Jackknife Multiplier Bootstrap: Finite Sample Approximations to the U-Process Suprem with Applications

By

Xiaohui Chen and Kengo Kato

April 2018

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION
DISCUSSION PAPER NO. 25

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION (CREPE)
THE UNIVERSITY OF TOKYO
http://www.crepe.e.u-tokyo.ac.jp/

JACKKNIFE MULTIPLIER BOOTSTRAP: FINITE SAMPLE APPROXIMATIONS TO THE U-PROCESS SUPREMUM WITH APPLICATIONS

XIAOHUI CHEN AND KENGO KATO

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 usable 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 contexts. We also discuss applications of the general approximation results to testing for qualitative features of nonparametric functions based on generalized local U-processes.

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 non-negative function on S^r such that

Date: First arXiv version: August 9, 2017. This version: August 14, 2017.

2010 *Mathematics Subject Classification.* 60F17, 62E17, 62F40, 62G10.

Key words and phrases. Gaussian approximation, multiplier bootstrap, jackknife, U-process, local U-process, local maximal inequality.

X. Chen is supported by NSF DMS-1404891 and UIUC Research Board Award RB17092. K. Kato is supported by Grant-in-Aid for Scientific Research (C) (15K03392) from the JSPS.

$H \geqslant \sup_{h\in\mathcal{H}} |h|$). 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.

A U-process is a collection of U-statistics indexed by a family of kernels. U-processes are powerful tools for a broad range of statistical applications such as testing for qualitative features of functions in nonparametric statistics [33, 22, 1], cross-validation for density estimation [38], and establishing limiting distributions of M-estimators [see, e.g., 3, 45, 46, 16]. 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 theory of empirical processes 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 [39, 3, 7, 16] (these references also cover degenerate U-processes where limiting processes are Gaussian chaoses rather than Gaussian processes). Under such classical settings, [4, 52] study limit theorems for bootstraps for U-processes; see also [5, 8, 2, 29, 28, 30, 50] as references on bootstraps for U-statistics. [23] introduce a notion of the local U-process, motivated by a density estimator of a function of several sample variables proposed by [21], and establish a version of UCLTs for local U-processes. More recently, [10] 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 [33, 22, 1]; see examples in Section 4 for details.

In this paper, we develop a general non-asymptotic theory for directly approximating the supremum Z_n 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 Section 2. Our Gaussian approximation result builds upon recent development in modern empirical process theory [13, 12, 14] and high-dimensional U-statistics [10]. As a significant departure from the existing literature [23, 3, 13, 14], our Gaussian approximation for U-processes has a multi-resolution nature, which neither parallelizes the theory of U-processes with fixed function classes nor that of empirical processes. In particular, unlike the U-processes with fixed function classes, the higher-order degenerate terms are not necessarily negligible compared with the Hájek projection (empirical) process (in the sense of the Hoeffding projections [27]) 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 Section 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 [12, 14] for empirical processes is not directly applicable to U-processes since the Hájek projection process also depends on P and hence it is unknown. Our second main contribution is to provide a fully data-dependent procedure for approximating the distribution of Z_n. Specifically, we propose a novel *jackknife multiplier bootstrap* (JMB) properly tailored to U-processes in Section 3. The key insight of the JMB is to replace the (unobserved) Hájek projection process associated with $\mathbb{U}_n$ by its jackknife estimate [cf. 9]. 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 Section 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, [22] 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 projection process and then apply Rio's coupling result [42], which is a Komlós-Major-Tusnády (KMT) [32] type strong approximation for empirical processes indexed by Vapnik-Červonenkis type classes of functions from an m-dimensional hyper-cube $[0,1]^m$ to $[-1,1]$ with bounded variations. See also [36, 31] 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 (see our examples in Section 4 for more discussions). Our regularity conditions on the function class and the underlying distribution to ensure validity of 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 particular, both Gaussian and bootstrap approximation results obtained in the present paper allow classes of functions with *unbounded* envelopes, provided suitable moment growth 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 [33, 22, 1]. In Section 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 [23]). 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 [26, 41]), 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 error in size of the JMB is decreasing polynomially fast in n. It is worth noting that [33], who develop a test for the conditional 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 error in size is decreasing polynomially fast in n [33, p.594]. The results of the present paper affirmatively answer 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 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 [22, 33, 1].

1.1. **Organization.** The rest of the paper is organized as follows. In Section 2, we derive non-asymptotic Gaussian approximation error bounds for the U-process supremum in the non-degenerate case. In Section 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 Section 4, we discuss applications of the general results developed in Sections 2 and 3 to testing for qualitative features of nonparametric functions based on generalized local U-processes. In Section 5, we prove new local maximal inequalities for U-processes that are key technical tools in the proofs for the results in the previous sections. In Section 6, we present the proofs for Sections 2–4. Appendix contains auxiliary technical results.

1.2. **Notation.** For a non-empty 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 pseudo-metric space (T, d), let $N(T, d, \varepsilon)$ denote the ε-covering number for (T, d) where $\varepsilon > 0$. See [48, Section 2.1] 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 well-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 Section 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. For our purpose, it is without loss of generality to assume that each $h \in \mathcal{H}$ is P^r-centered, i.e., $P^r h = \mathbb{E}[h(X_1, \dots, X_r)] = 0$. 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 Sections 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 $k = 1, \dots, r$, let $P^{r-k}h$ denote the function on S^k defined by

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

whenever the latter integral exists and is finite for every $(x_1, \dots, x_k) \in S^k$. 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 $k = 1, \dots, r$.

(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

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

such that the following hold:

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

where q appears in Condition (VC).

Some comments on the conditions are in order. Condition (PM) is made to avoid measurability complications. 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 [48, 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., [25, 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

$$\mathbb{E}[W_P(g)W_P(g')] = P(gg'),\ g, g' \in \mathcal{G}.$$

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 pseudo-metric $d_W(g, g') = (\mathbb{E}[(W(g) - W(g'))^2])^{1/2}, g, g' \in \mathcal{G}$, and W has sample paths almost surely uniformly d_W-continuous [48, 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 Sections 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, 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.

Condition (MT) assumes that $\inf_{g\in\mathcal{G}} \|g\|_{P,2} \geqslant \underline{\sigma}_{\mathfrak{g}} > 0$, which implies that the U-process is non-degenerate. In statistical applications, the function class $\mathcal{H}$ is often normalized such 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 $0 < c \leqslant \underline{\sigma}_{\mathfrak{g}} \leqslant \overline{\sigma}_{\mathfrak{g}} \leqslant C$ for some constants $0 < c < C$ independent of n; see Section 4 for details.

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 $X_1, \dots, X_n$, 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 commonly employed assumption is that the probability space is *rich enough*. For the sake of clarity, we will assume in Sections 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 $X_1, \dots, X_n$ are the coordinate projections of $(S^n, \mathcal{S}^n, P^n)$, multiplier random variables $\xi_1, \dots, \xi_n$ to be introduced in Section 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 $X_1, \dots, X_n$ and $\xi_1, \dots, \xi_n$, 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 $\gamma \in (0,1)$, there exists a random variable $\widetilde{Z}_n$ such that $\mathcal{L}(\widetilde{Z}_n) = \mathcal{L}(\sup_{g\in\mathcal{G}} W_P(g))$ and*

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

where $C, C' > 0$ 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.,

$$\rho(V, Y) := \sup_{t\in\mathbb{R}} |\mathbb{P}(V \leqslant t) - \mathbb{P}(Y \leqslant t)|.$$

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 all the conditions in Proposition 2.1. Then, there exists a constant $C > 0$ depending only on $r, \overline{\sigma}_{\mathfrak{g}}$ and $\underline{\sigma}_{\mathfrak{g}}$ such that*

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

In particular, if the function class $\mathcal{H}$ and the distribution P are independent of n, then

$$\rho(Z_n, \widetilde{Z}) = O(\{(\log n)^7/n\}^{1/8}).$$

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. (16) and (17) 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 Section 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 [13, 14].

