,i_k)\in I_{n,k}} \varepsilon_{i_1}^1 \cdots \varepsilon_{i_k}^k (P^{r-k} f)(X_{i_1}^1, \ldots, X_{i_k}^k)\right\|_{\mathcal{F}}\right] \\
&\lesssim \mathbb{E}\left[\left\|\frac{1}{|I_{n,k}|}\sum_{(i_1,\ldots,i_k)\in I_{n,k}} \varepsilon_{i_1} \cdots \varepsilon_{i_k} (P^{r-k} f)(X_{i_1}, \ldots, X_{i_k})\right\|_{\mathcal{F}}\right].
\end{aligned}$$

Conditionally on X_1^n,

$$R_{n,k}(f) := \frac{1}{\sqrt{|I_{n,k}|}} \sum_{(i_1,\ldots,i_k)\in I_{n,k}} \varepsilon_{i_1}\cdots\varepsilon_{i_k}(P^{r-k}f)(X_{i_1},\ldots,X_{i_k}),\ f\in\mathcal{F}$$

is a (homogeneous) Rademacher chaos process of order k. Denote by $\mathbb{P}_{I_{n,k}} = |I_{n,k}|^{-1}\sum_{(i_1,\ldots,i_k)\in I_{n,k}}\delta_{(X_{i_1},\ldots,X_{i_k})}$ the empirical distribution on all possible k-tuples of X_1^n; then Corollary 3.2.6 in [18] yields

$$\|R_{n,k}(f) - R_{n,k}(f')\|_{\psi_{2/k}|X_1^n} \lesssim \|P^{r-k}f - P^{r-k}f'\|_{\mathbb{P}_{I_{n,k}},2},\ \forall f, f'\in\mathcal{F},$$

where $\|\cdot\|_{\psi_{2/k}|X_1^n}$ denotes the Orlicz (quasi-)norm associated with $\psi_{2/k}(u) = e^{u^{2/k}} - 1$ evaluated conditionally on X_1^n. The $\|\cdot\|_{\psi_{2/k}|X_1^n}$-diameter of the function class $\mathcal{F}$ is at most $2\sigma_{I_{n,k}}$ with $\sigma_{I_{n,k}}^2 := \sup_{f\in\mathcal{F}}\|P^{r-k}f\|_{\mathbb{P}_{I_{n,k}},2}^2$. So, since the first moment is bounded by the $\psi_{2/k}$-(quasi)norm up to a constant that depends only on k (and hence r), by Corollary 5.1.8 in [18] together with Fubini's theorem and a change of variables, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\sqrt{|I_{n,k}|}}\sum_{(i_1,\ldots,i_k)\in I_{n,k}}\varepsilon_{i_1}\cdots\varepsilon_{i_k}(P^{r-k}f)(X_{i_1},\ldots,X_{i_k})\right\|_{\mathcal{F}}\right]\\
&\lesssim \mathbb{E}\left[\left\|\left\|\frac{1}{\sqrt{|I_{n,k}|}}\sum_{(i_1,\ldots,i_k)\in I_{n,k}}\varepsilon_{i_1}\cdots\varepsilon_{i_k}(P^{r-k}f)(X_{i_1},\ldots,X_{i_k})\right\|_{\mathcal{F}}\right\|_{\psi_{2/k}|X_1^n}\right]\\
&\lesssim \mathbb{E}\left[\int_0^{\sigma_{I_{n,k}}}\left[1+\log N(P^{r-k}\mathcal{F},\|\cdot\|_{\mathbb{P}_{I_{n,k}},2},\tau)\right]^{k/2}d\tau\right]\\
&= \mathbb{E}\left[\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2}\int_0^{\sigma_{I_{n,k}}/\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2}}\right.\\
&\qquad\left.\left[1+\log N(P^{r-k}\mathcal{F},\|\cdot\|_{\mathbb{P}_{I_{n,k}},2},\tau\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2})\right]^{k/2}d\tau\right]\\
&\leqslant \mathbb{E}\left[\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2}J_k(\sigma_{I_{n,k}}/\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2})\right].
\end{aligned}$$

The last inequality follows from the definition of J_k. Since $J_k(\sqrt{x/y})\sqrt{y}$ is jointly concave in $(x, y)\in[0,\infty)\times(0,\infty)$ by Lemma 5.2 (iv), Jensen's inequality yields

$$\begin{aligned}
n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] &\lesssim \|P^{r-k}F\|_{P^k,2}J_k(z),\\
\text{where } z &:= \sqrt{\mathbb{E}[\sigma_{I_{n,k}}^2]/\|P^{r-k}F\|_{P^k,2}^2}.
\end{aligned}\tag{21}$$

We shall bound $\mathbb{E}[\sigma^2_{I_{n,k}}]$. To this end, we will use Hoeffding's averaging [49, Section 5.1.6]. Let

$$S_{f,k}(x_1,\dots,x_n)=\frac{1}{m}\sum_{i=1}^{m}(P^{r-k}f)^2(x_{(i-1)k+1},\dots,x_{ik}),\ m=\lfloor n/k\rfloor.$$

Then, the U-statistic $\|P^{r-k}f\|^2_{\mathbb{P}_{I_{n,k}},2}=|I_{n,k}|^{-1}\sum_{I_{n,k}}(P^{r-k}f)^2(X_{i_1},\dots,X_{i_k})$ is the average of the variables $S_{f,k}(X_{j_1},\dots,X_{j_n})$ taken over all the permutations $j_1,\dots,j_n$ of $1,\dots,n$. Hence,

$$\mathbb{E}[\sigma^2_{I_{n,k}}]\leqslant\mathbb{E}\left[\sup_{f\in\mathcal{F}}S_{f,k}(X_1^n)\right]=\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}(P^{r-k}f)^2(X^{ik}_{(i-1)k+1})\right\|_{\mathcal{F}}\right]=:B_{n,k}$$

by Jensen's inequality, so that $z\leqslant\widetilde{z}:=\sqrt{B_{n,k}/\|P^{r-k}F\|^2_{P^k,2}}$. Since the blocks $X^{ik}_{(i-1)k+1}$, $i=1,\dots,m$ are i.i.d.,

$$\begin{aligned}
B_{n,k}&\leqslant_{(1)}\sigma_k^2+\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\left\{(P^{r-k}f)^2(X^{ik}_{(i-1)k+1})-\mathbb{E}[(P^{r-k}f)^2(X^{ik}_{(i-1)k+1})]\right\}\right\|_{\mathcal{F}}\right]\\
&\leqslant_{(2)}\sigma_k^2+2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i(P^{r-k}f)^2(X^{ik}_{(i-1)k+1})\right\|_{\mathcal{F}}\right]\\
&\leqslant_{(3)}\sigma_k^2+8\mathbb{E}\left[M_k\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i(P^{r-k}f)(X^{ik}_{(i-1)k+1})\right\|_{\mathcal{F}}\right]\\
&\leqslant_{(4)}\sigma_k^2+8\|M_k\|_{\mathbb{P},2}\sqrt{\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i(P^{r-k}f)(X^{ik}_{(i-1)k+1})\right\|^2_{\mathcal{F}}\right]},
\end{aligned}$$

where (1) follows from the triangle inequality, (2) follows from the symmetrization inequality [53, Lemma 2.3.1], (3) follows from the contraction principle [29, Corollary 3.2.2], and (4) follows from the Cauchy–Schwarz inequality. By (a version of) the Hoffmann-Jørgensen inequality to the empirical process [53, Proposition A.1.6],

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i(P^{r-k}f)(X^{ik}_{(i-1)k+1})\right\|^2_{\mathcal{F}}\right]}\\
&\quad\lesssim\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i(P^{r-k}f)(X^{ik}_{(i-1)k+1})\right\|_{\mathcal{F}}\right]+m^{-1}\|M_k\|_{\mathbb{P},2}.
\end{aligned}$$

The analysis of the expectation on the right hand side is rather standard. From the first half of the proof of Theorem 5.2 in [14] (or repeating the first half of this proof with

$r = k = 1$), we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varepsilon_i(P^{r-k}f)(X^{ik}_{(i-1)k+1})\right\|_{\mathcal{F}}\right] \lesssim \|P^{r-k}F\|_{P^k,2}\int_0^{\widetilde{z}}\sup_Q\sqrt{1+\log N(P^{r-k}\mathcal{F},\|\cdot\|_{Q,2},\tau\|P^{r-k}F\|_{Q,2})}d\tau.$$

Since the integral on the right hand side is bounded by $J_k(\widetilde{z})$, we have

$$B_{n,k} \lesssim \sigma_k^2 + n^{-1}\|M_k\|^2_{\mathbb{P},2} + n^{-1/2}\|M_k\|_{\mathbb{P},2}\|P^{r-k}F\|_{P^k,2}J_k(\widetilde{z}).$$

Therefore, we conclude that

$$\widetilde{z}^2 \lesssim \Delta^2 + \frac{\|M_k\|_{\mathbb{P},2}}{\sqrt{n}\|P^{r-k}F\|_{P^k,2}}J_k(\widetilde{z}), \quad \text{where } \Delta^2 := \frac{\sigma_k^2 \vee n^{-1}\|M_k\|^2_{\mathbb{P},2}}{\|P^{r-k}F\|^2_{P^k,2}}.$$

By Lemma 5.2 (i) and applying [54, Lemma 2.1] with $J(\cdot) = J_k(\cdot)$, $r = 1$, $A^2 = \Delta^2$, and $B^2 = \|M_k\|_{\mathbb{P},2}/(\sqrt{n}\|P^{r-k}F\|_{P^k,2})$, we have

$$J_k(z) \leqslant J_k(\widetilde{z}) \lesssim J_k(\Delta)\left[1 + J_k(\Delta)\frac{\|M_k\|_{\mathbb{P},2}}{\sqrt{n}\|P^{r-k}F\|_{P^k,2}\Delta^2}\right]. \tag{22}$$

Combining (21) and (22), we arrive at

$$n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \lesssim J_k(\Delta)\|P^{r-k}F\|_{P^k,2} + \frac{J_k^2(\Delta)\|M_k\|_{\mathbb{P},2}}{\sqrt{n}\Delta^2}. \tag{23}$$

We note that $\Delta \geqslant \delta_k$ and recall that $\delta_k = \sigma_k/\|P^{r-k}F\|_{P^k,2}$. Since the map $\delta \mapsto J_k(\delta)/\delta$ is non-increasing by Lemma 5.2 (iii), we have

$$J_k(\Delta) \leqslant \Delta\frac{J_k(\delta_k)}{\delta_k} = \max\left\{J_k(\delta_k), \frac{\|M_k\|_{\mathbb{P},2}J_r(\delta_k)}{\sqrt{n}\|P^{r-k}F\|_{P^k,2}\delta_k}\right\}.$$

In addition, since $J_k(\delta_k)/\delta_k \geqslant J_k(1) \geqslant 1$, we have

$$J_k(\Delta) \leqslant \max\left\{J_k(\delta_k), \frac{\|M_k\|_{\mathbb{P},2}J_k^2(\delta_k)}{\sqrt{n}\|P^{r-k}F\|_{P^k,2}\delta_k^2}\right\}.$$

Finally, since

$$\frac{J_k^2(\Delta)\|M_k\|_{\mathbb{P},2}}{\sqrt{n}\Delta^2} \leqslant \frac{J_k^2(\delta_k)\|M_k\|_{\mathbb{P},2}}{\sqrt{n}\delta_k^2},$$

the desired inequality (20) follows from (23). □

When the function class $\mathcal{F}$ is VC type, we may derive a more explicit bound on $n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}]$.

Corollary 5.3 (Local maximal inequalities for U-processes indexed by VC type classes) *If $\mathcal{F}$ is pointwise measurable and VC type with characteristics $A \geqslant (e^{2(r-1)}/16) \vee e$ and $v \geqslant 1$, then*

$$n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \lesssim \sigma_k \left\{ v \log(A\|P^{r-k}F\|_{P^k,2}/\sigma_k) \right\}^{k/2} + \frac{\|M_k\|_{\mathbb{P},2}}{\sqrt{n}} \left\{ v \log(A\|P^{r-k}F\|_{P^k,2}/\sigma_k) \right\}^{k} \tag{24}$$

for every $k = 1, \ldots, r$.

Remark 5.1 (i). Our maximal inequality (20) scales correctly with the order of degeneracy, namely, the bound on $\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}]$ scales as $n^{-k/2}$ if $\mathcal{F}$ is fixed with n; recall that the functions $\pi_k f$, $f \in \mathcal{F}$ are completely degenerate functions of k variables. In addition, our maximal inequality is "local" in the sense that the bound is able take into account the L^2-bound on functions $P^{r-k}f$, $f \in \mathcal{F}$, namely, the bound will yield a better estimate if we have an additional information that such an L^2-bound is small.

(ii). Giné and Mason [27, Theorem 8] establishes a different local maximal inequality for a U-process indexed by a VC type class with a bounded envelope. To be precise, they prove the following bound under the assumption that the envelope F is bounded by a constant M: there exist constants C_1 and C_2 depending only on r, A, v, and M such that

$$n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \leqslant C_1\sigma_r \left(\log \frac{A\|F\|_{P^r,2}}{\sigma_r} \right)^{k/2}, \quad k = 1, \ldots, r \tag{25}$$

whenever

$$n\sigma_r^2 \geqslant C_2 \log \left(\frac{2\|F\|_{P^r,2}}{\sigma_r} \right),$$

where σ_r is a positive constant satisfying $\sup_{f\in\mathcal{F}} \|f\|_{P^r,2} \leqslant \sigma_r \leqslant \|F\|_{P^r,2}$. Our Corollary 5.3 improves upon the bound (25) in several directions: 1) First, our bound (24) allows for an unbounded envelope while the bound (25) requires the envelope to be bounded. 2) Second, the constants C_1 and C_2 appearing in the bound (25) implicitly depend on the VC characteristics (A, v) and the L^∞-bound M on the envelope F, in addition to the order r, and so is not applicable to cases where the VC characteristics (A, v) and/or the L^∞-bound M change with n. On the other hand, the constant involved in our bound (24) depends only on r (recall that the notation $\lesssim$ in present section signifies that the left hand side is bounded by the right hand side up to a constant that depends only on r), and so is applicable to such cases. 3) Finally, our bound (24) is of

the multi-resolution nature in the sense that it depends on the L^2-bound on $P^{r-k}f$ for $f \in \mathcal{F}$ (i.e., σ_k) for each projection level $k = 1, \ldots, r$ rather than that on $f \in \mathcal{F}$ (i.e., σ_r), which allows us to obtain better rates of convergence for kernel type statistics than (25). In particular, σ_k for $k < r$ can be potentially much smaller than σ_r, which is indeed the case in the applications considered in Sect. 4. To be precise, for the function class $\{b_n^{m/2} c_n(\vartheta)^{-1} h_{n,\vartheta} : \vartheta \in \Theta\}$ appearing in Sect. 4, σ_k would be of order $b_n^{-m(k-1)/2}$ and so $\sigma_k \ll \sigma_r$ for $k < r$; see the proof of Theorem 4.2.

We also note that [2,26] derive sophisticated moment inequalities for U-statistics in Banach spaces. However, we find that their inequalities are difficult to apply in our setting.

(iii). Theorem 5.1 and Corollary 5.3 generalize Theorem 5.2 and Corollary 5.1 in [14] to U-processes. In fact, Theorem 5.1 and Corollary 5.3 reduce to Theorem 5.2 and Corollary 5.1 in [14] when $r = k = 1$, respectively.

Before proving Corollary 5.3, we first verify the following fact about VC type properties.

Lemma 5.4 *If $\mathcal{F}$ is VC type with characteristics (A, v), then for every $k = 1, \ldots, r-1$, $P^{r-k}\mathcal{F}$ is also VC type with characteristics $4\sqrt{A}$ and $2v$ for envelope $P^{r-k}F$, i.e.,*

$$\sup_Q N(P^{r-k}\mathcal{F}, \|\cdot\|_{Q,2}, \tau \|P^{r-k}F\|_{Q,2}) \leqslant (4\sqrt{A}/\tau)^{2v}, \ 0 < \forall \tau \leqslant 1.$$

Proof of Lemma 5.4 This follows from Lemma A.3 in Appendix A with $r = s = 2$. □

Proof of Corollary 5.3 For the notational convenience, put $A' = 4\sqrt{A}$ and $v' = 2v$. Then,

$$J_k(\delta) \leqslant \int_0^\delta (1 + v' \log(A'/\tau))^{k/2} d\tau \leqslant A'(v')^{k/2} \int_{A'/\delta}^\infty \frac{(1 + \log \tau)^{k/2}}{\tau^2} d\tau.$$

Integration by parts yields that for $c \geqslant e^{k-1}$,

$$\begin{aligned}\int_c^\infty \frac{(1 + \log \tau)^{k/2}}{\tau^2} d\tau &= \left[-\frac{(1 + \log \tau)^{k/2}}{\tau} \right]_c^\infty + \frac{k}{2} \int_c^\infty \frac{(1 + \log \tau)^{k/2}}{\tau^2 (1 + \log \tau)} d\tau \\ &\leqslant \frac{(1 + \log c)^{k/2}}{c} + \frac{1}{2} \int_c^\infty \frac{(1 + \log \tau)^{k/2}}{\tau^2} d\tau.\end{aligned}$$

Since $A'/\delta \geqslant A' \geqslant e^{r-1} \geqslant e^{k-1}$ for $0 < \delta \leqslant 1$, we conclude that

$$\int_{A/\delta'}^\infty \frac{(1 + \log \tau)^{k/2}}{\tau^2} d\tau \leqslant \frac{2\delta(1 + \log(A'/\delta))^{k/2}}{A'} \lesssim \frac{\delta(\log(A/\delta))^{k/2}}{A'}.$$

Combining Theorem 5.1, we obtain the desired inequality (24). □

The appearance of $\|P^{r-k}F\|_{P^k,2}/\sigma_k$ inside the log may be annoying in applications but there is a clever way to delete this term. Namely, choose $\sigma_k' = \sigma_k \vee (n^{-1/2}\|P^{r-k}F\|_{P^k,2})$ and apply Corollary 5.4 with σ_k replaced by σ_k'; then the bound for $n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}]$ is

$$\lesssim \sigma_k \{v \log(A \vee n)\}^{k/2} + \frac{\|P^{r-k}F\|_{P^k,2}}{\sqrt{n}} \{v \log(A \vee n)\}^{k/2} + \frac{\|M_k\|_{\mathbb{P},2}}{\sqrt{n}} \{v \log(A \vee n)\}^k .$$

Since $v \log(A \vee n) \geqslant 1$ by our assumption, the second term is bounded by the third term. We state the resulting bound as a separate corollary since this form would be most useful in (at least our) applications.

Corollary 5.5 *If $\mathcal{F}$ is pointwise measurable and VC type with characteristics $A \geqslant (e^{2(r-1)}/16) \vee e$ and $v \geqslant 1$, then,*

$$n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \lesssim \sigma_k \{v \log(A \vee n)\}^{k/2} + \frac{\|M_k\|_{\mathbb{P},2}}{\sqrt{n}} \{v \log(A \vee n)\}^k$$

for every $k = 1, \ldots, r$. Furthermore, $\|M_k\|_{\mathbb{P},2} \leqslant n^{1/q}\|P^{r-k}F\|_{P^k,q}$ for every $k = 1, \ldots, r$ and $q \in [2, \infty]$, where "$1/q$" for the $q = \infty$ case is interpreted as 0.

Proof of Corollary 5.5 The first half of the corollary is already proved. The latter half is trivial. □

If one is interested in bounding $\mathbb{E}[\|U_n^{(r)}(f) - P^r f\|_{\mathcal{F}}]$, then it suffices to apply (20) or (24) repeatedly for $k = 1, \ldots, r$. However, it is often the case that lower order Hoeffding projection terms are dominant, and for bounding higher order Hoeffding projection terms, it would suffice to apply the following simpler (but less sharp) maximal inequalities.

Corollary 5.6 (Alternative maximal inequalities for U-processes) *Let $p \in [1, \infty)$. Suppose that $\mathcal{F}$ is pointwise measurable and that $J_k(1) < \infty$ for $k = 1, \ldots, r$. Then, there exists a constant $C_{r,p}$ depending only on r, p such that*

$$n^{k/2}(\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}^p])^{1/p} \leqslant C_{r,p} J_k(1) \|P^{r-k}F\|_{P^k,2\vee p}$$

for every $k = 1, \ldots, r$. If $\mathcal{F}$ is VC type with characteristics $A \geqslant (e^{2(r-1)}/16) \vee e$ and $v \geqslant 1$, then $J_k(1) \lesssim (v \log A)^{k/2}$ for every $k = 1, \ldots, r$.

