Randomized Incomplete U-Statistics in High Dimensions

By

Xiaohui Chen and Kengo Kato

April 2018

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION
DISCUSSION PAPER NO. 26

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

RANDOMIZED INCOMPLETE U-STATISTICS IN HIGH DIMENSIONS

XIAOHUI CHEN AND KENGO KATO

Abstract. This paper studies inference for the mean vector of a high-dimensional U-statistic. In the era of Big Data, the dimension d of the U-statistic and the sample size n of the observations tend to be both large, and the computation of the U-statistic is prohibitively demanding. Data-dependent inferential procedures such as the empirical bootstrap for U-statistics is even more computationally expensive. To overcome such computational bottleneck, incomplete U-statistics obtained by sampling fewer terms of the U-statistic are attractive alternatives. In this paper, we introduce randomized incomplete U-statistics with sparse weights whose computational cost can be made independent of the order of the U-statistic. We derive non-asymptotic Gaussian approximation error bounds for the randomized incomplete U-statistics in high dimensions, namely in cases where the dimension d is possibly much larger than the sample size n, for both non-degenerate and degenerate kernels. In addition, we propose novel and generic bootstrap methods for the incomplete U-statistics that are computationally much less-demanding than existing bootstrap methods, and establish finite sample validity of the proposed bootstrap methods. The proposed bootstrap methods are illustrated on the application to nonparametric testing for the pairwise independence of a high-dimensional random vector under weaker assumptions than those appearing in the literature.

1. Introduction

Let $X_1, \dots, X_n$ be independent and identically distributed (i.i.d.) random variables taking values in a measurable space $(S, \mathcal{S})$ with common distribution P. Let $r \geqslant 2$ and $d \geqslant 1$ be given positive integers, and let $h = (h_1, \dots, h_d)^T : S^r \to \mathbb{R}^d$ be a fixed and jointly measurable function that is symmetric in its arguments, i.e., $h(x_1, \dots, x_r) = h(x_{i_1}, \dots, x_{i_r})$ for every permutation $i_1, \dots, i_r$ of $1, \dots, r$. Suppose that $\mathbb{E}[|h_j(X_1, \dots, X_r)|] < \infty$ for all $j = 1, \dots, d$, and consider inference on the mean vector $\theta = (\theta_1, \dots, \theta_d)^T = \mathbb{E}[h(X_1, \dots, X_r)]$. To this end, a commonly used statistic is the U-statistic with kernel h, i.e., the sample average of $h(X_{i_1}, \dots, X_{i_r})$ over all distinct r-tuples $(i_1, \dots, i_r)$ from $\{1, \dots, n\}$

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

where $I_{n,r} = \{(i_1, \dots, i_r) : 1 \leqslant i_1 < \dots < i_r \leqslant n\}$ and $|I_{n,r}| = n!/\{r!(n-r)!\}$ denotes the cardinality of $I_{n,r}$.

Date: First arXiv version: December 3, 2017. This version: December 18, 2017.

2010 *Mathematics Subject Classification.* 62E17, 62F40, 62H15.

Key words and phrases. Incomplete U-statistics, randomized inference, Gaussian approximation, bootstrap, divide and conquer, Bernoulli sampling, sampling with replacement.

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. This work is completed in part with the high-performance computing resource provided by the Illinois Campus Cluster Program at UIUC.

U-statistics are an important and general class of statistics, and applied in a wide variety of statistical problems; we refer to [28] as an excellent monograph on U-statistics. For univariate U-statistics ($d = 1$), the asymptotic distributions are derived in the seminal paper [21] for the non-degenerate case and in [35] for the degenerate case. There is also a large literature on bootstrap methods for univariate U-statistics [4, 6, 1, 24, 25, 15, 41]. A more recent interest lies in the high-dimensional case where d is much larger than n. [8] develops Gaussian and bootstrap approximations for non-degenerate U-statistics of order two in high dimensions, which extends the work [10, 12] from sample averages to U-statistics.

However, a major obstacle of inference using the complete U-statistic (1) is its computational intractability. Namely, the computation of the complete U-statistic (1) requires $O(n^r d)$ operations, and its computational cost can be prohibitively demanding even when n and d are moderately large, especially when the order of the U-statistic $r \geqslant 3$. For instance, the computation of a complete U-statistic with order 3 and dimension $d = 5000$ when the sample size is $n = 1000$ requires $\binom{n}{3} \times d \approx 0.8 \cdot 10^{12}$ (0.8 trillion) operations. In addition, the naive application of the empirical bootstrap for the U-statistic (1) requires even more operations, namely, $O(Bn^r d)$ operations, where B is the number of bootstrap repetitions.

This motivates us to study inference using *randomized incomplete U-statistics* with sparse weights instead of complete U-statistics. Specifically, we consider the Bernoulli sampling and sampling with replacement to construct random weights in Section 2. For a pre-specified *computational budget parameter* $N \leqslant |I_{n,r}|$, these sampling schemes randomly choose (on average) N indices from $I_{n,r}$, and the resulting incomplete U-statistics $U'_{n,N}$ are defined as the sample averages of $h(X_{i_1}, \ldots, X_{i_r})$ taken over the subset of chosen indices $(i_1, \ldots, i_r)$. Hence the computational cost of the incomplete U-statistics is reduced to $O(Nd)$, which can be much smaller than $n^r d$ as long as $N \ll n^r$ and can be made independent of the order of the U-statistic provided that N does not depend on r.

The goal of this paper is to develop computationally scalable and statistically correct inferential methods for the incomplete U-statistics with high-dimensional kernels and massive data, where d is possibly much larger than n but n can be also large. Specifically, we study distributional approximations to the randomized incomplete U-statistics in high dimensions. Our first main contribution is to derive Gaussian approximation error bounds for the incomplete U-statistics on the hyperrectangles in $\mathbb{R}^d$ for both non-degenerate and degenerate kernels. In Section 3, we show that the derived Gaussian approximation results display an interesting computational and statistical trade-off for non-degenerate kernels (see Remark 3.1), and reveal a fundamental difference between complete and randomized incomplete U-statistics for degenerate kernels (see Remark 3.2). The mathematical insight of introducing the random weights is to create the (conditional) independence for the terms in the U-statistic sum in order to obtain a Gaussian limit. Note that the Gaussian approximation results are often not directly applicable since the covariance matrices of the approximating Gaussian distributions depend on the underlying distribution P that is unknown in practice. Our second contribution is to propose fully data-dependent bootstrap methods for incomplete U-statistics that are computationally (much) less demanding than existing bootstrap methods for U-statistics [1, 8, 9]. Specifically, we introduce generic bootstraps for incomplete U-statistics in Section 4.1. Our generic bootstrap constructions are flexible enough to cover both non-degenerate and degenerate kernels, and meanwhile they take the computational concern into account for estimating the associated (and unobserved) Hájek projection. In particular, we propose two concrete estimation procedures for the Hájek projection: one is a deterministic construction based on the divide and conquer algorithm (Section 4.2), and another is a random construction based on a second randomization independent of everything else (Section 4.3). For both constructions, the overall computational complexity of the bootstrap methods can be made independent of the U-statisitic order r.

As a leading example to illustrate the usefulness of the inferential methods developed in the present paper, we consider testing for the pairwise independence of a high-dimensional random vector $X = (X^{(1)}, \ldots, X^{(p)})^T$, i.e., testing for the hypothesis that

$$H_0 : X^{(1)}, \ldots, X^{(p)} \text{ are pairwise independent.} \tag{2}$$

Let $X_1, \ldots, X_n$ be i.i.d. copies of X. Several nonparametric test statistics are proposed in the literature, including: Kendall's τ, Spearman's ρ, Hoeffding's D [22], Bergsma and Dassios' t^* [2], and the distance covariance [42], all of which can be expressed as functions of U-statistics.

Example 1.1 (Spearman's ρ)**.** Let Π_r be the collection of all possible permutations on $\{1, \ldots, r\}$. [21] shows that Spearman's rank correlation coefficient matrix ρ can be written as

$$\rho = \frac{n-2}{n+1}\widehat{\rho} + \frac{3}{n+1}\tau,$$

where $\widehat{\rho} = U_n^{(3)}(h^S)$ is the $p \times p$ matrix-valued U-statistic associated with the kernel

$$h^S(X_1, X_2, X_3) = \left(h^S_{j,k}(X_1, X_2, X_3)\right)_{1 \leqslant j,k \leqslant p} = \frac{1}{2} \sum_{\pi \in \Pi_3} \operatorname{sign}\left\{(X_{\pi(1)} - X_{\pi(2)})(X_{\pi(1)} - X_{\pi(3)})^T\right\},$$

and $\tau = (\tau_{j,k})_{1 \leqslant j,k \leqslant p} = U_n^{(2)}(h^K)$ is the $p \times p$ Kendall τ matrix with the kernel

$$h^K(X_1, X_2) = \operatorname{sign}\left\{(X_1 - X_2)(X_1 - X_2)^T\right\}.$$

Here, for a matrix $A = (a_{j,k})_{1 \leqslant j,k \leqslant p}$, $\operatorname{sign}\{A\}$ is the matrix of the same size as A whose (j,k)-th element is $\operatorname{sign}(a_{j,k}) = \mathbf{1}(a_{j,k} > 0) - \mathbf{1}(a_{j,k} < 0)$. It is seen that the leading term in Spearman's ρ is $\widehat{\rho}$, and so it is reasonable to reject the null hypothesis (2) if $\max_{1 \leqslant j < k \leqslant p} |\widehat{\rho}_{j,k}|$ is large. Precisely speaking, this test is testing for a weaker hypothesis that

$$H_0' : \mathbb{E}[\operatorname{sign}(X_1^{(j)} - X_2^{(j)})\operatorname{sign}(X_1^{(k)} - X_3^{(k)})] = 0 \text{ for all } 1 \leqslant j < k \leqslant p.$$

Example 1.2 (Bergsma and Dassios' t^*)**.** [2] propose a U-statistic $t^* = (t^*_{j,k})_{1 \leqslant j,k \leqslant p} = U_n^{(4)}(h^{BD})$ of order 4 with the kernel

$$h^{BD}(X_1, \ldots, X_4) = \frac{1}{24} \sum_{\pi \in \Pi_4} \phi(X_{\pi(1)}, \ldots, X_{\pi(4)})\phi(X_{\pi(1)}, \ldots, X_{\pi(4)})^T,$$

where $\phi(X_1, \ldots, X_4) = (\phi_j(X_1, \ldots, X_4))_{j=1}^p$ and

$$\begin{aligned}\phi_j(X_1, \ldots, X_4) = {} & \mathbf{1}(X_1^{(j)} \vee X_3^{(j)} < X_2^{(j)} \wedge X_4^{(j)}) + \mathbf{1}(X_1^{(j)} \wedge X_3^{(j)} > X_2^{(j)} \vee X_4^{(j)}) \\ & - \mathbf{1}(X_1^{(j)} \vee X_2^{(j)} < X_3^{(j)} \wedge X_4^{(j)}) - \mathbf{1}(X_1^{(j)} \wedge X_2^{(j)} > X_3^{(j)} \vee X_4^{(j)}).\end{aligned}$$

Here, $a \wedge b = \max\{a, b\}$ and $a \vee b = \min\{a, b\}$. Under the assumption that $(X^{(j)}, X^{(k)})$ has a bivariate distribution that is discrete or (absolutely) continuous, or a mixture of both, [2] show that $\mathbb{E}[t^*_{j,k}] = 0$ if and only if $X^{(j)}$ and $X^{(k)}$ are independent, and so it is reasonable to reject the null hypothesis (2) if $\max_{1 \leqslant j < k \leqslant p} |t^*_{j,k}|$ is large (or $\max_{1 \leqslant j < k \leqslant p} t^*_{j,k}$ is large, since in general $\mathbb{E}[t^*_{j,k}] \geqslant 0$).

Example 1.3 (Hoeffding's D)**.** [22] proposes a U-statistic $D = (D_{j,k})_{1 \leqslant j,k \leqslant p} = U_n^{(5)}(h^D)$ of order 5 with the kernel

$$h^D(X_1, \ldots, X_5) = \frac{1}{120} \sum_{\pi \in \Pi_5} \phi(X_{\pi(1)}, \ldots, X_{\pi(5)}) \phi(X_{\pi(1)}, \ldots, X_{\pi(5)})^T,$$

where $\phi(X_1, \ldots, X_5) = (\phi_j(X_1, \ldots, X_5))_{j=1}^p$ and

$$\phi_j(X_1, \ldots, X_5) = \frac{1}{4}[\mathbf{1}(X_1^{(j)} \geqslant X_2^{(j)}) - \mathbf{1}(X_1^{(j)} \geqslant X_3^{(j)})][\mathbf{1}(X_1^{(j)} \geqslant X_4^{(j)}) - \mathbf{1}(X_1^{(j)} \geqslant X_5^{(j)})].$$

Under the assumption that the joint distribution of $(X^{(j)}, X^{(k)})$ has continuous joint and marginal densities, [22] shows that $\mathbb{E}[D_{j,k}] = 0$ if and only if $X^{(j)}$ and $X^{(k)}$ are independent, and so it is reasonable to reject the null hypothesis (2) if $\max_{1 \leqslant j < k \leqslant p} |D_{j,k}|$ is large (or $\max_{1 \leqslant j < k \leqslant p} D_{j,k}$ is large, since in general $\mathbb{E}[D_{j,k}] \geqslant 0$). It is worth noting that Bergsma and Dassios' t^* is an improvement on Hoeffding's D since the former can characterize the pairwise independence under weaker assumptions on the distribution of X than the latter.

Note that h^S is non-degenerate, while h^{BD} and h^D are degenerate of order 1 under H_0. In Examples 1.1–1.3, to compute the test statistics, we have to compute U-statistics with dimension $d = p(p-1)/2$, which can be quite large. In addition, the orders of the U-statistics are at least 3, and so the computation of the test statistics is prohibitively demanding, not to mention the empirical bootstrap or subsampling for those U-statistics. It should be noted that there are efficient algorithms to reduce the computational costs for computing some of those U-statistics [cf. 30, Section 6.1], but such computational simplifications are case-by-case and not generically applicable, and more importantly they do not yield computationally tractable methods to approximate or estimate the sampling distributions of the U-statistics. The Gaussian and bootstrap approximation theorems developed in the present paper can be directly applicable to calibrating critical values for the max-type test statistics appearing in Examples 1.1–1.3, since $\{y = (y_1, \ldots, y_d)^T \in \mathbb{R}^d : \max_{1 \leqslant j \leqslant d} |y_j| \leqslant t\} = [-t, t]^d$ is a hyperrectangle in $\mathbb{R}^d$.

The above testing problem is motivated from recent papers by [30] and [19], which study testing for the null hypothesis

$$H_0'' : X^{(1)}, \ldots, X^{(p)} \text{ are mutually independent},$$

and develop tests based on functions of the U-statistics appearing in Examples 1.1–1.3. Note that H_0'' is a stronger hypothesis than H_0. Specifically, [30] consider tests statistics such as, e.g., $S_{\widehat{\rho}} = \sum_{1 \leqslant j < k \leqslant p} \widehat{\rho}_{j,k}^2 - 3\mu_{\widehat{\rho}}$ with $\mu_{\widehat{\rho}} = \mathbb{E}[\widehat{\rho}_{1,2}^2]$ under H_0'' and show that $nS_{\widehat{\rho}}/(9p\zeta_1^{\widehat{\rho}}) \xrightarrow{d} N(0,1)$ under H_0'' as $(n,p) \to \infty$ where $\zeta_1^{\widehat{\rho}} = \mathrm{Var}(\mathbb{E}[h_{1,2}^S(X_1, X_2, X_3) \mid X_1])$. On the other hand, [19] consider test statistics such as, e.g., $L_n = \max_{1 \leqslant j < k \leqslant p} |\widehat{\rho}_{j,k}|$ and show that $L_n^2/\mathrm{Var}(\widehat{\rho}_{1,2}) - 4 \log p + \log \log p$ converges in distribution to a Gumbel distribution as $n \to \infty$ and $p = p_n \to \infty$ under H_0'' provided that $\log p = o(n^{1/3})$ (precisely speaking, [19] rule out degenerate kernels). Importantly, compared with the tests developed in [30] and [19] based on analytical critical values, our bootstrap-based tests can directly detect the pairwise dependence for some pair of coordinates (or $\mathbb{E}[\mathrm{sign}(X_1^{(j)} - X_2^{(j)})\mathrm{sign}(X_1^{(k)} - X_3^{(k)})] \neq 0$ for some $1 \leqslant j < k \leqslant p$ for Spearman's ρ) rather than the non-mutual-independence and also work for non-continuous random vectors (see, e.g., [17] for interesting examples of pairwise independent but jointly dependent random variables; in particular, their examples include continuous random variables). In contrast, the derivations of the asymptotic null distributions in [30] and [19] critically depend on the mutual independence between the coordinates of X. In addition, they both assume that X is continuously distributed so that there are no ties in $X_1^{(j)}, \dots, X_n^{(j)}$ for each coordinate j, thereby ruling out discrete components. It is worth noting that the U-statistics appearing Examples 1.1–1.3 are rank-based, and so if X is continuous and H_0'' is true, then those U-statistics are pivotal, i.e., they have known (but difficult-to-compute) distributions, which is also a critical factor in their analysis; however, that is not the case under the weaker hypothesis of pairwise independence and without the continuity assumption on X.