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 [12] and [14], 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

$$\mathcal{G} \ni g \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \{g(X_i) - \overline{g}\}, \tag{5}$$

where $\xi_1, \dots, \xi_n$ are i.i.d. $N(0,1)$ random variables independent of the data $X_1^n := \{X_1, \dots, X_n\}$, 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 in the present paper. Let $\xi_1, \dots, \xi_n$ 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:

$$\mathbb{U}_n^\sharp(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \left[\frac{1}{|I_{n-1,r-1}|} \sum_{(i,i_2,\dots,i_r) \in I_{n,r}} h(X_i, X_{i_2}, \dots, X_{i_r}) - U_n(h) \right], \ h \in \mathcal{H}. \tag{6}$$

It is seen that $\{\mathbb{U}_n^\sharp(h) : h \in \mathcal{H}\}$ is a centered Gaussian process conditionally on the data X_1^n, and can be regarded as a version of the (infeasible) multiplier process (5) with each $g(X_i)$ replaced by a jackknife estimate. In fact, the multiplier process (5) can be alternatively represented as

$$\mathcal{H} \ni h \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \{(P^{r-1}h)(X_i) - \overline{P^{r-1}h}\}, \tag{7}$$

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 (6) 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 (6) replaces each $(P^{r-1}h)(X_i)$ in the infeasible multiplier process (7) 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 $\widetilde{Z}_n^\sharp$ (a real-valued random variable) 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. [13, 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*

$$\begin{gathered} \sigma_{\mathfrak{h}} K_n^{1/2} \leqslant \overline{\sigma}_{\mathfrak{g}} n^{1/2}, \ \nu_{\mathfrak{h}} K_n \leqslant \overline{\sigma}_{\mathfrak{g}} n^{3/4-1/q}, \ (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/2} K_n^{3/4} \leqslant \overline{\sigma}_{\mathfrak{g}} n^{3/4}, \\ b_{\mathfrak{h}} K_n^{3/2} \leqslant \overline{\sigma}_{\mathfrak{g}} n^{1-1/q}, \ \text{and } \chi_n \leqslant \overline{\sigma}_{\mathfrak{g}}. \end{gathered} \tag{8}$$

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

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

where $C, C' > 0$ are constants depending only on r, and

$$\begin{aligned} \varpi_n^\sharp := \varpi_n^\sharp(\gamma) := \frac{1}{\gamma^{3/2}} \Bigg\{ & \frac{\{(b_{\mathfrak{g}} \vee \sigma_{\mathfrak{h}})\overline{\sigma}_{\mathfrak{g}} K_n^{3/2}\}^{1/2}}{n^{1/4}} + \frac{b_{\mathfrak{g}} K_n}{n^{1/2-1/q}} + \frac{(\overline{\sigma}_{\mathfrak{g}} \nu_{\mathfrak{h}})^{1/2} K_n}{n^{3/8-1/(2q)}} \\ & + \frac{\overline{\sigma}_{\mathfrak{g}}^{1/2} (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/4} K_n^{7/8}}{n^{3/8}} + \frac{(\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{h}})^{1/2} K_n^{5/4}}{n^{1/2-1/(2q)}} + \overline{\sigma}_{\mathfrak{g}}^{1/2} \chi_n^{1/2} K_n^{1/2} \Bigg\}. \end{aligned} \tag{9}$$

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

It is not difficult to see 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,

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

The growth condition (8) is not serious restriction. 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, the growth condition (8) is a merely necessary condition for the coupling bound (9) to tend to zero.

The proof of Theorem 3.1 is lengthly 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 $C, c > 0$).

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

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

with the convention that $1/q = 0$ in the case of $q = \infty$. 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 (9). 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 complications and more restrictive moment conditions.

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

$$U_n^*(h) = \frac{1}{|I_{n,r}|} \sum_{(i_1,\ldots,i_r) \in I_{n,r}} \left\{ h(X_{i_1}^*, \ldots, X_{i_r}^*) - V_n(h) \right\}, \ h \in \mathcal{H},$$

where $X_1^*, \ldots, X_n^*$ are i.i.d. draws from the empirical distribution $n^{-1} \sum_{i=1}^n \delta_{X_i}$ 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 h (cf. [5, 2, 10]). Another example is the *randomly reweighted bootstrap*

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

where $w_1, \ldots, w_n$ is a sequence of random weights independent of $X_1^n = \{X_1, \ldots, X_n\}$ [29, 30, 17, 51]. The randomly reweighted bootstrap is a generalized bootstrap procedure, including: 1) the empirical bootstrap with multinomial weights; 2) the *Bayesian bootstrap* with $w_i = \eta_i / (n^{-1} \sum_{j=1}^n \eta_j)$ and $\eta_1, \ldots, \eta_n$ being i.i.d. exponential random variables with mean one (i.e., $(w_1, \ldots, w_n)$ follows a scaled Dirichlet distribution) [43, 34, 35, 52]. 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 $\{U_n^*(h) : h \in \mathcal{H}\}$ and the Bayesian bootstrap process

$\{U_n^\flat(h) : h \in \mathcal{H}\}$ (with Dirichlet weights) are known to have the same weak limit as the U-process $\{U_n(h) : h \in \mathcal{H}\}$, where the weak limit is the Gaussian process W_P in the non-degenerate case [4, 52]. The proposed multiplier process in (6) is closely connected to the empirical and 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. 10]. 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, [10] shows that the Gaussian multiplier, empirical and randomly reweighted bootstraps (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 $D_i = (X_i, V_i), i = 1, \dots, n$ 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

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

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

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

which we call, following [23], 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 conditional stochastic dominance)**.** 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 conditional stochastic dominance

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

Testing for the conditional stochastic dominance is an important topic in a variety of applied fields such as in economics [47, 6, 20]. For this problem, [33] consider a test for H_0 based on a local Kendall's tau statistic, inspired by [22]. Let $(X_i, Y_i), i = 1, \dots, n$ be i.i.d. copies of (X, Y). [33] consider the U-process

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

where $b_n \to 0$ is a sequence of bandwidths, and $\text{sign}(\cdot)$ is the sign function

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

They propose to reject the null hypothesis if

$$S_n = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{U_n(x,y)}{c_n(x)}$$

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. [33] 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 [40], but do not consider bootstrap approximations to S_n. It should be noted that [33] 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

$$Y = f(X) + \varepsilon, \ \mathbb{E}[\varepsilon \mid X] = 0,$$

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.

[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

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

It is not difficult to see that $\mathcal{D}_1, \ldots, \mathcal{D}_{m+2}$ are disjoint, and if, e.g., $(x_1, \ldots, x_{m+2}) \in \mathcal{D}_{m+2}$, then there exists a unique vector $(a_1, \ldots, a_{m+1}) \in \mathbb{R}^{m+1}$ with $0 < a_i < 1$ for all i and $\sum_{i=1}^{m+1} a_i = 1$ such that $x_{m+2} = \sum_{i=1}^{m+1} a_i x_i$.