Proof of Corollary 5.6 The last assertion follows from a similar computation to that in the proof of Corollary 5.3. Hence we focus here on the first assertion. The proof is a modification to the proof of Theorem 5.1 and we shall use the notation used in the proof.

The randomization theorem and Jensen's inequality yield that $n^{pk/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}^p]$ is bounded by

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{|I_{n,k}|}}\sum_{I_{n,k}}\varepsilon_{i_1}\cdots\varepsilon_{i_k}(P^{r-k}f)(X_{i_1},\ldots,X_{i_k})\right\|_{\mathcal{F}}^p\right],$$

up to a constant depending only on r, p, where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. Rademacher random variables independent of X_1^n. Denote by $\mathbb{E}_{|X_1^n}$ the conditional expectation given X_1^n. Since the L^p-norm is bounded from above by the $\psi_{2/k}$-(quasi-)norm up to a constant that depends only on k (and hence r) and p, we have

$$\begin{aligned}&\mathbb{E}_{|X_1^n}\left[\left\|\frac{1}{\sqrt{|I_{n,k}|}}\sum_{I_{n,k}}\varepsilon_{i_1}\cdots\varepsilon_{i_k}(P^{r-k}f)(X_{i_1},\ldots,X_{i_k})\right\|_{\mathcal{F}}^p\right]\\&\leqslant C\left\|\left\|\frac{1}{\sqrt{|I_{n,k}|}}\sum_{I_{n,k}}\varepsilon_{i_1}\cdots\varepsilon_{i_k}(P^{r-k}f)(X_{i_1},\ldots,X_{i_k})\right\|_{\mathcal{F}}\right\|_{\psi_{k/2}|X_1^n}^p\end{aligned}$$

for some constant C depending only on r and p. The entropy integral bound for Rademacher chaoses (see the proof of Theorem 5.1) yields that the right hand side is bounded by, after changing variables,

$$\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2}^p J_k^p\left(\sigma_{I_{n,k}}/\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2}\right)$$

up to a constant depending only on r, p. The desired result follows from bounding $\sigma_{I_{n,k}}/\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2}$ by 1 and observation that $\mathbb{E}[\|P^{r-k}F\|_{\mathbb{P}_{I_{n,k}},2}^p]\leqslant\|P^{r-k}F\|_{P^k,2\vee p}^p$ by Jensen's inequality. □

Remark 5.2 Corollary 5.6 is an extension of Theorem 2.14.1 in [53]. For $p=1$, Corollary 5.6 is often less sharp than Theorem 5.1 since $\sigma_k\leqslant\|P^{r-k}F\|_{P^k,2}$ and in some cases $\sigma_k\ll\|P^{r-k}F\|_{P^k,2}$. However, Corollary 5.6 is useful for directly bounding higher order moments of $\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}$. For the empirical process case (i.e., $k=1$), bounding higher order moments of the supremum is essentially reduced to bounding the first moment by the Hoffmann-Jørgensen inequality [53, Proposition A.1.6]. There is an analogous Hoffmann-Jørgensen type inequality for U-processes (see [18, Theorem 4.1.2]), but for $k\geqslant 2$, bounding higher order moments of $\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}$ using this Hoffmann-Jørgensen inequality combined with the local maximal inequality in Theorem 5.1 would be more involved.

6 Proofs for Sects. 2 and 3

In what follows, let $\mathcal{B}(\mathbb{R})$ denote the Borel σ-field on $\mathbb{R}$. For a set $B\subset\mathbb{R}$ and $\delta>0$, let B^δ denote the δ-enlargement of B, i.e., $B^\delta=\{x\in\mathbb{R}:\inf_{y\in B}|x-y|\leqslant\delta\}$.

6.1 Proofs for Sect. 2

We begin with stating the following lemma.

Lemma 6.1 *Work with the setup described in Sect. 2. Suppose that Conditions (PM), (VC), and (MT) hold. Let* $L_n := \sup_{g\in\mathcal{G}} n^{-1/2}\sum_{i=1}^n (g(X_i) - Pg)$ *and* $\widetilde{Z} := \sup_{g\in\mathcal{G}} W_P(g)$. *Then, there exist universal constants* $C, C' > 0$ *such that* $\mathbb{P}(L_n \in B) \leqslant \mathbb{P}(\widetilde{Z} \in B^{C\delta_n}) + C'(\gamma + n^{-1})$ *for every* $B \in \mathcal{B}(\mathbb{R})$, *where*

$$\delta_n = \frac{(\overline{\sigma}_{\mathfrak{g}}^2 b_{\mathfrak{g}} K_n^2)^{1/3}}{\gamma^{1/3} n^{1/6}} + \frac{b_{\mathfrak{g}} K_n}{\gamma n^{1/2-1/q}}. \tag{26}$$

In the case of $q = \infty$, *"*$1/q$*" is interpreted as 0.*

The proof is a minor modification to that of Theorem 2.1 in [15]. Differences are (1) Lemma 6.1 allows $q = \infty$, and constants C, C' to be independent of q; (2) the error bound δ_n contains $b_{\mathfrak{g}} K_n/(\gamma n^{1/2-1/q})$ instead of $b_{\mathfrak{g}} K_n/(\gamma^{1/q} n^{1/2-1/q})$; and (3) our definition of K_n is slightly different from theirs. For completeness, in "Appendix C.1", we provide a sketch of the proof for Lemma 6.1, which points out required modifications to the proof of Theorem 2.1 in [15].

Proof of Proposition 2.1 In view of the Strassen–Dudley theorem (see Theorem B.1), it suffices to verify that there exist constants C, C' depending only r such that

$$\mathbb{P}(Z_n \in B) \leqslant \mathbb{P}(\widetilde{Z} \in B^{C\varpi_n}) + C'(\gamma + n^{-1})$$

for every $B \in \mathcal{B}(\mathbb{R})$. In what follows, C, C' denote generic constants that depend only on r; their values may vary from place to place.

We shall follow the notation used in Sect. 5. Consider the Hoeffding decomposition for $U_n(h) = U_n^{(r)}(h)$: $U_n^{(r)}(h) - P^r h = rU_n^{(1)}(\pi_1 h) + \sum_{k=2}^r \binom{r}{k} U_n^{(k)}(\pi_k h)$, or

$$\mathbb{U}_n(h) = \sqrt{n}(U_n^{(r)}(h) - P^r h) = r\mathbb{G}_n(P^{r-1}h) + \sqrt{n}\sum_{k=2}^r \binom{r}{k} U_n^{(k)}(\pi_k h),$$

where $\mathbb{G}_n(P^{r-1}h) := n^{-1/2}\sum_{i=1}^n (P^{r-1}h(X_i) - P^r h)$ is the Hájek (empirical) process associated with $\mathbb{U}_n$. Recall that $\mathcal{G} = P^{r-1}\mathcal{H} = \{P^{r-1}h : h \in \mathcal{H}\}$, and let $L_n = \sup_{g\in\mathcal{G}} \mathbb{G}_n(g)$ and $R_n = \|\sqrt{n}\sum_{k=2}^r \binom{r}{k} U_n^{(k)}(\pi_k h)/r\|_{\mathcal{H}}$. Then, since $|Z_n - L_n| \leqslant R_n$, Markov's inequality and Lemma 6.1 yield that for every $B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
\mathbb{P}(Z_n \in B) &\leqslant \mathbb{P}(\{Z_n \in B\} \cap \{R_n \leqslant \gamma^{-1}\mathbb{E}[R_n]\}) + \mathbb{P}(R_n > \gamma^{-1}\mathbb{E}[R_n]) \\
&\leqslant \mathbb{P}(L_n \in B^{\gamma^{-1}\mathbb{E}[R_n]}) + \gamma \\
&\leqslant \mathbb{P}(\widetilde{Z} \in B^{C\delta_n + \gamma^{-1}\mathbb{E}[R_n]}) + C'(\gamma + n^{-1}),
\end{aligned} \tag{27}$$

where δ_n is given in (26).

It remains to bound $\mathbb{E}[R_n]$. To this end, we shall separately apply Corollary 5.5 for $k = 2$ and Corollary 5.6 for $k = 3, \ldots, r$. First, applying Corollary 5.5 to $\mathcal{F} = \mathcal{H}$ for $k = 2$ yields

$$n\mathbb{E}[\|U_n^{(2)}(\pi_2 h)\|_{\mathcal{H}}] \leqslant C\left(\sigma_{\mathfrak{h}} K_n + b_{\mathfrak{h}} K_n^2 n^{-1/2+1/q}\right).$$

Likewise, applying Corollary 5.6 to $\mathcal{F} = \mathcal{H}$ for $k = 3, \ldots, r$ yields

$$\sum_{k=3}^{r} \mathbb{E}[\|U_n^{(k)}(\pi_k h)\|_{\mathcal{H}}] \leqslant C \sum_{k=3}^{r} n^{-k/2} \|P^{r-k} H\|_{P^k,2} K_n^{k/2} = C n^{-1/2} \chi_n.$$

Therefore, we conclude that

$$\mathbb{E}[R_n] \leqslant C \sum_{k=2}^{r} n^{1/2} \mathbb{E}[\|U_n^{(k)}(\pi_k h)\|_{\mathcal{H}}] \leqslant C'\left(\sigma_{\mathfrak{h}} K_n n^{-1/2} + b_{\mathfrak{h}} K_n^2 n^{-1+1/q} + \chi_n\right). \tag{28}$$

Combining (27) with (28) leads to the conclusion of the proposition. □

Proof of Corollary 2.2 We begin with noting that we may assume that $b_{\mathfrak{g}} \leqslant n^{1/2}$, since otherwise the conclusion is trivial by taking $C \geqslant 1$. In this proof, the notation $\lesssim$ signifies that the left hand side is bounded by the right hand side up to a constant that depends only on $r, \overline{\sigma}_{\mathfrak{g}}$, and $\underline{\sigma}_{\mathfrak{g}}$. Let $\gamma \in (0, 1)$ and pick a version $\widetilde{Z}_{n,\gamma}$ of $\widetilde{Z}$ as in Proposition 2.1 ($\widetilde{Z}_{n,\gamma}$ may depend on γ). Proposition 2.1 together with [15, Lemma 2.1] yield that

$$\begin{aligned}\rho(Z_n, \widetilde{Z}) = \rho(Z_n, \widetilde{Z}_{n,\gamma}) &\leqslant \sup_{t\in\mathbb{R}} \mathbb{P}(|\widetilde{Z}_{n,\gamma} - t| \leqslant C\varpi_n) + C'(\gamma + n^{-1}) \\ &= \sup_{t\in\mathbb{R}} \mathbb{P}(|\widetilde{Z} - t| \leqslant C\varpi_n) + C'(\gamma + n^{-1}).\end{aligned}$$

Now, the anti-concentration inequality (see Lemma A.1 in "Appendix A") yields

$$\sup_{t\in\mathbb{R}} \mathbb{P}(|\widetilde{Z} - t| \leqslant C\varpi_n) \lesssim \varpi_n \left\{\mathbb{E}[\widetilde{Z}] + \sqrt{1 \vee \log(\underline{\sigma}_{\mathfrak{g}}/(C\varpi_n))}\right\}. \tag{29}$$

Since $\mathcal{G}$ is VC type with characteristics $4\sqrt{A}$ and $2v$ for envelope G (Lemma 5.4), by Lemma A.2, we have $N(\mathcal{G}, \|\cdot\|_{P,2}, \tau) \leqslant (16\sqrt{A}\|G\|_{P,2}/\tau)^{2v}$ for all $0 < \varepsilon \leqslant 1$. Hence, Dudley's entropy integral bound [29, Theorem 2.3.7] yields $\mathbb{E}[\widetilde{Z}] \lesssim (\overline{\sigma}_{\mathfrak{g}} \vee (n^{-1/2} b_{\mathfrak{g}})) K_n^{1/2} \lesssim K_n^{1/2}$ where the last inequality follows from the assumption that $b_{\mathfrak{g}} \leqslant n^{1/2}$. Since $\sqrt{1 \vee \log(\underline{\sigma}_{\mathfrak{g}}/(C\varpi_n))} \lesssim (K_n \vee \log(\gamma^{-1}))^{1/2}$, we conclude that

$$\rho(Z_n, \widetilde{Z}) \lesssim (K_n \vee \log(\gamma^{-1}))^{1/2} \varpi_n(\gamma) + \gamma + n^{-1}.$$

The desired result follows from balancing $K_n^{1/2} \varpi_n(\gamma)$ and γ. □

6.2 Proofs for Sect. 3

Proof of Theorem 3.1 In this proof we will assume that each $h \in \mathcal{H}$ is P^r-centered, i.e., $P^r h = 0$ for the rotational convenience. Recall that $\mathbb{P}_{|X_1^n}$ and $\mathbb{E}_{|X_1^n}$ denote the conditional probability and expectation given X_1^n, respectively. In view of the conditional version of the Strassen–Dudley theorem (see Theorem B.2), it suffices to find constants C, C' depending only on r, and an event $E \in \sigma(X_1^n)$ with $\mathbb{P}(E) \geqslant 1 - \gamma - n^{-1}$ on which

$$\mathbb{P}_{|X_1^n}(Z_n^\sharp \in B) \leqslant \mathbb{P}(\widetilde{Z} \in B^{C\varpi_n^\sharp}) + C'(\gamma + n^{-1}) \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

The proof of Theorem 3.1 is involved and divided into six steps. In what follows, let C denote a generic positive constant depending only on r; the value of C may change from place to place.

Step 1: Discretization For $0 < \varepsilon \leqslant 1$ to be determined later, let $N := N(\varepsilon) := N(\mathcal{G}, \|\cdot\|_{P,2}, \varepsilon\|G\|_{P,2})$. Since $\|G\|_{P,2} \leqslant b_{\mathfrak{g}}$, there exists an $\varepsilon b_{\mathfrak{g}}$-net $\{g_k\}_{k=1}^N$ for $(\mathcal{G}, \|\cdot\|_{P,2})$. By the definition of $\mathcal{G}$, each g_k corresponds to a kernel $h_k \in \mathcal{H}$ such that $g_k = P^{r-1}h_k$. The Gaussian process W_P extends to the linear hull of $\mathcal{G}$ in such a way that W_P has linear sample paths (e.g., see [29, Theorem 3.7.28]). Now, observe that

$$\begin{aligned} 0 &\leqslant \sup_{g\in\mathcal{G}} W_P(g) - \max_{1\leqslant j\leqslant N} W_P(g_j) \leqslant \|W_P\|_{\mathcal{G}_\varepsilon}, \\ 0 &\leqslant \sup_{h\in\mathcal{H}} \mathbb{U}_n^\sharp(h) - \max_{1\leqslant j\leqslant N} \mathbb{U}_n^\sharp(h_j) \leqslant \|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\varepsilon}, \end{aligned}$$

where $\mathcal{G}_\varepsilon = \{g - g' : g, g' \in \mathcal{G}, \|g - g'\|_{P,2} < 2\varepsilon b_{\mathfrak{g}}\}$ and $\mathcal{H}_\varepsilon = \{h - h' : h, h' \in \mathcal{H}, \|P^{r-1}h - P^{r-1}h'\|_{P,2} < 2\varepsilon b_{\mathfrak{g}}\}$.

Step 2: Construction of a high-probability event $E \in \sigma(X_1^n)$ We divide this step into several sub-steps.

(i). For a P-integrable function g on S, we will use the notation

$$\mathbb{G}_n(g) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{g(X_i) - Pg\}.$$

Consider the function class $\breve{\mathcal{G}} \cdot \breve{\mathcal{G}} = \{gg' : g, g' \in \breve{\mathcal{G}}\}$ with $\breve{\mathcal{G}} = \{g, g - Pg : g \in \mathcal{G}\}$. Recall that $\mathcal{G}$ with envelope G is VC type with characteristics $(4\sqrt{A}, 2v)$. The function class $\{g - Pg : g \in \mathcal{G}\}$ with envelope $\breve{G} := G + PG$ is VC type with characteristics $(4\sqrt{2A}, 2v + 1)$ from a simple calculation. Conclude that $\breve{\mathcal{G}}$ with envelope $\breve{G}$ is VC type with characteristics $(8\sqrt{2A}, 2v + 1)$, and by Lemma A.5, $\breve{\mathcal{G}} \cdot \breve{\mathcal{G}}$ with envelope $\breve{G}^2$ is VC type with characteristics $(16\sqrt{2A}, 4v + 2)$. For $g, g' \in \mathcal{G}$, $P(gg')^2 \leqslant$

$\sqrt{Pg^4}\sqrt{P(g')^4} \leqslant \overline{\sigma}_{\mathfrak{g}}^2 b_{\mathfrak{g}}^2$ by Condition (MT). Likewise,

$$\begin{aligned} P(g-Pg)^2(g'-Pg')^2 &\leqslant \sqrt{P(g-Pg)^4}\sqrt{P(g'-Pg')^4} \\ &\leqslant 8\sqrt{Pg^4+(Pg)^4}\sqrt{P(g')^4+(Pg')^4} \\ &\leqslant 16\sqrt{Pg^4}\sqrt{P(g')^4} \leqslant 16\overline{\sigma}_{\mathfrak{g}}^2 b_{\mathfrak{g}}^2. \end{aligned}$$

We also note that $\|\check{G}\|_{P,q} \leqslant 2\|G\|_{P,q} \leqslant 2b_{\mathfrak{g}}$. Hence, applying Corollary 5.5 with $\mathcal{F} = \check{\mathcal{G}} \cdot \check{\mathcal{G}}$, $r = k = 1$, and $q = q/2$ yields

$$n^{-1/2}\mathbb{E}[\|\mathbb{G}_n\|_{\check{\mathcal{G}}\cdot\check{\mathcal{G}}}] \leqslant C\left(\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{g}} K_n^{1/2} n^{-1/2} + b_{\mathfrak{g}}^2 K_n n^{-1+2/q}\right),$$

so that with probability at least $1-\gamma/3$,

$$n^{-1/2}\|\mathbb{G}_n\|_{\check{\mathcal{G}}\cdot\check{\mathcal{G}}} \leqslant C\gamma^{-1}\left(\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{g}} K_n^{1/2} n^{-1/2} + b_{\mathfrak{g}}^2 K_n n^{-1+2/q}\right) \tag{30}$$

by Markov's inequality.

(ii). Define

$$\Upsilon_n := \left\| \frac{1}{n}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - P^{r-1}h(X_i)\}^2 \right\|_{\mathcal{H}}. \tag{31}$$

We will show that

$$\begin{aligned} \mathbb{E}[\Upsilon_n] \leqslant C\Big\{&\sigma_{\mathfrak{h}}^2 K_n n^{-1} + \nu_{\mathfrak{h}}^2 K_n^2 n^{-3/2+2/q} + \sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{3/2} n^{-3/2} \\ &+ b_{\mathfrak{h}}^2 K_n^3 n^{-2+2/q} + \chi_n^2\Big\}. \end{aligned} \tag{32}$$

Together with Markov's inequality, we have that with probability at least $1-\gamma/3$,

$$\begin{aligned} \Upsilon_n \leqslant C\gamma^{-1}\Big\{&\sigma_{\mathfrak{h}}^2 K_n n^{-1} + \nu_{\mathfrak{h}}^2 K_n^2 n^{-3/2+2/q} + \sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{3/2} n^{-3/2} \\ &+ b_{\mathfrak{h}}^2 K_n^3 n^{-2+2/q} + \chi_n^2\Big\}. \end{aligned} \tag{33}$$

The proof of the inequality (32) is lengthy and deferred after the proof of the theorem.