To verify the finite sample performance of the proposed bootstrap methods for randomized incomplete U-statistics, we conduct simulation experiments in Section 5 on the leading example for nonparametric testing for the pairwise independence hypothesis in (2). Specifically, we consider to approximate the null distributions of the (leading term of) Spearman ρ and Bergsma-Dassios' t^* test statistics, and examine the cases where $n = 300, 500, 1000$ and $p = 30, 50, 100$ (and hence $d = p(p-1)/2 = 435, 1225, 4950$). Statistically, we observe that the empirical rejection probability of the null hypothesis with the critical values calibrated by our bootstrap methods is very close to the nominal size for (almost) all setups. Computationally, we find that the (log-)running time for our bootstrap methods scales linearly with the (log-)sample size, and in addition, the slope coefficient matches very well with the computational complexity of the bootstrap methods. Therefore, the simulation results demonstrate a promising agreement between the empirical evidences and our theoretical analysis.

1.1. **Existing literature.** Incomplete U-statistics were first considered in [5], and the asymptotic distributions of incomplete U-statistics (for fixed d) are derived in [7] and [26]; see also Section 4.3 in [28] for a review on incomplete U-statistics. Closely related to the present paper is [26], which establishes the asymptotic properties of univariate incomplete U-statistics based on sampling with and without replacement and Bernoulli sampling. To the best of our knowledge, the present paper is the first paper that establishes approximation theorems for the distributions of randomized incomplete U-statistics in high dimensions. See also Remark 3.4 for more detailed comparisons with [26]. Incomplete U-statistics can be viewed as an special case of weighted U-statistics, and there is a large literature on limit theorems for weighted U-statistics; see [37, 33, 31, 34, 23, 20] and references therein. These references focus on the univariate case and do not cover the high-dimensional case. There are few references that study data-dependent inferential procedures for incomplete U-statistics that take computational considerations into account. An exception is [3], which proposes several inferential methods for univariate (generalized) incomplete U-statistics, but do not develop formal asymptotic justifications for these methods. It is also interesting to note that incomplete U-statistics have gained renewed interests in the recent statistics and machine learning literatures [13, 32], although the focuses of these references are substantially different from ours.

From a technical point of view, this paper builds on recent development of Gaussian and bootstrap approximation theorems for averages of independent high-dimensional random vectors [10, 12] and for high-dimensional U-statistics of order two [8]. Importantly, however, developing Gaussian approximations for the randomized incomplete U-statistics in high dimensions requires a novel proof-strategy that combines iterative conditioning arguments and applications of Berry-Esseen type bounds, and extending some of results in [8] to cover general order incomplete U-statistics. In addition, these references do not consider bootstrap methods for incomplete U-statistics that take computational considerations into account.

1.2. **Organization.** The rest of the paper is organized as follows. In Section 2, we introduce randomized incomplete U-statistics with sparse weights generated from the Bernoulli sampling and sampling with replacement. In Section 3, we derive non-asymptotic Gaussian approximation error bounds for the randomized incomplete U-statistics in high dimensions for both non-degenerate and degenerate kernels. In Section 4, we first propose generic bootstrap methods for the incomplete U-statistics and then incorporate the computational budget constraint by two concrete estimates of the Hájek projection: one deterministic estimate by the divide and conquer, and one randomized estimate by incomplete U-statistics of a lower order. Simulation examples are provided in Section 5 and Appendix B. In Section 6, we leave some additional discussions including extensions of the present paper. All the technical proofs are gathered in Appendix A.

1.3. **Notation.** For a hyperrectangle $R = \prod_{j=1}^{d}[a_j, b_j]$ in $\mathbb{R}^d$, a constant $c > 0$, and a vector $y = (y_1, \dots, y_d)^T \in \mathbb{R}^d$, we use the notation $[cR + y] = \prod_{j=1}^{d}[ca_j + y_j, cb_j + y_j]$. For vectors $y = (y_1, \dots, y_d)^T, z = (z_1, \dots, z_d)^T \in \mathbb{R}^d$, the notation $y \leqslant z$ means that $y_j \leqslant z_j$ for all $j = 1, \dots, d$. For $a, b \in \mathbb{R}$, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a finite set J, $|J|$ denotes the cardinality of J. Let $|\cdot|_\infty$ denote the max-norm for vectors and matrices, i.e., for a matrix $A = (a_{ij})$, $|A|_\infty = \max_{i,j} |a_{ij}|$. "Constants" refer to finite, positive, and non-random numbers.

For $0 < \beta < \infty$, let ψ_β be the function on $[0, \infty)$ defined by $\psi_\beta(x) = e^{x^\beta} - 1$, and for a real-valued random variable ξ, define $\|\xi\|_{\psi_\beta} = \inf\{C > 0 : \mathbb{E}[\psi_\beta(|\xi|/C)] \le 1\}$. For $\beta \in [1, \infty)$, $\|\cdot\|_{\psi_\beta}$ is an Orlicz norm, while for $\beta \in (0, 1)$, $\|\cdot\|_{\psi_\beta}$ is not a norm but a quasi-norm, i.e., there exists a constant C_β depending only on β such that $\|\xi_1 + \xi_2\|_{\psi_\beta} \le C_\beta(\|\xi_1\|_{\psi_\beta} + \|\xi_2\|_{\psi_\beta})$ (indeed, there is a norm equivalent to $\|\cdot\|_{\psi_\beta}$ obtained by linearizing ψ_β in a neighborhood of the origin; cf. Lemma A.2 ahead).

For a generic random variable Y, let $\mathbb{P}_{|Y}(\cdot)$ and $\mathbb{E}_{|Y}[\cdot]$ denote the conditional probability and expectation given Y, respectively. For a given probability space $(\mathcal{X}, \mathcal{A}, Q)$ and a measurable function f on $\mathcal{X}$, we use the notation $Qf = \int f dQ$ whenever the latter integral is well-defined. For a jointly measurable symmetric function f on S^r and $k = 1, \dots, r$, let $P^{r-k}f$ denote the function on S^k defined by $P^{r-k}f(x_1, \dots, x_k) = \int \cdots \int f(x_1, \dots, x_k, x_{k+1}, \dots, x_r) dP(x_{k+1}) \cdots dP(x_r)$ whenever the integral exists and is finite for every $(x_1, \dots, x_k) \in S^k$. For given $1 \leqslant k \leqslant \ell \leqslant n$, we use the notation $X_k^\ell = (X_k, \dots, X_\ell)$. Throughout the paper, we assume that $n \geqslant 4 \vee r$ and $d \geqslant 3$.

2. Randomized incomplete U-statistics

In this paper, to construct sparsely weighted U-statistics, we shall use random sparse weights. For $\iota = (i_1, \ldots, i_r) \in I_{n,r}$, let us write $X_\iota = (X_{i_1}, \ldots, X_{i_r})$, and observe that the complete U-statistic (1) can be written as

$$U_n = \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} h(X_\iota).$$

Now, let $N := N_n$ be an integer such that $0 < N \leqslant |I_{n,r}|$, and let $p_n = N/|I_{n,r}|$. Instead of taking the average over all possible ι in $I_{n,r}$, we will take the average over a subset of about N indices chosen randomly from $I_{n,r}$. In the present paper, we study Bernoulli sampling and sampling with replacement.

2.1. Bernoulli sampling.

Generate i.i.d. $\mathsf{Ber}(p_n)$ random variables $\{Z_\iota : \iota \in I_{n,r}\}$ with success probability p_n, i.e., $Z_\iota, \iota \in I_{n,r}$ are i.i.d. with $\mathbb{P}(Z_\iota = 1) = 1 - \mathbb{P}(Z_\iota = 0) = p_n$. Consider the following weighted U-statistic with random weights

$$U'_{n,N} = \frac{1}{\widehat{N}} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota), \tag{3}$$

where $\widehat{N} = \sum_{\iota \in I_{n,r}} Z_\iota$ is the number of non-zero weights. We call $U'_{n,N}$ the randomized incomplete U-statistic based the Bernoulli sampling. Note that $\widehat{N}$ follows $\mathsf{Bin}(|I_{n,r}|, p_n)$, the binomial distribution with parameters $(|I_{n,r}|, p_n)$. Hence $\mathbb{E}[\widehat{N}] = |I_{n,r}|p_n = N$ and the computation of the incomplete U-statistic (3) only requires $O(Nd)$ operations on average. In addition, by Bernstein's inequality (cf. Lemma 2.2.9 in [39]),

$$\mathbb{P}\left(|\widehat{N}/N - 1| > \sqrt{2t/N} + 2t/(3N) \right) \leqslant 2e^{-t} \tag{4}$$

for every $t > 0$, and hence $\widehat{N}$ concentrates around its mean N. Therefore, we can view N as a *computational budget parameter* and p_n as a *sparsity design parameter* for the incomplete U-statistic.

The reader may wonder that generating $|I_{n,r}| \approx n^r$ Bernoulli random variables is computationally demanding, but there is no need to do so. In fact, we can equivalently compute the randomized incomplete U-statistic in (3) as follows.

1. Generate $\widehat{N} \sim \mathsf{Bin}(|I_{n,r}|, p_n)$.
2. Choose indices $\iota_1, \ldots, \iota_{\widehat{N}}$ randomly without replacement from $I_{n,r}$.
3. Compute $U'_{n,N} = \widehat{N}^{-1} \sum_{j=1}^{\widehat{N}} h(X_{\iota_j})$.

In fact, define $Z_\iota = 1$ if ι is one of $\iota_1, \ldots, \iota_{\widehat{N}}$, and $Z_\iota = 0$ otherwise; then, it is not difficult to see that $\{Z_\iota : \iota \in I_{n,r}\}$ are i.i.d. $\mathsf{Ber}(p_n)$ random variables. So, we can think of the Bernoulli sampling as a sampling without replacement with a random sample size.

Remark 2.1 (Comments on the random normalization)**.** Interestingly, changing the normalization in (3) *does* affect approximating distributions to the resulting incomplete U-statistic. Namely, if we change $\widehat{N}$ to N in (3), i.e., $\breve{U}'_{n,N} = N^{-1} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota)$, then we have different approximating distributions unless $\theta = 0$. In general, changing $\widehat{N}$ to N in (3) results in the approximating Gaussian distributions with larger covariance matrices, and hence it is recommended to use $U'_{n,N}$ rather than $\breve{U}'_{n,N}$. See also Remark 3.3 ahead.

2.2. **Sampling with replacement.** Conditionally on $X_1^n = (X_1, \ldots, X_n)$, let $X^*_{\iota_j}, j = 1, \ldots, N$ be i.i.d. draws from the empirical distribution $|I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} \delta_{X_\iota}$ (δ_{X_ι} denotes the point mass at X_ι). Let

$$U'_{n,N} = \frac{1}{N} \sum_{j=1}^{N} h(X^*_{\iota_j}) \tag{5}$$

be the incomplete U-statistic obtained by sampling with replacement. We call $U'_{n,N}$ the randomized incomplete U-statistic based on sampling with replacement. Note that $U'_{n,N}$ can be written as a weighted U statistic. Indeed, for each $\iota \in I_{n,r}$, let Z_ι denote the number of times that X_ι is redrawn in the sample $\{X^*_{\iota_1}, \ldots, X^*_{\iota_N}\}$. Then, the vector $Z = (Z_\iota)_{\iota \in I_{n,r}}$ (ordered in an arbitrary way) follows a multinomial distribution with parameters N and probabilities $1/|I_{n,r}|, \ldots, 1/|I_{n,r}|$ independent of X_1^n, and $U'_{n,N}$ can be written as

$$U'_{n,N} = \frac{1}{N} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota). \tag{6}$$

Hence, we can think of $U'_{n,N}$ as a statistic of $X_1, \ldots, X_n$ and $Z_\iota, \iota \in I_{n,r}$, but we will use both representations (5) and (6) interchangeably in the subsequent analysis.

Remark 2.2. All the theoretical results presented below apply to incomplete U-statistics based on either the Bernoulli sampling or sampling with replacement. Both sampling schemes will be covered in a unified way.

3. Gaussian approximations

In this section, we will derive Gaussian approximation results for the incomplete U-statistics (3) and (5) on the hyperrectangles in $\mathbb{R}^d$. Let $\mathcal{R}$ denote the class of (closed) hyperrectangles in $\mathbb{R}^d$, i.e., $\mathcal{R}$ consists sets of the form $\prod_{j=1}^d [a_j, b_j]$ where $-\infty \leqslant a_j \leqslant b_j \leqslant \infty$ for $j = 1, \ldots, d$ with the convention that $[a_j, b_j] = (-\infty, b_j]$ for $a_j = -\infty$ and $[a_j, b_j] = [a_j, \infty)$ for $b_j = \infty$. For the expository purpose, we mainly focus on the non-degenerate case where $\min_{1 \leqslant j \leqslant d} \mathrm{Var}(\mathbb{E}[h_j(X_1, \ldots, X_r) \mid X_1])$ is bounded away from zero in the following discussion. However, our Gaussian approximation results also cover the degenerate case (cf. Theorem 3.3).

Before stating the formal results, we begin with the intuition behind the Gaussian approximation results. Consider the Bernoulli sampling. Decompose the difference $U'_{n,N} - \theta$ as

$$U'_{n,N} - \theta = \frac{N}{\widehat{N}} \cdot \frac{1}{N} \sum_{\iota \in I_{n,r}} Z_\iota \{h(X_\iota) - \theta\} = \frac{N}{\widehat{N}} (A_n + \sqrt{1 - p_n} B_n),$$

where A_n and B_n are defined by

$$A_n = U_n - \theta \quad \text{and} \quad B_n = \frac{1}{N} \sum_{\iota \in I_{n,r}} \frac{(Z_\iota - p_n)}{\sqrt{1 - p_n}} \{h(X_\iota) - \theta\}.$$

For the notational convenience, we write $W_n = A_n + \sqrt{1-p_n} B_n$. For any hyperrectangle $R \in \mathcal{R}$, observe that

$$\mathbb{P}(\sqrt{n} W_n \in R) = \mathbb{P}\left(\sqrt{N} B_n \in \left[\frac{1}{\sqrt{\alpha_n(1-p_n)}} R - \sqrt{\frac{N}{1-p_n}} A_n \right] \right),$$

where $\alpha_n = n/N$. Since A_n is $\sigma(X_1^n)$-measurable, conditionally on X_1^n, A_n can be treated as a constant. On the other hand, conditionally on X_1^n, $\sqrt{N} B_n = (N(1-p_n))^{-1/2} \sum_{\iota \in I_{n,r}} (Z_\iota - p_n)\{h(X_\iota) - \theta\} = |I_{n,r}|^{-1/2} \sum_{\iota \in I_{n,r}} (p_n(1-p_n))^{-1/2} (Z_\iota - p_n)\{h(X_\iota) - \theta\}$ is the sum of independent random vectors with mean zero whose (conditional) covariance matrix is

$$\frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} \{h(X_\iota) - \theta\}\{h(X_\iota) - \theta\}^T =: \widehat{\Gamma}_h.$$

Subject to suitable moment conditions, $\widehat{\Gamma}_h$ can be approximated by $\Gamma_h := P^r(h-\theta)(h-\theta)^T$ under the max-norm. Hence, letting $\gamma_B = N(0, \Gamma_h)$, we have

$$\mathbb{P}_{|X_1^n}\left(\sqrt{N} B_n \in \left[\frac{1}{\sqrt{\alpha_n(1-p_n)}} R - \sqrt{\frac{N}{1-p_n}} A_n \right] \right) \approx \gamma_B \left(\left[\frac{1}{\sqrt{\alpha_n(1-p_n)}} R - \sqrt{\frac{N}{1-p_n}} A_n \right] \right),$$

and by Fubini,

$$\mathbb{P}\left(\sqrt{N} B_n \in \left[\frac{1}{\sqrt{\alpha_n(1-p_n)}} R - \sqrt{\frac{N}{1-p_n}} A_n \right] \right) \approx \mathbb{E}\left[\gamma_B \left(\left[\frac{1}{\sqrt{\alpha_n(1-p_n)}} R - \sqrt{\frac{N}{1-p_n}} A_n \right] \right) \right].$$

The right hand side of the last expression can be written as

$$\mathbb{P}\left(\sqrt{1-p_n} Y_B \in [\alpha_n^{-1/2} R - \sqrt{N} A_n] \right) = \mathbb{P}\left(\sqrt{n} A_n \in [R - \sqrt{\alpha_n(1-p_n)} Y_B] \right)$$

for $Y_B \sim N(0, \Gamma_h)$ independent of X_1^n. Next, since U_n is a complete U-statistic, conditionally on Y_B, we have (cf. [8])

$$\mathbb{P}_{|Y_B}\left(\sqrt{n} A_n \in [R - \sqrt{\alpha_n(1-p_n)} Y_B] \right) \approx \gamma_A \left([R - \sqrt{\alpha_n(1-p_n)} Y_B] \right),$$

where $\gamma_A = N(0, r^2 \Gamma_g)$ and $\Gamma_g = P(g-\theta)(g-\theta)^T$ with $g = P^{r-1} h$. Hence,

$$\mathbb{P}(\sqrt{n} W_n \in R) \approx \mathbb{E}\left[\gamma_A \left([R - \sqrt{\alpha_n(1-p_n)} Y_B] \right) \right] = \mathbb{P}\left(Y_A + \sqrt{\alpha_n(1-p_n)} Y_B \in R \right)$$

for $Y_A \sim N(0, r^r \Gamma_g)$ independent of Y_B. Since $Y_A + \sqrt{\alpha_n(1-p_n)} Y_B \sim N(0, r^2 \Gamma_g + \alpha_n(1-p_n)\Gamma_h)$, $\alpha_n p_n = n/|I_{n,r}| \approx 0$, and $N/\widehat{N} \approx 1$, it is expected that the distribution of $\sqrt{n}(U'_{n,N} - \theta)$ can be approximated by $N(0, r^2 \Gamma_g + \alpha_n \Gamma_h)$ on the hyperrectangles. Similar arguments carry through the sampling with replacement case as well.