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 $(a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_{m+2})$ 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 $(a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_{m+2})$ are in fact functions of x_i's. It is not difficult to see that $\mathcal{D}$ is symmetric (i.e., its indicator function is symmetric), and $w(v_1, \ldots, v_{m+2})$ is well-defined and symmetric in its arguments.

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

$$U_n(x) = \frac{1}{|I_{n,m+2}|} \sum_{(i_1, \ldots, i_{m+2}) \in I_{n,m+2}} \varphi(V_{i_1}, \ldots, V_{i_{m+2}}) \prod_{k=1}^{m+2} L_{b_n}(x - X_{i_k}), \tag{10}$$

where $\varphi(v_1, \ldots, v_{m+2}) = 1\{(x_1, \ldots, x_{m+2}) \in \mathcal{D}\}\mathrm{sign}(w(v_1, \ldots, v_{m+2}))$. It is seen that U_n is a U-process of order $(m+2)$. To test concavity and convexity of f, [1] propose to reject the hypotheses if

$$\overline{S}_n = \sup_{x \in \mathcal{X}} \frac{U_n(x)}{c_n(x)} \quad \text{and} \quad \underline{S}_n = \inf_{x \in \mathcal{X}} \frac{U_n(x)}{c_n(x)},$$

are large, 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 also be written as the supremum of a U-process by replacing φ by $-\varphi$, 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 $\{x_1, \ldots, x_G\}$. [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 [42] 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, [22] 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

$$S_n = \sup_{x \in \mathcal{X}} \frac{\check{U}_n(x)}{c_n(x)},$$

is large, where $\mathcal{X}$ is a subset of the support of X,

$$\check{U}_n(x) = \frac{1}{n(n-1)} \sum_{1 \leqslant i \neq j \leqslant n} \operatorname{sign}(Y_j - Y_i) \operatorname{sign}(X_i - X_j) L_{b_n}(x - X_i) L_{b_n}(x - X_j), \tag{11}$$

and $c_n(x) > 0$ is a suitable normalizing constant. [22] 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, [22] derive a Gumbel limiting distribution for a properly scaled version of S_n, but do not study bootstrap approximations to S_n.

Remark 4.1 (Alternative tests for concavity or convexity of f)**.** Instead of the original localized simplex statistic (10) 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 $\widetilde{\varphi}(v_1,\dots,v_{m+2}) = 1\{(x_1,\dots,x_{m+2}) \in \mathcal{D}\}w(v_1,\dots,v_{m+2})$, 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.

Remark 4.2 (Comments on [15])**.** [15] considers testing monotonicity of the regression function f without the assumption that the error term ε is independent of X. [15] studies, e.g., U-statistics given by replacing $\operatorname{sign}(Y_j - Y_i)$ in (11) 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 [11]. It should be noted that [15] considers more general test statistics and studies multi-step procedures to improve on powers of his tests.

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

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

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 $\operatorname{Cov}_P(\cdot,\cdot)$ and $\operatorname{Var}_P(\cdot)$ 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, [22, 33] 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 [42] when deriving limiting null distributions of their test statistics. Rio's coupling is a powerful KMT [32] 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. 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 [38, 24, 25]. 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 $A \leqslant C_1 \log n$ 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 [22, 33] where $m = 1$ and $q = \infty$. For example, [22] assume $nb_n^2/(\log n)^4 \to \infty$ and $b_n \log n \to 0$ for size control. 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 Section 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 $x - b_n x_{2:r} = (x - b_n x_2, \dots, x - b_n x_r)$. 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

$$\widehat{c}_n(\vartheta) = \sqrt{\frac{b_n^m}{n} \sum_{i=1}^n \left\{ U_{n-1,-i}^{(r-1)}(\delta_{D_i} h_{n,\vartheta}) - U_n(h_{n,\vartheta}) \right\}^2}, \ \vartheta \in \Theta. \tag{12}$$

The following lemma verifies that the jackknife estimate (12) 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 [22, 33, 1].

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 (12). 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 a 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

$$\mathbb{E}[W_{P,n}(\vartheta)W_{P,n}(\vartheta')] = b_n^m \mathrm{Cov}_P(P^{r-1}h_{n,\vartheta}, P^{r-1}h_{n,\vartheta'})/\{c_n(\vartheta)c_n(\vartheta')\} \tag{13}$$

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

$$\begin{aligned} &\sup_{t\in\mathbb{R}} \left|\mathbb{P}(\widehat{S}_n \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t)\right| \leqslant Cn^{-c}, \\ &\mathbb{P}\left\{\sup_{t\in\mathbb{R}} \left|\mathbb{P}_{|D_1^n}(\widehat{S}_n^\sharp \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t)\right| > Cn^{-c}\right\} \leqslant Cn^{-c}, \end{aligned} \tag{14}$$

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 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

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

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 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 [33] for testing conditional stochastic monotonicity and [19] 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 $\kappa_n = \bar{b}_n/\underline{b}_n$ 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 $2m(r-1)\bar{b}_n \leqslant \zeta/2$.

(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*

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

for $(\vartheta, b), (\vartheta', b') \in \Theta\times\mathcal{B}_n$, and the result (14) continues to hold with $\widetilde{S}_n := \sup_{(\vartheta,b)\in\Theta\times\mathcal{B}_n} W_{P,n}(\vartheta, b)$.

If $\underline{b}_n = \bar{b}_n = b_n$ (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). Note that such "uniform-in-bandwidth" type results are not covered in [22, 33, 1].

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 [16]. 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. For a symmetric measurable function $f : S^r \to \mathbb{R}$ and $k = 1, \ldots, r$, we define $P^{r-k}f : S^k \to \mathbb{R}$ by

$$P^{r-k}f(x_1, \ldots, x_k) = \int \cdots \int f(x_1, \ldots, x_k, x_{k+1}, \ldots, x_r)dP(x_{k+1}) \cdots dP(x_r),$$

where we assume that the integral exists and is finite for every $(x_1, \ldots, x_k) \in S^k$.

Definition 5.1 (Kernel degeneracy)**.** 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 $f : S^r \to \mathbb{R}$ ($\mathcal{F}$ need not be P^r-centered). 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

$$U_n^{(r)}(f) = \frac{1}{|I_{n,r}|} \sum_{(i_1,\dots,i_r)\in I_{n,r}} f(X_{i_1},\dots,X_{i_r}), \ f \in \mathcal{F}, \tag{15}$$

where $I_{n,r} = \{(i_1,\dots,i_r) : 1 \leqslant i_j \leqslant n, i_j \neq i_k \text{ if } 1 \leqslant j \neq k \leqslant r\}$ and $|I_{n,r}| = n!/(n-r)!$ denotes the cardinality of $I_{n,r}$. For each $k = 1,\dots,r$, the *Hoeffding projection* (with respect to P) is defined by

$$(\pi_k f)(x_1,\dots,x_k) := (\delta_{x_1} - P)\cdots(\delta_{x_k} - P)P^{r-k}f. \tag{16}$$

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

$$U_n^{(r)}(f) - P^r f = \sum_{k=1}^r \binom{r}{k} U_n^{(k)}(\pi_k f). \tag{17}$$

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 $\|PF^{r-k}\|_{P^k,2} > 0$ (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 $X_{(i-1)k+1}^{ik} = (X_{(i-1)k+1},\dots,X_{ik})$.