(iii). We shall bound $\mathbb{E}[\|U_n(h) - P^r h\|_{\mathcal{H}}^2]$. Applying Corollary 5.6 to $\mathcal{H}$ for $k = 2, \ldots, r$ yields

$$\sum_{k=2}^r \mathbb{E}[\|U_n^{(k)}(\pi_k h)\|_{\mathcal{H}}^2] \leqslant C\left(b_{\mathfrak{h}}^2 K_n^2 n^{-2} + n^{-1}\chi_n^2\right).$$

Next, since $U_n^{(1)}(\pi_1 h), h \in \mathcal{H}$ is an empirical process, we may apply the Hoffmann-Jørgensen inequality [53, Proposition A.1.6] to deduce that

$$\begin{aligned}\mathbb{E}[\|U_n^{(1)}(\pi_1 h)\|_{\mathcal{H}}^2] &\leqslant C\left\{(\mathbb{E}[\|U_n^{(1)}(\pi_1 h)\|_{\mathcal{H}}])^2 + b_{\mathfrak{g}}^2 n^{-2+2/q}\right\} \\ &\leqslant C\left(\overline{\sigma}_{\mathfrak{g}}^2 K_n n^{-1} + b_{\mathfrak{g}}^2 K_n^2 n^{-2+2/q} + b_{\mathfrak{g}}^2 n^{-2+2/q}\right) \\ &\leqslant C\left(\overline{\sigma}_{\mathfrak{g}}^2 K_n n^{-1} + b_{\mathfrak{g}}^2 K_n^2 n^{-2+2/q}\right),\end{aligned}$$

where the second inequality follows from Corollary 5.5. Since $\overline{\sigma}_{\mathfrak{g}} \leqslant \sigma_{\mathfrak{h}}$ and $b_{\mathfrak{g}} \leqslant b_{\mathfrak{h}}$,

$$\mathbb{E}[\|U_n(h) - P^r h\|_{\mathcal{H}}^2] \leqslant C\left(\sigma_{\mathfrak{h}}^2 K_n n^{-1} + b_{\mathfrak{h}}^2 K_n^2 n^{-2+2/q} + n^{-1}\chi_n^2\right),$$

so that by Markov's inequality, with probability at least $1 - \gamma/3$,

$$\|U_n(h) - P^r h\|_{\mathcal{H}}^2 \leqslant C\gamma^{-1}\left(\sigma_{\mathfrak{h}}^2 K_n n^{-1} + b_{\mathfrak{h}}^2 K_n^2 n^{-2+2/q} + n^{-1}\chi_n^2\right). \tag{34}$$

(iv). Let $\mathbb{P}_{I_{n,r}} = |I_{n,r}|^{-1}\sum_{(i_1,\ldots,i_r)\in I_{n,r}} \delta_{(X_{i_1},\ldots,X_{i_r})}$ denote the empirical distribution on all possible r-tuples of X_1^n. Then Markov's inequality yields that with probability at least $1 - n^{-1}$,

$$\|H\|_{\mathbb{P}_{I_{n,r}},2} \leqslant n^{1/2}\|H\|_{P^r,2}. \tag{35}$$

Now, define the event E by the the intersection of the events (30), (33), (34), and (35). Then, $E \in \sigma(X_1^n)$ and $\mathbb{P}(E) \geqslant 1 - \gamma - n^{-1}$.

Step 3: Bounding the discretization error for W_P By the Borell-Sudakov-Tsirel'son inequality (cf. [29, Theorem 2.5.8]), we have

$$\mathbb{P}\left(\|W_P\|_{\mathcal{G}_\varepsilon} \geqslant \mathbb{E}[\|W_P\|_{\mathcal{G}_\varepsilon}] + 2\varepsilon b_{\mathfrak{g}}\sqrt{2\log n}\right) \leqslant n^{-1}.$$

From a standard calculation, $N(\mathcal{G}_\varepsilon, \|\cdot\|_{P,2}, \tau) \leqslant N^2(\mathcal{G}, \|\cdot\|_{P,2}, \tau/2)$. Since $\mathcal{G}$ is VC type with characteristics $4\sqrt{A}$ and $2v$ for envelope G, by Lemma A.2, we have $N(\mathcal{G}, \|\cdot\|_{P,2}, \tau\|G\|_{P,2}) \leqslant C(16\sqrt{A}/\tau)^{2v}$, so that $N(\mathcal{G}_\varepsilon, \|\cdot\|_{P,2}, \tau) \leqslant (32\sqrt{A}b_{\mathfrak{g}}/\tau)^{4v}$. Now, Dudley's entropy integral bound [53, Corollary 2.2.8] yields

$$\mathbb{E}[\|W_P\|_{\mathcal{G}_\varepsilon}] \leqslant C(\varepsilon b_{\mathfrak{g}})\sqrt{v\log(A/\varepsilon)}.$$

Choosing $\varepsilon = 1/n^{1/2}$, we have

$$\mathbb{E}[\|W_P\|_{\mathcal{G}_\varepsilon}] \leqslant Cb_{\mathfrak{g}}n^{-1/2}\sqrt{v\log(An^{1/2})} \leqslant Cb_{\mathfrak{g}}K_n^{1/2}n^{-1/2}.$$

Since $\log n \leqslant K_n$, we conclude that

$$\mathbb{P}\left(\|W_P\|_{\mathcal{G}_\varepsilon} \geqslant Cb_{\mathfrak{g}}K_n^{1/2}n^{-1/2}\right) \leqslant n^{-1}.$$

Step 4: Bounding the discretization error for $\mathbb{U}_n^\sharp$. Since $\{\mathbb{U}_n^\sharp(h) : h \in \mathcal{H}\}$ is a centered Gaussian process conditionally on X_1^n, applying the Borell-Sudakov-Tsirel'son inequality conditionally on X_1^n, we have

$$\mathbb{P}_{|X_1^n}\left(\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\varepsilon} \geqslant \mathbb{E}_{|X_1^n}[\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\varepsilon}] + \sqrt{2\Sigma_n \log n}\right) \leqslant n^{-1},$$

where $\Sigma_n := \|n^{-1}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - U_n(h)\}^2\|_{\mathcal{H}_\varepsilon}$ with $\varepsilon = 1/n^{1/2}$.

We begin with bounding Σ_n. For any $h \in \mathcal{H}_\varepsilon$, $n^{-1}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - U_n(h)\}^2$ is bounded by $n^{-1}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h)\}^2$ since the average of $U_{n-1,-i}^{(r-1)}(\delta_{X_i}h)$, $i = 1, \ldots, n$ is $U_n(h)$ and the variance is bounded by the second moment. Further, the term $n^{-1}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h)\}^2$ is bounded by

$$\begin{aligned}&\frac{2}{n}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - P^{r-1}h(X_i)\}^2 \\ &\quad + \frac{2}{n}\sum_{i=1}^n \{(P^{r-1}h(X_i))^2 - P(P^{r-1}h)^2\} + 2P(P^{r-1}h)^2.\end{aligned} \tag{36}$$

The last term on the right hand side of (36) is bounded by $8(\varepsilon b_{\mathfrak{g}})^2$. The supremum of the first term on $\mathcal{H}_\varepsilon$ is bounded by $8\Upsilon_n$ since $\mathcal{H}_\varepsilon \subset \{h - h' : h, h' \in \mathcal{H}\}$ [the notation Υ_n is defined in (31)]. For the second term, observe that $\{(P^{r-1}h)^2 : h \in \mathcal{H}_\varepsilon\} \subset \{(g-g')^2 : g, g' \in \mathcal{G}\}$, $(g-g')^2 - P(g-g')^2 = (g^2 - Pg^2) + 2(gg' - Pgg') + ((g')^2 - P(g')^2)$, and $\{g^2 : g \in \mathcal{G}\} \subset \check{\mathcal{G}} \cdot \check{\mathcal{G}}$, so that the supremum of the second term on the right hand side of (36) is bounded by $8n^{-1/2}\|\mathbb{G}_n\|_{\check{\mathcal{G}}\cdot\check{\mathcal{G}}}$. Therefore, recalling that we have chosen $\varepsilon = 1/n^{1/2}$, we conclude that

$$\begin{aligned}\Sigma_n &\leqslant 8(\varepsilon b_{\mathfrak{g}})^2 + 8n^{-1/2}\|\mathbb{G}_n\|_{\check{\mathcal{G}}\cdot\check{\mathcal{G}}} + 8\Upsilon_n \\ &\leqslant C\gamma^{-1}\Big\{\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{g}} K_n^{1/2} n^{-1/2} + b_{\mathfrak{g}}^2 K_n n^{-1+2/q} + \sigma_{\mathfrak{h}}^2 K_n n^{-1} \\ &\quad + \nu_{\mathfrak{h}}^2 K_n^2 n^{-3/2+2/q} + \sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{3/2} n^{-3/2} + b_{\mathfrak{h}}^2 K_n^3 n^{-2+2/q} + \chi_n^2\Big\}\end{aligned}$$

on the event E.

Next, we shall bound $\mathbb{E}_{|X_1^n}[\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\varepsilon}]$ on the event E. Since $\mathcal{H}$ is VC type with characteristics (A, v), we have

$$N(\mathcal{H}_\varepsilon, \|\cdot\|_{\mathbb{P}_{I_{n,r}},2}, 2\tau\|H\|_{\mathbb{P}_{I_{n,r}},2}) \leqslant N^2(\mathcal{H}, \|\cdot\|_{\mathbb{P}_{I_{n,r}},2}, \tau\|H\|_{\mathbb{P}_{I_{n,r}},2}) \leqslant (A/\tau)^{2v}.$$

In addition, since

$$
\begin{aligned}
d^2(h,h') &:= \mathbb{E}_{|X_1^n}[\{\mathbb{U}_n^\sharp(h) - \mathbb{U}_n^\sharp(h')\}^2] \\
&= \frac{1}{n}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - U_n(h) - U_{n-1,-i}^{(r-1)}(\delta_{X_i}h') + U_n(h')\}^2 \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - U_{n-1,-i}^{(r-1)}(\delta_{X_i}h')\}^2 \leqslant \|h-h'\|^2_{\mathbb{P}_{I_{n,r}},2},
\end{aligned}
$$

where the last inequality follows from Jensen's inequality, and since a weaker pseudometric induces a smaller covering number, we have

$$
N(\mathcal{H}_\varepsilon, d, 2\tau\|H\|_{\mathbb{P}_{I_{n,r}},2}) \leqslant N(\mathcal{H}_\varepsilon, \|\cdot\|_{\mathbb{P}_{I_{n,r}},2}, 2\tau\|H\|_{\mathbb{P}_{I_{n,r}},2}) \leqslant (A/\tau)^{2v}.
$$

Hence, using $2\left[(n^{-(r-1)/2}\|H\|_{P^r,2}) \vee \Sigma_n^{1/2}\right]$ as a bound on the d-diameter of $\mathcal{H}_\varepsilon$, we have by Dudley's entropy integral bound

$$
\begin{aligned}
\mathbb{E}_{|X_1^n}[\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\varepsilon}] &\leqslant C\int_0^{(n^{-(r-1)/2}\|H\|_{P^r,2})\vee\Sigma_n^{1/2}} \sqrt{v\log(A\|H\|_{\mathbb{P}_{I_{n,r},2}}/\tau)}d\tau \\
&\leqslant C\left((n^{-(r-1)/2}\|H\|_{P^r,2}) \vee \Sigma_n^{1/2}\right) \\
&\quad \sqrt{v\log(A\|H\|_{\mathbb{P}_{I_{n,r},2}}/(n^{-(r-1)/2}\|H\|_{P^r,2}))} \\
&\leqslant C\left((n^{-(r-1)/2}\|H\|_{P^r,2}) \vee \Sigma_n^{1/2}\right)\sqrt{v\log(An^{r/2})}
\end{aligned}
$$

on the event E (we have used $\|H\|_{\mathbb{P}_{I_{n,k},2}} \leqslant n^{1/2}\|H\|_{P^r,2}$ on E). Since $n^{-(r-1)/2}\|H\|_{P^r,2} \leqslant \chi_n$, we have

$$
\begin{aligned}
\mathbb{E}_{|X_1^n}[\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\varepsilon}] &\leqslant C(\chi_n \vee \Sigma_n^{1/2})K_n^{1/2} \\
&\leqslant C\gamma^{-1/2}\Big\{(\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{g}} K_n^{3/2})^{1/2} n^{-1/4} + b_{\mathfrak{g}} K_n n^{-1/2+1/q} \\
&\quad + \sigma_{\mathfrak{h}} K_n n^{-1/2} + \nu_{\mathfrak{h}} K_n^{3/2} n^{-3/4+1/q} + (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/2} K_n^{5/4} n^{-3/4} \\
&\quad + b_{\mathfrak{h}} K_n^2 n^{-1+1/q} + \chi_n K_n^{1/2}\Big\}
\end{aligned}
$$

on the event E. Hence, we conclude that

$$
\mathbb{P}_{|X_1^n}(\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\varepsilon} \geqslant C\delta_n^{(1)}) \leqslant n^{-1}
$$

on the event E, where

$$
\begin{aligned}
\delta_n^{(1)} = \frac{1}{\gamma^{1/2}}\Big\{&\frac{(\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{g}} K_n^{3/2})^{1/2}}{n^{1/4}} + \frac{b_{\mathfrak{g}} K_n}{n^{1/2-1/q}} + \frac{\sigma_{\mathfrak{h}} K_n}{n^{1/2}} + \frac{\nu_{\mathfrak{h}} K_n^{3/2}}{n^{3/4-1/q}} \\
&+ \frac{(\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/2} K_n^{5/4}}{n^{3/4}} + \frac{b_{\mathfrak{h}} K_n^2}{n^{1-1/q}} + \chi_n K_n^{1/2}\Big\}.
\end{aligned}
$$

Step 5: Gaussian comparison Let $Z_n^{\sharp,\varepsilon} := \max_{1\leqslant j\leqslant N} \mathbb{U}_n^{\sharp}(h_j)$ and $\widetilde{Z}^{\varepsilon} := \max_{1\leqslant j\leqslant N} W_P(g_j)$. Observe that the covariance between $\mathbb{U}_n^{\sharp}(h_k)$ and $\mathbb{U}_n^{\sharp}(h_\ell)$ conditionally on X_1^n is

$$
\begin{aligned}
\widehat{C}_{k,\ell} &:= \frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_k) - U_n(h_k)\}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_\ell) - U_n(h_\ell)\} \\
&= \frac{1}{n}\sum_{i=1}^{n} U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_k)U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_\ell) - U_n(h_k)U_n(h_\ell) \\
&= \frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_k) - P^{r-1}h_k(X_i)\}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_\ell) - P^{r-1}h_\ell(X_i)\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_k) - P^{r-1}h_k(X_i)\}P^{r-1}h_\ell(X_i) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_\ell) - P^{r-1}h_\ell(X_i)\}P^{r-1}h_k(X_i) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(P^{r-1}h_k(X_i))(P^{r-1}h_\ell(X_i)) - U_n(h_k)U_n(h_\ell).
\end{aligned}
$$

Recall that $g_k = P^{r-1}h_k$ for each k. Replacing h_k by $h_k - P^r h_k$ in the above expansion, we have

$$
\begin{aligned}
&\left|\widehat{C}_{k,\ell} - P(g_k - Pg_k)(g_\ell - Pg_\ell)\right| \\
&\leqslant \left[\frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_k) - P^{r-1}h_k(X_i)\}^2\right]^{1/2} \\
&\quad \left[\frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_\ell) - P^{r-1}h_\ell(X_i)\}^2\right]^{1/2} \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_k) - P^{r-1}h_k(X_i)\}^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\{g_\ell(X_i) - Pg_\ell\}^2\right]^{1/2} \\
&\quad + \left[\frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_\ell) - P^{r-1}h_\ell(X_i)\}^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\{g_k(X_i) - Pg_k\}^2\right]^{1/2} \\
&\quad + n^{-1/2}|\mathbb{G}_n\left((g_k - Pg_k)(g_\ell - Pg_\ell)\right)| + |(U_n(h_k) - P^r h_k)(U_n(h_\ell) - P^r h_\ell)|,
\end{aligned}
$$

where we have used the Cauchy–Schwarz inequality. Since $n^{-1}\sum_{i=1}^{n}\{g(X_i) - Pg\}^2$ is decomposed as $P(g - Pg)^2 + n^{-1/2}\mathbb{G}_n((g - Pg)^2)$ and the supremum of the latter

on $\mathcal{G}$ is bounded by $\overline{\sigma}_{\mathfrak{g}}^2 + n^{-1/2}\|\mathbb{G}_n\|_{\breve{\mathcal{G}}.\breve{\mathcal{G}}}$, we have

$$\begin{aligned}\Delta_n &:= \max_{1\leqslant k,\ell\leqslant N} \left|\widehat{C}_{k,\ell} - P(g_k - Pg_k)(g_\ell - Pg_\ell)\right| \\ &\leqslant \Upsilon_n + 2\overline{\sigma}_{\mathfrak{g}}\Upsilon_n^{1/2} + 2n^{-1/4}\Upsilon_n^{1/2}\|\mathbb{G}_n\|_{\breve{\mathcal{G}}.\breve{\mathcal{G}}}^{1/2} + n^{-1/2}\|\mathbb{G}_n\|_{\breve{\mathcal{G}}.\breve{\mathcal{G}}} + \|U_n(h) - P^r h\|_{\mathcal{H}}^2 \\ &\leqslant 2\Upsilon_n + 2\overline{\sigma}_{\mathfrak{g}}\Upsilon_n^{1/2} + 2n^{-1/2}\|\mathbb{G}_n\|_{\breve{\mathcal{G}}.\breve{\mathcal{G}}} + \|U_n(h) - P^r h\|_{\mathcal{H}}^2,\end{aligned}$$

where the second inequality follows from the inequality $2ab \leqslant a^2 + b^2$ for $a, b \in \mathbb{R}$. Now, Condition (9) ensures that

$$\begin{aligned}&\Upsilon_n \bigvee (\overline{\sigma}_{\mathfrak{g}}\Upsilon_n^{1/2}) \bigvee \|U_n(h) - P^r h\|_{\mathcal{H}}^2 \\ &\quad\leqslant C\gamma^{-1}\overline{\sigma}_{\mathfrak{g}}\left\{\sigma_{\mathfrak{h}} K_n^{1/2} n^{-1/2} + \nu_{\mathfrak{h}} K_n n^{-3/4+1/q} + (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/2} K_n^{3/4} n^{-3/4}\right. \\ &\qquad\left. + b_{\mathfrak{h}} K_n^{3/2} n^{-1+1/q} + \chi_n\right\}\end{aligned}$$

on the event E, so that

$$\begin{aligned}\Delta_n &\leqslant C\gamma^{-1}\left[(b_{\mathfrak{g}} \vee \sigma_{\mathfrak{h}})\overline{\sigma}_{\mathfrak{g}} K_n^{1/2} n^{-1/2} + b_{\mathfrak{g}}^2 K_n n^{-1+2/q}\right. \\ &\qquad\left. + \overline{\sigma}_{\mathfrak{g}}\left\{\nu_{\mathfrak{h}} K_n n^{-3/4+1/q} + (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/2} K_n^{3/4} n^{-3/4} + b_{\mathfrak{h}} K_n^{3/2} n^{-1+1/q} + \chi_n\right\}\right] \\ &=: \overline{\Delta}_n.\end{aligned}$$

Therefore, the Gaussian comparison inequality of [15, Theorem 3.2] yields that on the event E,

$$\mathbb{P}_{|X_1^n}(Z_n^{\sharp,\varepsilon} \in B) \leqslant \mathbb{P}(\widetilde{Z}^\varepsilon \in B^\eta) + C\eta^{-1}\overline{\Delta}_n^{1/2} K_n^{1/2} \quad \forall B \in \mathcal{B}(\mathbb{R}),\ \forall \eta > 0.$$

Step 6: Conclusion Let

$$\begin{aligned}\delta_n^{(2)} := \frac{1}{\gamma^{1/2}}\left\{\frac{\{(b_{\mathfrak{g}} \vee \sigma_{\mathfrak{h}})\overline{\sigma}_{\mathfrak{g}} K_n^{3/2}\}^{1/2}}{n^{1/4}} + \frac{b_{\mathfrak{g}} K_n}{n^{1/2-1/q}} + \frac{(\overline{\sigma}_{\mathfrak{g}}\nu_{\mathfrak{h}})^{1/2} K_n}{n^{3/8-1/(2q)}}\right. \\ \left. + \frac{\overline{\sigma}_{\mathfrak{g}}^{1/2}(\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/4} K_n^{7/8}}{n^{3/8}} + \frac{(\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{h}})^{1/2} K_n^{5/4}}{n^{1/2-1/(2q)}} + \overline{\sigma}_{\mathfrak{g}}^{1/2}\chi_n^{1/2} K_n^{1/2}\right\}.\end{aligned}$$

Then, from Steps 1–5, we have for every $B \in \mathcal{B}(\mathbb{R})$ and $\eta > 0$,

$$\begin{aligned}\mathbb{P}_{|X_1^n}(Z_n^\sharp \in B) &\leqslant \mathbb{P}_{|X_1^n}(Z_n^{\sharp,\varepsilon} \in B^{C\delta_n^{(1)}}) + n^{-1} \\ &\leqslant \mathbb{P}(\widetilde{Z}^\varepsilon \in B^{C\delta_n^{(1)}+\eta}) + C\eta^{-1}\delta_n^{(2)} + n^{-1} \\ &\leqslant \mathbb{P}(\widetilde{Z} \in B^{C\delta_n^{(1)}+\eta+Cb_{\mathfrak{g}}K_n^{1/2}n^{-1/2}}) + C\eta^{-1}\delta_n^{(2)} + 2n^{-1}.\end{aligned}$$

Choosing $\eta = \gamma^{-1}\delta_n^{(2)}$ leads to the conclusion of the theorem. □

It remains to prove the inequality (32).