Now, we turn to stating the formal Gaussian approximation results. We assume the following conditions. Let $\underline{\sigma} > 0$ and $D_n \geqslant 1$ be given constants, and recall that $g = (g_1, \ldots, g_d)^T = P^{r-1} h$. Suppose that

(C1) $P^r |h_j|^{2+k} \leqslant D_n^k$ for all $j = 1, \ldots, d$ and $k = 1, 2$.

(C2) $\|h_j(X_1^r)\|_{\psi_1} \leqslant D_n$ for all $j = 1, \ldots, d$.

In addition, suppose that either one of the following conditions holds:

(C3-ND) $P(g_j - \theta_j)^2 \geqslant \underline{\sigma}^2$ for all $j = 1, \ldots, d$.

(C3-D) $P^r(h_j-\theta_j)^2 \geqslant \underline{\sigma}^2$ for all $j=1,\dots,d$.

Conditions (C1) and (C2) are adapted from [12] and [8]. Condition (C2) assumes the kernel h to be sub-exponential, which in particular covers the bounded kernel. In principle, it is possible to extend our analysis under milder moment conditions on the kernel h, but this would result in more involved error bounds. For the sake of clear presentation, we work with Condition (C2). By Jensen's inequality, Conditions (C1) and (C2) imply that $P|g_j|^{2+k} \leqslant D_n^k$ for all j and for $k=1,2$, and $\|g_j(X_1)\|_{\psi_1} \leqslant D_n$ for all j. In addition, Condition (C1) implies that $P^r h_j^2 \leqslant 1+P|h_j|^3 \leqslant 1+D_n$ for all j. Condition (C3-ND) implies that the kernel h is non-degenerate. In the degenerate case, we will require Condition (C3-D) to derive Gaussian approximations.

In all what follows, we assume that

$$p_n = N/|I_{n,r}| \leqslant 1/2$$

without further mentioning. The value $1/2$ has no special meaning; we can allow $p_n \leqslant c$ for any constant $c \in (0,1)$, and in that case, the constants appearing in the following theorems depend in addition on c. Since we are using randomization for the purpose of computational reduction, we are mainly interested in the case where $N \ll |I_{n,r}|$, and the assumption that p_n is bounded away from 1 is immaterial.

The following theorem derives bounds on the Gaussian approximation to the randomized incomplete U-statistics on the hyperrectangles in the case where the kernel h is non-degenerate. Recall that $\alpha_n = n/N, p_n = N/|I_{n,r}|, \theta = P^r h = Pg, \Gamma_g = P(g-\theta)(g-\theta)^T$, and $\Gamma_h = P^r(h-\theta)(h-\theta)^T$.

Theorem 3.1 (Gaussian approximation under non-degeneracy)**.** *Suppose that Conditions (C1), (C2), and (C3-ND) hold. Then there exists a constant C depending only on $\underline{\sigma}$ and r such that*

$$\begin{aligned}
&\sup_{R\in\mathcal{R}} \left|\mathbb{P}\left\{\sqrt{n}(U'_{n,N}-\theta) \in R\right\} - \mathbb{P}(Y \in R)\right| \\
&= \sup_{R\in\mathcal{R}} \left|\mathbb{P}\{\sqrt{N}(U'_{n,N}-\theta) \in R\} - \mathbb{P}(\alpha_n^{-1/2} Y \in R)\right| \leqslant C\left(\frac{D_n^2 \log^7(dn)}{n \wedge N}\right)^{1/6},
\end{aligned} \tag{7}$$

where $Y \sim N(0, r^2\Gamma_g + \alpha_n \Gamma_h)$.

Theorem 3.1 shows that the distribution of $\sqrt{n}(U'_{n,N}-\theta)$ can be approximated by the Gaussian distribution $N(0, r^2\Gamma_g+\alpha_n\Gamma_h)$ on the hyperrectangles provided that $D_n^2\log^7(dn) \ll n \wedge N$, from which we deduce that the Gaussian approximation on the hyperrectangles holds for $U'_{n,N}$ even when $d \gg n$. In the cases where $N \gg n$ (i.e., $\alpha_n \ll 1$) and $N \ll n$ (i.e, $\alpha_n \gg 1$), the approximating distribution can be simplified to $N(0, r^2\Gamma_g)$ and $N(0,\Gamma_h)$, respectively.

Corollary 3.2. *Suppose that Conditions (C1), (C2), and (C3-ND) hold. Then there exists a constant C depending only on $\underline{\sigma}$ and r such that*

$$\sup_{R\in\mathcal{R}} \left|\mathbb{P}\left\{\sqrt{n}(U'_{n,N}-\theta)\in R\right\} - \gamma_A(R)\right| \leqslant C\left\{\left(\frac{nD_n\log^2 d}{N}\right)^{1/3} + \left(\frac{D_n^2\log^7(dn)}{n\wedge N}\right)^{1/6}\right\},$$

where $\gamma_A = N(0, r^2\Gamma_g)$, and

$$\sup_{R\in\mathcal{R}} \left|\mathbb{P}\left\{\sqrt{N}(U'_{n,N}-\theta)\in R\right\} - \gamma_B(R)\right| \leqslant C\left\{\left(\frac{ND_n\log^2 d}{n}\right)^{1/3} + \left(\frac{D_n^2\log^7(dn)}{n\wedge N}\right)^{1/6}\right\},$$

where $\gamma_B = N(0, \Gamma_h)$.

Remark 3.1 (Comments on the computational and statistical trade-off for the randomized incomplete U-statistics with non-degenerate kernels)**.** Theorem 3.1 and Corollary 3.2 reveal an interesting phase transition phenomenon between the computational complexity and the statistical efficiency for the randomized incomplete U-statistics. Suppose that $n \wedge N \gg D_n^2 \log^7(dn)$ and $\underline{\sigma}$ is bounded away from zero. First, if the computational budget parameter N is *superlinear* in the sample size n (i.e., $N \gg nD_n \log^2 d$), then both the incomplete U-statistic $\sqrt{n}(U'_{n,N} - \theta)$ and its complete version $\sqrt{n}(U_n - \theta)$ can be approximated by the same Gaussian distribution $\gamma_A = N(0, r^2\Gamma_g)$ (cf. [8] for $r = 2$ case). Second, if N is on the same order of n, then the scaling factor of $U'_{n,N}$ remains the same as U_n, namely, $\sqrt{n}$. However, the approximating Gaussian distribution for $\sqrt{n}(U'_{n,N} - \theta)$ has a larger covariance matrix than that for $\sqrt{n}(U_n - \theta)$ in the sense that $\alpha_n\Gamma_h$ is positive semi-definite. In this case, we sacrifice the statistical efficiency for the sake of keeping the computational cost linear in n. Third, if we further reduce the computational budget parameter N to be *sublinear* in n (i.e., $N \ll n/(D_n \log^2 d)$), then the scaling factor of $U'_{n,N}$ changes from $\sqrt{n}$ to $\sqrt{N}$, and the distribution of $U'_{n,N}$ is approximated by $N(\theta, N^{-1}\Gamma_h)$ on the hyperrectangles. Hence, the decay rate of the covariance matrix of the approximating Gaussian distribution is now N^{-1}, which is slower than the n^{-1} rate for the previous two cases.

Next, we consider the case where the kernel h is degenerate, i.e., $P(g_j - \theta_j)^2 = 0$ for all $j = 1, \ldots, d$. We consider the case where the kernel h is degenerate of order $k-1$ for some $k = 2, \ldots, r$, i.e., $P^{r-k+1}h(x_1, \ldots, x_{k-1}) = P^r h$ for all $(x_1, \ldots, x_{k-1}) \in S^{k-1}$. Even in such cases, a Gaussian approximation holds for $\sqrt{N}(U'_{n,N} - \theta)$ on the hyperrectangles provided that $N \ll n^k$ up to log factors. More precisely, we obtain the following theorem.

Theorem 3.3 (Gaussian approximation under degeneracy)**.** *Suppose the kernel h is degenerate of order $k-1$ for some $k = 2, \ldots, r$. In addition, suppose that Conditions (C1), (C2), and (C3-D) hold. Then there exists a constant C depending only on $\underline{\sigma}$ and r such that*

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}\left\{ \sqrt{N}(U'_{n,N} - \theta) \in R \right\} - \gamma_B(R) \right| \leqslant C \left\{ \left(\frac{N D_n^2 \log^{k+3} d}{n^k} \right)^{1/4} + \left(\frac{D_n^2 (\log n) \log^5(dn)}{n} \right)^{1/6} + \left(\frac{D_n^2 \log^7(dn)}{N} \right)^{1/6} \right\}, \tag{8}$$

where $\gamma_B = N(0, \Gamma_h)$.

Remark 3.2 (Comments on the Gaussian approximation under degeneracy)**.** In the degenerate case, for the Gaussian approximation to hold, we must have $N \ll n^k$ (more precisely, $N \ll n^k/(D_n^2 \log^{k+3} d)$), which is an indispensable condition even for the $d = 1$ case. To see this, consider the Bernoulli sampling case (similar arguments apply to the sampling with replacement case) and recall that $\sqrt{N}W_n = \sqrt{N}A_n + \sqrt{N(1-p_n)}B_n$, where $A_n = U_n - \theta$ and $B_n = U'_{n,N} - U_n$. According to Theorem 12.10 in [38], $n^{k/2}A_n$ converges in distribution to a Gaussian chaos of order k. Hence, in order to approximate $\sqrt{N}(U'_{n,N} - \theta) \approx \sqrt{N}W_n$ by a Gaussian distribution, it is necessary that $\sqrt{N}A_n$ is stochastically vanishing, which leads to the condition $N \ll n^k$.

It is worth noting that Theorem 3.3 reveals a fundamental difference between complete and randomized incomplete U-statistics with the degenerate kernel. Namely, in the degenerate case, the complete U-statistic $n^{k/2}(U_n - \theta)$ is known to have a non-Gaussian limiting distribution when d is fixed, while thanks to the randomizations, our incomplete U-statistics $\sqrt{N}(U'_{n,N} - \theta)$ can be approximated by the Gaussian distribution, and in addition the Gaussian approximation can hold even when $d \gg n$. On one hand, the rate of convergence of the incomplete U-statistics is $N^{-1/2}$ and is slower than that of the complete U-statistic, namely, $n^{-k/2}$. So in that sense we are sacrificing the rate of convergence by using the incomplete U-statistics instead of the complete U-statistic, although the rate $N^{-1/2}$ can be arbitrarily close to $n^{-k/2}$ up to log factors. On the other hand, the approximating Gaussian distribution for the incomplete U-statistics is easy to estimate by using a multiplier bootstrap developed in Section 4. The multiplier bootstrap developed in Section 4 is computationally much less demanding than e.g., the empirical bootstraps for complete (degenerate) U-statistics [cf. 6, 1], and can consistently estimate the approximating Gaussian distribution γ_B on the hyperrectangles even when $d \gg n$; see Theorem 4.1. To the best of our knowledge, there is no existing work that formally derives Gaussian chaos approximations to degenerate U-statistics in high dimensions where $d \gg n$, and in addition such non-Gaussian approximating distributions appear to be more difficult to estimate in high dimensions. Hence, in the degenerate case, the randomizations not only reduce the computational cost but also provide more tractable alternatives to make statistical inference on θ in high dimensions.

Remark 3.3 (Effect of deterministic normalization in the Bernoulli sampling case)**.** In the Bernoulli sampling case, consider the deterministic normalization, i.e, $\breve{U}'_{n,N} = N^{-1} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota)$, instead of the random one, i.e., $U'_{n,N} = \widehat{N}^{-1} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota)$. Then, in the non-degenerate case, the distribution of $\sqrt{n}(\breve{U}'_{n,N} - \theta)$ can be approximated by $N(0, r^2\Gamma_g + \alpha_n P^r hh^T)$, and in the degenerate case, $\sqrt{N}(\breve{U}'_{n,N} - \theta)$ can be approximated by $N(0, P^r hh^T)$ (provided that $N \ll n^k$ for the degenerate case). To see this, observe that $\breve{U}'_{n,N} - \theta = (U_n - \theta) + N^{-1} \sum_{\iota \in I_{n,r}} (Z_\iota - p_n) h(X_\iota)$, and the distribution of $N^{-1} \sum_{\iota \in I_{n,r}} (Z_\iota - p_n) h(X_\iota)$ can be approximated by $N(0, (1 - p_n) P^r hh^T)$. Since $P^r hh^T$ is larger than Γ_h unless $\theta = 0$ (in the sense that $P^r hh^T - \Gamma_h = \theta\theta^T$ is positive semi-definite), the approximating Gaussian distributions have larger covariance matrices for $\breve{U}'_{n,N}$ than those for $U'_{n,N}$, and hence it is in general recommended to use the random normalization rather than the deterministic one.

Remark 3.4 (Comparisons with [26] for $d = 1$)**.** The Gaussian approximation results established in Theorems 3.1, 3.3, and Corollary 3.2 can be considered as (partial) extensions of Theorem 1 and Corollary 1 in [26] to high dimensions. [26] focuses on the univariate case ($d = 1$) and derives the asymptotic distributions of randomized incomplete U-statistics based on sampling without replacement, sampling with replacement, and Bernoulli sampling ([26] considers the deterministic normalization for the Bernoulli sampling case). For the illustrative purpose, consider sampling with replacement. Suppose that $p_n \to p \in [0, 1]$ and the kernel h is degenerate of order $k - 1$ for some $k = 1, \ldots, r$ (the $k = 1$ case corresponds to a non-degenerate kernel). Then Theorem 1 in [26] shows that $(n^{k/2}(U_n - \theta), N^{1/2}(U'_{n,N} - U_n)) \xrightarrow{d} (V, W)$, where V is a Gaussian chaos of order k (in particular, $V \sim N(0, r^2 P(g - \theta)^2)$ if $k = 1$) and $W \sim N(0, P^r(h - \theta)^2)$ such that V

and W are independent. Hence, provided that $n^k/N \to \alpha \in [0,\infty]$, $n^{k/2}(U'_{n,N} - \theta) \xrightarrow{d} V + \alpha W$ if $\alpha < \infty$ and $\sqrt{N}(U'_{n,N} - \theta) \xrightarrow{d} W$ if $\alpha = \infty$. The present paper focuses on the cases where the approximating distributions are Gaussian (i.e., the cases where $k = 1$ and α is finite, or $k \geqslant 2$ and $\alpha = \infty$), but covers high-dimensional kernels and derives explicit and non-asymptotic Gaussian approximation error bounds that are not obtained in [26]. In addition, the proof strategy of our Gaussian approximation results differs substantially from that of [26]. [26] shows the convergence of the joint characteristic function of $(n^{k/2}(U_n - \theta), N^{1/2}(U'_{n,N} - U_n))$ to obtain his Theorem 1, but the characteristic function approach is not very useful to derive explicit error bounds on distributional approximations in high dimensions. Instead, our proofs iteratively use conditioning arguments combined with Berry-Esseen type bounds, as briefly discussed in the beginning of this section.

Finally, we expect that the results of the present paper can be extended to the case where $k \geqslant 2$ and α is finite; in that case, the approximating distribution to $n^{k/2}(U'_{n,N} - \theta)$ will be non-Gaussian and the technical analysis will be more involved in high dimensions. We leave the analysis of this case as a future research topic.

4. Bootstrap approximations

The Gaussian approximation results developed in the previous section are often not directly applicable in statistical applications since the covariance matrix of the approximating Gaussian distribution, $r^2\Gamma_g + \alpha_n\Gamma_h$ (or Γ_h in the degenerate case), is unknown to us. In this section, we develop data-dependent procedures to further approximate or estimate the $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$ distribution (or the $N(0,\Gamma_h)$ distribution in the degenerate case) that are computationally (much) less-demanding than existing bootstrap methods for U-statistics such as the empirical bootstrap.

4.1. Generic bootstraps for incomplete U-statistics.

Let $\mathcal{D}_n = \{X_1, \dots, X_n\} \cup \{Z_\iota : \iota \in I_{n,r}\}$. For the illustrative purpose, consider to estimate the $N(0, r^2\Gamma + \alpha_n\Gamma_h)$ distribution and let $Y \sim N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$. The basic idea of our approach is as follows. Since $Y \overset{d}{=} Y_A + \alpha_n^{1/2} Y_B$ where $Y_A \sim N(0, r^2\Gamma_g)$ and $Y_B \sim N(0,\Gamma_h)$ are independent, to approximate the distribution of Y, it is enough to construct data-dependent random vectors $U^\sharp_{n,A}$ and $U^\sharp_{n,B}$ such that, conditionally on $\mathcal{D}_n$, (i) $U^\sharp_{n,A}$ and $U^\sharp_{n,B}$ are independent, and (ii) the conditional distributions of $U^\sharp_{n,A}$ and $U^\sharp_{n,B}$ are computable and "close" enough to $N(0, r^2\Gamma_g)$ and $N(0,\Gamma_h)$, respectively. Then, the conditional distribution of $U^\sharp_n = U^\sharp_{n,A} + \alpha_n^{1/2} U^\sharp_{n,B}$ should be close to $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$ and hence to the distribution of $\sqrt{n}(U'_{n,N} - \theta)$. Of course, if the target distribution is $N(0, r^2\Gamma_g)$ or $N(0,\Gamma_h)$, then it is enough to simulate the conditional distribution of $U^\sharp_{n,A}$ or $U^\sharp_{n,B}$ alone, respectively.