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

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

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. Note that $P^{r-k}F$ is an envelope for $P^{r-k}\mathcal{F}$. To avoid measurablity complications, we will assume that $\mathcal{F}$ is pointwise measurable. It is not difficult to see from the dominated convergence theorem that, 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 $k = 1,\dots,r$ are all pointwise measurable.

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 poinwise measurable and that $J_r(1) < \infty$. Let $\delta_k = \sigma_k/\|P^{r-k}F\|_{P^k,2}$ for $k = 1,\dots,r$. Then*

$$n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \lesssim J_k(\delta_k)\|P^{r-k}F\|_{P^k,2} + \frac{J_k^2(\delta_k)\|M_k\|_{\mathbb{P},2}}{\delta_k^2\sqrt{n}} \tag{18}$$

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

In view of Lemma A.2 (and approximating P^{r-k} by finitely discrete distributions), the assumption that $J_r(1) < \infty$ ensures that $J_k(1) < \infty$ (and hence $J_k(\delta) < \infty$ for every $\delta > 0$) for $k = 1, \ldots, r-1$. 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 $J_r(1) < \infty$. 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 [13, Lemma A.2], and hence omitted. $\square$

Proof of Theorem 5.1. It suffices to prove (18) when $\|P^{r-k}F\|_{P^k,2} > 0$ and $J_k(1) < \infty$, since otherwise there is nothing to prove (recall that we have assumed that $P^rF^2 < \infty$, 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. From the randomization theorem for U-processes [16, Theorem 3.5.3], we have that

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

where the second inequality follows from Jensen's inequality. 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 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 [16] yields that

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

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. So, Corollary 5.1.8 in [16] together with Fubini's theorem yield that

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

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 that

$$
n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \lesssim \|P^{r-k}F\|_{P^k,2} J_k(z), \quad \text{where } z := \sqrt{\mathbb{E}[\sigma_{I_{n,k}}^2]/\|P^{r-k}F\|_{P^k,2}^2}. \tag{19}
$$

Now, we shall bound $\mathbb{E}[\sigma_{I_{n,k}}^2]$. To this end, we will use Hoeffding's averaging [cf. 44, 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 $\|P^{r-k}f\|_{\mathbb{P}_{I_{n,k}},2}^2 = |I_{n,k}|^{-1}\sum_{I_{n,k}}(P^{r-k}f)^2(X_{i_1},\dots,X_{i_k})$ 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,

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

so that $z \leqslant \widetilde{z} := \sqrt{B_{n,k}/\|P^{r-k}F\|^2_{P^k,2}}$. Observe that, 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 [48, Lemma 2.3.1], (3) follows from the contraction principle [25, Corollary 3.2.2], and (4) follows from the Cauchy-Schwarz inequality. By the Hoffmann-Jørgensen inequality [48, Proposition A.1.6],

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

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 [13] (or repeating the first half of this proof with $r = k = 1$), we have that

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

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

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

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 [49, 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

$$J_k(z) \leqslant J_k(\widetilde{z}) \lesssim J_k(\Delta)\left[1 + J_k(\Delta)\frac{\|M_k\|_{\mathbb{P},2}}{\sqrt{n}\|P^{r-k}F\|_{P^k,2}\Delta^2}\right]. \tag{20}$$

Combining (19) and (20), we arrive at

$$n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \lesssim J_k(\Delta)\|P^{r-k}F\|_{P^k,2} + \frac{J_k^2(\Delta)\|M_k\|_{\mathbb{P},2}}{\sqrt{n}\Delta^2}. \tag{21}$$

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 (18) follows from (21). □

In the case where 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*

$$\begin{aligned} &n^{k/2}\mathbb{E}[\|U_n^{(k)}(\pi_k f)\|_{\mathcal{F}}] \\ &\quad\lesssim \sigma_k\left\{v\log(A\|P^{r-k}F\|_{P^k,2}/\sigma_k)\right\}^{k/2} + \frac{\|M_k\|_{\mathbb{P},2}}{\sqrt{n}}\left\{v\log(A\|P^{r-k}F\|_{P^k,2}/\sigma_k)\right\}^{k} \end{aligned} \tag{22}$$

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

Remark 5.1. (i). [23, Theorem 8] establishes a local maximal inequality for a U-process indexed by a VC type class with a bounded envelope. The bound in [23, Theorem 8] is uniform over all Hoeffding projection levels $k = 1, \ldots, r$. In contrast, our Corollary 5.3 is sharper than Theorem 8 in [23] in the sense that the bound in (22) is of the multi-resolution nature, which allows us to obtain better rates of convergence for kernel type statistics. In particular, σ_k^2 (or $\|M_k\|_{\mathbb{P},2}$) can be potentially much smaller than σ_r^2 (or $\|M_r\|_{\mathbb{P},2}$), which is indeed the case in the applications considered in Section 4. Furthermore, Corollary 5.3 allows the envelope F to be unbounded.

(ii). Theorem 5.1 and Corollary 5.3 generalize Theorem 5.2 and Corollary 5.1 in [13] to U-processes. In fact, Theorem 5.1 and Corollary 5.3 reduce to Theorem 5.2 and Corollary 5.1 in [13] 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 characteristic A, v, then for every $k = 1, \dots, r-1$, $P^{r-k}\mathcal{F}$ 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. Using an approximation argument [cf. 48, Problem 2.5.1], for every (not necessarily finitely discrete) probability measure R on $(S^r, \mathcal{S}^r)$ such that $RF^2 < \infty$, we have that

$$N(\mathcal{F}, \|\cdot\|_{R,2}, \tau\|F\|_{R,2}) \leqslant (4A/\tau)^v, \ 0 < \forall \tau \leqslant 1.$$

Hence, applying Lemma A.2 in Appendix A with $r = s = 2$, for every finitely discrete distribution Q on S^k, we have that

$$N(P^{r-k}\mathcal{F}, \|\cdot\|_{Q,2}, \tau\|P^{r-k}F\|_{Q,2}) \leqslant (16A/\tau^2)^v = (4\sqrt{A}/\tau)^{2v}, \ 0 < \forall \tau \leqslant 1.$$

This completes the proof. □

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 (22). □

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 $n^{k/2}\mathbb{E}[\|U_n^{(k)}(f)\|_{\mathcal{F}}]$ is

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

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. $\square$

If one is interested in bounding $\mathbb{E}[\|U_n^{(r)}(f) - P^r f\|_{\mathcal{F}}]$, then it suffices to apply (18) or (22) repeatedly for $k = 1, \ldots, r$. However, it is often the case that lower order Hoeffding projection terms are dominating, 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 $p \in [2, \infty)$. Suppose that $\mathcal{F}$ is pointwise measurable and that $J_r(1) < \infty$. 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 $\psi_{2/k}$-(quasi-)norm bounds the L^p-norm from above up to a constant that depends only on k (and hence r)

and p,

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

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 of 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. Now, 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 [48]. 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. There is an analogous Hoffmann-Jørgensen type inequality for U-processes [see 16, 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 Sections 2–4

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 Section 2.

We begin with stating the following lemma.

Lemma 6.1. *Work with the setup described in Section 2. Suppose that Conditions (PM), (VC), and (MT) hold. Let $L_n := \sup_{g\in\mathcal{G}} n^{-1/2}\sum_{i=1}^n g(X_i)$ 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*

$$\delta_n = \frac{(\overline{\sigma}_{\mathfrak{g}}^2 b_{\mathfrak{g}} K_n^2)^{1/3}}{\gamma^{1/3}n^{1/6}} + \frac{b_{\mathfrak{g}}K_n}{\gamma n^{1/2-1/q}}. \tag{23}$$

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 [14]. 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, we provide a sketch of the proof for Lemma 6.1, which points out required modifications to the proof of Theorem 2.1 in [14].