Proof of the inequality (32) For a P^{r-1}-integrable symmetric function f on S^{r-1}, $U^{(r-1)}_{n-1,-i}(f)$ is a U-statistic of order $r-1$ and its first projection term is

$$\frac{r-1}{n-1}\sum_{j=1,\neq i}^{n}\{P^{r-2}f(X_j) - P^{r-1}f\} =: S_{n-1,-i}(f).$$

Consider the following decomposition:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{U^{(r-1)}_{n-1,-i}(\delta_{X_i}h) - P^{r-1}h(X_i)\}^2 \\
&\quad\leqslant \frac{2}{n}\sum_{i=1}^{n}\{S_{n-1,-i}(\delta_{X_i}h)\}^2 \\
&\qquad + \frac{2}{n}\sum_{i=1}^{n}\{U^{(r-1)}_{n-1,-i}(\delta_{X_i}h) - P^{r-1}(\delta_{X_i}h) - S_{n-1,-i}(\delta_{X_i}h)\}^2.
\end{aligned} \tag{37}$$

Consider the second term. By Corollary A.4, for given $x \in S$, $\delta_x\mathcal{H} = \{\delta_x x : h \in \mathcal{H}\}$ is VC type with characteristics (A, v) for envelope $\delta_x H$. Hence, we apply Corollary 5.6 conditionally on X_i and deduce that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\left\|U^{(r-1)}_{n-1,-i}(\delta_{X_i}h) - P^{r-1}(\delta_{X_i}h) - S_{n-1,-i}(\delta_{X_i}h)\right\|^2_{\mathcal{H}} \;\middle|\; X_i\right]\right] \\
&\quad\leqslant C\sum_{k=2}^{r-1} n^{-k}\mathbb{E}\left[\|P^{r-k-1}(\delta_x H)\|^2_{P^k,2}|_{x=X_i}\right]K_n^k = C\sum_{k=2}^{r-1} n^{-k}\|P^{r-k-1}H\|^2_{P^{k+1},2}K_n^k.
\end{aligned}$$

Since $\sum_{k=2}^{r-1} n^{-k}\|P^{r-k-1}H\|^2_{P^{k+1},2}K_n^k = \sum_{k=3}^{r} n^{-(k-1)}\|P^{r-k}H\|^2_{P^k,2}K_n^{k-1} \leqslant C\chi_n^2$, the expectation of the supremum on $\mathcal{H}$ of the second term on the right hand side of (37) is at most $C\chi_n^2$.

For the first term, observe that

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\{S_{n-1,-i}(\delta_{X_i}h)\}^2 \\
&\quad= \frac{(r-1)^2}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i}\Big\{(P^{r-2}h)(X_i,X_j)(P^{r-2}h)(X_i,X_k) \\
&\qquad - (P^{r-2}h)(X_i,X_j)(P^{r-1}h)(X_i) \\
&\qquad - (P^{r-2}h)(X_i,X_k)(P^{r-1}h)(X_i) + (P^{r-1}h)^2(X_i)\Big\}.
\end{aligned}$$

Let $\mathcal{F} = \{P^{r-2}h : h \in \mathcal{H}\}$ and $F = P^{r-2}H$, and observe that for $f \in \mathcal{F}$,

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i} \Big\{ f(X_i, X_j) f(X_i, X_k) - f(X_i, X_j)(Pf)(X_i) \\
&\quad - f(X_i, X_k)(Pf)(X_i) + (Pf)^2(X_i) \Big\} \\
&\quad = n(n-1)\{P^2 f^2 - P(Pf)^2\} \\
&\qquad + \sum_{(i,j) \in I_{n,2}} \Big\{ f^2(X_i, X_j) - 2f(X_i, X_j)(Pf)(X_i) + (Pf)^2(X_i) \\
&\quad - P^2 f^2 + P(Pf)^2 \Big\} \\
&\qquad + \sum_{(i,j,k) \in I_{n,3}} \Big\{ f(X_i, X_j) f(X_i, X_k) - f(X_i, X_j)(Pf)(X_i) \\
&\quad - f(X_i, X_k)(Pf)(X_i) + (Pf)^2(X_i) \Big\}.
\end{aligned}$$

Since $P^2 f^2 - P(Pf)^2 \leqslant \sigma_{\mathfrak{h}}^2$, we focus on bounding the suprema of the last two terms. The second term is proportional to a non-degenerate U-statistic of order 2, and the third term is proportional to a degenerate U-statistic of order 3. Define the function classes

$$\begin{aligned}
\mathcal{F}_1 &:= \Big\{ (x_1, x_2) \mapsto f^2(x_1, x_2) - 2f(x_1, x_2)(Pf)(x_1) + (Pf)^2(x_1) : f \in \mathcal{F} \Big\}, \\
\mathcal{F}_2^0 &:= \left\{ (x_1, x_2, x_3) \mapsto \begin{cases} f(x_1, x_2) f(x_1, x_3) - f(x_1, x_2)(Pf)(x_1) \\ \quad - f(x_1, x_3)(Pf)(x_1) + (Pf)^2(x_1) \end{cases} : f \in \mathcal{F} \right\}, \\
\mathcal{F}_2 &:= \Big\{ (x_2, x_3) \mapsto \mathbb{E}[f(X_1, x_2, x_3)] : f \in \mathcal{F}_2^0 \Big\}, \\
\mathcal{F}_3 &:= \Big\{ (x_1, x_2, x_3) \mapsto f(x_1, x_2, x_3) - \mathbb{E}[f(X_1, x_2, x_3)] : f \in \mathcal{F}_2^0 \Big\},
\end{aligned}$$

together with their envelopes

$$\begin{aligned}
F_1(x_1, x_2) &:= F^2(x_1, x_2) + 2F(x_1, x_2)(PF)(x_1) + (PF)^2(x_1), \\
F_2^0(x_1, x_2, x_3) &:= F(x_1, x_2)F(x_1, x_3) + F(x_1, x_2)(PF)(x_1) \\
&\quad + F(x_1, x_3)(PF)(x_1) + (PF)^2(x_1), \\
F_2(x_2, x_3) &:= \mathbb{E}[F_2^0(X_1, x_2, x_3)], \\
F_3(x_1, x_2, x_3) &:= F_2^0(x_1, x_2, x_3) + F_2(x_2, x_3),
\end{aligned}$$

respectively. Lemma 5.4 yields that $\mathcal{F}$ is VC type with characteristics $(4\sqrt{A}, 2v)$ for envelope F, and Corollary A.1 (i) in [14] together with Lemma 5.4 yield that $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ are VC type with characteristics bounded by CA, Cv for envelopes F_1, F_2, F_3, respectively. Functions in $\mathcal{F}_1$ are not symmetric, but after symmetrization we may apply Corollaries 5.5 and 5.6 for $k = 1$ and $k = 2$, respectively. Together

with the Jensen and Cauchy–Schwarz inequalities, we deduce that

$$\begin{aligned}
&\mathbb{E}[\|U_n^{(2)}(f) - P^2 f\|_{\mathcal{F}_1}] \\
&\leqslant C\left\{\sup_{f\in\mathcal{F}}\|f^2\|_{P^2,2}K_n^{1/2}n^{-1/2} + \|F^2\|_{P^2,q/2}K_n n^{-1+2/q} + \|F^2\|_{P^2,2}K_n n^{-1}\right\} \\
&\leqslant C\left(\sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{1/2} n^{-1/2} + b_{\mathfrak{h}}^2 K_n n^{-1+2/q}\right),
\end{aligned}$$

where we have used $\|P^{r-2}h\|_{P^2,4}^4 \leqslant \sigma_{\mathfrak{h}}^2 b_{\mathfrak{h}}^2$ for $h \in \mathcal{H}$ by Condition (MT).

Next, observe that $\|U_n^{(3)}(f)\|_{\mathcal{F}_2^0} \leqslant \|U_n^{(2)}(f)\|_{\mathcal{F}_2} + \|U_n^{(3)}(f)\|_{\mathcal{F}_3}$. Since for $f \in \mathcal{F}_2^0$, $\mathbb{E}[f(x_1, X_2, X_3)] = \mathbb{E}[f(X_1, x_2, X_3)] = \mathbb{E}[f(X_1, X_2, x_3)] = \mathbb{E}[f(x_1, X_2, x_3)] = \mathbb{E}[f(x_1, x_2, X_3)] = 0$ for all $x_1, x_2, x_3 \in S$, both $U_n^{(2)}(f)$, $f \in \mathcal{F}_2$ and $U_n^{(3)}(f)$, $f \in \mathcal{F}_3$ are completely degenerate. So, applying Corollary 5.5 to $\mathcal{F}_2$ and $\mathcal{F}_3$ after symmetrization, combined with the Jensen and Cauchy–Schwarz inequalities, we deduce that

$$\begin{aligned}
\mathbb{E}[\|U_n^{(3)}(f)\|_{\mathcal{F}_2^0}] &\leqslant C\Bigg\{\sup_{f\in\mathcal{F}}\|f^{\odot 2}\|_{P^2,2}K_n n^{-1} + \|F^{\odot 2}\|_{P^2,q/2}K_n^2 n^{-3/2+2/q} \\
&\quad + \sup_{f\in\mathcal{F}}\|f^2\|_{P^2,2}K_n^{3/2}n^{-3/2} + \|F^2\|_{P^2,q/2}K_n^3 n^{-2+2/q}\Bigg\} \\
&\leqslant C\Bigg\{\sup_{f\in\mathcal{F}}\|f^{\odot 2}\|_{P^2,2}K_n n^{-1} + \|F^{\odot 2}\|_{P^2,q/2}K_n^2 n^{-3/2+2/q} \\
&\quad + \sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{3/2} n^{-3/2} + b_{\mathfrak{h}}^2 K_n^3 n^{-2+2/q}\Bigg\}
\end{aligned}$$

where recall that $f^{\odot 2}(x_1, x_2) := f_P^{\odot 2}(x_1, x_2) := \int f(x_1, x) f(x, x_2) dP(x)$ for a symmetric measurable function f on S^2. For $f \in \mathcal{F}$, observe that by the Cauchy–Schwarz inequality,

$$\begin{aligned}
\|f^{\odot 2}\|_{P^2,2}^2 &= \iint\left(\int f(x_1, x) f(x, x_2) dP(x)\right)^2 dP(x_1) dP(x_2) \\
&\leqslant \left(\iint f^2(x_1, x_2) dP(x_1) dP(x_2)\right)^2 = \|f\|_{P^2,2}^4 \leqslant \sigma_{\mathfrak{h}}^4.
\end{aligned}$$

On the other hand, $\|F^{\odot 2}\|_{P^2,q/2} = \nu_{\mathfrak{h}}^2$ by the definition of $\nu_{\mathfrak{h}}$. Therefore, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\left\|n^{-1}\sum_{i=1}^n\{S_{n-1,-i}(\delta_{X_i}h)\}^2\right\|_{\mathcal{H}}\right] \\
&\leqslant C\left\{\sigma_{\mathfrak{h}}^2 K_n n^{-1} + \nu_{\mathfrak{h}}^2 K_n^2 n^{-3/2+2/q} + \sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{3/2} n^{-3/2} + b_{\mathfrak{h}}^2 K_n^3 n^{-2+2/q} + \chi_n^2\right\}.
\end{aligned}$$

This completes the proof. □

Proof of Corollary 3.2 This follows from the discussion before Theorem 3.1 combined with the anti-concentration inequality (Lemma A.1), and optimization with respect to γ. It is without loss of generality to assume that $\eta_n \leqslant \overline{\sigma}_{\mathfrak{g}}^{1/2}$ since otherwise the result is trivial by taking C or C' large enough, and hence Condition (9) is automatically satisfied. □

Acknowledgements The authors would like to thank the anonymous referees and an Associate Editor for their constructive comments that improve the quality of this paper.

Appendix A. Supporting lemmas

This appendix collects some supporting lemmas that are repeatedly used in the main text.

Lemma A.1 (An anti-concentration inequality for the Gaussian supremum) *Let $(S, \mathcal{S}, P)$ be a probability space, and let $\mathcal{G} \subset L^2(P)$ be a P-pre-Gaussian class of functions. Denote by W_P a tight Gaussian random variable in $\ell^\infty(\mathcal{G})$ with mean zero and covariance function $\mathbb{E}[W_P(g)W_P(g')] = \mathrm{Cov}_P(g, g')$ for all $g, g' \in \mathcal{G}$ where $\mathrm{Cov}_P(\cdot, \cdot)$ denotes the covariance under P. Suppose that there exist constants $\underline{\sigma}, \overline{\sigma} > 0$ such that $\underline{\sigma}^2 \leqslant \mathrm{Var}_P(g) \leqslant \overline{\sigma}^2$ for all $g \in \mathcal{G}$. Then for every $\varepsilon > 0$,*

$$\sup_{t\in\mathbb{R}} \mathbb{P}\left\{\left|\sup_{g\in\mathcal{G}} W_P(g) - t\right| \leqslant \varepsilon\right\} \leqslant C_\sigma \varepsilon \left\{\mathbb{E}\left[\sup_{g\in\mathcal{G}} W_P(g)\right] + \sqrt{1 \vee \log(\underline{\sigma}/\varepsilon)}\right\},$$

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

Proof See Lemma A.1 in [14]. □

Lemma A.2 *Let $\mathcal{F}$ be a class of real-valued measurable functions on a measurable space $(\mathcal{X}, \mathcal{A})$ with finite measurable envelope F. Then for any probability measure R on $(\mathcal{X}, \mathcal{A})$ such that $RF^2 < \infty$, we have*

$$N(\mathcal{F}, \|\cdot\|_{R,2}, 4\varepsilon\|F\|_{R,2}) \leqslant \sup_Q N(\mathcal{F}, \|\cdot\|_{Q,2}, \varepsilon\|F\|_{Q,2})$$

for every $0 < \varepsilon \leqslant 1$, where $\sup_Q$ is taken over all finitely discrete distributions on $\mathcal{X}$.

Proof This follows from approximating R by a finitely discrete distribution. See Problem 2.5.1 in [53]. □

Lemma A.3 *Let $(\mathcal{X}, \mathcal{A})$, $(\mathcal{Y}, \mathcal{C})$ be measurable spaces and let $\mathcal{F}$ be a class of real-valued jointly measurable functions on $\mathcal{X} \times \mathcal{Y}$ with finite measurable envelope F. Let R be a probability measure on $(\mathcal{Y}, \mathcal{C})$ and for a jointly measurable function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, define $\overline{f} : \mathcal{X} \to \mathbb{R}$ by $\overline{f}(x) := \int f(x, y)dR(y)$ whenever the latter*

integral is defined and finite for every $x \in \mathcal{X}$. Suppose that $\overline{F}$ is everywhere finite and let $\overline{\mathcal{F}} = \{\overline{f} : f \in \mathcal{F}\}$. Then, for every $r, s \in [1, \infty)$,

$$\sup_Q N(\overline{\mathcal{F}}, \|\cdot\|_{Q,r}, 2\varepsilon\|\overline{F}\|_{Q,r}) \leqslant \sup_{Q'} N(\mathcal{F}, \|\cdot\|_{Q',s}, \varepsilon^r\|F\|_{Q',s}/4)$$

where $\sup_Q$ and $\sup_{Q'}$ are taken over all finitely discrete distributions on $\mathcal{X}$ and $\mathcal{X} \times \mathcal{Y}$, respectively.

Proof This follows from Lemma A.2 in [25] combined with Lemma A.2. □

If $R = \delta_y$ for some $y \in \mathcal{Y}$, then $\|\delta_y f\|_{Q,r}^r = \|f\|_{Q\times\delta_y,r}^r$ (with $\delta_y f(x) = f(x, y)$) and $Q \times \delta_y$ is finitely discrete if Q is so. Hence, we have the following corollary.

Corollary A.4 *Under the setting of Lemma A.3, for every $y \in \mathcal{Y}$ and $r \in [1, \infty)$,*

$$\sup_Q N(\delta_y F, \|\cdot\|_{Q,r}, \varepsilon\|\delta_y F\|_{Q,r}) \leqslant \sup_{Q'} N(\mathcal{F}, \|\cdot\|_{Q',r}, \varepsilon\|F\|_{Q',r}).$$

Lemma A.5 *Let $\mathcal{F}$ and $\mathcal{G}$ be function classes on a set $\mathcal{X}$ with finite envelopes F and G, respectively. If $\mathcal{F} \cdot \mathcal{G}$ stands for the class of pointwise products of functions from $\mathcal{F}$ and $\mathcal{G}$, then for any $r \in [1, \infty)$,*

$$\begin{aligned}&\sup_Q N(\mathcal{F} \cdot \mathcal{G}, \|\cdot\|_{Q,r}, 2\varepsilon\|FG\|_{Q,r}) \\ &\quad\leqslant \sup_Q N(\mathcal{F}, \|\cdot\|_{Q,r}, \varepsilon\|F\|_{Q,r}) \sup_Q N(\mathcal{G}, \|\cdot\|_{Q,r}, \varepsilon\|G\|_{Q,r}),\end{aligned}$$

where $\sup_Q$ is taken over all finitely discrete distributions on $\mathcal{X}$.

Proof See Lemma A.1 in [25] or [53, Section 2.10.3]. □

Appendix B. Strassen–Dudley theorem and its conditional version

In this appendix, we state the Strassen–Dudley theorem together with its conditional version due to [42]. These results play fundamental roles in the proofs of Proposition 2.1 and Theorem 3.1. In what follows, let (S, d) be a Polish metric space equipped with its Borel σ-field $\mathcal{B}(S)$. For any set $A \subset S$ and $\delta > 0$, let $A^\delta = \{x \in S : \inf_{y\in A} d(x, y) \leqslant \delta\}$. We first state the Strassen–Dudley theorem.

Theorem B.1 (Strassen–Dudley) *Let X be an S-valued random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ which admits a uniform random variable on $(0, 1)$ independent of X. Let $\alpha, \beta > 0$ be given constants, and let G be a Borel probability measure on S such that $\mathbb{P}(X \in A) \leqslant G(A^\alpha) + \beta$ for all $A \in \mathcal{B}(S)$. Then there exists an S-valued random variable Y such that $\mathcal{L}(Y)(:= \mathbb{P} \circ Y^{-1}) = G$ and $\mathbb{P}(d(X, Y) > \alpha) \leqslant \beta$.*

For a proof of the Strassen–Dudley theorem, we refer to [20]. Next, we state a conditional version of the Strassen–Dudley theorem due to [42, Theorem 4].

Theorem B.2 (Conditional version of Strassen–Dudley) *Let X be an S-valued random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and let $\mathcal{G}$ be a countably generated sub σ-field of $\mathcal{A}$. Suppose that there is a uniform random variable on $(0,1)$ independent of $\mathcal{G} \vee \sigma(X)$, and let $\Omega \times \mathcal{B}(S) \ni (\omega, A) \mapsto G(A \mid \mathcal{G})(\omega)$ be a regular conditional distribution given $\mathcal{G}$, i.e., for each fixed $A \in \mathcal{B}(S)$, $G(A \mid \mathcal{G})$ is measurable with respect to $\mathcal{G}$ and for each fixed $\omega \in \Omega$, $G(\cdot \mid \mathcal{G})(\omega)$ is a probability measure on $\mathcal{B}(S)$. If*

$$\mathbb{E}^* \left[\sup_{A \in \mathcal{B}(S)} \{ \mathbb{P}(X \in A \mid \mathcal{G}) - G(A^\alpha \mid \mathcal{G}) \} \right] \leqslant \beta,$$

then there exists an S-valued random variable Y such that the conditional distribution of Y given $\mathcal{G}$ is identical to $G(\cdot \mid \mathcal{G})$, and $\mathbb{P}(d(X,Y) > \alpha) \leqslant \beta$.