Construction of $U^\sharp_{n,B}$ is straightforward; in fact, it is enough to apply the (Gaussian) multiplier bootstrap to $\sqrt{Z_\iota}h(X_\iota), \iota \in I_{n,r}$.

Construction of $U^\sharp_{n,B}$.

1. Generate i.i.d. $N(0,1)$ variables $\{\xi'_\iota : \iota \in I_{n,r}\}$ independent of the data $\mathcal{D}_n$.
2. Construct
$$U^\sharp_{n,B} = \frac{1}{\sqrt{\widehat{N}}} \sum_{\iota \in I_{n,r}} \xi'_\iota \sqrt{Z_\iota}\{h(X_\iota) - U'_{n,N}\},$$
where $\widehat{N}$ is replaced by N for the sampling with replacement case.

In the Bernoulli sampling case, $U^{\sharp}_{n,B}$ reduces to $U^{\sharp}_{n,B} = \widehat{N}^{-1/2}\sum_{j=1}^{\widehat{N}} \xi'_{\iota_j}\{h(X_{\iota_j}) - U'_{n,N}\}$, while in the sampling with replacement case, simulating $U^{\sharp}_{n,B}$ can be equivalently implemented by simulating $U^{\sharp}_{n,B} = N^{-1/2}\sum_{j=1}^{N} \eta_j\{h(X^*_{\iota_j}) - U'_{n,N}\}$ for $\eta_1,\dots,\eta_N \sim N(0,1)$ i.i.d. independent of $X^*_{\iota_1},\dots,X^*_{\iota_N}$; in fact the distribution of $U^{\sharp}_{n,B}$ in the latter definition (conditionally on $X^*_{\iota_1},\dots.X^*_{\iota_N}$) is Gaussian with mean zero and covariance matrix $N^{-1}\sum_{j=1}^{N}\{h(X^*_{\iota_j}) - U'_{n,N}\}\{h(X^*_{\iota_j}) - U'_{n,N}\}^T$, which is identical to the conditional distribution of $U^{\sharp}_{n,B}$ in the original definition. In either case, in practice, we only need to generate (on average) N multiplier variables. The following theorem establishes conditions under which the conditional distribution of $U^{\sharp}_{n,B}$ is able to consistently estimate the $N(0,\Gamma_h)$ $(=\gamma_B)$ distribution on the hyperrectangles with polynomial error rates.

Theorem 4.1 (Validity of $U^{\sharp}_{n,B}$). *Suppose that Conditions (C1), (C2), and (C3-D) hold. In addition, suppose that*

$$\frac{D_n^2 \log^5(dn)}{n \wedge N} \leqslant C_1 n^{-\zeta} \tag{9}$$

for some constants $0 < C_1 < \infty$ *and* $\zeta \in (0,1)$. *Then there exists a constant* C *depending only on* $\underline{\sigma}, r,$ *and* C_1 *such that*

$$\sup_{R\in\mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(U^{\sharp}_{n,B} \in R) - \gamma_B(R) \right| \leqslant C n^{-\zeta/8}$$

with probability at least $1 - Cn^{-\zeta/8}$.

In the degenerate case, the approximating distribution is $\gamma_B = N(0,\Gamma_h)$. So, in that case, we can approximate the distribution of $\sqrt{N}(U'_{n,N} - \theta)$ on the hyperrectangles by the conditional distribution of $U^{\sharp}_{n,B}$, which can be simulated by drawing multiplier variables many times. We call simulating $U^{\sharp}_{n,B}$ the *multiplier bootstrap under degeneracy* (MB-DG). On average, the computational cost of the MB-DG is $O(BNd)$ where B denotes the number of bootstrap iterations, which can be independent of the order of the U-statistic provided that N is so. In the remainder of this section, we will focus on the non-degenerate case.

In contrast to $U^{\sharp}_{n,B}$, construction of $U^{\sharp}_{n,A}$ is more involved. We might be tempted to apply the multiplier bootstrap to the Hájek projection, $rn^{-1}\sum_{i_1=1}^{n} g(X_{i_1})$, but the function $g = P^{r-1}h$ is unknown so the direct application of the multiplier bootstrap to the Hájek projection is infeasible. Instead, we shall construct estimates of $g(X_{i_1})$ for $i_1 \in \{1,\dots,n\}$ or a subset of $\{1,\dots,n\}$, and then apply the multiplier bootstrap to the estimated Hájek projection. Generically, construction of $U^{\sharp}_{n,A}$ is as follows.

Generic construction of $U^{\sharp}_{n,A}$.

1. Choose a subset S_1 of $\{1,\dots,n\}$ and generate i.i.d. $N(0,1)$ variables $\{\xi_{i_1} : i_1 \in S_1\}$ independent of the data $\mathcal{D}_n$ and $\{\xi'_\iota : \iota \in I_{n,r}\}$. Let $n_1 = |S_1|$.
2. For each $i_1 \in S_1$, construct an estimate $\widehat{g}^{(i_1)}$ of g based on X_1^n.
3. Construct

$$U^{\sharp}_{n,A} = \frac{r}{\sqrt{n_1}} \sum_{i_1 \in S_1} \xi_{i_1}\{\widehat{g}^{(i_1)}(X_{i_1}) - \breve{g}\},$$

where $\breve{g} = n_1^{-1}\sum_{i_1\in S_1} \widehat{g}^{(i_1)}(X_{i_1})$.

Step 1 chooses a subset S_1 to reduce the computational cost of the resulting bootstrap. Construction of estimates $\widehat{g}^{(i_1)}, i_1 \in S_1$ can be flexible. For instance, the estimates $\widehat{g}^{(i_1)}, i_1 \in S_1$ may depend on another randomization independent of everything else. In Sections 4.2 and 4.3, we will consider deterministic and random constructions of $\widehat{g}^{(i_1)}, i_1 \in S_1$, respectively.

It is worth noting that the jackknife multiplier bootstrap (JMB) developed in [8] (for the $r = 2$ case) and [9] (for the general r case) is a special case of $U_{n,A}^{\sharp}$ where $S_1 = \{1, \ldots, n\}$ and $\widehat{g}^{(i_1)}(X_{i_1})$ is realized by its jackknife estimate, i.e., by the U-statistic with kernel $(x_2, \ldots, x_r) \mapsto h(X_{i_1}, x_2, \ldots, x_r)$ for the sample without the i_1-th observation. Nevertheless, the bottleneck is that the computation of the jackknife estimates of $g(X_{i_1}), i_1 = 1, \ldots, n$ requires $O(n^r d)$ operations and hence implementing the JMB can be computationally demanding.

Now, consider $U_n^{\sharp} = U_{n,A}^{\sharp} + \alpha_n^{1/2} U_{n,B}^{\sharp}$. We call simulating $U_n^{\sharp}$ the *multiplier bootstrap under non-degeneracy* (MB-NDG). The following theorem establishes conditions under which the conditional distribution of $U_n^{\sharp}$ is able to consistently estimate the $N(0, r^2\Gamma_g + \alpha_n \Gamma_h)$ distribution on the hyperrectangles with polynomial error rates. Define

$$\widehat{\Delta}_{A,1} := \max_{1 \leqslant j \leqslant d} \frac{1}{n_1} \sum_{i_1 \in S_1} \{\widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1})\}^2,$$

which quantifies the errors of the estimates $\widehat{g}^{(i_1)}, i_1 \in S_1$.

Theorem 4.2 (Generic bootstrap validity under non-degeneracy)**.** *Let $U_n^{\sharp} = U_{n,A}^{\sharp} + \alpha_n^{1/2} U_{n,B}^{\sharp}$. Suppose that Conditions (C1), (C2), and (C3-ND) hold. In addition, suppose that*

$$\frac{D_n^2 \log^5(dn)}{n_1 \wedge N} \leqslant C_1 n^{-\zeta} \quad \text{and} \quad \mathbb{P}\left(\overline{\sigma}_g^2 \widehat{\Delta}_{A,1} \log^4 d > C_1 n^{-3\zeta/4}\right) \leqslant C_1 n^{-\zeta/8} \tag{10}$$

for some constants $0 < C_1 < \infty$ and $\zeta \in (0,1)$, where $\overline{\sigma}_g := \max_{1 \leqslant j \leqslant d} \sqrt{P(g_j - \theta_j)^2}$. Then there exists a constant C depending only on $\underline{\sigma}, r,$ and C_1 such that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(U_n^{\sharp} \in R) - \mathbb{P}(Y \in R) \right| \leqslant C n^{-\zeta/8} \tag{11}$$

with probability at least $1 - Cn^{-\zeta/8}$, where $Y \sim N(0, r^2\Gamma_g + \alpha_n \Gamma_h)$. If the estimates $g^{(i_1)}, i_1 \in S_1$ depend on an additional randomization independent of $\mathcal{D}_n, \{\xi_{i_1} : i_1 \in S_1\}$, and $\{\xi'_\iota : \iota \in I_{n,r}\}$, then the result (11), with $\mathcal{D}_n$ replaced by the augmentation of $\mathcal{D}_n$ with variables used in the additional randomization, holds with probability at least $1 - Cn^{-\zeta/8}$.

The second part of Condition (10) is a high-level condition on the estimation accuracy of $\widehat{g}^{(i_1)}, i_1 \in S_1$. In Sections 4.2 and 4.3, we will verify the second part of Condition (10) for deterministic and random constructions of $\widehat{g}^{(i_1)}, i_1 \in S_1$. Note that the bootstrap distribution is taken with respect to the multiplier variables $\{\xi_{i_1} : i_1 \in S_1\}$ and $\{\xi'_\iota : \iota \in I_{n,r}\}$, and so if the estimation step for g depends on an additional randomization, then the variables used in the additional randomization have to be generated outside the bootstrap iterations.

In the case where the approximating distribution can be simplified to $\gamma_A = N(0, r^2\Gamma_g)$, then it is sufficient to estimate $N(0, r^2\Gamma_g)$ by the conditional distribution of $U_{n,A}^{\sharp}$.

Corollary 4.3 (Validity of $U^\sharp_{n,A}$)**.** *Suppose that all the conditions in Theorem 4.2 hold. Then there exists a constant C depending only on $\underline{\sigma}, r,$ and C_1 such that*

$$\sup_{R\in\mathcal{R}}\left|\mathbb{P}_{|\mathcal{D}_n}(U^\sharp_{n,A}\in R)-\gamma_A(R)\right|\leqslant Cn^{-\zeta/8} \tag{12}$$

with probability at least $1-Cn^{-\zeta/8}$. If the estimates $g^{(i_1)}, i_1\in S_1$ depend on an additional randomization independent of $\mathcal{D}_n, \{\xi_{i_1}: i_1\in S_1\}$, and $\{\xi'_\iota:\iota\in I_{n,r}\}$, then the result (12), with $\mathcal{D}_n$ replaced by the augmentation of $\mathcal{D}_n$ with variables used in the additional randomization, holds with probability at least $1-Cn^{-\zeta/8}$.

Remark 4.1 (Comments on the partial bootstrap simplification under non-degeneracy)**.** When the distribution of $\sqrt{N}(U'_{n,N}-\theta)$ can be simplified to $\gamma_B=N(0,\Gamma_B)$, it is also possible to use the partial bootstrap $U^\sharp_{n,B}$ to estimate $N(0,\Gamma_B)$. In this case, we must take N to be sublinear in n (i.e., $N\ll n/(D_n\log^2 d)$) to ensure the Gaussian approximation validity (cf. Remark 3.1). However, we do not recommend this simplification because the decay rate of the covariance matrix of the approximating Gaussian distribution $N(\theta,N^{-1}\Gamma_B)$ to $U'_{n,N}$ is N^{-1}, which is slower than the n^{-1} rate for the linear and superlinear cases. In particular, this implies a power loss in the testing problems if the critical values are calibrated by $U^\sharp_{n,B}$.

The rest of this section is devoted to concrete constructions of estimates $\widehat{g}^{(i_1)}, i_1\in S_1$.

4.2. **Divide and conquer estimation.** We first propose a deterministic construction of $\widehat{g}^{(i_1)}, i_1\in S_1$ via the divide and conquer (DC) algorithm [43].

1. For each $i_1\in S_1$, choose K *disjoint* subsets $S^{(i_1)}_{2,k}, k=1,\dots,K$ with common size $L\geq r-1$ from $\{1,\dots,n\}\setminus\{i_1\}$.
2. For each $i_1\in S_1$, estimate g by computing U-statistics with kernel $(x_2,\dots,x_r)\mapsto h(x,x_2,\dots,x_r)$ applied to the subsamples $\{X_i: i\in S^{(i_1)}_{2,k}\}, k=1,\dots,K$, and taking the average of those U-statistics of order $r-1$, i.e.,

$$\widehat{g}^{(i_1)}(x)=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_{L,r-1}|}\sum_{\substack{i_2,\dots,i_r\in S^{(i_1)}_{2,k}\\ i_2<\cdots<i_r}}h(x,X_{i_2},\dots,X_{i_r}).$$

The DC algorithm can be viewed as an estimation procedure for g via incomplete U-statistics of order $r-1$ with a *block diagonal* sampling scheme (up to a permutation on the indices). We call simulating $U^\sharp_n$ with the DC algorithm as the MB-NDG-DC. In Section 4.3, we will propose a different estimation procedure for g via randomized incomplete U-statistics of order $r-1$ based on an additional Bernoulli sampling. As a practical guidance to implement the DC algorithm, we suggest to choose $S_1=\{1,\dots,n\}, L=r-1$, and $K=\lfloor(n-1)/L\rfloor$ consecutive blocks, which are the parameter values used in our simulation examples in Section 5. In this case, the DC algorithm turns out to be calculating Hoeffding's averages of the U-statistics of order $r-1$, which requires $O(nd)$ operations for each i_1. In contrast, the JMB constructs $\widehat{g}^{(i_1)}$ by complete U-statistics of order $r-1$, which requires $O(n^{r-1}d)$ operations for each i_1. Since the estimation step for g can be done outside the bootstrap iterations, the overall computational cost of the MB-NDG-DC is

$O((BN + n_1KL + Bn_1)d) = O(n^2d + B(N + n)d)$ (where B denotes the number of bootstrap iterations), which is independent of the order of the U-statistic. In addition, if we choose to only simulate $U^{\sharp}_{n,A}$, then the computational cost is $O(n^2d+Bnd)$, since the $O(BNd)$ computations came from simulating $U^{\sharp}_{n,B}$. We can certainly make the computational cost even smaller by taking n_1 and K smaller than n. For instance, if we choose n_1 and K in such a way that $n_1K = O(n)$ and $L = r-1$, then the overall computational cost is reduced to $O(nd+B(N+n)d) = O(B(N+n)d)$ (or $O(Bnd)$ if we only simulate $U^{\sharp}_{n,A}$). In general, choosing smaller n_1 and K would sacrifice the statistical accuracy of the resulting bootstrap, but if $O(n^2d)$ computations are difficult to implement, then choosing smaller n_1 and K would be a reasonable option.

Our MB-NDG-DC differs from the the Bag of Little Bootstraps (BLB) proposed in [27], which is another generically scalable bootstrap method for large datasets based on the DC algorithm. Specifically, tailored to the U-statistic $U_n := U_n^{(r)}(h)$ with kernel h, let $Q_n := Q_n(P)$ be the distribution of U_n and $\lambda(Q_n(P)) = \lambda(Q_n(P), P)$ be a quality assessment of U_n (cf. Chapter 6.5 in [29]). For instance, $\lambda(Q_n(P))$ can be the 95%-quantile of the distribution of $\max_{1\leqslant j\leqslant d} \sqrt{n}(U_{n,j}-\theta_j)$. A natural estimate of $\lambda(Q_n(P))$ is the plug-in estimate $\lambda(Q_n(\mathbb{P}_n))$, where $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution of $X_1, \dots, X_n$. Typically, $\lambda(Q_n(\mathbb{P}_n))$ is computationally difficult to compute, even for a moderate sample size n. The BLB first divides the original sample $\{X_1, \dots, X_n\}$ into K subsets $\mathcal{I}_1, \dots, \mathcal{I}_K$ of size L uniformly at random. Denote by $\mathbb{P}^{(k)}_{n,L} = L^{-1}\sum_{i\in\mathcal{I}_k} \delta_{X_i}$ the empirical distribution of $\{X_i\}_{i\in\mathcal{I}_k}$. Then, on each subset $\mathcal{I}_k, k = 1, \dots, K$, the BLB repeatedly resamples n points i.i.d. from $\mathbb{P}^{(k)}_{n,L}$, computes the U-statistic with kernel h for each resample, forms the empirical distribution $\mathbb{Q}^*_{n,k}$ of the computed U-statistics, and approximates $\lambda(Q_n(\mathbb{P}^{(k)}_{n,L}))$ by $\lambda(\mathbb{Q}^*_{n,k})$. Finally the BLB takes the average $K^{-1}\sum_{k=1}^K \lambda(\mathbb{Q}^*_{n,k})$ as an estimate of $\lambda(Q_n(P))$. The computational cost of the BLB is $O(BKL^rd) = O(BnL^{r-1}d)$. Note that asymptotic validity of the BLB requires that $L \to \infty$ (cf. Theorem 1 of [27]), so that $\mathbb{P}^{(k)}_{n,L}$ is close enough to P. Therefore, in order for the BLB to approach the population quality assessment value $\lambda(Q_n(P))$, its computational complexity has to depend on the order r of the U-statistic. On the contrary, our MB-NDG-DC applies the DC algorithm to estimation of the Hájek projection and the overall computational cost is $O(n^2d + B(N + n)d)$, which does not depend on r. In particular, the computational cost of the MB-NDG-DC is $O(n^2d + Bnd)$ if we choose N to be on the same order of n.