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 Section 5. Consider the Hoeffding decomposition for $U_n(h) = U_n^{(r)}(h)$: $U_n^{(r)}(h) = rU_n^{(1)}(P^{r-1}h) + \sum_{k=2}^r \binom{r}{k} U_n^{(k)}(\pi_k h)$, or

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

where $\mathbb{G}_n(P^{r-1}h) := n^{-1/2}\sum_{i=1}^n (P^{r-1}h)(X_i)$ 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 $B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
\mathbb{P}(Z_n \in B) &\leqslant \mathbb{P}(\{Z_n \in B\} \cap \{R_n \leqslant \gamma^{-1}\mathbb{E}[R_n]\}) + \mathbb{P}(R_n > \gamma^{-1}\mathbb{E}[R_n]) \\
&\leqslant \mathbb{P}(L_n \in B^{\gamma^{-1}\mathbb{E}[R_n]}) + \gamma \\
&\leqslant \mathbb{P}(\widetilde{Z} \in B^{C\delta_n + \gamma^{-1}\mathbb{E}[R_n]}) + C'(\gamma + n^{-1}),
\end{aligned} \tag{24}$$

where δ_n is given in (23).

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 $k = 3, \dots, r$. First, applying Corollary 5.5 to $\mathcal{F} = \mathcal{H}$ for $k = 2$ yields that

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

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

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

Therefore, we conclude that

$$\mathbb{E}[R_n] \leqslant C\sum_{k=2}^r n^{1/2}\mathbb{E}[\|U_n^{(k)}(\pi_k h)\|_{\mathcal{H}}] \leqslant C'\left(\sigma_{\mathfrak{h}} K_n n^{-1/2} + b_{\mathfrak{h}} K_n^2 n^{-1+1/q} + \chi_n\right). \tag{25}$$

Combining (24) with (25) 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 $\widetilde{Z}_n$ of $\widetilde{Z}$ as in Proposition 2.1 ($\widetilde{Z}_n$ may depend on γ). Proposition 2.1 together with [14, Lemma 2.1] yield that

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

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

$$\sup_{t \in \mathbb{R}} \mathbb{P}(|\widetilde{Z} - t| \leqslant C\varpi_n) \lesssim \varpi_n \left\{ \mathbb{E}[\widetilde{Z}] + \sqrt{1 \vee \log(\underline{\sigma}_{\mathfrak{g}}/(C\varpi_n))} \right\}. \tag{26}$$

Since $\mathcal{G}$ is VC type with characteristics $4\sqrt{A}$ and $2v$ for envelope G (Lemma 5.4), using an approximation argument, we have that

$$N(\mathcal{G}, \|\cdot\|_{P,2}, \tau) \leqslant (16\sqrt{A}\|G\|_{P,2}/\tau)^{2v}, \ 0 < \forall \tau \leqslant 1.$$

Hence, Dudley's entropy integral bound [25, Theorem 2.3.7] yields that $\mathbb{E}[\widetilde{Z}] \lesssim (\overline{\sigma}_{\mathfrak{g}} \vee (n^{-1/2} b_{\mathfrak{g}})) K_n^{1/2} \lesssim K_n^{1/2}$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 Section 3.**

Proof of Theorem 3.1. 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 $g_k = P^{r-1}h_k$. Invoke that the Gaussian process W_P can be extended to the linear hull of $\mathcal{G}$ in such a way that W_P has linear sample paths [e.g., see 25, Theorem 3.7.28]. Now, observe that

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

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 $\mathcal{G} \cdot \mathcal{G} = \{gg' : g, g' \in \mathcal{G}\}$. Recall that $\mathcal{G}$ is VC type with characteristics $4\sqrt{A}$ and $2v$ for envelope G, and by Corollary A.1 (i) in [13] ("$\sqrt{2}$" there should read "2"), it follows that $\mathcal{G} \cdot \mathcal{G}$ is VC type with characteristics $8\sqrt{A}$ and $4v$ for envelope G^2. Observe that 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). Hence, applying Corollary 5.5 with $\mathcal{F} = \mathcal{G} \cdot \mathcal{G}, r = k = 1$, and $q = q/2$ yields that

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

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

$$n^{-1/2}\|\mathbb{G}_n\|_{\mathcal{G}\cdot\mathcal{G}} \leqslant C\gamma^{-1}\left(\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{g}} K_n^{1/2} n^{-1/2} + b_{\mathfrak{g}}^2 K_n n^{-1+2/q}\right) \tag{27}$$

by Markov's inequality.

(ii). Define

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

We will show that

$$\mathbb{E}[\Upsilon_n] \leqslant C\left\{\sigma_{\mathfrak{h}}^2 K_n n^{-1} + \nu_{\mathfrak{h}}^2 K_n^2 n^{-3/2+2/q} + \sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{3/2} n^{-3/2} + b_{\mathfrak{h}}^2 K_n^3 n^{-2+2/q} + \chi_n^2\right\}. \tag{28}$$

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

$$\Upsilon_n \leqslant C\gamma^{-1}\left\{\sigma_{\mathfrak{h}}^2 K_n n^{-1} + \nu_{\mathfrak{h}}^2 K_n^2 n^{-3/2+2/q} + \sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{3/2} n^{-3/2} + b_{\mathfrak{h}}^2 K_n^3 n^{-2+2/q} + \chi_n^2\right\}. \tag{29}$$

The proof of the inequality (28) is lengthly and deferred after the proof of the theorem.

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

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

Next, since $U_n^{(1)}(\pi_1 h), h \in \mathcal{H}$ is an empirical process, we may apply the Hoffmann-Jørgensen inequality 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 $b_{\mathfrak{g}} \leqslant b_{\mathfrak{h}}$,

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

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

$$\|U_n\|_{\mathcal{H}}^2 \leqslant C\gamma^{-1}\left(\sigma_{\mathfrak{h}}^2 K_n n^{-1} + b_{\mathfrak{h}}^2 K_n^2 n^{-2+2/q} + n^{-1}\chi_n^2\right). \tag{30}$$

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

$$\|H\|_{\mathbb{P}_{I_{n,r}},2} \leqslant n^{1/2}\|H\|_{P^r,2}. \tag{31}$$

Now, define the event E by the the intersection of the events (27), (29), (30), and (31). 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. 25, Theorem 2.5.8], we have that

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

Note that $N(\mathcal{G}_\varepsilon, \|\cdot\|_{P,2}, \tau) \leqslant N^2(\mathcal{G}, \|\cdot\|_{P,2}, \tau/2)$. Since $\mathcal{G}$ is VC type with characteristics $4\sqrt{A}$ and $2v$ for envelope G, using an approximation argument, we have that $N(\mathcal{G}, \|\cdot\|_{P,2}, \tau\|G\|_{P,2}) \leqslant C(16\sqrt{A}/\tau)^{2v}$, 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 [48, Corollary 2.2.8] yields that

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

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

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

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

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

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 that

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

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

We first bound 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 $U_{n-1,-i}^{(r-1)}, i = 1,\dots,n$ 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

$$\begin{aligned}&\frac{3}{n}\sum_{i=1}^n \{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - P^{r-1}h(X_i)\}^2 \\ &\quad + \frac{3}{n}\sum_{i=1}^n \{(P^{r-1}h(X_i))^2 - P(P^{r-1}h)^2\} + 3P(P^{r-1}h)^2.\end{aligned} \tag{32}$$