Remark B.1 (i) The map $(\omega, A) \mapsto \mathbb{P}(X \in A \mid \mathcal{G})(\omega)$ should be understood as a regular conditional distribution (which is guaranteed to exist since X takes values in a Polish space). (ii) $\mathbb{E}^*$ denotes the outer expectation.

For completeness, we provide a self-contained proof of Theorem B.2, since [42] do not provide its direct proof.

Proof of Theorem B.2 Since $\mathcal{G}$ is countably generated, there exists a real-valued random variable W such that $\mathcal{G} = \sigma(W)$. For $n = 1, 2, \ldots$ and $k \in \mathbb{Z}$, let $D_{n,k} = \{k/2^n \leqslant W < (k+1)/2^n\}$. For each n, $\{D_{n,k} : k \in \mathbb{Z}\}$ forms a partition of Ω. Pick any D from $\{D_{n,k} : n = 1, 2, \ldots; k \in \mathbb{Z}\}$; let $\mathbb{P}_D = \mathbb{P}(\cdot \mid D)$ and $G(\cdot \mid D) = \int G(\cdot \mid \mathcal{G}) d\mathbb{P}_D$. Then, the Strassen–Dudley theorem yields that there exists an S-valued random variable Y_D such that $\mathbb{P}_D \circ Y_D^{-1} = G(\cdot \mid D)$ and $\mathbb{P}_D(d(X, Y_D) > \alpha) \leqslant \varepsilon(D) := \sup_{A \in \mathcal{B}(S)} \{\mathbb{P}_D(X \in A) - G(A^\alpha \mid D)\}$.

For each $n = 1, 2, \ldots$, let $Y_n = \sum_{k \in \mathbb{Z}} Y_{D_{n,k}} 1_{D_{n,k}}$, and observe that

$$\mathbb{P}(d(X, Y_n) > \alpha) = \sum_k \mathbb{P}_{D_{n,k}}(d(X, Y_{D_{n,k}}) > \alpha) \mathbb{P}(D_{n,k}) \leqslant \sum_k \varepsilon(D_{n,k}) \mathbb{P}(D_{n,k}).$$

Let M be any (proper) random variable such that $M \geqslant \sup_{A \in \mathcal{B}(S)} \{\mathbb{P}(X \in A \mid \mathcal{G}) - G(A^\alpha \mid \mathcal{G})\}$, and observe that

$$\mathbb{P}_D(X \in A) - G(A^\alpha \mid D) = \mathbb{E}^{\mathbb{P}_D}[\mathbb{P}(X \in A \mid \mathcal{G}) - G(A^\alpha \mid \mathcal{G})] \leqslant \mathbb{E}^{\mathbb{P}_D}[M],$$

where the notation $\mathbb{E}^{\mathbb{P}_D}$ denotes the expectation under $\mathbb{P}_D$. So,

$$\sum_k \varepsilon(D_{n,k}) \mathbb{P}(D_{n,k}) \leqslant \sum_k \mathbb{E}^{\mathbb{P}_{D_{n,k}}}[M] \mathbb{P}(D_{n,k}) = \mathbb{E}[M],$$

and taking infimum with respect to M yields that the left hand side is bounded by β.

Next, we shall verify that $\{\mathcal{L}(Y_n) : n \geqslant 1\}$ is uniformly tight. In fact,

$$\mathbb{P}(Y_n \in A) = \sum_k \mathbb{P}(\{Y_{D_{n,k}} \in A\} \cap D_{n,k}) = \sum_k \mathbb{P}_{D_{n,k}}(Y_{D_{n,k}} \in A) \mathbb{P}(D_{n,k})$$

$$= \sum_k G(A \mid D_{n,k})\mathbb{P}(D_{n,k}) = \mathbb{E}[G(A \mid \mathcal{G})],$$

and since any Borel probability measure on a Polish space is tight by Ulam's theorem, $\{\mathcal{L}(Y_n) : n \geqslant 1\}$ is uniformly tight. This implies that the family of joint laws $\{\mathcal{L}(X, W, Y_n) : n \geqslant 1\}$ is uniformly tight and hence has a weakly convergent subsequence by Prohorov's theorem. Let $\mathcal{L}(X, W, Y_{n'}) \xrightarrow{w} Q$ (the notation $\xrightarrow{w}$ denotes weak convergence), and observe that the marginal law of Q on the "first two" coordinates, $S \times \mathbb{R}$, is identical to $\mathcal{L}(X, W)$.

We shall verify that there exists an S-valued random variable Y such that $\mathcal{L}(X, W, Y) = Q$. Since S is polish, there exists a unique regular conditional distribution, $\mathcal{B}(S) \times (S \times \mathbb{R}) \ni (A, (x, w)) \mapsto Q_{x,w}(A) \in [0, 1]$, for Q given the first two coordinates. By the Borel isomorphism theorem [20, Theorem 13.1.1], there exists a bijective map π from S onto a Borel subset of $\mathbb{R}$ such that π and π^{-1} are Borel measurable. Pick and fix any $(x, w) \in S \times \mathbb{R}$, and observe that $Q_{x,w} \circ \pi^{-1}$ extends to a Borel probability measure on $\mathbb{R}$. Denote by $F_{x,w}$ the distribution function of $Q_{x,w} \circ \pi^{-1}$, and let $F^{-1}_{x,w}$ denotes its quantile function. Let U be a uniform random variable on $(0, 1)$ (defined on $(\Omega, \mathcal{A}, \mathbb{P})$) independent of (X, W). Then $F^{-1}_{x,w}(U)$ has law $Q_{x,w} \circ \pi^{-1}$, and hence $Y = \pi^{-1} \circ F^{-1}_{X,W}(U)$ is the desired random variable.

Now, for any bounded continuous function f on S, observe that, whenever $N \geqslant n$, $\mathbb{E}[f(Y_N)1_{D_{n,k}}] = \int_{D_{n,k}} \int f(y)G(dy \mid \mathcal{G})d\mathbb{P}$, which implies that the conditional distribution of Y given $\mathcal{G}$ is identical to $G(\cdot \mid \mathcal{G})$. Finally, the Portmanteau theorem yields $\mathbb{P}(d(X, Y) > \alpha) \leqslant \liminf_{n'} \mathbb{P}(d(X, Y_{n'}) > \alpha) \leqslant \beta$. This completes the proof. □

Appendix C. Additional proofs for the main text

C.1. Proof of Lemma 6.1

We begin with noting that $\mathcal{G}$ is VC type with characteristics $4\sqrt{A}$ and $2v$ for envelope G. The rest of the proof is almost the same as that of Theorem 2.1 in [15] with $B(f) \equiv 0$ (up to adjustments of the notation), but we now allow $q = \infty$. To avoid repetitions, we only point out required modifications. In what follows, we will freely use the notation in the proof of [15, Theorem 2.1], but modify K_n to $K_n = v \log(A \vee n)$, and C refers to a universal constant whose value may vary from place to place. In Step 1, change ε to $\varepsilon = 1/n^{1/2}$. For this choice, $\log N(\mathcal{F}, e_P, \varepsilon b) \leqslant C \log(Ab/(\varepsilon b)) = C \log(A/\varepsilon) \leqslant CK_n$, and Dudley's entropy integral bound yields that $\mathbb{E}[\|G_P\|_{\mathcal{F}_\varepsilon}] \leqslant C\varepsilon b\sqrt{\log(Ab/(\varepsilon b))} \leqslant Cb\sqrt{K_n/n}$ (there is a slip in the estimate of $\mathbb{E}[\|G_P\|_{\mathcal{F}_\varepsilon}]$ in [15], namely, "Ab/ε" inside the log should read "$Ab/(\varepsilon b)$", which of course does not affect the proof under their definition of K_n). Combining the Borell-Sudakov-Tsirel'son inequality yields that $\mathbb{P}\{\|G_P\|_{\mathcal{F}_\varepsilon} > Cb\sqrt{K_n/n}\} \leqslant 2n^{-1}$. In Step 3, Corollary 5.5 in the present paper (with $r = k = 1$) yields that $\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon}] \leqslant C(b\sqrt{K_n/n} + bK_n/n^{1/2-1/q}) \leqslant CbK_n/n^{1/2-1/q}$, which is valid even when $q = \infty$. Then, instead of applying their Lemma 6.1, we apply Markov's inequality to deduce that

$$\mathbb{P}\left\{\|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon} > CbK_n/(\gamma n^{1/2-1/q})\right\} \leqslant \gamma.$$

In Step 4, instead of their equation (14), we have

$$\mathbb{P}(Z^\varepsilon \in B) \leqslant \mathbb{P}(\widetilde{Z}^\varepsilon \in B^{C_7\delta}) + C\left(\frac{b\sigma^2 K_n^2}{\delta^3\sqrt{n}} + \frac{M_{n,X}(\delta)K_n^2}{\delta^3\sqrt{n}} + \frac{1}{n}\right) \quad \forall B \in \mathcal{B}(\mathbb{R})$$

whenever $\delta \geqslant 2c\sigma^{-1/2}(\log N)^{3/2} \cdot (\log n)$ for some universal constant c (C_7 comes from their Theorem 3.1 and is universal). Finally, in Step 5, take

$$\delta = C'\left\{\frac{(b\sigma^2 K_n^2)^{1/3}}{\gamma^{1/3}n^{1/6}} + \frac{2bK_n}{\gamma n^{1/2-1/q}}\right\}$$

for some large but universal constant $C' > 1$. Under the assumption that $K_n^3 \leqslant n$, this choice ensures that $\delta \geqslant 2c\sigma^{-1/2}(\log N)^{3/2} \cdot (\log n)$, and

$$\frac{b\sigma^2 K_n^2}{\delta^3\sqrt{n}} \leqslant \frac{1}{(C')^3 n}.$$

It remains to bound $M_{n,X}(\delta)$. For finite q, their Step 4 shows that

$$\frac{M_{n,X}(\delta)K_n^2}{\delta^3\sqrt{n}} \leqslant \frac{2^q b^q K_n^2 (\log N)^{q-3}}{\delta^q n^{q/2-1}}.$$

Since $\log N \leqslant C'' K_n$ for some universal constant C'', the right hand side is bounded by

$$\frac{\gamma^q (C'')^{q-3}}{(C')^q K_n}.$$

Since K_n is bounded from below by a universal positive constant (by assumption), and $\gamma \in (0, 1)$, by taking $C' > C''$, the above term is bounded by γ up to a universal constant.

Now, consider the $q = \infty$ case. In that case, $\max_{1\leqslant j\leqslant N} |\widetilde{X}_{1j}| \leqslant 2b$ almost surely and $\delta\sqrt{n}/\log N \geqslant 2C'b/(C''\gamma) > 2b$ provided that $C' > C''$. Hence $M_{n,X}(\delta) = 0$ in that case. These modifications lead to the desired conclusion. □

C.1. Proofs for Sect. 4

We first prove Theorem 4.2 and Corollary 4.3, and then prove Lemma 4.1 and Theorem 4.4.

Proof of Theorem 4.2 In what follows, the notation $\lesssim$ signifies that the left hand side is bounded by the right hand side up to a constant that depends only on $r, m, \zeta, c_1, c_2, C_1, L$. We also write $a \simeq b$ if $a \lesssim b$ and $b \lesssim a$. In addition, let c, C, C'

denote generic constants depending only on $r, m, \zeta, c_1, c_2, C_1, L$; their values may vary from place to place. We divide the rest of the proof into three steps.

Step 1 Let

$$S_n^\sharp := \sup_{\vartheta\in\Theta} \frac{b_n^{m/2}}{c_n(\vartheta)\sqrt{n}} \sum_{i=1}^n \xi_i \left[U_{n-1,-i}^{(r-1)}(\delta_{D_i} h_{n,\vartheta}) - U_n(h_{n,\vartheta}) \right].$$

In this step, we shall show that the result (15) holds with $\widehat{S}_n$ and $\widehat{S}_n^\sharp$ replaced by S_n and $S_n^\sharp$, respectively.

We first verify Conditions (PM), (VC), (MT), and (5) for the function class

$$\mathcal{H}_n = \left\{ b_n^{m/2} c_n(\vartheta)^{-1} h_{n,\vartheta} : \vartheta \in \Theta \right\}$$

with a symmetric envelope

$$H_n(d_{1:r}) = b_n^{-(r-1/2)m} c_1^{-1} \|L\|_{\mathbb{R}^m}^r \overline{\varphi}(v_{1:r}) \prod_{i=1}^r 1_{\mathcal{X}^{\zeta/2}}(x_i) \prod_{1\leqslant i<j\leqslant r} 1_{[-2,2]^m}(b_n^{-1}(x_i - x_j)).$$

Condition (PM) follows from our assumption. For Condition (VC), that $\mathcal{H}_n$ is VC type with characteristics (A', v') satisfying $\log A' \lesssim \log n$ and $v' \lesssim 1$ follows from a slight modification of the proof of Lemma 3.1 in [25]. The latter part follows from our assumption. Condition (VC) guarantees the existence of a tight Gaussian random variable $\mathcal{W}_{P,n}(g)$, $g \in P^{r-1}\mathcal{H}_n =: \mathcal{G}_n$ in $\ell^\infty(\mathcal{G}_n)$ with mean zero and covariance function $\mathbb{E}[\mathcal{W}_{P,n}(g)\mathcal{W}_{P,n}(g')] = \mathrm{Cov}_P(g, g')$ for $g, g' \in \mathcal{G}_n$. Let $W_{P,n}(\vartheta) = \mathcal{W}_{P,n}(g_{n,\vartheta})$ for $\vartheta \in \Theta$ where $g_{n,\vartheta} = b_n^{m/2} c_n(\vartheta)^{-1} P^{r-1} h_{n,\vartheta}$. It is seen that $W_{P,n}(\vartheta)$, $\vartheta \in \Theta$ is a tight Gaussian random variable in $\ell^\infty(\Theta)$ with mean zero and covariance function (14).

Next, we determine the values of parameters $\underline{\sigma}_\mathfrak{g}, \overline{\sigma}_\mathfrak{g}, b_\mathfrak{g}, \sigma_\mathfrak{h}, b_\mathfrak{h}, \chi_n, \nu_\mathfrak{h}$ for the function class $\mathcal{H}_n$. We will show in Step 3 that we may choose

$$\underline{\sigma}_\mathfrak{g} \simeq 1,\ \overline{\sigma}_\mathfrak{g} \simeq 1,\ b_\mathfrak{g} \simeq b_n^{-m/2},\ \sigma_\mathfrak{h} \simeq b_n^{-m/2},\ b_\mathfrak{h} \simeq b_n^{-3m/2}, \tag{38}$$

and bound $\nu_\mathfrak{h}$ and χ_n as

$$\nu_\mathfrak{h} \lesssim b_n^{-m(1-1/q)},\ \chi_n \lesssim (\log n)^{3/2}/(n b_n^{3m/2}). \tag{39}$$

Given these choices and bounds, Corollaries 2.2 and 3.2 yield that

$$\begin{aligned} &\sup_{t\in\mathbb{R}} \left| \mathbb{P}(S_n \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t) \right| \leqslant C n^{-c} \text{ and} \\ &\mathbb{P}\left\{ \sup_{t\in\mathbb{R}} \left| \mathbb{P}_{|D_1^n}(S_n^\sharp \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t) \right| > C n^{-c} \right\} \leqslant C n^{-c}. \end{aligned} \tag{40}$$

Step 2 Observe that

$$|\widehat{S}_n - S_n| \leqslant \sup_{\vartheta\in\Theta}\left|\frac{c_n(\vartheta)}{\widehat{c}_n(\vartheta)} - 1\right| \|\sqrt{n}U_n\|_{\mathcal{H}_n} \quad \text{and}$$
$$|\widehat{S}_n^\sharp - S_n^\sharp| \leqslant \sup_{\vartheta\in\Theta}\left|\frac{c_n(\vartheta)}{\widehat{c}_n(\vartheta)} - 1\right| \|\mathbb{U}_n^\sharp\|_{\mathcal{H}_n}. \tag{41}$$

We shall bound $\sup_{\vartheta\in\Theta}|c_n(\vartheta)/\widehat{c}_n(\vartheta) - 1|$, $\|\sqrt{n}U_n\|_{\mathcal{H}_n}$, and $\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_n}$.

Choose n_0 by the smallest n such that $C_1 n^{-c_2} \leqslant 1/2$; it is clear that n_0 depends only on c_2 and C_1. It suffices to prove (15) for $n \geqslant n_0$, since for $n < n_0$, the result (15) becomes trivial by taking C sufficiently large. So let $n \geqslant n_0$. Then Condition (T8) ensures that with probability at least $1 - C_1 n^{-c_2}$, $\inf_{\vartheta\in\Theta}\widehat{c}_n(\vartheta)/c_n(\vartheta) \geqslant 1/2$. Since $|a^{-1} - 1| \leqslant 2|a - 1|$ for $a \geqslant 1/2$, Condition (T8) also ensures that

$$\mathbb{P}\left\{\sup_{\vartheta\in\Theta}\left|\frac{c_n(\vartheta)}{\widehat{c}_n(\vartheta)} - 1\right| > Cn^{-c}\right\} \leqslant Cn^{-c}. \tag{42}$$

Next, we shall bound $\|\sqrt{n}U_n\|_{\mathcal{H}_n}$ and $\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_n}$. Given (38) and (39), and in view of the fact that the covering number of $\mathcal{H}_n \cup (-\mathcal{H}_n) := \{h, -h : h \in \mathcal{H}_n\}$ is at most twice that of $\mathcal{H}_n$, applying Corollaries 2.2 and 3.2 to the function class $\mathcal{H}_n \cup (-\mathcal{H}_n)$, we deduce that

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}(\|\sqrt{n}U_n\|_{\mathcal{H}_n} \leqslant t) - \mathbb{P}(\|\mathcal{W}_{P,n}\|_{\mathcal{G}_n} \leqslant t)\right| \leqslant Cn^{-c} \text{ and}$$
$$\mathbb{P}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{P}_{|D_1^n}(\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_n} \leqslant t) - \mathbb{P}(\|\mathcal{W}_{P,n}\|_{\mathcal{G}_n} \leqslant t)\right| > Cn^{-c}\right\} \leqslant Cn^{-c}.$$

(Theorem 3.7.28 in [29] ensures that the Gaussian process $\mathcal{W}_{P,n}$ extends to the symmetric convex hull of $\mathcal{G}_n$ in such a way that $\mathcal{W}_{P,n}$ has linear, bounded, and uniformly continuous (with respect to the intrinsic pseudometric) sample paths; in particular, $\{\mathcal{W}_{P,n}(g) : g \in \mathcal{G}_n \cup (-\mathcal{G}_n)\}$ is a tight Gaussian random variable in $\ell^\infty(\mathcal{G}_n \cup (-\mathcal{G}_n))$ with mean zero and covariance function $\mathbb{E}[\mathcal{W}_{P,n}(g)\mathcal{W}_{P,n}(g')] = \mathrm{Cov}_P(g, g')$ for $g, g' \in \mathcal{G}_n \cup (-\mathcal{G}_n)$ and $\sup_{g\in\mathcal{G}_n\cup(-\mathcal{G}_n)}\mathcal{W}_n(g) = \|\mathcal{W}_{P,n}\|_{\mathcal{G}_n}$.) Dudley's entropy integral bound and the Borell-Sudakov-Tsirel'son inequality yield that $\mathbb{P}\{\|\mathcal{W}_{P,n}\|_{\mathcal{G}_n} > C(\log n)^{1/2}\} \leqslant 2n^{-1}$, so that

$$\mathbb{P}\{\|\sqrt{n}U_n\|_{\mathcal{H}_n} > C(\log n)^{1/2}\} \leqslant Cn^{-c} \text{ and}$$
$$\mathbb{P}\left\{\mathbb{P}_{|D_1^n}\{\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_n} > C(\log n)^{1/2}\} > Cn^{-c}\right\} \leqslant Cn^{-c}. \tag{43}$$

Now, the desired result (15) follows from combining (40)–(43) and the anti-concentration inequality (Lemma A.1). In fact, the anti-concentration inequality yields

$$\sup_{t\in\mathbb{R}}\mathbb{P}(|\widetilde{S}_n - t| \leqslant Cn^{-c}) \leqslant C'n^{-c}(\log n)^{1/2}. \tag{44}$$