The following proposition provides conditions for validity of the multiplier bootstrap equipped with the DC estimation (MB-NDG-DC).

Proposition 4.4 (Validity of bootstrap with DC estimation)**.** *Suppose that Conditions (C1), (C2), and (C3-ND) hold. In addition, suppose that*

$$\frac{D_n^2\log^5(dn)}{n_1 \wedge N} \leqslant C_1 n^{-\zeta} \quad \textit{and} \quad \frac{\overline{\sigma}_g^2 D_n^2 \log^7 d}{KL}\left(1 + \frac{\log^2 d}{K^{1-1/q}}\right) \leqslant C_1 n^{-7\zeta/8} \tag{13}$$

for some constants $0 < C_1 < \infty, q \in [1,\infty)$, and $\zeta \in (0,1)$. Then, there exists a constant C depending only on $\underline{\sigma}, r, q$, and C_1 such that each of the results (11) and (12) holds with probability at least $1 - Cn^{-\zeta/8}$.

For instance, consider to take $N = n, S_1 = \{1, \dots, n\}, L = r - 1$, and $K = \lfloor (n-1)/L \rfloor$, and suppose that $D_n^2 \log^7(dn) \leqslant n^{1-\zeta}$ for some $\zeta \in (0,1)$. Then, by Theorem 3.1 and Proposition 4.4, there exists a constant C depending only on $\overline{\sigma}_g, \underline{\sigma}$, and r such that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}(U'_{n,N} - \theta) \in R) - \mathbb{P}_{|\mathcal{D}_n}(U_n^\sharp \in R) \right| \leqslant C n^{-\zeta/8} \tag{14}$$

with probability at least $1 - Cn^{-\zeta/8}$. Hence, the conditional distribution of the MB-NDG-DC approaches uniformly on the hyperrectangles in $\mathbb{R}^d$ to the distribution of the randomized incomplete U-statistic at a polynomial rate in the sample size.

4.3. **Random sampling estimation.** Next, we propose a random construction of $\widehat{g}^{(i_1)}, i_1 \in S_1$ based on an additional Bernoulli sampling. For each $i_1 = 1, \ldots, n$, let $I_{n-1,r-1}(i_1) = \{(i_2, \ldots, i_r) : 1 \leqslant i_2 < \cdots < i_r \leqslant n, i_j \neq i_1 \ \forall j \neq 1\}$. In addition, define $\sigma_{i_1} : \{1, \ldots, n-1\} \to \{1, \ldots, n\} \setminus \{i_1\}$ as follows: if $\{1, \ldots, n\} \setminus \{i_1\} = \{j_1, \ldots, j_{n-1}\}$ with $j_1 < \cdots < j_{n-1}$, then $\sigma_{i_1}(\ell) = j_\ell$ for $\ell = 1, \ldots, n-1$. For the notational convenience, for $\iota' = (i_2, \ldots, i_r) \in I_{n-1,r-1}$, we write $\sigma_{i_1}(\iota') = (\sigma_{i_1}(i_2), \ldots, \sigma_{i_1}(i_r)) \in I_{n-1,r-1}(i_1)$.

Now, consider the following randomized procedure to construct $\widehat{g}^{(i_1)}, i_1 \in S_1$.

1. Let $0 < M = M_n \leqslant |I_{n-1,r-1}|$ be a positive integer, and generate i.i.d. $\mathsf{Ber}(\vartheta_n)$ random variables $\{Z'_{\iota'} : \iota' = (i_2, \ldots, i_r) \in I_{n-1,r-1}\}$ independent of $\mathcal{D}_n, \{\xi_{i_1} : i_1 \in S_1\}$, and $\{\xi'_\iota : \iota \in I_{n,r}\}$, where $\vartheta_n = M/|I_{n-1,r-1}|$.
2. For each $i_1 \in S_1$, construct $\widehat{g}^{(i_1)}(x) = M^{-1} \sum_{\iota' \in I_{n-1,r-1}} Z'_{\iota'} h(x, X_{\sigma_{i_1}(\iota')})$.

The resulting bootstrap method is called the *multiplier bootstrap under non-degeneracy with random sampling* (MB-NDG-RS). Equivalently, the above procedure can be implemented as follows:

1. Generate $\widehat{M} \sim \mathsf{Bin}(|I_{n-1,r-1}|, \vartheta_n)$.
2. Sample $\iota'_1, \ldots, \iota'_{\widehat{M}}$ randomly without replacement from $I_{n-1,r-1}$.
3. Construct $\widehat{g}^{(i_1)}(x) = M^{-1} \sum_{j=1}^{\widehat{M}} h(x, X_{\sigma_{i_1}(\iota'_j)})$ for each $i_1 \in S_1$.

So, on average, the computational cost to construct $\widehat{g}^{(i_1)}, i_1 \in S_1$ by the random sampling estimation is $O(n_1 M d)$, and the overall computational cost of the MB-NDG-RS is $O(n_1 M d + B(N + n_1)d)$ (or $O(n_1 M d + B n_1 d)$ if we only simulate $U^\sharp_{n,A}$). As a practical guidance to implement the random sampling estimation, we suggest to choose $S_1 = \{1, \ldots, n\}$ and M proportional to $n - 1$, which are the parameter values used in our simulation examples in Section 5. Then the overall computational cost of the MB-NDG-RS is $O(n^2 d + B(N + n)d)$ (or $O(n^2 d + Bnd)$ if we only simulate $U^\sharp_{n,A}$), which is independent of the order of the U-statistic. In addition, the computational cost can be made even smaller, e.g., can be reduced to $O(B(N + n)d)$ by choosing n_1 and M in such a way that $n_1 M = O(n)$ (or $O(Bnd)$ if we only simulate $U^\sharp_{n,A}$), which would be a reasonable option if $O(n^2 d)$ computations are difficult to implement.

Proposition 4.5 (Validity of bootstrap with Bernoulli sampling estimation). *Suppose that Conditions (C1), (C2), and (C3-ND) hold. In addition, suppose that*

$$\frac{D_n^2 \log^5(dn)}{n_1 \wedge N} \leqslant C_1 n^{-\zeta} \quad \textit{and} \quad \frac{\overline{\sigma}_g^2 D_n^2 \log^7(dn)}{n \wedge M} \leqslant C_1 n^{-7\zeta/8} \tag{15}$$

for some constants $0 < C_1 < \infty$ *and* $\zeta \in (0, 1)$. *Then, there exists a constant* C *depending only on* $\underline{\sigma}, r$, *and* C_1 *such that each of the results (11) and (12), with* $\mathcal{D}_n$ *replaced by* $\mathcal{D}'_n = \mathcal{D}_n \cup \{Z'_{\iota'} : \iota' \in I_{n-1,r-1}\}$, *holds with probability at least* $1 - Cn^{-\zeta/8}$.

For instance, consider to take $N = n, S_1 = \{1, \dots, n\}$, and M proportional to $n-1$, and suppose that $D_n^2 \log^7(dn) \leqslant n^{1-\zeta}$ for some $\zeta \in (0,1)$. Then, by Theorem 3.1 and Proposition 4.5, the result (14) holds with probability at least $1 - Cn^{-\zeta/8}$.

Remark 4.2 (Alternative options for random sampling estimation)**.** In construction of $\widehat{g}^{(i_1)}$, instead of normalization by M, we may use normalization by $\widehat{M}$, namely, $\widehat{M}^{-1}\sum_{j=1}^{\widehat{M}} h(x, X_{\sigma_{i_1}(\iota'_j)})$ for $\widehat{g}^{(i_1)}(x)$. In view of the concentration inequality for $\widehat{M}$ (cf. equation (4)), it is not difficult to see that the same conclusion of Proposition 4.5 holds for $\widehat{g}^{(i_1)}(x) = \widehat{M}^{-1}\sum_{j=1}^{\widehat{M}} h(x, X_{\sigma_{i_1}(\iota'_j)})$.

Next, alternatively to the Bernoulli sampling, we may use sampling with replacement to construct $\widehat{g}^{(i_1)}$, which can be implemented as follows: 1) sample $\iota'_1, \dots, \iota'_M$ randomly with replacement from $I_{n-1,r-1}$ (independently of everything else); and 2) construct $\widehat{g}^{(i_1)}(x) = M^{-1}\sum_{j=1}^{M} h(x, X_{\sigma_{i_1}(\iota'_j)})$ for $i_1 \in S_1$. For each $i_1 \in S_1$, conditionally on X_1^n, $X_{\sigma_{i_1}(\iota'_j)}, j = 1, \dots, M$ are i.i.d. draws from the empirical distribution $|I_{n-1,r-1}|^{-1}\sum_{\iota' \in I_{n-1,r-1}(i_1)} \delta_{X_{\iota'}}$. Mimicking the proof of Proposition 4.5, it is not difficult to see that the conclusion of the proposition holds for the estimation of g via sampling with replacement under the condition (15) (here $Z'_{\iota'}$ is the number of times that ι' is redrawn in the sample $\{\iota'_1, \dots, \iota'_M\}$, for which $\widehat{g}^{(i_1)}(x)$ can be expressed as $\widehat{g}^{(i_1)}(x) = M^{-1}\sum_{\iota' \in I_{n-1,r-1}} Z'_{\iota'} h(x, X_{\sigma_{i_1}(\iota')})$).

5. Numerical examples

In this section, we provide some numerical examples to verify the validity of our proposed bootstrap algorithms (i.e., MB-DG, MB-NDG-DC, MB-NDG-RS) for approximating the distributions of incomplete U-statistics. In particular, we examine the statistical accuracy and computational running time of the bootstrap algorithms in the leading example of testing for the pairwise independence of a high-dimensional vector. We consider two test statistics: Spearman's ρ and Bergsma-Dassios' t^*. Under H_0 in (2), the leading term $\widehat{\rho}$ of Spearman's ρ is non-degenerate while Bergsma-Dassios' t^* is degenerate of order 1. Slightly abusing notation, we will use $\widehat{\rho}$ as Spearman's ρ statistic throughout this section. We consider tests of the forms

$$\max_{1 \leqslant j < k \leqslant p} |\widehat{\rho}_{j,k}| > c \Rightarrow \text{reject } H_0 \quad \text{and} \quad \max_{1 \leqslant j < k \leqslant p} |t^*_{j,k}| > c \Rightarrow \text{reject } H_0,$$

where the critical values are calibrated by the bootstrap methods. For Spearman's ρ, we use $U_n^\sharp$ for MB-NDG-DC and MB-NDG-RS. For Bergsma-Dassios' t^*, we use $U_{n,B}^\sharp$ for MB-DG. In addition, we also test the performance of the *partial* versions of MB-NDG-DC and MB-NDG-RS (i.e., $U_{n,A}^\sharp$; cf. Corollary 4.3) for Spearman's ρ statistic when its distribution can be approximated by $\gamma_A = N(0, r^2\Gamma_g)$ (cf. Corollary 3.2).

5.1. **Simulation setup.** We simulate i.i.d. data from the non-central t-distribution with $\nu = 3$ degrees of freedom and non-centrality parameter $\mu = 2$. This data generating process implies H_0. We consider $n = 300, 500, 1000$ and $p = 30, 50, 100$ (so the number of the free parameters is $d = p(p-1)/2 = 435, 1225, 4950$). For each setup (n,p), we fix the bootstrap sample size $B = 200$ and report the empirical rejection probabilities of the bootstrap tests averaged over 2,000 simulations. For Spearman's ρ, we apply the MB-NDG-DC and MB-NDG-RS (full version $U_n^\sharp$) and set the computational budget parameter value $N = 2n$. In addition, we implement the MB-NDG-DC with the parameter values suggested in Section 4.2 (i.e., $S_1 = \{1, \dots, n\}, L = r-1$, and

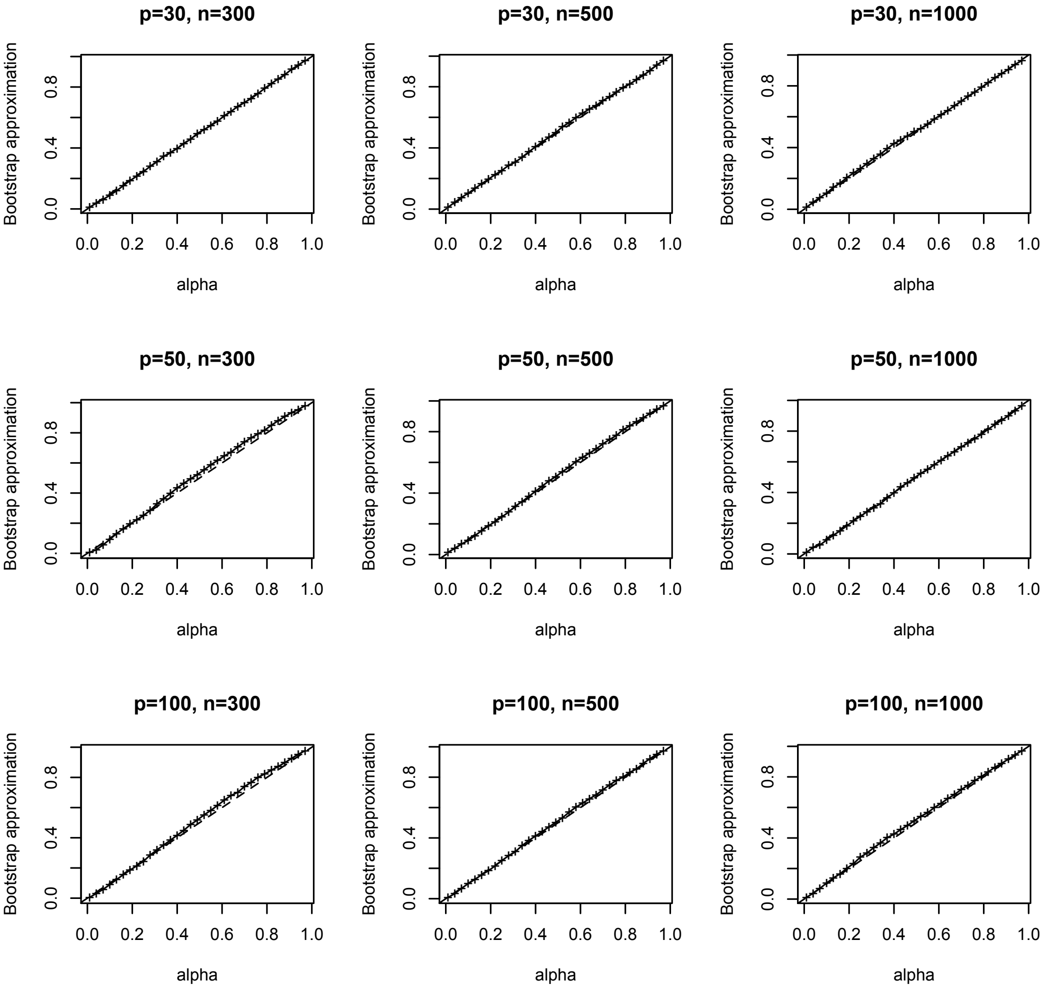

FIGURE 1. Bootstrap approximation $U_n^\sharp$ for Spearman's ρ test statistic with the divide and conquer estimation (MB-NDG-DC). Plot of the nominal size α versus the empirical rejection probability $\widehat{R}(\alpha)$.

$K = \lfloor (n-1)/L \rfloor$), and the MB-NDG-RS with the parameter values suggested in Section 4.3 (i.e., $S_1 = \{1, \ldots, n\}$ and $M = 2(n-1)$). For Bergsma-Dassios' t^*, we apply the MB-DG $U_{n,B}^\sharp$ with $N = n^{4/3}$. Moreover, we also apply the partial versions of MB-NDG-DC and MB-NDG-RS $U_{n,A}^\sharp$ with $N = 4n^{3/2}$. Note that these computational budget parameter values are chosen to minimize the error bounds in the corresponding Gaussian and bootstrap approximations. We only report the simulation results for the randomized incomplete U-statistic with the Bernoulli sampling since the simulation results for the sampling with replacement case are qualitatively similar.