The last term on the right hand side of (32) is bounded by $12(\varepsilon b_{\mathfrak{g}})^2$. The supremum of the first term on $\mathcal{H}_\varepsilon$ is bounded by $12\Upsilon_n$ since $\mathcal{H}_\varepsilon \subset \{h - h' : h, h' \in \mathcal{H}\}$ (the notation Υ_n appears in Step 2-(ii)). 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 $\{g^2 : g \in \mathcal{G}\} \subset \mathcal{G}\cdot\mathcal{G}$, so that the supremum of the second term on the right hand side of (32) is bounded by $12n^{-1/2}\|\mathbb{G}_n\|_{\mathcal{G}\cdot\mathcal{G}}$. Therefore, recalling that we have chosen $\varepsilon = 1/n^{1/2}$, we conclude that

$$\begin{aligned}\Sigma_n &\leqslant 12(\varepsilon b_{\mathfrak{g}})^2 + 12n^{-1/2}\|\mathbb{G}_n\|_{\mathcal{G}\cdot\mathcal{G}} + 12\Upsilon_n \\ &\leqslant C\gamma^{-1}\Bigg\{\overline{\sigma}_{\mathfrak{g}} b_{\mathfrak{g}} K_n^{1/2} n^{-1/2} + b_{\mathfrak{g}}^2 K_n n^{-1+2/q} + \sigma_{\mathfrak{h}}^2 K_n n^{-1} \\ &\qquad + \nu_{\mathfrak{h}}^2 K_n^2 n^{-3/2+2/q} + \sigma_{\mathfrak{h}} b_{\mathfrak{h}} K_n^{3/2} n^{-3/2} + b_{\mathfrak{h}}^2 K_n^3 n^{-2+2/q} + \chi_n^2\Bigg\}\end{aligned}$$

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, it is not difficult to see that

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

In addition, since

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

where the last inequality follows from Jensen's inequality, it follows that

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

Hence, Dudley's entropy integral bound yields that

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

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

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

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, so that

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

where we have used the Cauchy-Schwarz inequality. Since $n^{-1}\sum_{i=1}^n g^2(X_i), g \in \mathcal{G}$ is decomposed as $Pg^2 + n^{-1/2}\mathbb{G}_n(g^2)$ and the supremum of the latter on $\mathcal{G}$ is bounded by $\overline{\sigma}_{\mathfrak{g}}^2 + n^{-1/2}\|\mathbb{G}_n\|_{\mathcal{G}\cdot\mathcal{G}}$, we have that

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

where the second inequality follows from the inequality $2ab \leqslant a^2 + b^2$ for $a, b \in \mathbb{R}$. Now, the growth condition (8) ensures that

$$\Upsilon_n \bigvee (\overline{\sigma}_{\mathfrak{g}} \Upsilon_n^{1/2}) \bigvee \|U_n\|_{\mathcal{H}}^2 \leqslant C\gamma^{-1}\overline{\sigma}_{\mathfrak{g}} \Bigg\{ \sigma_{\mathfrak{h}} K_n^{1/2} n^{-1/2} + \nu_{\mathfrak{h}} K_n n^{-3/4+1/q} + (\sigma_{\mathfrak{h}} b_{\mathfrak{h}})^{1/2} K_n^{3/4} n^{-3/4} + b_{\mathfrak{h}} K_n^{3/2} n^{-1+1/q} + \chi_n \Bigg\}$$

on the event E, so that

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

Therefore, the Gaussian comparison inequality of [14, 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

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

Then, from Steps 1–5, we have that 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 (28).

Proof of the inequality (28). For a P^{r-1}-integrable symmetric function f on S^{r-1}, $U_{n-1,-i}^{(r-1)}(f)$ 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:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - P^{r-1}h(X_i)\}^2 \\
&\leqslant \frac{2}{n}\sum_{i=1}^{n}\{S_{n-1,-i}(\delta_{X_i}h)\}^2 + \frac{2}{n}\sum_{i=1}^{n}\{U_{n-1,-i}^{(r-1)}(\delta_{X_i}h) - P^{r-1}(\delta_{X_i}h) - S_{n-1,-i}(\delta_{X_i}h)\}^2.
\end{aligned} \tag{33}
$$

Consider the second term. By Lemma A.2 in Appendix A, for given $x \in S$, $\delta_x\mathcal{H} = \{\delta_x x : h \in \mathcal{H}\}$ is VC type with characteristics $2\sqrt{A}$ and $2v$ for envelope $\delta_x H$. Hence, we apply Corollary 5.6 conditionally on X_i and deduce that

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

Since $\sum_{k=2}^{r-1} n^{-k}\|P^{r-k-1}H\|_{P^{k+1},2}^2 K_n^k = \sum_{k=3}^{r} n^{-(k-1)}\|P^{r-k}H\|_{P^k,2}^2 K_n^{k-1} \leqslant C\chi_n^2$, the expectation of the supremum on $\mathcal{H}$ of the second term on the right hand side of (33) is at most $C\chi_n^2$.

For the first term, observe that

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

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

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

Since $P^2f^2 - P(Pf)^2 \leqslant \sigma_{\mathfrak{h}}^2$, 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

$$
\mathcal{F}_1 := \left\{(x_1,x_2) \mapsto f^2(x_1,x_2) - 2f(x_1,x_2)(Pf)(x_1) + (Pf)^2(x_1) : f \in \mathcal{F}\right\},
$$

$$\mathcal{F}_2^0 := \left\{ (x_1, x_2, x_3) \mapsto \left\{ \begin{array}{l} f(x_1,x_2)f(x_1,x_3) - f(x_1,x_2)(Pf)(x_1) \\ -f(x_1,x_3)(Pf)(x_1) + (Pf)^2(x_1) \end{array} \right\} : f \in \mathcal{F} \right\},$$

$$\mathcal{F}_2 := \left\{ (x_2,x_3) \mapsto \mathbb{E}[f(X_1,x_2,x_3)] : f \in \mathcal{F}_2^0 \right\},$$

$$\mathcal{F}_3 := \left\{ (x_1,x_2,x_3) \mapsto f(x_1,x_2,x_3) - \mathbb{E}[f(X_1,x_2,x_3)] : f \in \mathcal{F}_2^0 \right\},$$

together with their envelopes

$$F_1(x_1,x_2) := F^2(x_1,x_2) + 2F(x_1,x_2)(PF)(x_1) + (PF)^2(x_1),$$

$$F_2^0(x_1,x_2,x_3) := F(x_1,x_2)F(x_1,x_3) + F(x_1,x_2)(PF)(x_1) + F(x_1,x_3)(PF)(x_1) + (PF)^2(x_1),$$

$$F_2(x_2,x_3) := \mathbb{E}[F_2^0(X_1,x_2,x_3)],$$

$$F_3(x_1,x_2,x_3) := F_2^0(x_1,x_2,x_3) + F_2(x_2,x_3),$$

respectively. Lemma A.2 yields that $\mathcal{F}$ is VC type with characteristics $2\sqrt{A}, 2v$ for envelope F, and Corollary A.1 (i) in [13] together with Lemma A.2 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 that $\|P^{r-2}h\|_{P^2,4}^4 \leqslant \sigma_{\mathfrak{h}}^2 b_{\mathfrak{h}}^2$ 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

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

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

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

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. Note that 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 the growth condition (8) is automatically satisfied. □

6.3. **Proofs for Section 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. For the notational convenience, we will assume that each $h_{n,\vartheta}$ is P^r-centered; otherwise replace $h_{n,\vartheta}$ by $h_{n,\vartheta} - P^r h_{n,\vartheta}$, and the proof below applies to the non-centered case as well. 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 (14) 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), and (MT) 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' with $\log A' \lesssim \log n, v' \lesssim 1$ follows from a slight modification of the proof of Lemma 3.1 in [22]. 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 (13).