Hence, combining the bounds (40)–(44), we have for every $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(\widehat{S}_n \leqslant t) &\leqslant \mathbb{P}(S_n \leqslant t + Cn^{-c}) + Cn^{-c} \\ &\leqslant \mathbb{P}(\widetilde{S}_n \leqslant t + Cn^{-c}) + Cn^{-c} \\ &\leqslant \mathbb{P}(\widetilde{S}_n \leqslant t) + Cn^{-c},\end{aligned}$$

and likewise $\mathbb{P}(\widehat{S}_n \leqslant t) \geqslant \mathbb{P}(\widetilde{S}_n \leqslant t) - Cn^{-c}$. Similarly, we have

$$\mathbb{P}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{P}_{|D_1^n}(\widehat{S}_n^{\natural} \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t)\right| > Cn^{-c}\right\} \leqslant Cn^{-c}.$$

Step 3 It remains to verify (38) and (39). First, that we may choose $\underline{\sigma}_{\mathfrak{g}} \simeq 1$ follows from Conditions (T6) and (T7). For $\varphi \in \Phi$ and $k = 1, \dots, r-1$, let

$$\varphi_{[r-k]}(v_{1:k}, x_{k+1:r}) = \mathbb{E}[\varphi(v_{1:k}, V_{k+1:r}) \mid X_{k+1:r} = x_{k+1:r}] \prod_{j=k+1}^{r} p(x_j),$$

and define $\overline{\varphi}_{[r-k]}$ similarly. Then, for $k = 1, \dots, r$,

$$\begin{aligned}(P^{r-k}h_{n,\vartheta})(d_{1:k}) = \left(\prod_{j=1}^{k} L_{b_n}(x - x_j)\right) \int_{[-1,1]^{m(r-k)}} \varphi_{[r-k]}(v_{1:k}, x - b_n x_{k+1:r}) \\ \left(\prod_{j=k+1}^{r} L(x_j)\right) dx_{k+1:r},\end{aligned}$$

where $x - b_n x_{k+1:r} = (x - b_n x_{k+1}, \dots, x - b_n x_r)$. Likewise, we have

$$\begin{aligned}(P^{r-k}H_n)(d_{1:k}) \lesssim b_n^{-(k-1/2)m} \left(\prod_{i=1}^{k} 1_{\mathcal{X}^{s/2}}(x_i)\right) \left(\prod_{1\leqslant i<j\leqslant k} 1_{[-2,2]^m}(b_n^{-1}(x_i - x_j))\right) \\ \times \int_{[-2,2]^{m(r-k)}} \overline{\varphi}_{[r-k]}(v_{1:k}, x_1 - b_n x_{k+1:r}) dx_{k+1:r}.\end{aligned}$$

Suppose first that q is finite and let $\ell \in [2, q]$. Observe that by Jensen's inequality,

$$\begin{aligned}\|P^{r-k}h_{n,\vartheta}\|_{P^k,\ell}^{\ell} &\leqslant C^{\ell} b_n^{-(\ell-1)mk} \int_{[-1,1]^{mr}} \mathbb{E}\left[\overline{\varphi}^{\ell}(V_{1:r}) \mid X_{1:r} = x - b_n x_{1:r}\right] \\ &\qquad \left(\prod_{j=1}^{k} p(x - b_n x_j)\right) dx_{1:r} \\ &\leqslant C^{\ell} b_n^{-(\ell-1)mk} \int_{[-1,1]^{mr}} \mathbb{E}\left[\overline{\varphi}^{\ell}(V_{1:r}) \mid X_{1:r} = x - b_n x_{1:r}\right] dx_{1:r} \\ &\leqslant C^{\ell} b_n^{-(\ell-1)mk},\end{aligned}$$

so that $\sup_{h\in\mathcal{H}_n}\|P^{r-k}h\|_{P^k,\ell}\lesssim b_n^{-m[(k-1/2)-k/\ell]}$. Hence, we may choose $\overline{\sigma}_{\mathfrak{g}}\simeq 1$ and $\sigma_{\mathfrak{h}}\simeq b_n^{-m/2}$. Similarly, Jensen's inequality and the symmetry of $\overline{\varphi}$ yield that

$$\begin{aligned}\|P^{r-k}H_n\|_{P^k,\ell}^{\ell}&\leqslant C^{\ell}b_n^{-(k-1/2)m\ell+m(k-1)}\times\int_{\mathcal{X}^{\zeta/2}\times[-2,2]^{m(r-1)}}\\&\quad\mathbb{E}\left[\overline{\varphi}^{\ell}(V_{1:r})\mid X_1=x_1,X_{2:r}=x_1-b_nx_{2:j}\right]p(x_1)\\&\quad\prod_{j=2}^{k}p(x_1-b_nx_j)dx_{1:r}\\&\leqslant C^{\ell}b_n^{-(k-1/2)m\ell+m(k-1)}\int_{\mathcal{X}^{\zeta/2}\times[-2,2]^{m(r-1)}}\\&\quad\mathbb{E}\left[\overline{\varphi}^{\ell}(V_{1:r})\mid X_1=x_1,X_{2:r}=x_1-b_nx_{2:j}\right]dx_{1:r}\\&\leqslant C^{\ell}b_n^{-(k-1/2)m\ell+m(k-1)},\end{aligned}$$

so that $\|P^{r-k}H_n\|_{P^k,\ell}\lesssim b_n^{-m[(1-1/\ell)k-(1/2-1/\ell)]}$. Hence, we may choose $b_{\mathfrak{g}}\simeq b_n^{-m/2}$, $b_{\mathfrak{h}}\simeq b_n^{-3m/2}$, and bound χ_n as

$$\chi_n\lesssim\sum_{k=3}^{r}n^{-(k-1)/2}(\log n)^{k/2}b_n^{-mk/2}\lesssim\frac{(\log n)^{3/2}}{nb_n^{3m/2}}.$$

Similar calculations yield that

$$\begin{aligned}\|(P^{r-2}H_n)^{\odot 2}\|_{P^2,q/2}^{q/2}&\leqslant C^qb_n^{-m(q-1)}\int_{\mathcal{X}^{\zeta/2}\times[-2,2]^{m(r-1)}}\\&\quad\mathbb{E}\left[\overline{\varphi}^{q}(V_{1:r})\mid X_1=x_1,X_{2:r}=x_1-b_nx_{2:j}\right]dx_{1:r}\\&\leqslant C^qb_n^{-m(q-1)}.\end{aligned}$$

Hence, $\nu_{\mathfrak{h}}\lesssim b_n^{-m(1-1/q)}$.

It is not difficult to verify that (38) and (39) hold in the $q=\infty$ case as well under the convention that $1/q=0$ for $q=\infty$. This completes the proof. □

Proof of Corollary 4.3 Let $\eta_n:=Cn^{-c}$ where the constants c,C are those given in Theorem 4.2. Denote by $q_{\widetilde{S}_n}(\alpha)$ the α-quantile of $\widetilde{S}_n$. Define the event

$$\mathcal{E}_n:=\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{P}_{|D_1^n}(\widehat{S}_n^{\sharp}\leqslant t)-\mathbb{P}(\widetilde{S}_n\leqslant t)\right|\leqslant\eta_n\right\},$$

whose probability is at least $1-\eta_n$. On this event,

$$\begin{aligned}\mathbb{P}_{|D_1^n}\left\{\widehat{S}_n^{\sharp}\leqslant q_{\widetilde{S}_n}(\alpha+\eta_n)\right\}&\geqslant\mathbb{P}\left\{\widetilde{S}_n\leqslant q_{\widetilde{S}_n}(\alpha+\eta_n)\right\}-\eta_n\\&=\alpha+\eta_n-\eta_n=\alpha,\end{aligned}$$

where the second equality follows from the fact that the distribution function of $\widetilde{S}_n$ is continuous (cf. Lemma A.1). This shows that the inequality $q_{\widehat{S}_n^\sharp}(\alpha) \leqslant q_{\widetilde{S}_n}(\alpha+\eta_n)$ holds on the event $\mathcal{E}_n$, so that

$$\begin{aligned}\mathbb{P}\left\{\widehat{S}_n \leqslant q_{\widehat{S}_n^\sharp}(\alpha)\right\} &\leqslant \mathbb{P}\left\{\widehat{S}_n \leqslant q_{\widetilde{S}_n}(\alpha+\eta_n)\right\} + \mathbb{P}(\mathcal{E}_n^c) \\ &\leqslant \mathbb{P}\left\{\widetilde{S}_n \leqslant q_{\widetilde{S}_n}(\alpha+\eta_n)\right\} + 2\eta_n \\ &= \alpha + 3\eta_n.\end{aligned}$$

The above discussion presumes that $\alpha+\eta_n < 1$, but if $\alpha+\eta_n \geqslant 1$, then the last inequality is trivial. Likewise, we have $\mathbb{P}\left\{\widehat{S}_n \leqslant q_{\widehat{S}_n^\sharp}(\alpha)\right\} \geqslant \alpha - 3\eta_n$. This completes the proof. □

Proof of Lemma 4.1 We begin with noting that

$$\left|\frac{\widehat{c}_n(\vartheta)}{c_n(\vartheta)} - 1\right| \leqslant \left|\frac{\widehat{c}_n^2(\vartheta)}{c_n^2(\vartheta)} - 1\right| \leqslant \frac{1}{n}\sum_{i=1}^n \left[\{U_{n-1,-i}^{(r-1)}(\delta_{D_i}\check{h}_{n,\vartheta}) - U_n(\check{h}_{n,\vartheta})\}^2 - 1\right],$$

where $\check{h}_{n,\vartheta} = b_n^{m/2} c_n(\vartheta)^{-1} h_{n,\vartheta}$. We note that $\mathrm{Var}_P(P^{r-1}\check{h}_{n,\vartheta}) = 1$ by the definition of $c_n(\vartheta)$. Recall from the proof of Theorem 4.2 that the function class $\mathcal{H}_n = \{\check{h}_{n,\vartheta} : \vartheta \in \Theta\}$ is VC type with characteristics (A', v') satisfying $\log A' \lesssim \log n$ and $v' \lesssim 1$ for envelope H_n. Now, from Step 5 in the proof of Theorem 3.1 applied with $\mathcal{H} = \mathcal{H}_n$, we have for every $\gamma \in (0,1)$, with probability at least $1 - \gamma - n^{-1}$,

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^n \left[\{U_{n-1,-i}^{(r-1)}(\delta_{D_i}h) - U_n(h)\}^2 - 1\right]\right\|_{\mathcal{H}_n} \\ &\quad\leqslant C\gamma^{-1}\Bigg[(b_{\mathfrak{g}} \vee \sigma_{\mathfrak{h}})\overline{\sigma}_{\mathfrak{g}} K_n^{1/2} n^{-1/2} + b_{\mathfrak{g}}^2 K_n n^{-1+2/q} \\ &\qquad + \overline{\sigma}_{\mathfrak{g}}\left\{\nu_{\mathfrak{h}} K_n n^{-3/4+1/q} + (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/2} K_n^{3/4} n^{-3/4} + b_{\mathfrak{h}} K_n^{3/2} n^{-1+1/q} + \chi_n\right\}\Bigg]\end{aligned}$$

for some constant C depending only on r. The desired result follows from the choices of parameters $\overline{\sigma}_{\mathfrak{g}}$, $b_{\mathfrak{g}}$, $\sigma_{\mathfrak{h}}$, $b_{\mathfrak{h}}$, χ_n, and $\nu_{\mathfrak{h}}$ given in the proof of Theorem 4.2 together with choosing $\gamma = n^{-c}$ for some constant c sufficiently small but depending only on $r, m, \zeta, c_1, c_2, C_1, L$. □

Proof of Theorem 4.4 The proof follows from similar arguments to those in the proof of Theorem 4.2, so we only highlight the differences. Define the function class

$$\mathcal{H}_n = \left\{b^{m/2} c_n(\vartheta, b)^{-1} h_{\vartheta,b} : \vartheta \in \Theta, b \in \mathcal{B}_n\right\}$$

with a symmetric envelope

$$H_n(d_{1:r}) = \underline{b}_n^{-(r-1/2)m} c_1^{-1} \|L\|_{\mathbb{R}^m}^r \overline{\varphi}(v_{1:r}) \prod_{i=1}^r 1_{\mathcal{X}^{\zeta/2}}(x_i) \prod_{1 \leqslant i < j \leqslant r} 1_{[-2,2]^m}(\overline{b}_n^{-1}(x_i - x_j)).$$

Recall that we assume $q = \infty$ in this theorem. In view of the calculations in the proof of Theorem 4.2, we may choose

$$\underline{\sigma}_{\mathfrak{g}} \simeq 1,\ \overline{\sigma}_{\mathfrak{g}} \simeq 1,\ b_{\mathfrak{g}} \simeq \kappa_n^{m(r-1)} \underline{b}_n^{-m/2},\ \sigma_{\mathfrak{h}} \simeq \underline{b}_n^{-m/2},\ b_{\mathfrak{h}} \simeq \kappa_n^{m(r-2)} \underline{b}_n^{-3m/2},$$

and bound $\nu_{\mathfrak{h}}$ and χ_n as

$$\nu_{\mathfrak{h}} \lesssim \kappa_n^{m/2} \underline{b}_n^{-m},\ \chi_n \lesssim \frac{\kappa_n^{m(r-2)} (\log n)^{3/2}}{n \underline{b}_n^{3m/2}}.$$

Given these choices and bounds, the conclusion of the theorem follows from repeating the proof of Theorem 4.2. □

Appendix D. Conditional UCLT for JMB

In this section we prove the conditional UCLT for the JMB when the function class $\mathcal{H}$ and the distribution P are independent of n under a metric entropy condition. We obey the notation used in Sects. 2 and 3 but since we consider a limit theorem we assume that the probability space is $(\Omega, \mathcal{A}, \mathbb{P}) = (S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}}, P^{\mathbb{N}}) \times (\Xi, \mathcal{C}, R)$ and $X_1, X_2, \ldots$ are the coordinate projections of $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}}, P^{\mathbb{N}})$. To formulate the conditional UCLT, recall that weak convergence in $\ell^\infty(\mathcal{H})$ is "metrized" by the bounded Lipschitz distance: for arbitrary maps $\mathbb{X}_n : \Omega \to \ell^\infty(\mathcal{H})$ and a tight Borel measurable map $\mathbb{X} : \Omega \to \ell^\infty(\mathcal{H})$, $\mathbb{X}_n$ converge weakly to $\mathbb{X}$ if and only if

$$d_{BL}(\mathbb{X}_n, \mathbb{X}) := \sup_{f \in BL_1} |\mathbb{E}^*[f(\mathbb{X}_n)] - \mathbb{E}[f(\mathbb{X})]| \to 0,$$

where $BL_1 = \{f : \ell^\infty(\mathcal{H}) \to \mathbb{R} : |f| \leqslant 1, |f(x) - f(y)| \leqslant \|x - y\|_{\mathcal{H}}\ \forall x, y \in \ell^\infty(\mathcal{H})\}$; see [53, p. 73]. If the function class $\mathcal{G} = P^{r-1}\mathcal{H} = \{P^{r-1}h : h \in \mathcal{H}\}$ is P-pre-Gaussian, then there exists a tight Gaussian random variable W_P in $\ell^\infty(\mathcal{G})$ with mean zero and covariance function $\mathbb{E}[W_P(g)W_P(g')] = \mathrm{Cov}_P(g, g')$. Set $\mathbb{W}_P(h) = W_P \circ P^{r-1}(h)$, which is a tight Gaussian random variable in $\ell^\infty(\mathcal{H})$ with mean zero and covariance function $\mathbb{E}[\mathbb{W}_P(h)\mathbb{W}_P(h')] = \mathrm{Cov}_P(P^{r-1}h, P^{r-1}h')$. We will show that conditionally on $X_1^\infty = \{X_1, X_2, \ldots\}$, $\mathbb{U}_n^\sharp$ converges weakly to $\mathbb{W}_P$ in probability in the sense that

$$d_{BL|X_1^\infty}(\mathbb{U}_n^\sharp, \mathbb{W}_P) := \sup_{f \in BL_1} |\mathbb{E}_{|X_1^\infty}[f(\mathbb{U}_n^\sharp)] - \mathbb{E}[f(\mathbb{W}_P)]|$$

converges to zero in outer probability under regularity conditions ($\mathbb{E}_{|X_1^\infty}$ denotes the conditional expectation given X_1^∞). Since the map $(\xi_1, \ldots, \xi_n) \mapsto n^{-1/2} \sum_{i=1}^n \xi_i [U_{n-1,-i}^{(r-1)}(\delta_{X_i} \cdot) - U_n(\cdot)]$ is continuous from $\mathbb{R}^n$ into $\ell^\infty(\mathcal{H})$, the multiplier process $\mathbb{U}_n^\sharp$ induces a Borel measurable map into $\ell^\infty(\mathcal{H})$ for fixed X_1^∞. For an arbitrary map $Y : \Omega \to \mathbb{R}$, let Y^* denote the measurable cover [53, lemma 1.2.1].

Theorem D.1 (Conditional UCLT for JMB) *Let $\mathcal{H}$ be a fixed pointwise measurable class of symmetric measurable functions on S^r with symmetric envelope $H \in L^2(P^r)$ such that $\int_0^1 \sqrt{\lambda(\varepsilon)} d\varepsilon < \infty$ with $\lambda(\varepsilon) = \sup_Q \log N(\mathcal{H}, \|\cdot\|_{Q,2}, \varepsilon \|H\|_{Q,2})$. Then $\mathcal{G} = P^{r-1}\mathcal{H} = \{P^{r-1}h : h \in \mathcal{H}\}$ is P-pre-Gaussian, $d_{BL}(\mathbb{U}_n/r, \mathbb{W}_P) \to 0$, and $d_{BL|X_1^\infty}(\mathbb{U}_n^\sharp, \mathbb{W}_P)^* \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.*

Theorem D.1 should be compared with Theorem 2.1 in [5] that establishes a conditional UCLT for the empirical bootstrap for a non-degenerate U-process under the same metric entropy condition. Interestingly, however, our moment condition on the envelope H is weaker than their condition (2.3), which, if $r = 2$, requires $\mathbb{E}[H(X_1, X_1)] < \infty$ in addition to $\mathbb{E}[H^2(X_1, X_2)] < \infty$. This comes from the difference in how to estimate the Hajék projection; our JMB estimates the Hajék projection by a jackknife U-statistic, while the empirical bootstrap estimates it by a V-statistic (see Remark 3.1).

If we are interested in $\sup_{h \in \mathcal{H}} \mathbb{U}_n(h)/r$, then the result of Theorem D.1 implies that

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\sup_{h \in \mathcal{H}} \mathbb{U}_n(h)/r \leqslant t \right) - \mathbb{P}\left(\sup_{g \in \mathcal{G}} W_P(g) \leqslant t \right) \right| \to 0 \quad \text{and}$$

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}_{|X_1^\infty}\left(\sup_{h \in \mathcal{H}} \mathbb{U}_n^\sharp(h) \leqslant t \right) - \mathbb{P}\left(\sup_{g \in \mathcal{G}} W_P(g) \leqslant t \right) \right| \xrightarrow{\mathbb{P}} 0$$

as long as the distribution function of $\sup_{g \in \mathcal{G}} W_P(g)$ is continuous, which is true if $\inf_{g \in \mathcal{G}} \mathrm{Var}_P(g) > 0$ (cf. Lemma A.1). When the function class $\mathcal{H}$ is centrally symmetric (i.e., $-h \in \mathcal{H}$ whenever $h \in \mathcal{H}$) so that $\sup_{h \in \mathcal{H}} \mathbb{U}_n(h) = \|\mathbb{U}_n\|_{\mathcal{H}}$, $\sup_{g \in \mathcal{G}} W_P(g) = \|W_P\|_{\mathcal{G}}$, and $\sup_{h \in \mathcal{H}} \mathbb{U}_n^\sharp(h) = \|\mathbb{U}_n^\sharp\|_{\mathcal{H}}$, then the distribution function of $\|W_P\|_{\mathcal{G}}$ is continuous under a much less restrictive assumption that $\mathrm{Var}_P(g) > 0$ for some $g \in \mathcal{G}$. Indeed, from Theorem 11.1 in [17], the distribution of $\|W_P\|_{\mathcal{G}}$ is (absolutely) continuous on (ℓ_0, ∞) with $\ell_0 \geqslant 0$ being the left endpoint of the support of $\|W_P\|_{\mathcal{G}}$, but from [37, p. 57–58], $\ell_0 = 0$. This implies that, unless $\|W_P\|_{\mathcal{G}} = 0$ almost surely, the distribution function of $\|W_P\|_{\mathcal{G}}$ does not have a jump at $\ell_0 = 0$ (as $\mathbb{P}(\|W_P\|_{\mathcal{G}} = 0) = 0$) and so is everywhere continuous on $\mathbb{R}$.