5.2. **Simulation results.** We first examine the statistical accuracy of the bootstrap tests in terms of size for $U_n^\sharp$ for Spearman's ρ and $U_{n,B}^\sharp$ for Bergsma-Dassios' t^*. Due to the space concern, we report the simulation results of the partial bootstrap $U_{n,A}^\sharp$ for Spearman's ρ in Appendix B. For

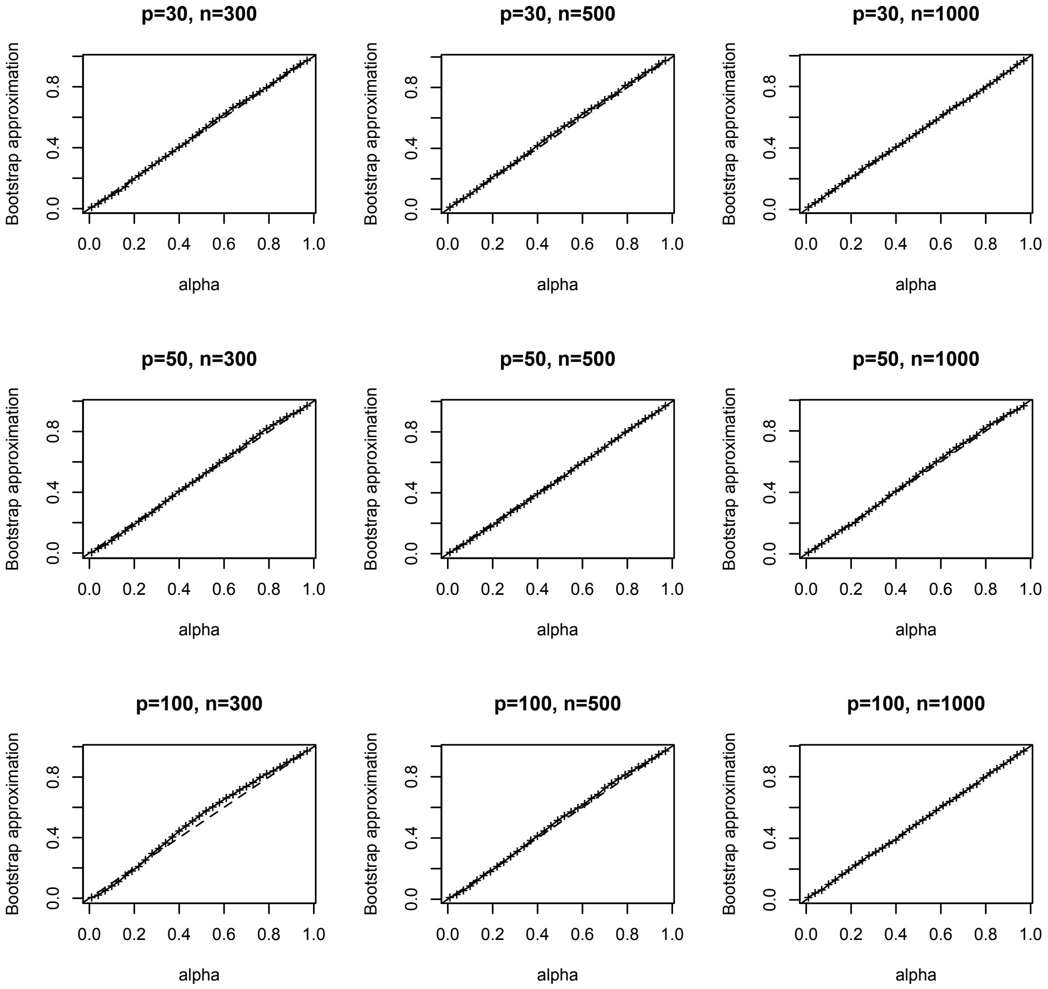

Figure 2. Bootstrap approximation $U_n^\sharp$ for Spearman's ρ test statistic with the random sampling estimation (MB-NDG-RS). Plot of the nominal size α versus the empirical rejection probability $\widehat{R}(\alpha)$.

each nominal size $\alpha \in (0, 1)$, we denote by $\widehat{R}(\alpha)$ the empirical rejection probability of the null hypothesis, where the critical values are calibrated by our bootstrap methods. Figures 1, 2, and 3 display the plots of $\widehat{R}(\alpha)$ versus α for MB-NDG-DC (Spearman's ρ), MB-NDG-RS (Spearman's ρ), and MB-DG (Bergsma-Dassios' t^*), respectively. Clearly, the bootstrap approximations becomes more accurate as n increases. In particular, it is worth noting that the bootstrap approximations work quite well on the right tail, which is relevant in the testing application.

Next, we report the computer running time of the bootstrap tests. Figure 7 displays the computer running time versus the sample size, both on the log-scale. It is observed that the (log-)running time for the bootstrap methods scales linearly with the (log-)sample size. We further fit a linear model of the (log-)running time against the (log-)sample size (with the intercept term) for each p.

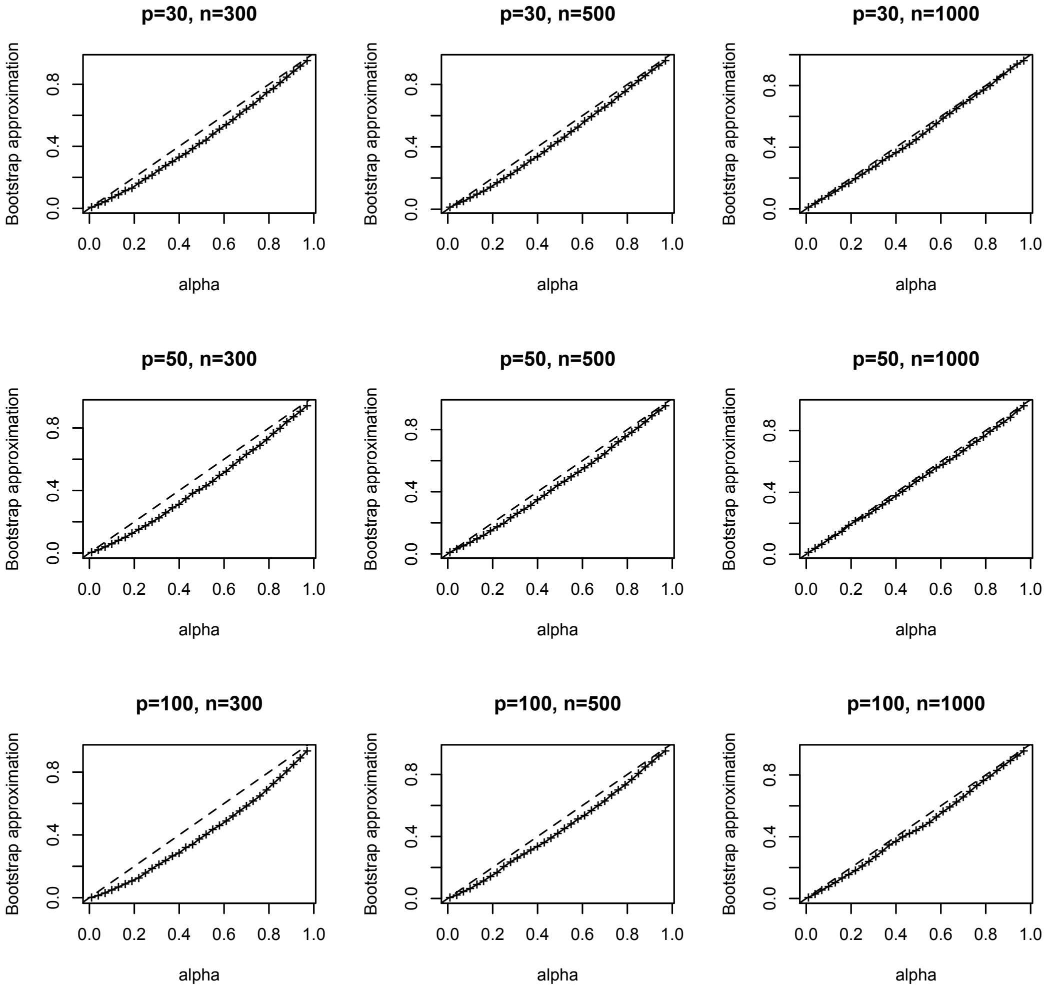

FIGURE 3. Bootstrap approximation $U^{\sharp}_{n,B}$ for Bergsma-Dassios' t^* test statistic (MB-DG). Plot of the nominal size α versus the empirical rejection probability $\widehat{R}(\alpha)$.

For Spearman's ρ, the slope coefficient for $p = (30, 50, 100)$ is $(1.820, 1.863, 1.819)$ in the case MB-NDG-DC, and $(1.987, 1.874, 1.918)$ in the case MB-NDG-RS. In either case, the slope coefficient is close to the theoretic value 2. Recall that the computational complexity for MB-NDG-DC and MB-NDG-RS is the same as $O((n + B)nd)$ for the suggested parameter values. For n larger than B, the computational cost is approximately quadratic in n for each p. For Bergsma-Dassios' t^*, the slope coefficient for $p = (30, 50, 100)$ is $(1.314, 1.318, 1.316)$, which matches very well to the the exponent $4/3$ of the computational budget parameter value $N = n^{4/3}$. In addition, the running time lines are in parallel with each other. This also makes sense because the computational costs of all the bootstrap methods are linear in d (and thus quadratic in p) and the increase of p only affects the intercept on the log scale.

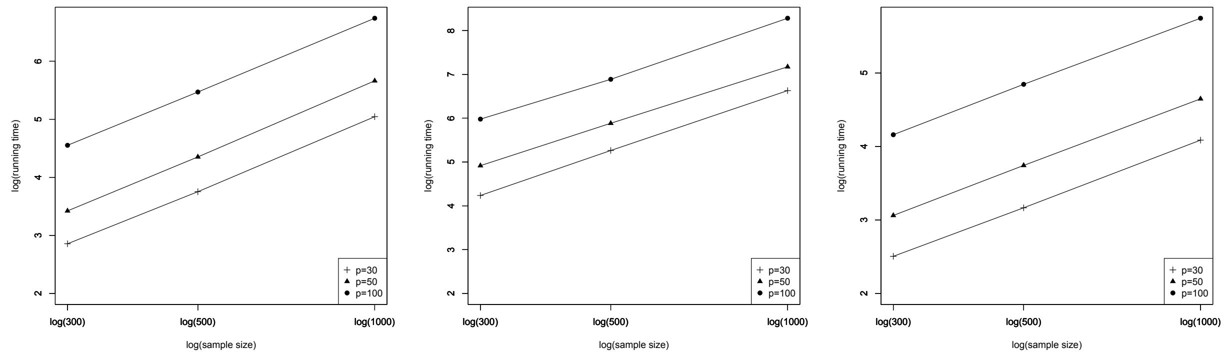

FIGURE 4. Computer running time of the bootstrap versus the sample size on the log-scale. Left: bootstrap $U_n^\sharp$ for Spearman's ρ with the divide and conquer estimation (MB-NDG-DC). Middle: bootstrap $U_n^\sharp$ for Spearman's ρ with the random sampling estimation (MB-NDG-RS). Right: bootstrap $U_{n,B}^\sharp$ for Bergsma-Dassios' t^* (MB-DG).

6. Discussions

In this paper, we have derived the Gaussian and bootstrap approximation results for incomplete U-statistics with random and sparse weights in high dimensions. Specifically, we have considered two sampling schemes: Bernoulli sampling and sampling with replacement, both subject to a computational budget parameter to construct the random weights. On one hand, the sparsity in the design makes the computation of the incomplete U-statistics tractable. On the other hand, the randomness of the weights opens the possibility for us to obtain unified Central Limit Theorem (CLT) type behaviors for both non-degenerate and degenerate kernels, thus revealing the fundamental difference between complete and randomized incomplete U-statistics. Building upon the Gaussian approximation results, we have developed novel bootstrap methods for incomplete U-statistics that take computational considerations into account, and established finite sample error bounds for the proposed bootstrap methods. We end this paper with discussions on two extensions.

6.1. **Extension to normalized U-statistics.**

In applications to, e.g., testing problems, if the variances of the coordinates of $U'_{n,N}$ are heterogeneous, it would be natural to normalize the incomplete U-statistic $U'_{n,,N}$ in such a way that all the coordinates have approximately unit variance, and use a max-type test statistic of $U'_{n,N}$. Often, the coordinatewise variances are unknown and have to be estimated. From Theorems 3.1 and 3.3, in the non-degenerate case the approximate variance of the j-th coordinate of $\sqrt{n}(U'_{n,N} - \theta)$ is $\sigma_j^2 := \sigma_{A,j}^2 + \alpha_n \sigma_{B,j}^2$, where $\sigma_{A,j}^2 := r^2 P(g_j - \theta_j)^2$ and $\sigma_{B,j}^2 := P^r(h_j - \theta_j)^2$, while in the degenerate case, the approximate variance of the j-th coordinate of $\sqrt{N}(U'_{n,N} - \theta)$ is $\sigma_{B,j}^2$. So, the problem boils down to estimating $\sigma_{A,j}^2$ and $\sigma_{B,j}^2$. To this end, we propose the following estimators: recall the setup in Section 4.1 and define

$$\widehat{\sigma}_{A,j}^2 := \frac{r^2}{n_1} \sum_{i_1 \in S_1} \{\widehat{g}_j^{(i_1)}(X_{i_1}) - \breve{g}_j\}^2 \quad \text{and} \quad \widehat{\sigma}_{B,j}^2 := \frac{1}{\widehat{N}} \sum_{\iota \in I_{n,r}} Z_\iota \{h_j(X_\iota) - U'_{n,N,j}\}^2,$$

where $\widehat{N}$ is replaced by N in the definition of $\widehat{\sigma}^2_{B,j}$ for the sampling with replacement case. These estimators are the (j,j)-elements of the conditional covariance matrices of $U^{\sharp}_{n,A}$ and $U^{\sharp}_{n,B}$, respectively. Note that the computational cost to construct $\widehat{\sigma}^2_{B,j}, j = 1, \dots, d$ is (on average) $O(Nd)$, while that of $\widehat{\sigma}^2_{A,j}, j = 1, \dots, d$ is $O(n^2 d)$ if the DC estimation with the parameter values suggested in Section 4.2 is used for estimation of g. Then the proofs of Theorems 4.1 and 4.2 immediately imply the following lemma.

Lemma 6.1 (Variance estimation). *(i) Suppose that Conditions (C1), (C2), and (C3-ND) hold, and in addition suppose that Condition (10) holds for some constants* $0 < C_1 < \infty$ *and* $\zeta \in (0,1)$. *Then there exists a constant* C *depending only on* $\underline{\sigma}, r,$ *and* C_1 *such that* $\max_{1 \leqslant j \leqslant d} |\widehat{\sigma}^2_{A,j}/\sigma^2_{A,j} - 1| \leqslant Cn^{-3\zeta/8}/\log^2 d$ *with probability at least* $1 - Cn^{-\zeta/8}$.

(ii) Suppose that Conditions (C1), (C2), and (C3-D) hold, and in addition suppose that Condition (9) holds for some constants $0 < C_1 < \infty$ *and* $\zeta \in (0,1)$. *Then there exists a constant* C *depending only on* $\underline{\sigma}, r,$ *and* C_1 *such that* $\max_{1 \leqslant j \leqslant d} |\widehat{\sigma}^2_{B,j}/\sigma^2_{B,j} - 1| \leqslant Cn^{-3\zeta/8}/\log^2 d$ *with probability at least* $1 - Cn^{-\zeta/8}$.

Now, let $\Lambda_A = \mathrm{diag}\{\sigma^2_{A,1}, \dots, \sigma^2_{A,d}\}, \Lambda_B = \mathrm{diag}\{\sigma^2_{B,1}, \dots, \sigma^2_{B,d}\}, \widehat{\Lambda}_A = \mathrm{diag}\{\widehat{\sigma}^2_{A,1}, \dots, \widehat{\sigma}^2_{A,d}\}, \Lambda_B = \mathrm{diag}\{\widehat{\sigma}^2_{B,1}, \dots, \widehat{\sigma}^2_{B,d}\}, \Lambda = \mathrm{diag}\{\sigma^2_1, \dots, \sigma^2_d\} = \Lambda_A + \alpha_n \Lambda_B$, and $\widehat{\Lambda} = \mathrm{diag}\{\widehat{\sigma}^2_1, \dots, \widehat{\sigma}^2_d\} = \widehat{\Lambda}_A + \alpha_n \widehat{\Lambda}_B$. We consider to approximate the distributions of $\sqrt{n}\widehat{\Lambda}^{-1/2}(U'_{n,N} - \theta)$ in the non-degenerate case and $\sqrt{N}\widehat{\Lambda}_B^{-1/2}(U'_{n,N} - \theta)$ in the degenerate case. Recall the setup in Section 4.1.

Corollary 6.2 (Gaussian and bootstrap approximations to normalized incomplete U-statistics). *(i) Suppose that Conditions (C1), (C2), and (C3-ND) hold, and in addition suppose that Condition (10) together with* $D_n^2(\log^7(dn))/(n \wedge N) \leqslant C_1 n^{-3\zeta/4}$ *hold for some constants* $0 < C_1 < \infty$ *and* $\zeta \in (0,1)$. *Then there exists a constant* C *depending only on* $\underline{\sigma}, r,$ *and* C_1 *such that*

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}(U'_{n,N} - \theta) \in R) - \mathbb{P}(\Lambda^{-1/2}Y \in R) \right| \leqslant Cn^{-\zeta/8} \quad \textit{and}$$

$$\mathbb{P}\left\{ \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(\widehat{\Lambda}^{-1/2}U^{\sharp}_n \in R) - \mathbb{P}(\Lambda^{-1/2}Y \in R) \right| > Cn^{-\zeta/8} \right\} \leqslant Cn^{-\zeta/8},$$

where $Y \sim N(0, r^2\Gamma_g + \alpha_n \Gamma_h)$.