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

$$\underline{\sigma}_{\mathfrak{g}} \simeq 1,\ \overline{\sigma}_{\mathfrak{g}} \simeq 1,\ b_{\mathfrak{g}} \simeq b_n^{-m/2},\ \sigma_{\mathfrak{h}} \simeq b_n^{-m/2},\ b_{\mathfrak{h}} \simeq b_n^{-3m/2}, \tag{34}$$

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

$$\nu_{\mathfrak{h}} \lesssim b_n^{-m(1-1/q)},\ \chi_n \lesssim (\log n)^{3/2}/(n b_n^{3m/2}). \tag{35}$$

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

$$\begin{aligned} &\sup_{t \in \mathbb{R}} \left| \mathbb{P}(S_n \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t) \right| \leqslant C n^{-c}, \text{ and} \\ &\mathbb{P}\left\{ \sup_{t \in \mathbb{R}} \left| \mathbb{P}_{|D_1^n}(S_n^\sharp \leqslant t) - \mathbb{P}(\widetilde{S}_n \leqslant t) \right| > C n^{-c} \right\} \leqslant C n^{-c}. \end{aligned} \tag{36}$$

<u>Step 2</u>. Observe that

$$|\widehat{S}_n - S_n| \leqslant \sup_{\vartheta \in \Theta} \left| \frac{c_n(\vartheta)}{\widehat{c}_n(\vartheta)} - 1 \right| \|\sqrt{n} U_n\|_{\mathcal{H}_n} \quad \text{and} \quad |\widehat{S}_n^\sharp - S_n^\sharp| \leqslant \sup_{\vartheta \in \Theta} \left| \frac{c_n(\vartheta)}{\widehat{c}_n(\vartheta)} - 1 \right| \|\mathbb{U}_n^\sharp\|_{\mathcal{H}_n}. \tag{37}$$

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 (14) for $n \geqslant n_0$, since for $n < n_0$, the result (14) 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 $|a^{-1} - 1| \leqslant 4|a - 1|$ for $a \geqslant 1/2$, Condition (T8) also ensures that

$$\mathbb{P}\left\{ \sup_{\vartheta \in \Theta} \left| \frac{c_n(\vartheta)}{\widehat{c}_n(\vartheta)} - 1 \right| > C n^{-c} \right\} \leqslant C n^{-c}. \tag{38}$$

Next, we shall bound $\|\sqrt{n} U_n\|_{\mathcal{H}_n}$ and $\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_n}$. Given (34) and (35), 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 [25] ensures that the Gaussian process $\mathcal{W}_{P,n}$ can extended 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 pseudo-metric) 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

$$\begin{aligned} &\mathbb{P}\{\|\sqrt{n}U_n\|_{\mathcal{H}_n} > C(\log n)^{1/2}\} \leqslant Cn^{-c}, \text{ and} \\ &\mathbb{P}\left\{\mathbb{P}_{|D_1^n}\{\|\mathbb{U}_n^\sharp\|_{\mathcal{H}_n} > C(\log n)^{1/2}\} > Cn^{-c}\right\} \leqslant Cn^{-c}. \end{aligned} \tag{39}$$

Now, the desired result (14) follows from combining (36)–(39) and the anti-concentration inequality (Lemma A.1). In fact, the anti-concentration inequality yields that

$$\sup_{t\in\mathbb{R}} \mathbb{P}(|\widetilde{S}_n - t| \leqslant Cn^{-c}) \leqslant C'n^{-c}(\log n)^{1/2}. \tag{40}$$

Hence, combining the bounds (36)–(39) and (40), we have that 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 that

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

Step 3. It remains to verify (34) and (35). First, that we may choose $\underline{\sigma}_{\mathfrak{g}} \simeq 1$ follows from Conditions (T6) and (T7). For $\varphi \in \Phi$ and $k = 1, \ldots, r-1$, 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 $k = 1, \ldots, r$,

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

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

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

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

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

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

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

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

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

Similar calculations yield that

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

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

It is not difficult to verify that (34) and (35) 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 $q_{\widehat{S}_n}(\alpha)$ and $q_{\widetilde{S}_n}(\alpha)$ the α-quantiles of $\widehat{S}_n$ and $\widetilde{S}_n$, respectively. 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 distribution presumes that $\alpha+\eta_n<1$, but if $\alpha+\eta_n \geqslant 1$, then the last inequality is trivial. Likewise, we have that

$$\mathbb{P}\left\{\widehat{S}_n \leqslant q_{\widehat{S}_n^\sharp}(\alpha)\right\} \geqslant \alpha - 3\eta_n.$$

This completes the proof. □

Proof of Lemma 4.1. As in the proof of Theorem 4.2, we will assume that each $h_{n,\vartheta}$ is P^r-centered. We begin with noting that

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

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

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

for some constant C depending only on r. The desired result follows from the choices of parameters $\overline{\sigma}_{\mathfrak{g}}, b_{\mathfrak{g}}, \sigma_{\mathfrak{h}}, b_{\mathfrak{h}}, \chi_n$, 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 A. Supporting lemmas

This appendix collects two 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 [13]. □

Lemma A.2. *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 $\mathcal{X} \times \mathcal{Y}$ equipped with finite 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 well-defined and is 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 finite $r, s \geqslant 1$,

$$\sup_Q N(\overline{\mathcal{F}}, \|\cdot\|_{Q,r}, 2\varepsilon \|\overline{F}\|_{Q,r}) \leqslant \sup_Q N(\mathcal{F}, \|\cdot\|_{Q \times R, s}, \varepsilon^r \|F\|_{Q \times R})$$

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

Proof. See Lemma A.1 in [22]. □

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 [37]. 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 [18]. Next, we state a conditional version of the Strassen-Dudley theorem due to [37, 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*

$$\mathbb{E}^* \left[\sup_{A \in \mathcal{B}(S)} \{\mathbb{P}(X \in A \mid \mathcal{G}) - G(A^\alpha \mid \mathcal{G})\} \right] \leqslant \beta, \tag{41}$$

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 [37] 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 $n = 1, 2, \dots$ 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 $\{D_{n,k} : n = 1, 2, \dots; k \in \mathbb{Z}\}$; 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 $\mathcal{L}(Y_D) = G(\cdot \mid D)$ and $\mathbb{P}_D(d(X, Y_D) > \alpha) \leqslant \varepsilon(D) := \sup_{A \in \mathcal{B}(S)} \{\mathbb{P}_D(A) - G(A^\alpha \mid D)\}$.