Proof of Theorem D.1 The first two results are essentially implied by the proof of Theorem 4.9 in [4] but we include their proofs for completeness. By changing H to $H \vee 1$ if necessary, we may assume $\|G\|_{P,2} > 0$ (recall $G = P^{r-1}H$), which implies $\|H\|_{P,2} > 0$. By Jensen's inequality, $\|P^{r-1}h\|_{P,2} \leqslant \|h\|_{P^r,2}$ and so we have

$$N(\mathcal{G}, \|\cdot\|_{P,2}, \tau \|H\|_{P^r,2}) \leqslant N(\mathcal{H}, \|\cdot\|_{P^r,2}, \tau \|H\|_{P^r,2}).$$

The right hand side is bounded by $\sup_Q N(\mathcal{H}, \|\cdot\|_{Q,2}, \tau\|H\|_{Q,2}/4)$ by Lemma A.2. Conclude that

$$\int_0^1 \sqrt{\log N(\mathcal{G}, \|\cdot\|_{P,2}, \tau\|H\|_{P^r,2})}d\tau < \infty,$$

which implies by Dudley's criterion for sample continuity that $\mathcal{G}$ is P-pre-Gaussian (to be precise we have to verify $\int_0^1 \sqrt{\log N(\{g - Pg : g \in \mathcal{G}\}, \|\cdot\|_{P,2}, \tau)}d\tau < \infty$ but this is immediate). The convergence of marginals of $\mathbb{U}_n/r$ to $\mathbb{W}_P$ follows from the multidimensional CLT for U-statistics. To conclude $d_{BL}(\mathbb{U}_n/r, \mathbb{W}_P) \to 0$, it suffices to show the asymptotic equicontinuity condition

$$\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|h-h'\|_{P^r,2}<\delta\|H\|_{P^r,2}}|\mathbb{U}_n(h-h')| > \eta\right) = 0 \tag{45}$$

holds for every $\eta > 0$. We defer the proof of (45) after the proof of the theorem.

To prove the last result of the theorem, let $e_P(h, h') = \|P^{r-1}(h - h')\|_{P,2}$ and for given $\delta > 0$ let $\{h_1, \ldots, h_{N(\delta)}\}$ be a $(\delta\|G\|_{P,2})$-net of $(\mathcal{H}, e_P)$. Let $\pi_\delta : \mathcal{H} \to \{h_1, \ldots, h_{N(\delta)}\}$ be a map such that for each $h \in \mathcal{H}$, $e_P(h, \pi_\delta(h)) \leqslant \delta\|G\|_{P,2}$. Define $\mathbb{U}_{n,\delta}^\sharp := \mathbb{U}_n^\sharp \circ \pi_\delta$ and $\mathbb{W}_{P,\delta} := \mathbb{W}_P \circ \pi_\delta$. For any $f \in BL_1$, we have

$$\begin{aligned}|\mathbb{E}_{|X_1^\infty}[f(\mathbb{U}_n^\sharp)] - \mathbb{E}[f(\mathbb{W}_P)]| \leqslant\ & |\mathbb{E}_{|X_1^\infty}[f(\mathbb{U}_n^\sharp)] - \mathbb{E}_{|X_1^\infty}[f(\mathbb{U}_{n,\delta}^\sharp)]| \\ &+ |\mathbb{E}_{|X_1^\infty}[f(\mathbb{U}_{n,\delta}^\sharp)] - \mathbb{E}[f(\mathbb{W}_{P,\delta})]| \\ &+ |\mathbb{E}[f(\mathbb{W}_{P,\delta})] - \mathbb{E}[f(\mathbb{W}_P)]|.\end{aligned} \tag{46}$$

The third term on the right hand side of (46) is bounded by $\mathbb{E}[2 \wedge \|\mathbb{W}_{P,\delta} - \mathbb{W}_P\|_{\mathcal{H}}]$ and by construction $\mathbb{W}_P$ has sample paths almost surely uniformly e_P-continuous, so that $\mathbb{E}[2 \wedge \|\mathbb{W}_{P,\delta} - \mathbb{W}_P\|_{\mathcal{H}}] \to 0$ as $\delta \downarrow 0$ by the dominated convergence theorem. Since $\mathbb{U}_{n,\delta}^\sharp$ can be identified with a Gaussian vector of dimension $N(\delta)$ conditionally on X_1^∞, by Lemma 3.7.46 in [29], the second term on the right hand side of (46) is bounded by

$$c(\delta)\max_{1\leqslant j,k\leqslant N(\delta)}|\widehat{C}_{j,k} - \mathrm{Cov}_P(P^{r-1}h_j, P^{r-1}h_k)|^{1/3}$$

for some constant $c(\delta)$ that depends only on δ, where

$$\widehat{C}_{j,k} = n^{-1}\sum_{i=1}^n\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_j) - U_n(h_j)\}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h_k) - U_n(h_k)\}.$$

From Step 5 of the proof of Theorem 3.1 and using the notation in the proof, we have

$$\max_{1\leqslant j,k\leqslant N(\delta)}|\widehat{C}_{j,k} - \mathrm{Cov}_P(P^{r-1}h_j, P^{r-1}h_k)|$$

$$\leqslant 2\Upsilon_n + 2\|G\|_{P,2}\Upsilon_n^{1/2} + 2n^{-1/2}\|\mathbb{G}_n\|_{\check{\mathcal{G}}\cdot\check{\mathcal{G}}} + \|U_n(h) - P^r h\|_{\mathcal{H}}^2.$$

From the UCLT for the U-process established in the first paragraph, the last term on the right hand side is $o_{\mathbb{P}}(1)$. The function class $\check{\mathcal{G}}\cdot\check{\mathcal{G}}$ is weak P-Glivenko-Cantelli by Lemmas A.3 and A.5 together with Theorem 2.4.3 in [53], which implies that $n^{-1/2}\|\mathbb{G}_n\|_{\check{\mathcal{G}}\cdot\check{\mathcal{G}}} = o_{\mathbb{P}}(1)$. From Lemma D.3 below, we also have $\Upsilon_n = o_{\mathbb{P}}(1)$.

Finally, the first term on the right hand side of (46) is bounded by

$$\varepsilon + 2\mathbb{P}_{|X_1^\infty}(\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\delta} > \varepsilon)$$

for any $\varepsilon > 0$, where $\mathcal{H}_\delta = \{h - h' : h, h' \in \mathcal{H}, e_P(h, h') < 2\delta\|G\|_{P,2}\}$. Let $\Sigma_{n,\delta} := \|n^{-1}\sum_{i=1}^n\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - U_n(h)\}^2\|_{\mathcal{H}_\delta}$. By Markov's inequality,

$$\mathbb{P}_{|X_1^\infty}(\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\delta} > \varepsilon) \leqslant \frac{\mathbb{E}_{|X_1^\infty}[\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\delta}]}{\varepsilon}.$$

From Step 5 of the proof of Theorem 3.1,

$$N(\mathcal{H}_\delta, d, 2\tau\|H\|_{\mathbb{P}_{I_{n,r},2}}) \leqslant N^2(\mathcal{H}, \|\cdot\|_{\mathbb{P}_{I_{n,r},2}}, \tau\|H\|_{\mathbb{P}_{I_{n,r},2}})$$

with $d(h, h') = \{\mathbb{E}_{|X_1^\infty}[\{\mathbb{U}_n^\sharp(h) - \mathbb{U}_n^\sharp(h')\}^2]\}^{1/2}$. Hence by Dudley's entropy integral bound, we have

$$\mathbb{E}_{|X_1^\infty}[\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\delta}] \lesssim \int_0^{\Sigma_{n,\delta}^{1/2}} \sqrt{1 + \lambda(\tau/\|H\|_{\mathbb{P}_{I_{n,r},2}})}d\tau$$

up to a constant independent of n and δ, and $\|H\|_{\mathbb{P}_{I_{n,r},2}}^2 = |I_{n,r}|^{-1}\sum_{I_{n,r}} H^2(X_{i_1}, \ldots, X_{i_r}) = \|H\|_{P^r,2}^2 + o_{\mathbb{P}}(1)$ by the law of large numbers for U-statistics [18, Theorem 4.1.4]. From Step 4 of the proof of Theorem 3.1,

$$\Sigma_{n,\delta} \leqslant 8(\delta\|G\|_{P,2})^2 + 8n^{-1/2}\|\mathbb{G}_n\|_{\check{\mathcal{G}}\cdot\check{\mathcal{G}}} + 8\Upsilon_n,$$

and the last two terms on the right hand side are $o_{\mathbb{P}}(1)$ while the first term can be arbitrarily small by taking δ sufficiently small. This implies that for any $\eta > 0$,

$$\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\mathbb{P}\left(\mathbb{P}_{|X_1^\infty}(\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_\delta} > \varepsilon) > \eta\right) = 0.$$

Putting everything together, we conclude $d_{BL|X_1^\infty}(\mathbb{U}_n^\sharp, \mathbb{W}_P)^* \xrightarrow{\mathbb{P}} 0$, completing the proof. □

Lemma D.2 *Under the assumption of Theorem D.1, the asymptotic equicontinuity condition (45) holds.*

Proof of Lemma D.2 For $\delta \in (0,1]$, let $\mathcal{H}'_\delta = \{h - h' : \|h - h'\|_{P^r,2} < \delta \|H\|_{P^r,2}\}$. By Markov's inequality, it suffices to show that

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbb{E}[\|\mathbb{U}_n\|_{\mathcal{H}'_\delta}] = 0.$$

We use Hoeffding's averaging [49, Section 5.1.6] to bound the expectation. Let

$$S_f(x_1, \ldots, x_n) = \frac{1}{m} \sum_{i=1}^{m} f(x_{(i-1)r+1}, \ldots, x_{ir}) \text{ with } m = \lfloor n/r \rfloor.$$

Then we have

$$U_n(h) = \frac{1}{n!} \sum_{j_1, \ldots, j_n} S_h(X_{j_1}, \ldots, X_{j_n}),$$

where $\sum_{j_1,\ldots,j_n}$ are taken over all permutations $j_1, \ldots, j_n$ of $1, \ldots, n$. By Jensen's inequality, $\mathbb{E}[\|\mathbb{U}_n\|_{\mathcal{H}'_\delta}]$ is bounded by $\sqrt{n}\mathbb{E}[\|S_h(X_1, \ldots, X_n) - P^r h\|_{\mathcal{H}'_\delta}]$. Since

$$S_h(X_1, \ldots, X_n) - P^r h = \frac{1}{m} \sum_{i=1}^{m} (h(X_{(i-1)r+1}, \ldots, X_{ir}) - P^r h)$$

and since $(X_{(i-1)r+1}, \ldots, X_{ir})$, $i = 1, \ldots, m$ are i.i.d., we can apply Theorem 5.2 in [14] to conclude that

$$\mathbb{E}[\|\mathbb{U}_n\|_{\mathcal{H}'_\delta}] \lesssim \|H\|_{P^r,2} J(\delta, \mathcal{H}'_\delta, 2H) + \frac{\|M_r\|_{\mathbb{P},2} J^2(\delta, \mathcal{H}'_\delta, 2H)}{\delta^2 \sqrt{m}}$$

up to a constant that depends only on r, where $M_r = \max_{1 \leqslant i \leqslant m} H(X_{(i-1)r+1}, \ldots, X_{ir})$ and the J function is defined in [14]. From a standard calculation, $J(\delta, \mathcal{H}'_\delta, 2H) \lesssim J(\delta, \mathcal{H}, H) = \int_0^\delta \sqrt{1 + \lambda(\tau)} d\tau$ up to a universal constant and $\|M_r\|_{\mathbb{P},2} = o(\sqrt{m})$ by $H \in L^2(P^r)$ [53, Problem 2.3.4]. Hence we conclude

$$\limsup_{n \to \infty} \mathbb{E}[\|\mathbb{U}_n\|_{\mathcal{H}'_\delta}] \lesssim \|H\|_{P^r,2} J(\delta, \mathcal{H}, H)$$

up to a constant that depends only on r, and by the dominated convergence theorem the right hand side is $o(1)$ as $\delta \downarrow 0$. This completes the proof. □

Lemma D.3 *Under the assumption of Theorem D.1, we have* $\mathbb{E}[\Upsilon_n] = O(n^{-1})$ *where* Υ_n *is defined in (31).*

Proof of Lemma D.3 We begin with noting that

$$\mathbb{E}[\Upsilon_n] \leqslant \mathbb{E}\left[\mathbb{E}\left[\left\|U^{(r-1)}_{n-1,-n}(\delta_{X_n} h) - P^{r-1}(\delta_{X_n} h)\right\|^2_{\mathcal{H}} \,\middle|\, X_n\right]\right].$$

By Hoeffding's averaging [49, Section 5.1.6],

$$U_{n-1,-n}^{(r-1)}(f) = \frac{1}{(n-1)!} \sum_{j_1,\ldots,j_{n-1}} T_f(X_{j_1}, \ldots, X_{j_{n-1}}),$$

where $\sum_{j_1,\ldots,j_{n-1}}$ is taken over all permutations $j_1, \ldots, j_{n-1}$ of $1, \ldots, n-1$, and

$$T_f(x_1, \ldots, x_{n-1}) = \frac{1}{m} \sum_{i=1}^{m} f(x_{(i-1)(r-1)+1}, \ldots, x_{i(r-1)}) \text{ with } m = \lfloor (n-1)/(r-1) \rfloor.$$

By Jensen's inequality,

$$\begin{aligned} &\mathbb{E}\left[\left\| U_{n-1,-n}^{(r-1)}(\delta_{X_n} h) - P^{r-1}(\delta_{X_n} h) \right\|_{\mathcal{H}}^2 \,\middle|\, X_n \right] \\ &\leqslant \mathbb{E}\left[\left\| T_{\delta_{X_n} h}(X_1, \ldots, X_{n-1}) - P^{r-1}(\delta_{X_n} h) \right\|_{\mathcal{H}}^2 \,\middle|\, X_n \right]. \end{aligned}$$

By Corollary A.4 and the condition of Theorem D.1, for given $x \in S$,

$$\int_0^1 \sqrt{\sup_Q \log N(\delta_x \mathcal{H}, \|\cdot\|_{Q,2}, \tau \|\delta_x H\|_{Q,2})} \leqslant \int_0^1 \sqrt{\lambda(\tau)} d\tau < \infty.$$

Hence, applying Theorem 2.14.1 in [53] conditionally on X_n, we have

$$\mathbb{E}\left[\left\| T_{\delta_{X_n} h}(X_1, \ldots, X_{n-1}) - P^{r-1}(\delta_{X_n} h) \right\|_{\mathcal{H}}^2 \,\middle|\, X_n \right] \lesssim n^{-1} \|\delta_{X_n} H\|_{P^{r-1},2}^2$$

up to a constant independent of n. Since $\mathbb{E}[\|\delta_{X_n} H\|_{P^{r-1},2}^2] = \|H\|_{P^r,2}^2$, we obtain the desired conclusion by Fubini's theorem. □

Appendix E. Gaussian approximation for suprema of *U*-processes indexed by general function classes

In this section we derive Gaussian approximation error bounds for the U-process supremum indexed by general function classes. We obey the notation used in Sects. 2, 3 and 5. We make the following assumptions on the function class $\mathcal{H}$ and the distribution P.

(A1) The function class $\mathcal{H}$ is pointwise measurable.
(A2) The envelope H satisfies that $H \in L^3(P^r)$.
(A3) The class $\mathcal{G} = P^{r-1}\mathcal{H} = \{P^{r-1}h : h \in \mathcal{H}\}$ is P-pre-Gaussian, i.e., there exists a tight Gaussian random variable W_p in $\ell^\infty(\mathcal{G})$ with mean zero and covariance function $\mathbb{E}[W_P(g) W_P(g')] = \mathrm{Cov}(g(X_1), g'(X_1))$ for all $g, g' \in \mathcal{G}$.

Conditions (A1)–(A3) are parallel with the corresponding conditions in [14]. Condition (A1) is the same as Condition (PM) in Sect. 2. Condition (A3) is a high-level assumption that is implied by Condition (VC) in Sect. 2.

For $\varepsilon > 0$, define $\mathcal{N}_n(\varepsilon) = \log(N(\mathcal{G}, \|\cdot\|_{P,2}, \varepsilon\|G\|_{P,2}) \vee n)$ with $G = P^{r-1}H$. Under Condition (A3), $\mathcal{G}$ is totally bounded for the intrinsic pseudometric induced by $\|\cdot\|_{P,2}$ and $\mathcal{N}_n(\varepsilon)$ is finite for every $\varepsilon \in (0, 1]$. In addition, the Gaussian process W_P extends to the linear hull of $\mathcal{G}$ in such a way that W_P has linear sample paths (see e.g., Theorem 3.7.28 in [29]). For $\varepsilon \in (0, 1]$, $\gamma \in (0, 1)$, and $\kappa > 0$, define

$$\begin{aligned}\Delta_n(\varepsilon, \gamma, \kappa) :=& \gamma^{-1}\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{G}_\varepsilon}] + \mathbb{E}[\|W_P\|_{\mathcal{G}_\varepsilon}] \\ &+ \sqrt{\log(1/\gamma)}\varepsilon\|G\|_{P,2} + n^{-1/6}\gamma^{-1/3}\kappa\mathcal{N}_n^{2/3}(\varepsilon) \\ &+ n^{-1/4}\gamma^{-1/2}(\mathbb{E}\|\mathbb{G}_n\|_{\breve{\mathcal{G}}\cdot\breve{\mathcal{G}}})^{1/2}\mathcal{N}_n^{1/2}(\varepsilon) \\ &+ n^{1/2}\gamma^{-1}\sum_{k=2}^{r}\mathbb{E}[\|U_n^{(k)}(\pi_k h)\|_{\mathcal{H}}], \\ \delta_n(\varepsilon, \gamma, \kappa) :=& \frac{1}{5}P\left[(\breve{G}/\kappa)^3 1(\breve{G}/\kappa > c\gamma^{-1/3}n^{1/3}\mathcal{N}_n(\varepsilon)^{-1/3})\right],\end{aligned}$$

where $\mathcal{G}_\varepsilon = \{g - g' : g, g' \in \mathcal{G}, \|g - g'\|_{P,2} < 2\varepsilon\|G\|_{P,2}\}$, $\breve{\mathcal{G}}\cdot\breve{\mathcal{G}} = \{gg' : g, g' \in \breve{\mathcal{G}}\}$, $\breve{\mathcal{G}} = \{g, g - Pg : g \in \mathcal{G}\}$, and $\breve{G} = G + PG$. Here $c > 0$ is some universal constant. Below is an abstract (yet general) version of the Gaussian coupling bound.

Proposition E.1 (Abstract Gaussian coupling bound) *Let $Z_n = \sup_{h\in\mathcal{H}}\mathbb{U}_n(h)/r$. Suppose that Conditions (A1)–(A3) hold. Let $\kappa > 0$ be any positive constant such that $\kappa^3 \geqslant \mathbb{E}[\|n^{-1}\sum_{i=1}^n |g(X_i) - Pg|^3\|_{\mathcal{G}}]$. Then, for every $n \geqslant r + 1$, $\varepsilon \in (0, 1]$, and $\gamma \in (0, 1)$, one can construct a random variable $\widetilde{Z}_n = \widetilde{Z}_{n,\varepsilon,\gamma,\kappa}$ such that $\mathcal{L}(\widetilde{Z}_n) = \mathcal{L}(\sup_{g\in\mathcal{G}} W_P(g))$ and*

$$\mathbb{P}\left(|Z_n - \widetilde{Z}_n| > C_1\Delta_n(\varepsilon, \gamma, \kappa)\right) \leqslant \gamma\{1 + \delta_n(\varepsilon, \gamma, \kappa)\} + \frac{C_2\log n}{n},$$

where $C_1 = C_{1,r}$ is a constant depending only on r and C_2 is a universal constant.