(ii) Suppose that Conditions (C1), (C2), and (C3-D) hold, and in addition suppose that Condition (9) holds for some constants $0 < C_1 < \infty$ *and* $\zeta \in (0,1)$. *Then there exists a constant* C *depending only on* $\underline{\sigma}, r,$ *and* C_1 *such that*

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(\widehat{\Lambda}_B^{-1/2}U^{\sharp}_{n,B} \in R) - \gamma^{\dagger}_B(R) \right| \leqslant Cn^{-\zeta/8}$$

with probability at least $1 - Cn^{-\zeta/8}$, *where* $\gamma^{\dagger}_B = N(0, \Lambda_B^{-1/2}\Gamma_h\Lambda_B^{-1/2})$. *If, in addition, the kernel* h *is degenerate of order* $k - 1$ *for some* $k = 2, \dots, r$, *and if* $ND_n^2(\log^{k+3} d)/n^k \leqslant C_1 n^{-\zeta/2}$ *and* $D_n^2(\log^7(dn))/N \leqslant C_1 n^{-3\zeta/4}$, *then there exists a constant* C' *depending only on* $\underline{\sigma}, r,$ *and* C_1 *such that*

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{N}\widehat{\Lambda}_B^{-1/2}(U'_{n,N} - \theta) \in R) - \gamma^{\dagger}_B(R) \right| \leqslant C'n^{-\zeta/8}.$$

6.2. **Incomplete U-statistics with increasing orders.** Finally, it is interesting to note a connection of incomplete U-statistics with machine learning. The recent paper by [32] studies asymptotic theory for one-dimensional incomplete U-statistics with increasing orders (i.e., $r = r_n \to \infty$). Specifically, they use sampling with replacement and establish asymptotic normality for the non-degenerate case. Their motivation is coming from uncertainty quantification for subbagging and (subsampled) random forests, which, from a mathematical point of view, are defined as *infinite order* U-statistics [16] where the order of the U-statistics corresponds to the subsample size for a single tree and so $r = r_n \to \infty$. Since exact computation of subbagging and random forests is in most cases intractable, a common practice is to choose a smaller number of subsamples randomly. Building on the asymptotic normality result, [32] develop pointwise confidence intervals for subbagging and random forests; see also [40] for related results. Extending the results of [32] to high dimensions enables us to develop methods to construct simultaneous confidence bands for subbagging and random forests and hence would be an interesting venue for future research. Such extension is by no means trivial since the constants appearing in the error bounds developed in the present paper depend on the order r in complicated ways.

Appendix A. Proofs

A.1. **Preliminary lemmas.** This section collects some useful lemmas that will be used in the subsequent proofs. We will freely use the following maximal inequalities for the ψ_β-norms.

Lemma A.1 (Maximal inequalities for the ψ_β-norms)**.** *Let $\xi_1, \dots, \xi_k$ be real-valued random variables such that $\|\xi_i\|_{\psi_\beta} < \infty$ for all $i = 1, \dots, k$ for some $0 < \beta < \infty$, where $k \geqslant 2$. Then*

$$\left\| \max_{1 \leqslant i \leqslant k} |\xi_i| \right\|_{\psi_\beta} \leqslant C_\beta (\log k)^{1/\beta} \max_{1 \leqslant i \leqslant k} \|\xi_i\|_{\psi_\beta},$$

where C_β is a constant that depends only on β.

Proof of Lemma A.1. For $\beta \geqslant 1$, the lemma follows from Lemma 2.2.2 in [39]. For $\beta \in (0,1)$, ψ_β is not convex and so we can not directly apply Lemma 2.2.2 in [39], but apply the lemma for the norm equivalent to $\| \cdot \|_{\psi_\beta}$ obtained by linearizing ψ_β in a neighborhood of the origin; see Lemma A.2 below. □

Lemma A.2 (Norm equivalent to $\| \cdot \|_{\psi_\beta}$)**.** *Let $\beta \in (0,1)$, and take $x_\beta > 0$ large enough so that the function*

$$\widetilde{\psi}_\beta(x) = \begin{cases} \psi_\beta(x) & \text{if } x \geqslant x_\beta \\ \frac{\psi_\beta(x_\beta)}{x_\beta} x & \text{if } 0 \leqslant x \leqslant x_\beta \end{cases}$$

is convex. Then there exists a constant $1 < C_\beta < \infty$ depending only on β such that

$$C_\beta^{-1} \|\xi\|_{\widetilde{\psi}_\beta} \leqslant \|\xi\|_{\psi_\beta} \leqslant C_\beta \|\xi\|_{\widetilde{\psi}_\beta}$$

for every real-valued random variable ξ.

Proof of Lemma A.2. This seems to be well-known, but we include a proof of the lemma since we could not find a right reference. 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 β. We first show that $\|\xi\|_{\widetilde{\psi}_\beta} \lesssim \|\xi\|_{\psi_\beta}$. To this end, we may assume that $\|\xi\|_{\psi_\beta} = 1$, i.e., $\mathbb{E}[\psi_\beta(|\xi|)] = 1$, and show that $\|\xi\|_{\widetilde{\psi}_\beta} \lesssim 1$. By Taylor's theorem, we have $\psi_\beta(x) \gtrsim x$ and $\psi_\beta(x/C) \leqslant C^{-\beta}\psi_\beta(x)$ for $C > 1$, so that

$$\mathbb{E}[\widetilde{\psi}_\beta(|\xi/C|)] \lesssim \mathbb{E}[|\xi/C|] + \mathbb{E}[\psi_\beta(|\xi/C|)] \lesssim C^{-\beta}.$$

This implies that $\|\xi\|_{\widetilde{\psi}_\beta} \lesssim 1$. Next, suppose that $\|\xi\|_{\widetilde{\psi}_\beta} = 1$ and we show that $\|\xi\|_{\psi_\beta} \lesssim 1$. By convexity of $\widetilde{\psi}_\beta$, we have $\mathbb{E}[\widetilde{\psi}_\beta(|\xi/C|)] \leqslant C^{-1}$ for $C > 1$. Combining the fact that $\psi_\beta(x/C) \leqslant C^{-\beta}\psi_\beta(x_\beta)$ for $0 \leqslant x \leqslant x_\beta$ and $C > 1$, we have

$$\mathbb{E}[\psi_\beta(|\xi/C|)] \leqslant C^{-\beta}\psi_\beta(x_\beta) + \mathbb{E}[\widetilde{\psi}_\beta(|\xi/C|)] \lesssim C^{-\beta},$$

which implies that $\|\xi\|_{\psi_\beta} \lesssim 1$. This completes the proof. □

Lemma A.3 (Useful maximal inequalities for U-statistics)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$ with common distribution P, and let $h = (h_1, \dots, h_d)^T : S^r \to \mathbb{R}^d$ be a symmetric and jointly measurable function such that $\|h_j(X_1^r)\|_{\psi_\beta} < \infty$ for all $j = 1, \dots, d$ for some $\beta \in (0, 1]$. Consider the associated U-statistic $U_n(h) = |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h(X_\iota)$ with kernel h, and let $\mathsf{Z} = \max_{1 \leqslant j \leqslant d} |U_n(h_j) - P^r h_j|$. In addition, let*

$$\breve{\mathsf{Z}} = \max_{1 \leqslant j \leqslant d} \left| \sum_{i=1}^m \{h_j(X_{(i-1)r+1}^{ir}) - P^r h_j\} \right|, \text{ and}$$
$$\mathsf{M} = \max_{1 \leqslant i \leqslant m} \max_{1 \leqslant j \leqslant d} |h_j(X_{(i-1)r+1}^{ir}) - P^r h_j|,$$

where $m = \lfloor n/r \rfloor$ is the integer part of n/r. Then, for every $\eta \in (0, 1]$ and $\delta > 0$, there exists a constant C depending only on β, η, and δ such that

$$\mathbb{P}\left(m\mathsf{Z} \geqslant (1+\eta)\mathbb{E}[\breve{\mathsf{Z}}] + t\right) \leqslant \exp\left(-\frac{t^2}{2(1+\delta)m\sigma^2}\right) + 3\exp\left\{-\left(\frac{t}{C\|\mathsf{M}\|_{\psi_\beta}}\right)^\beta\right\}$$

for every $t > 0$, where $\sigma^2 = \max_{1 \leqslant j \leqslant d} P^r(h_j - P^r h_j)^2$.

Proof of Lemma A.3. The proof essentially follows from that of Lemma E.1 in [8], and so we only point out required modifications. The difference is that in Lemma E.1 in [8], $\breve{\mathsf{Z}}$ is defined as $\max_{1 \leqslant j \leqslant d} |\sum_{i=1}^m \{\overline{h}_j(X_{(i-1)r+1}^{ir}) - P^r \overline{h}_j\}|$ where $\overline{h}$ is to be defined below. Without loss of generality, we may assume that $t \geqslant C_1 \|\mathsf{M}\|_{\psi_\beta}$ for some sufficiently large constant C_1 that depends only on β, η and δ. For $\tau = 8\mathbb{E}[\mathsf{M}]$, let $\overline{h}(x_1, \dots, x_r) = h(x_1, \dots, x_r)\mathbf{1}(|h(x_1, \dots, x_r)|_\infty \leqslant \tau)$ and $\underline{h} = h - \overline{h}$. In addition, define $V_\ell(x_1, \dots, x_n), T_\ell, \ell = 1, 2$ as in the proof of Lemma E.1 in [8]. Then, $\mathsf{Z} \leqslant T_1 + T_2$, and since $h = \overline{h} + \underline{h}$ and hence $\overline{h} = h + (-\underline{h})$, we have $\mathbb{E}[|W_1(X_1^n)|_\infty] \leqslant \mathbb{E}[\breve{\mathsf{Z}}] + \mathbb{E}[|W_2(X_1^n)|_\infty]$, so that $\mathbb{E}[\breve{\mathsf{Z}}] \geqslant \mathbb{E}[|W_1(X_1^n)|_\infty] - \mathbb{E}[|W_2(X_1^n)|_\infty]$. Hence, for every $\eta > 0$ and $\varepsilon \in (0, 1)$,

$$\begin{aligned} &\mathbb{P}\left(\mathsf{Z} \geqslant (1+\eta)\mathbb{E}[\breve{\mathsf{Z}}] + t\right) \\ &\quad \leqslant \mathbb{P}\left(T_1 \geqslant (1+\eta)(\mathbb{E}[|W_1(X_1^n)|_\infty] - \mathbb{E}[|W_2(X_1^n)|_\infty]) + (1-\varepsilon)t\right) + \mathbb{P}(T_2 \geqslant \varepsilon t). \end{aligned}$$

Choose $\varepsilon = \varepsilon(\delta) < 1/2$ small enough so that $(1-2\varepsilon)^{-2}(1+\delta/2) \leqslant 1+\delta$. From the proof of Lemma E.1 in [8], we have $\mathbb{E}[|W_2(X_1^n)|_\infty] \leqslant C_2\|\mathsf{M}\|_{\psi_\beta}$ for some constant C_2 that depends only on β. By choosing C_1 sufficiently large, we have $(1+\eta)C_2\|\mathsf{M}\|_{\psi_\beta} \leqslant \varepsilon t$, so that

$$\mathbb{P}\left(\mathsf{Z} \geqslant (1+\eta)\mathbb{E}[\breve{\mathsf{Z}}] + t\right) \leqslant \mathbb{P}\left(T_1 \geqslant (1+\eta)\mathbb{E}[|W_1(X_1^n)|_\infty] + (1-2\varepsilon)t\right) + \mathbb{P}(T_2 \geqslant \varepsilon t).$$

The rest of the proof is analogous to the proof of Lemma E.1 in [8] and hence omitted. □

Lemma A.4 (Gaussian comparison on hyperrectangles)**.** *Let Y, W be centered Gaussian random vectors in $\mathbb{R}^d$ with covariance matrices $\Sigma^Y = (\Sigma^Y_{j,k})_{1\leqslant j,k\leqslant d}$, $\Sigma^W = (\Sigma^W_{j,k})_{1\leqslant j,k\leqslant d}$, respectively, and let $\Delta = |\Sigma^Y - \Sigma^W|_\infty$. Suppose that $\min_{1\leqslant j\leqslant d}\Sigma^Y_{j,j} \bigvee \min_{1\leqslant j\leqslant d}\Sigma^W_{j,j} \geqslant \underline{\sigma}^2$ for some constant $\underline{\sigma} > 0$. Then*

$$\sup_{R\in\mathcal{R}} |\mathbb{P}(Y \in R) - \mathbb{P}(W \in R)| \leqslant C\Delta^{1/3}\log^{2/3} d,$$

where C is a constant that depends only on $\underline{\sigma}$.

Proof of Lemma A.4. The proof is implicit in the proof of Theorem 4.1 in [12]. □

A.2. **Proofs for Section 3.** Observe that $P|h_j - \theta_j|^{2+k} \leqslant 2^{1+k}D_n^k$ by Jensen's inequality for all j and $k = 1, 2$, and $\|h_j(X_1^r) - \theta_j\|_{\psi_1} \leqslant (1 + 1/\log 2)D_n$ for all j. So, in view of the identity $U'_{n,N} - \theta = \widehat{N}^{-1}\sum_{\iota\in I_{n,r}} Z_\iota\{h(X_\iota) - \theta\}$ where $\widehat{N}$ is replaced by N for the sampling with replacement case, it is without loss of generality to assume that

$$\theta = P^r h = 0$$

by replacing h with $h - \theta$.

Throughout 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 $\underline{\sigma}$ and r. In addition, let C denote a generic constant that depends only on $\underline{\sigma}$ and r; its value may change from place to place.

Proof of Theorem 3.1. It is not difficult to see that the equality of the first two terms in (7) holds since $n = N\alpha_n$. So it suffices to prove the second line in (7). In this proof, without loss of generality, we may assume that

$$D_n^2 \log^7(dn) \leqslant c_1(N \wedge n) \tag{16}$$

for some sufficiently small constant c_1 depending only on $\underline{\sigma}$ and r, since otherwise the conclusion of the theorem is trivial by taking C in (7) sufficiently large. In addition, for the notational convenience, let

$$\varpi_n := \left(\frac{D_n^2\log^7(dn)}{n \wedge N}\right)^{1/6}.$$

Bernoulli sampling case. First, consider the Bernoulli sampling. The proof is divided into several steps.

Step 1. Recall the decomposition $W_n = (\widehat{N}/N)U'_{n,N} = U_n + N^{-1}\sum_{\iota\in I_{n,r}}(Z_\iota - p_n)h(X_\iota) = A_n + \sqrt{1-p_n}B_n$, and observe that $\sqrt{N}B_n = |I_{n,r}|^{-1/2}\sum_{\iota\in I_{n,r}}(p_n(1-p_n))^{-1/2}(Z_\iota - p_n)h(X_\iota)$. Let $\widehat{Y}$ be a random vector in $\mathbb{R}^d$ such that $\widehat{Y} \mid X_1^n \sim N(0, \widehat{\Gamma}_h)$ where $\widehat{\Gamma}_h = |I_{n,r}|^{-1}\sum_{\iota\in I_{n,r}} h(X_\iota)h(X_\iota)^T$. In this step, we shall show that with probability at least $1 - Cn^{-1}$,

$$\rho^{\mathcal{R}}_{|X_1^n}(\sqrt{N}B_n, \widehat{Y}) := \sup_{R\in\mathcal{R}}\left|\mathbb{P}_{|X_1^n}(\sqrt{N}B_n \in R) - \mathbb{P}_{|X_1^n}(\widehat{Y} \in R)\right| \leqslant C\varpi_n.$$

The proof of Step 1 is lengthy and divided into six sub-steps.

Step 1.1. We first derive a generic upper bound on $\rho^{\mathcal{R}}_{|X_1^n}(\sqrt{N}B_n, \widehat{Y})$. Let $\widehat{Y}_\iota, \iota \in I_{n,r}$ be random vectors in $\mathbb{R}^d$ independent conditionally on X_1^n such that $\widehat{Y}_\iota \mid X_1^n \sim N(0, h(X_\iota)h(X_\iota)^T)$ for $\iota \in I_{n,r}$. Observe that conditionally on X_1^n, $\widehat{Y} \stackrel{d}{=} |I_{n,r}|^{-1/2} \sum_{\iota \in I_{n,r}} \widehat{Y}_\iota$. Define

$$\widehat{L}_n = \max_{1 \leqslant j \leqslant d} \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} (p_n(1-p_n))^{-3/2} |h_j(X_\iota)|^3 \mathbb{E}[|Z_\iota - p_n|^3].$$

Further, for $\phi \geqslant 1$, define

$$\begin{aligned}
\widehat{M}_{n,X}(\phi) &= \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} \mathbb{E}_{|X_1^n} \left[\max_{1 \leqslant j \leqslant d} \left| \frac{(Z_\iota - p_n) h_j(X_\iota)}{\sqrt{p_n(1-p_n)}} \right|^3 \mathbf{1} \left(\max_{1 \leqslant j \leqslant d} \left| \frac{(Z_\iota - p_n) h_j(X_\iota)}{\sqrt{p_n(1-p_n)}} \right| > \frac{\sqrt{|I_{n,r}|}}{4\phi \log d} \right) \right], \\
\widehat{M}_{n,Y}(\phi) &= \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} \mathbb{E}_{|X_1^n} \left[\max_{1 \leqslant j \leqslant d} |\widehat{Y}_{\iota,j}|^3 \mathbf{1} \left(\max_{1 \leqslant j \leqslant d} |\widehat{Y}_{\iota,j}| > \frac{\sqrt{|I_{n,r}|}}{4\phi \log d} \right) \right],
\end{aligned}$$

and $\widehat{M}_n(\phi) = \widehat{M}_{n,X}(\phi) + \widehat{M}_{n,Y}(\phi)$. Let $\overline{L}_n$ and $\overline{M}_n$ be constants whose values will be determined later.