For each $n = 1, 2, \dots$, 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,

$$\begin{aligned}\mathbb{P}(Y_n \in A) &= \sum_k \mathbb{P}(\{Y_{D_{n,k}} \in A\} \cap D_{n,k}) = \sum_k \mathbb{P}_{D_{n,k}}(Y_{D_{n,k}} \in A)\mathbb{P}(D_{n,k}) \\ &= \sum_k G(A \mid D_{n,k})\mathbb{P}(D_{n,k}) = \mathbb{E}[G(A \mid \mathcal{G})],\end{aligned}$$

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 [18, 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 $F_{x,w}^{-1}$ 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 $F_{x,w}^{-1}(U)$ has law $Q_{x,w} \circ \pi^{-1}$, and hence $Y = \pi^{-1} \circ F_{X,W}^{-1}(U)$ 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 that

$$\mathbb{P}(d(X,Y) > \alpha) \leqslant \liminf_{n'} \mathbb{P}(d(X, Y_{n'}) > \alpha) \leqslant \beta.$$

This completes the proof. $\square$

Appendix C. 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 [14] 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 [14, 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 [14], 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 that

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

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. □

References

[1] Jason Abrevaya and Wei Jiang. A nonparametric approach to measuring and testing curvature. *Journal of Business & Economic Statistics*, 23(1):1–19, 2005.

[2] Miguel Arcones and Evarist Giné. On the bootstrap of U- and V-statistics. *Annals of Statistics*, 20(2):655–674, 1992.

[3] Miguel Arcones and Evarist Giné. Limit theorems for U-processes. *Annals of Probability*, 21(3):1495–1542, 1993.

[4] Miguel Arcones and Evarist Giné. U-processes indexed by Vapnik-Červonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. *Stochastic Processes and Their Applications*, 52(1):17–38, 1994.

[5] Peter J. Bickel and David A. Freedman. Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6):1196–1217, 1981.

[6] Richard Blundell, Amanda Gosling, Hidehiko Ichimura, and Costas Meghir. Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, 75(2):323–363, 2007.

[7] Yu. V. Borovskikh. *U-Statistics in Banach Spaces.* V.S.P. Intl Science, 1996.

[8] J. Bretagnolle. Lois limits du Bootstrap de certaines functionnelles. *Annales de l'Institut Henri Poincaré Section B*, XIX(3):281–296, 1983.

[9] Herman Callaert and Noël Veraverbeke. The order of the normal approximation for a Studentized U-statistic. *Annals of Statistics*, 9(1):360–375, 1981.

[10] Xiaohui Chen. Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Annals of Statistics, to appear*, 2017.

[11] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41(6):2786–2819, 2013.

[12] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Anti-concentration and honest, adaptive confidence bands. *Annals of Statistics*, 42(5):1787–1818, 2014.

[13] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *Annals of Statistics*, 42(4):1564–1597, 2014.

[14] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. *Stochastic Processes and Their Applications*, 126(12):3632–3651, 2016.

[15] Denis Chetverikov. Testing regression monotonicity in econometric models. *arXiv:1212.6757*, 2012.

[16] Victor de la Peña and Evarist Giné. *Decoupling: From Dependence to Independence.* Springer, 1999.

[17] Herold Dehling and Thomas Mikosch. Random quadratic forms and the bootstrap for U-statistics. *Journal of Multivariate Analysis*, 51(2):392–413, 1994.

[18] Richard M. Dudley. *Real Analysis and Probability.* Cambridge University Press, 2002.

[19] Uwe Einmahl and David M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33(3):1380–1403, 2005.

[20] Glenn Ellison and Sara Fisher Ellison. Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration. *American Economic Journal: Microeconomics*, 3(1):1–36, 2011.

[21] Edward W. Frees. Estimating densities of functions of observations. *Journal of American Statistical Association*, 89(426):517–525, 1994.

[22] Subhashis Ghosal, Arusharka Sen, and Aad van der Vaart. Testing monotonicity of regression. *Annals of Statistics*, 28(4):1054–1082, 2000.

[23] Evarist Giné and David M. Mason. On local U-statistic processes and the estimation of densities of functions of several sample variables. *Annals of Statistics*, 35(3):1105–1145, 2007.

[24] Evarist Giné and Richard Nickl. Uniform limit theorems for wavelet density estimators. *Annals of Probability*, 37(4):1605–1646, 2009.

[25] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

[26] Peter Hall. On convergence rates of suprema. *Probability Theory and Related Fields*, 89(4):447–455, 1991.

[27] Wassily Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[28] Marie Hušková and Paul Jansen. Generalized bootstrap for studentized U-statistics: a rank statistic approach. *Statistics and Probability Letters*, 16(3):225–233, 1993.

[29] Marie Huškova and Paul Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *Annals of Statistics*, 21(4):1811–1823, 1993.

[30] Paul Janssen. Weighted bootstrapping of U-statistics. *Journal of Statistical Planning and Inference*, 38(1):31–42, 1994.

[31] Vladmir I. Koltchinskii. Komlos-Major-Tusnády approximation for the general empirical process and Haar expansions of classes of functions. *Journal of Theoretical Probability*, 7(1):73–118, 1994.

[32] J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent rv's and the sample df. I. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, 32(1-2):111–131, 1975.

[33] Sokbae Lee, Oliver Linton, and Yoon-Jae Whang. Testing for stochastic monotonicity. *Econometrica*, 77(2):585–602, 2009.

[34] Albert Y. Lo. A large sample study of the Bayesian bootstrap. *Annals of Statistics*, 15(1):360–375, 1987.

[35] David M. Mason and Micheal A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Annals of Statistics*, 20(3):1611–1624, 1992.

[36] Pascal Massart. Strong approximation for multivariate empirical and related processes, via KMT constructions. *Annals of Probability*, 17(1):266–291, 1989.

[37] Ditlev Monrad and Walter Philipp. Nearby variables with nearby conditional laws and a strong approximation theorem for Hilbert space valued martingales. *Probability Theory and Related Fields*, 88(3):381–404, 1991.

[38] Deborah Nolan and David Pollard. U-processes: rates of convergence. *Annals of Statistics*, 15(2):780–799, 1987.

[39] Deborah Nolan and David Pollard. Functional limit theorems for U-processes. *Annals of Probability*, 16(3):1291–1298, 1988.

[40] Vladimir I. Piterberg. *Asymptotic Methods in the Theory of Gaussian Processes and Fields*. American Mathematical Society, 1996.

[41] Sidney I. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer-Verlag, 1987.

[42] Emmanuel Rio. Local invariance principles and their application to density estimation. *Probability Theory and Related Fields*, 98(1):21–45, 1994.

[43] Donald B. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9(1):130–134, 1981.

[44] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

[45] Robert P. Sherman. Limiting distribution of the maximal rank correlation estimator. *Econometrica*, 61(1):123–137, 1993.

[46] Robert P. Sherman. Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics*, 22(1):439–459, 1994.

[47] G. Solon. Intergenerational income mobility in the United States. *American Economic Review*, 82(3):393–408, 1992.

[48] Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.

[49] Aad van der Vaart and Jon A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011.

[50] Qiying Wang and Bin-Ying Jing. Weighted bootstrap for U-statistics. *Journal of Multivariate Analysis*, 91(2):177–198, 2004.

[51] Chung-Sing Weng. On a second-order asymptotic property of the Bayesian bootstrap mean. *Annals of Statistics*, 17(2):705–710, 1989.

[52] Dixin Zhang. Bayesian bootstraps for U-processes, hypothesis tests and convergence of Dirichlet U-processes. *Statistica Sinica*, 11(2):463–478, 2001.

(X. Chen) Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, IL 61874 USA.

E-mail address: `xhchen@illinois.edu`

(K. Kato) Graduate School of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

E-mail address: `kkato@e.u-tokyo.ac.jp`