The proposition should be considered as an extension of Theorem 2.1 in [14] to the U-process. To apply the above proposition, we need to derive bounds on

$$\begin{aligned}&\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{G}_\varepsilon}],\ \mathbb{E}[\|W_P\|_{\mathcal{G}_\varepsilon}],\ \mathbb{E}\left[\left\|n^{-1}\sum_{i=1}^n |g(X_i) - Pg|^3\right\|_{\mathcal{G}}\right], \\ &\mathbb{E}[\|\mathbb{G}_n\|_{\breve{\mathcal{G}}\cdot\breve{\mathcal{G}}}],\ \text{and } \mathbb{E}[\|U_n^{(k)}(\pi_k h)\|_{\mathcal{H}}, k = 2, \ldots, r,\end{aligned} \tag{47}$$

which can be derived under some moment conditions on H and by using the uniform entropy integrals $J_k(\delta), k = 1, \ldots, r$ defined in (19) (cf. Lemma 2.2 in [14] and our Theorem 5.1), where the latter can be simplified in terms of the VC characteristics (A, v) for a VC type function class (cf. the proof of Corollary 5.3).

Proof of Proposition E.1 The proof is based on a modification to that of Theorem 2.1 in [14]. In this proof C denotes a generic universal constant; the value of C may change from place to place. Let $\{g_k\}_{k=1}^N$ be a minimal $\varepsilon\|G\|_{P,2}$-net of $(\mathcal{G}, \|\cdot\|_{P,2})$ with $N := N(\mathcal{G}, \|\cdot\|_{P,2}, \varepsilon\|G\|_{P,2})$. By the definition of $\mathcal{G}$, each g_k corresponds to a kernel $h_k \in \mathcal{H}$ such that $g_k = P^{r-1}h_k$. Recall the Hoeffding decomposition $\mathbb{U}_n(h) = r\mathbb{G}_n(P^{r-1}h) + \sqrt{n}\sum_{k=2}^r \binom{r}{k} U_n^{(k)}(\pi_k h)$, where $\mathbb{G}_n(P^{r-1}h) = n^{-1/2}\sum_{i=1}^n (P^{r-1}h(X_i) - P^r h)$. Let $L_n = \sup_{g\in\mathcal{G}} \mathbb{G}_n(g)$ and $R_n = \|r^{-1}\sqrt{n}\sum_{k=2}^r \binom{r}{k} U_n^{(k)}(\pi_k h)\|_{\mathcal{H}}$. Then $|Z_n - L_n| \leqslant R_n$. Define

$$L_n^\varepsilon = \max_{1\leqslant j\leqslant N} \mathbb{G}_n(g_j),\ \widetilde{Z} = \sup_{g\in\mathcal{G}} W_P(g),\ \widetilde{Z}^\varepsilon = \max_{1\leqslant j\leqslant N} W_P(g_j).$$

We note that $|L_n - L_n^\varepsilon| \leqslant \|\mathbb{G}_n\|_{\mathcal{G}_\varepsilon}$ and $|\widetilde{Z} - \widetilde{Z}^\varepsilon| \leqslant \|W_P\|_{\mathcal{G}_\varepsilon}$. By Corollary 4.1 in [14], we have for every $B \in \mathcal{B}(\mathbb{R})$ and $\delta > 0$,

$$\mathbb{P}(L_n^\varepsilon \in B) - \mathbb{P}(\widetilde{Z}^\varepsilon \in B^{16\delta}) \leqslant C\delta^{-2}\{T_1 + \delta^{-1}(T_2 + T_3)\mathcal{N}_n(\varepsilon)\}\mathcal{N}_n(\varepsilon) + Cn^{-1}\log n,$$

where

$$\begin{aligned}
T_1 =& n^{-1} \\
&\mathbb{E}\left[\max_{1\leqslant j,k\leqslant N}\left|\sum_{i=1}^n (g_j(X_i) - Pg_j)(g_k(X_i) - Pg_k) - P(g_j - Pg_j)(g_k - Pg_k)\right|\right], \\
T_2 =& n^{-3/2}\mathbb{E}\left[\max_{1\leqslant j\leqslant N}\sum_{i=1}^n |g_j(X_i) - Pg_j|^3\right], \\
T_3 =& n^{-1/2} \\
&\mathbb{E}\left[\max_{1\leqslant j\leqslant N} |g_j(X_1) - Pg_j|^3 \cdot 1\left(\max_{1\leqslant j\leqslant N} |g_j(X_1) - Pg_j| > \delta\sqrt{n}\mathcal{N}_n(\varepsilon)^{-1}\right)\right].
\end{aligned}$$

Observe that $T_1 \leqslant n^{-1/2}\mathbb{E}[\|\mathbb{G}_n\|_{\breve{\mathcal{G}}\cdot\breve{\mathcal{G}}}]$, $T_2 \leqslant n^{-1/2}\kappa^3$, and $T_3 \leqslant n^{-1/2}P[\breve{G}^3 1(\breve{G} > \delta\sqrt{n}\mathcal{N}_n(\varepsilon)^{-1})]$. Thus choosing

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

we have

$$\mathbb{P}(L_n^\varepsilon \in B) \leqslant \mathbb{P}(\widetilde{Z}^\varepsilon \in B^{16\delta}) + \frac{2\gamma}{5} + \frac{\gamma}{5}\kappa^{-3}P[\breve{G}^3 1(\breve{G} > \delta\sqrt{n}\mathcal{N}_n(\varepsilon)^{-1})] + \frac{C\log n}{n}.$$

Since $\delta \geqslant c\gamma^{-1/3}n^{-1/6}\kappa\mathcal{N}_n^{2/3}(\varepsilon)$, we have

$$P[\breve{G}^3 1(\breve{G} > \delta\sqrt{n}\mathcal{N}_n(\varepsilon)^{-1})] \leqslant P[\breve{G}^3 1(\breve{G}/\kappa > c\gamma^{-1/3}n^{1/3}\mathcal{N}_n(\varepsilon)^{-1/3})].$$

Conclude that with $\eta_n = (\gamma/5)P[(\check{G}/\kappa)^3 1(\check{G}/\kappa > c\gamma^{-1/3}n^{1/3}\mathcal{N}_n(\varepsilon)^{-1/3})]$,

$$\mathbb{P}(L_n^\varepsilon \in B) \leqslant \mathbb{P}(\widetilde{Z}^\varepsilon \in B^{16\delta}) + \frac{2\gamma}{5} + \eta_n + \frac{C\log n}{n}.$$

Next, we will bound $\|\mathbb{G}_n\|_{\mathcal{G}_\varepsilon}$ and $\|W_P\|_{\mathcal{G}_\varepsilon}$. By Markov's inequality, with probability at least $1-\gamma/5$,

$$\|\mathbb{G}_n\|_{\mathcal{G}_\varepsilon} \leqslant 5\gamma^{-1}\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{G}_\varepsilon}] =: a.$$

Further, by the Borell–Sudakov–Tsirel'son inequality (see Theorem 2.5.8 in [29]), with probability at least $1-\gamma/5$, we have

$$\|W_P\|_{\mathcal{G}_\varepsilon} \leqslant \mathbb{E}[\|W_P\|_{\mathcal{G}_\varepsilon}] + 2\varepsilon\|G\|_{P,2}\sqrt{2\log(5/\gamma)} =: b.$$

Therefore, for every $B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}\mathbb{P}(Z_n \in B) &\leqslant \mathbb{P}(L_n \in B^{5\gamma^{-1}\mathbb{E}[R_n]}) + \frac{\gamma}{5} \leqslant \mathbb{P}(L_n^\varepsilon \in B^{a+5\gamma^{-1}\mathbb{E}[R_n]}) + \frac{2\gamma}{5}\\ &\leqslant \mathbb{P}(\widetilde{Z}^\varepsilon \in B^{a+16\delta+5\gamma^{-1}\mathbb{E}[R_n]}) + \frac{4\gamma}{5} + \eta_n + \frac{C\log n}{n}\\ &\leqslant \mathbb{P}(\widetilde{Z} \in B^{a+b+16\delta+5\gamma^{-1}\mathbb{E}[R_n]}) + \gamma + \eta_n + \frac{C\log n}{n}.\end{aligned}$$

The conclusion of the proposition follows from the Strassen–Dudley theorem (see Theorem B.1). □

Appendix F. Alternative tests for concavity/convexity and monotonicity of regression functions

We will obey the setting of Example 4.2.

F.1. Alternative tests for concavity/convexity of regression function *f*

Instead of the original localized simplex statistic (11) proposed in [1], we may consider the following modified version:

$$\widetilde{U}_n(x) = \frac{1}{|I_{n,m+2}|}\sum_{(i_1,\ldots,i_{m+2})\in I_{n,m+2}} \widetilde{\varphi}(V_{i_1},\ldots,V_{i_{m+2}})\prod_{k=1}^{m+2} L_{b_n}(x - X_{i_k}),$$

where $\widetilde{\varphi}(v_1,\ldots,v_{m+2}) = 1\{(x_1,\ldots,x_{m+2}) \in \mathcal{D}\}w(v_1,\ldots,v_{m+2})$, and test concavity or convexity of f if the scaled supremum or infimum of $\widetilde{U}_n$ is large or small, respectively. These alternative tests will work without the symmetry assumption on the conditional distribution of ε, which is maintained in [1]. Our results below also cover these alternative tests.

F.2. Alternative tests for monotonicity of regression function f

Chetverikov [16] considers testing monotonicity of the regression function f without the assumption that the error term ε is independent of X. Chetverikov [16] studies, e.g., U-statistics given by replacing $\text{sign}(Y_j - Y_i)$ in (12) by $Y_j - Y_i$, and the test statistic defined by taking the maximum of such U-statistics over a discrete set of design points and bandwidths whose cardinality may grow with the sample size (indeed, the cardinality can be much larger than the sample size). His analysis is conditional on X_i's, and he cleverly avoids U-process machineries and applies directly high-dimensional Gaussian and bootstrap approximation theorems developed in [12]. It should be noted that [16] considers more general test statistics and studies multi-step procedures to improve on powers of his tests.

Another related test for regression monotonicity is based on the local linear rank statistics [21]. Let $R_{mk}(i) = \sum_{j=m+1}^{k} 1(Y_j \leqslant Y_i)$ be the local rank of Y_i among $Y_{m+1}, \ldots, Y_k$. In [21], Dümbgen considers a test for monotone trend of f (with fixed design points $X_1, \ldots, X_n$) via the local linear rank statistics

$$T_{mk} = \sum_{i=m+1}^{k} \beta\left(\frac{i-m}{k-m+1}\right) q\left(\frac{R_{mk}(i)}{k-m+1}\right), \quad 0 \leqslant m < k \leqslant n,$$

where β and q are functions on $(0, 1)$ such that: 1) $\beta(1-u) = -\beta(u)$ and $q(1-u) = -q(u)$ for $u \in (0, 1)$; 2) $\beta(\cdot)$ and $q(\cdot)$ are nondecreasing on $(0, 1)$. Then [21] proposes the multiscale test statistic

$$T = \max_{0 \leqslant m < k \leqslant n} (s_{k-m}|T_{mk}| - c_{k-m}),$$

where s_i and c_i are properly chosen nonnegative numbers. For the special case of the Wilcoxon score function $q(u) = 2u - 1$ and $\beta(u) = q(u)$, one can write

$$T_{mk} = \frac{2}{(k-m+1)^2} \sum_{m < i < j \leqslant k} (j-i)\text{sign}(Y_j - Y_i).$$

The statistic T_{mk} is related to our test statistic $\check{U}_n(x)$ with $L(u) = 1(u \in [-1, 1])$, namely T_{mk} and $\check{U}_n(x)$ are (local) U-statistics with kernels $(j-i)\text{sign}(Y_j - Y_i)$ and $\text{sign}(X_i - X_j)\text{sign}(Y_j - Y_i)$, respectively. Thus for a given sequence of bandwidths b_n, our monotonicity test based on the U-process $\check{U}_n(x)$ can be viewed as a single-scale test T_{mk} with $(k-m)/n = 2b_n$ in Dümbgen's sense. In particular, both T_{0n} and $\check{U}_n(x)$ with $b_n = 1$ quantify the monotonicity on the global scale. In addition, the "uniform-in-bandwidth" type results for our U-process approach in Sect. 4.1 can be viewed as the multiscale analog T of T_{mk} with the Wilcoxon score function. Nevertheless, since [21] considers the fixed design points, T_{mk} is a local U-statistic on Y_i's and $\check{U}_n(x)$ is a local U-statistic on (X_i, Y_i)'s. Our analysis (which requires a Lebesgue density on X) is not directly applicable for the local linear rank statistics of [21].

References

1. Abrevaya, J., Jiang, W.: A nonparametric approach to measuring and testing curvature. J. Bus. Econ. Stat. **23**(1), 1–19 (2005)
2. Adamczak, R.: Moment inequalities for U-statistics. Ann. Probab. **34**(6), 2288–2314 (2006)
3. Arcones, M., Giné, E.: On the bootstrap of U- and V-statistics. Ann. Stat. **20**(2), 655–674 (1992)
4. Arcones, M., Giné, E.: Limit theorems for U-processes. Ann. Probab. **21**(3), 1495–1542 (1993)
5. Arcones, M., Giné, E.: U-processes indexed by Vapnik–Červonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. Stoch. Process. Appl. **52**(1), 17–38 (1994)
6. Bickel, P.J., Freedman, D.A.: Some asymptotic theory for the bootstrap. Ann. Stat. **9**(6), 1196–1217 (1981)
7. Blundell, R., Gosling, A., Ichimura, H., Meghir, C.: Changes in the distribution of male and female wages accounting for employment composition using bounds. Econometrica **75**(2), 323–363 (2007)
8. Borovskikh, Y.V.: U-Statistics in Banach Spaces. V.S.P. Intl Science, Zeist (1996)
9. Bretagnolle, J.: Lois limits du Bootstrap de certaines functionnelles. Annales de l'Institut Henri Poincaré Section B **XIX**(3), 281–296 (1983)
10. Callaert, H., Veraverbeke, N.: The order of the normal approximation for a Studentized U-statistic. Ann. Stat. **9**(1), 360–375 (1981)
11. Chen, X.: Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. Ann. Stat. **46**(2), 642–678 (2018)
12. Chernozhukov, V., Chetverikov, D., Kato, K.: Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. Ann. Stat. **41**(6), 2786–2819 (2013)
13. Chernozhukov, V., Chetverikov, D., Kato, K.: Anti-concentration and honest, adaptive confidence bands. Ann. Stat. **42**(5), 1787–1818 (2014)
14. Chernozhukov, V., Chetverikov, D., Kato, K.: Gaussian approximation of suprema of empirical processes. Ann. Stat. **42**(4), 1564–1597 (2014)
15. Chernozhukov, V., Chetverikov, D., Kato, K.: Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. Stoch. Process. Appl. **126**(12), 3632–3651 (2016)
16. Chetverikov, D.: Testing regression monotonicity in econometric models. arXiv:1212.6757 (2012)
17. Davydov, Y., Lifshits, M., Smorodina, N.: Local Properties of Distributions of Stochastic Functions (Transaction of Mathematical Monographs, Vol. 173). American Mathematical Society, New York (1998)
18. de la Peña, V., Giné, E.: Decoupling: From Dependence to Independence. Springer, Berlin (1999)
19. Dehling, H., Mikosch, T.: Random quadratic forms and the bootstrap for U-statistics. J. Multivar. Anal. **51**(2), 392–413 (1994)
20. Dudley, R.M.: Real Analysis and Probability. Cambridge University Press, Cambridge (2002)
21. Dümbgen, L.: Application of local rank tests to nonparametric regression. J. Nonparametric Stat. **14**(5), 511–537 (2002)
22. Einmahl, U., Mason, D.M.: Uniform in bandwidth consistency of kernel-type function estimators. Ann. Stat. **33**(3), 1380–1403 (2005)
23. Ellison, G., Ellison, S.F.: Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration. Am. Econ. J. Microecon. **3**(1), 1–36 (2011)
24. Frees, E.W.: Estimating densities of functions of observations. J. Am. Stat. Assoc. **89**(426), 517–525 (1994)
25. Ghosal, S., Sen, A., van der Vaart, A.: Testing monotonicity of regression. Ann. Stat. **28**(4), 1054–1082 (2000)
26. Giné, E., Latała, R., Zinn, J.: Exponential and moment inequalities for U-statistics. High Dimensional Probability II. Springer, Berlin (2000)
27. Giné, E., Mason, D.M.: On local U-statistic processes and the estimation of densities of functions of several sample variables. Ann. Stat. **35**(3), 1105–1145 (2007)
28. Giné, E., Nickl, R.: Uniform limit theorems for wavelet density estimators. Ann. Probab. **37**(4), 1605–1646 (2009)
29. Giné, E., Nickl, R.: Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge University Press, Cambridge (2016)
30. Hall, P.: On convergence rates of suprema. Probab. Theory Relat. Fields **89**(4), 447–455 (1991)

31. Hoeffding, W.: A class of statistics with asymptotically normal distributions. Ann. Math. Stat. **19**(3), 293–325 (1948)
32. Huškova, M., Janssen, P.: Consistency of the generalized bootstrap for degenerate U-statistics. Ann. Stat. **21**(4), 1811–1823 (1993)
33. Hušková, M., Janssen, P.: Generalized bootstrap for studentized U-statistics: a rank statistic approach. Stat. Probab. Lett. **16**(3), 225–233 (1993)
34. Janssen, P.: Weighted bootstrapping of U-statistics. J. Stat. Plann. Inference **38**(1), 31–42 (1994)
35. Koltchinskii, V.I.: Komlos–Major–Tusnády approximation for the general empirical process and Haar expansions of classes of functions. J. Theor. Probab. **7**(1), 73–118 (1994)
36. Komlós, J., Major, P., Tusnády, G.: An approximation of partial sums of independent rv's and the sample df. I. Z. Wahrscheinlichkeitstheor. Verw. Geb. **32**(1–2), 111–131 (1975)
37. Ledoux, M., Talagrand, M.: Probability in Banach Spaces: Isoperimetry and Processes. Springer, New York (1991)
38. Lee, S., Linton, O., Whang, Y.-J.: Testing for stochastic monotonicity. Econometrica **77**(2), 585–602 (2009)
39. Albert, Y.L.: A large sample study of the Bayesian bootstrap. Ann. Stat. **15**(1), 360–375 (1987)
40. Mason, D.M., Newton, M.A.: A rank statistics approach to the consistency of a general bootstrap. Ann. Stat. **20**(3), 1611–1624 (1992)
41. Massart, P.: Strong approximation for multivariate empirical and related processes, via KMT constructions. Ann. Probab. **17**(1), 266–291 (1989)
42. Monrad, D., Philipp, W.: Nearby variables with nearby conditional laws and a strong approximation theorem for Hilbert space valued martingales. Probab. Theory Relat. Fields **88**(3), 381–404 (1991)
43. Nolan, D., Pollard, D.: U-processes: rates of convergence. Ann. Stat. **15**(2), 780–799 (1987)
44. Nolan, D., Pollard, D.: Functional limit theorems for U-processes. Ann. Probab. **16**(3), 1291–1298 (1988)
45. Piterberg, V.I.: Asymptotic Methods in the Theory of Gaussian Processes and Fields. American Mathematical Society, New York (1996)
46. Resnick, S.I.: Extreme Values, Regular Variation, and Point Processes. Springer, Berlin (1987)
47. Rio, E.: Local invariance principles and their application to density estimation. Probab. Theory Relat. Fields **98**(1), 21–45 (1994)
48. Rubin, D.B.: The Bayesian bootstrap. Ann. Stat. **9**(1), 130–134 (1981)
49. Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)
50. Sherman, R.P.: Limiting distribution of the maximal rank correlation estimator. Econometrica **61**(1), 123–137 (1993)
51. Sherman, R.P.: Maximal inequalities for degenerate U-processes with applications to optimization estimators. Ann. Stat. **22**(1), 439–459 (1994)
52. Solon, G.: Intergenerational income mobility in the United States. Am. Econ. Rev. **82**(3), 393–408 (1992)
53. van der Vaart, A., Wellner, J.A.: Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, Berlin (1996)
54. van der Vaart, A., Wellner, J.A.: A local maximal inequality under uniform entropy. Electron. J. Stat. **5**, 192–203 (2011)
55. Wang, Q., Jing, B.-Y.: Weighted bootstrap for U-statistics. J. Multivar. Anal. **91**(2), 177–198 (2004)
56. Zhang, D.: Bayesian bootstraps for U-processes, hypothesis tests and convergence of Dirichlet U-processes. Stat. Sin. **11**(2), 463–478 (2001)