Then, Theorem 2.1 in [12] (applied conditionally on X_1^n) yields that there exists a constant C_2 depending only on $\underline{\sigma}$ such that for

$$\phi_n = C_2 \left(\frac{\overline{L}_n^2 \log^4 d}{|I_{n,r}|} \right)^{-1/6},$$

we have that

$$\rho^{\mathcal{R}}_{|X_1^n}(\sqrt{N}B_n, \widehat{Y}) \leqslant C \left\{ \left(\frac{\overline{L}_n^2 \log^7 d}{|I_{n,r}|} \right)^{1/6} + \frac{\overline{M}_n}{\overline{L}_n} \right\}$$

on the event $\mathcal{E}_n := \{\widehat{M}_n(\phi_n) \leqslant \overline{M}_n\} \cap \{\widehat{L}_n \leqslant \overline{L}_n\} \cap \{\min_{1 \leqslant j \leqslant d} \widehat{\Gamma}_{h,jj} \geqslant \underline{\sigma}^2/2\}$. In Steps 1.2–1.4, we will bound $\widehat{L}_n$ and $\widehat{M}_n(\phi_n)$, and in Step 1.5, we will evaluate the probability that $\min_{1 \leqslant j \leqslant d} \widehat{\Gamma}_{h,jj} \geqslant \underline{\sigma}^2/2$. In Step 1.6, we will derive an explicit bound on $\rho^{\mathcal{R}}_{|X_1^n}(\sqrt{N}B_n, \widehat{Y})$ that holds with probability at least $1 - Cn^{-1}$.

Step 1.2: Bounding $\widehat{L}_n$. Since $p_n \leqslant 1/2$ and $\mathbb{E}[|Z_\iota - p_n|^3] = p_n(1-p_n)\{p_n^2 + (1-p_n)^2\} \leqslant Cp_n$, $\widehat{L}_n$ is bounded from above by $Cp_n^{-1/2}$ times $\max_{1 \leqslant j \leqslant d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} |h_j(X_\iota)|^3 =: \mathsf{Z}_1$. Let $m = \lfloor n/r \rfloor, \check{\mathsf{Z}}_1 = \max_{1 \leqslant j \leqslant d} \sum_{i=1}^m |h_j(X_{(i-1)r+1}^{ir})|^3$, and $\mathsf{M}_1 = \max_{1 \leqslant i \leqslant m} \max_{1 \leqslant j \leqslant d} |h_j(X_{(i-1)r+1}^{ir})|$. Then, Lemma E.3 in [8] yields that

$$\mathbb{P}\left(m\mathsf{Z}_1 \geqslant 2\mathbb{E}[\check{\mathsf{Z}}_1] + C\|\mathsf{M}_1^3\|_{\psi_{1/3}} t^3 \right) \leqslant 3e^{-t}$$

for every $t > 0$. Further, since the blocks $X_{(i-1)r+1}^{ir}, i = 1, \ldots, m$ are i.i.d., Lemma 9 in [11] yields that

$$\mathbb{E}[\check{\mathsf{Z}}_1] \lesssim \max_{1 \leqslant j \leqslant d} \sum_{i=1}^m \mathbb{E}\left[|h_j(X_{(i-1)r+1}^{ir})|^3 \right] + \mathbb{E}[\mathsf{M}_1^3] \log d \lesssim mD_n + \mathbb{E}[\mathsf{M}_1^3] \log d.$$

Since $\mathbb{E}[\mathsf{M}_1^3] \lesssim \|\mathsf{M}_1^3\|_{\psi_{1/3}} = \|\mathsf{M}_1\|_{\psi_1}^3 \lesssim D_n^3 \log^3(dn)$, we have

$$\mathbb{P}\left(\widehat{L}_n \geqslant Cp_n^{-1/2} D_n \{1 + n^{-1} D_n^2 \log^4(dn) + t^3 n^{-1} D_n^2 \log^3(dn)\} \right) \leqslant 3e^{-t}.$$

Since $D_n^2 \log^7(dn) \leqslant c_1 n$, by choosing $\overline{L}_n = Cp_n^{-1/2} D_n$ and $t = \log n$, we conclude that $\mathbb{P}(\widehat{L}_n \geqslant \overline{L}_n) \leqslant 3n^{-1}$.

Step 1.3: Bounding $\widehat{M}_{n,X}(\phi_n)$. We begin with noting that

$$\widehat{M}_{n,X}(\phi) \leqslant \frac{C}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} p_n^{-1/2} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)|^3 \mathbf{1}\left(\max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| > \frac{\sqrt{N}}{4\phi \log d} \right).$$

Since $\| \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \|_{\psi_1} \lesssim D_n \log(dn)$, we have that

$$\max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \leqslant C D_n \log^2(dn)$$

with probability at least $1 - 2n^{-1}$. Now, since

$$\frac{\sqrt{N}}{4\phi_n \log d} \gtrsim \left(\frac{D_n N}{\log d} \right)^{1/3} \geqslant c_1^{-1/3} D_n \log^2(dn),$$

by choosing c_1 in (16) sufficiently small, we have that $\widehat{M}_{n,X}(\phi_n) = 0$ with probability at least $1 - 2n^{-1}$.

Step 1.4: Bounding $\widehat{M}_{n,Y}(\phi_n)$. Suppose that

$$\max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \leqslant C D_n \log^2(dn),$$

which holds with probability at least $1 - 2n^{-1}$ by Step 1.3. Recall that $\|\xi\|_{\psi_1} \leqslant (1+e)\|\xi\|_{\psi_2}$ for every real-valued random variable. (For completeness, we provide its proof: assume $\|\xi\|_{\psi_2} = 1$, and observe that $\mathbb{E}[e^{|\xi|}] \leqslant e + \mathbb{E}[e^{\xi^2}] \leqslant e + 2$, so that $\mathbb{E}[\psi_1(|\xi|)] \leqslant 1 + e$. The desired result follows from the observation that $\psi_1(x/C) \leqslant C^{-1}\psi_1(x)$ for $C > 1$.) Conditionally on X_1^n, since $\widehat{Y}_{\iota,j} \mid X_1^n \sim N(0, h_j^2(X_\iota))$ for every $\iota \in I_{n,r}$, we have $\| \max_{1 \leqslant j \leqslant d} |\widehat{Y}_{\iota,j}| \|_{\psi_1} \leqslant (1+e) \| \max_{1 \leqslant j \leqslant d} |\widehat{Y}_{\iota,j}| \|_{\psi_2} \lesssim \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \log^{1/2} d$, so that

$$\mathbb{P}_{|X_1^n}\left(\max_{1 \leqslant j \leqslant d} |\widehat{Y}_{\iota,j}| \geqslant t \right) \leqslant 2 \exp\left(-\frac{t}{C \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \log^{1/2} d} \right)$$

for every $t > 0$. Hence, it follows from Lemma C.1 in [12] that

$$\begin{aligned}
&\mathbb{E}_{|X_1^n}\left[\max_{1 \leqslant j \leqslant d} |\widehat{Y}_{\iota,j}|^3 \mathbf{1}\left(\max_{1 \leqslant j \leqslant d} |\widehat{Y}_{\iota,j}| > \frac{\sqrt{|I_{n,r}|}}{4\phi_n \log d} \right) \right] \\
&\lesssim \left(\frac{\sqrt{|I_{n,r}|}}{\phi_n \log d} + \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \log^{1/2} d \right)^3 \exp\left(-\frac{\sqrt{|I_{n,r}|}}{C\phi_n \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \log^{3/2} d} \right) \\
&\lesssim (n^{r/2} + D_n \log^{5/2}(dn))^3 \exp\left(-\frac{|I_{n,r}|^{1/3}}{C D_n^{2/3} \log^{17/6}(dn)} \right) \\
&\lesssim n^{3r/2} \exp\left(-\frac{n^{2/3}}{C D_n^{2/3} \log^{17/6}(dn)} \right) \leqslant n^{3r/2} \exp\left(-\frac{n^{1/3}}{C \log^{1/2}(dn)} \right) \\
&\leqslant n^{3r/2} \exp(-n^{11/42}/C) \leqslant n^{3r/2} e^{-n^{1/4}/C},
\end{aligned}$$

where we have used the assumption (16). Therefore, we conclude that $\widehat{M}_{n,Y}(\phi_n) \leqslant C n^{3r/2} e^{-n^{1/4}/C}$ with probability at least $1 - 2n^{-1}$.

Step 1.5: Bounding $|\widehat{\Gamma}_h - \Gamma_h|_\infty$. Let $\mathsf{Z}_2 = |\widehat{\Gamma}_h - \Gamma_h|_\infty$, and observe that

$$\begin{aligned}
\max_{1 \leqslant j,k \leqslant d} \|h_j(X_\iota) h_k(X_\iota)\|_{\psi_{1/2}} &\leqslant \max_{1 \leqslant j,k \leqslant d} \|h_j^2(X_\iota)/2 + h_k^2(X_\iota)/2\|_{\psi_{1/2}} \\
&\lesssim \max_{1 \leqslant j,k \leqslant d} (\|h_j^2(X_\iota)\|_{\psi_{1/2}} + \|h_k^2(X_\iota)\|_{\psi_{1/2}}) \lesssim \max_{1 \leqslant j \leqslant d} \|h_j^2(X_\iota)\|_{\psi_{1/2}} \\
&= \max_{1 \leqslant j \leqslant d} \|h_j(X_\iota)\|_{\psi_1}^2 \leqslant D_n^2,
\end{aligned}$$

and $\max_{1 \leqslant j,k \leqslant d} P^r (h_j h_k)^2 = \max_{1 \leqslant j \leqslant d} P^r h_j^4 \leqslant D_n^2$. Hence, Lemma A.3 yields that

$$\mathbb{P}\left(m\mathsf{Z}_2 \geqslant 2\mathbb{E}[\breve{\mathsf{Z}}_2] + t \right) \leqslant e^{-t^2/(3mD_n^2)} + 3\exp\left\{ -\left(\frac{t}{C\|\mathsf{M}_2\|_{\psi_{1/2}}} \right)^{1/2} \right\},$$

where $m = \lfloor n/r \rfloor$, and $\breve{\mathsf{Z}}_2$ and M_2 are defined by

$$\begin{aligned}
\breve{\mathsf{Z}}_2 &= \max_{1 \leqslant j,k \leqslant d} \left| \sum_{i=1}^m \{ h_j(X_{(i-1)r+1}^{ir}) h_k(X_{(i-1)r+1}^{ir}) - P^r h_j h_k \} \right| \quad \text{and} \\
\mathsf{M}_2 &= \max_{1 \leqslant j,k \leqslant d} \max_{1 \leqslant i \leqslant m} \left| h_j(X_{(i-1)r+1}^{ir}) h_k(X_{(i-1)r+1}^{ir}) - P^r h_j h_k \right|.
\end{aligned}$$

Observe that $\|\mathsf{M}_2\|_{\psi_{1/2}} \lesssim D_n^2 \log^2(dn)$. In addition, Lemma 8 in [11] yields that

$$\mathbb{E}[\breve{\mathsf{Z}}_2] \lesssim \sqrt{mD_n^2 \log d} + \sqrt{\mathbb{E}[\mathsf{M}_2^2]} \log d \lesssim D_n \sqrt{n \log d} + D_n^2 \log^3(dn).$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(\mathsf{Z}_2 \geqslant C\{ n^{-1/2} D_n \log^{1/2} d + n^{-1} D_n^2 \log^3(dn) \} + t \right) \\
&\leqslant \exp\left(-\frac{nt^2}{3rD_n^2} \right) + 3\exp\left(-\frac{(nt)^{1/2}}{CD_n \log(dn)} \right).
\end{aligned}$$

Choosing $t = Cn^{-1/2} D_n (\log n)^{1/2} \bigvee Cn^{-1} D_n^2 (\log n)^2 \log^2(dn)$ for large enough C leads to

$$\mathbb{P}\left(\mathsf{Z}_2 \geqslant C\{ n^{-1/2} D_n \log^{1/2}(dn) + n^{-1} D_n^2 (\log n) \log^3(dn) \} \right) \leqslant Cn^{-1}.$$

Choosing c_1 in (16) small enough, we conclude that $|\widehat{\Gamma}_h - \Gamma_h|_\infty \leqslant \underline{\sigma}^2/2$ and hence $\min_{1 \leqslant j \leqslant d} \widehat{\Gamma}_{h,jj} \geqslant \underline{\sigma}^2/2$ with probability at least $1 - Cn^{-1}$.

Step 1.6. In view of Steps 1.1-1.5, choosing $\overline{L}_n = Cp_n^{-1/2} D_n$ and $\overline{M}_n = Cn^{3r/2} e^{-n^{1/4}/C}$, we have $\mathbb{P}(\mathcal{E}_n) \geqslant 1 - Cn^{-1}$. Hence,

$$\rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N} B_n, \widehat{Y}) \leqslant C \left\{ \left(\frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \frac{p_n^{1/2}}{D_n} n^{3r/2} e^{-n^{1/4}/C} \right\} \lesssim \varpi_n$$

with probability at least $1 - Cn^{-1}$.

Step 2: Gaussian comparison. In this step, we shall show that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|X_1^n}(\widehat{Y} \in R) - \gamma_B(R) \right| \leqslant C\varpi_n$$

with probability at least $1 - Cn^{-1}$, where $\gamma_B = N(0, \Gamma_h)$. First, the Gaussian comparison inequality (Lemma A.4) yields that the left hand side is bounded by $C\overline{\Delta}^{1/3} \log^{2/3} d$ on the event $\{|\widehat{\Gamma}_h - \Gamma_h|_\infty \leqslant$

$\overline{\Delta}\}$. From Step 1.5, $|\widehat{\Gamma}_h - \Gamma_h|_\infty \leqslant C\{n^{-1/2}D_n \log^{1/2}(dn) + n^{-1}D_n^2(\log n)\log^3(dn)\}$ with probability at least $1 - Cn^{-1}$, so that

$$\sup_{R\in\mathcal{R}} \left|\mathbb{P}_{|X_1^n}(\widehat{Y} \in R) - \gamma_B(R)\right| \leqslant C\left\{\left(\frac{D_n^2\log^5(dn)}{n}\right)^{1/6} + \left(\frac{D_n^2(\log n)\log^5(dn)}{n}\right)^{1/3}\right\} \lesssim \varpi_n$$

with probability at least $1 - Cn^{-1}$.

Step 3: Gaussian approximation to A_n. Recall that $\Gamma_g = Pgg^T$ since $\theta = 0$. In this step, we shall show that

$$\sup_{R\in\mathcal{R}} \left|\mathbb{P}(\sqrt{n}A_n \in R) - \gamma_A(R)\right| \leqslant C\varpi_n, \tag{17}$$

where $\gamma_A = N(0, r^2\Gamma_g)$. The Hoeffding decomposition yields that

$$A_n = \sum_{k=1}^r \binom{r}{k} U_n^{(k)}(\pi_k h) = rU_n^{(1)}(\pi_1 h) + \underbrace{\sum_{k=2}^r \binom{r}{k} U_n^{(k)}(\pi_k h)}_{=:\mathsf{R}_n},$$

where $(\pi_k h)(x_1,\dots,x_k) = (\delta_{x_1} - P)\cdots(\delta_{x_k} - P)P^{r-k}h$ is the Hoeffding projection at level k; see, e.g., [14], p.137. Since $rU_n^{(1)}(\pi_1 h) = rn^{-1}\sum_{i=1}^n g(X_i)$ is the average of centered independent random vectors with covariance matrix $r^2\Gamma_g$, Proposition 2.1 in [12] yields that

$$\sup_{R\in\mathcal{R}} \left|\mathbb{P}(r\sqrt{n}U_n^{(1)}(\pi_1 h) \in R) - \gamma_A(R)\right| \le C\varpi_n$$

under our assumption. It remains to bound the effect of the remainder term R_n. To this end, we make use of Corollary 5.6 in [9], which yields that

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant d} |U_n^{(k)}(\pi_k h_j)|\right] \lesssim n^{-k/2}(\log^{k/2} d)\sqrt{P^r\left(\max_{1\leqslant j\leqslant d} h_j^2\right)} \lesssim n^{-k/2}D_n\log^{k/2+1} d$$

for every $k = 2,\dots,r$. Hence,

$$\mathbb{E}\left[|\mathsf{R}_n|_\infty\right] \lesssim D_n\sum_{k=2}^r n^{-k/2}\log^{k/2+1} d \lesssim n^{-1}D_n\log^2 d.$$

Now, for $R = \prod_{j=1}^d [a_j, b_j]$, let $a = (a_1,\dots,a_d)^T$ and $b = (b_1,\dots,b_d)^T$, and for $t > 0$, we use the convention that $b + t = (b_1 + t,\dots,b_d + t)^T$. Observe that

$$\begin{aligned}
\mathbb{P}(\sqrt{n}A_n \in R) &= \mathbb{P}(\{-\sqrt{n}A_n \leqslant -a\} \cap \{\sqrt{n}A_n \leqslant b\}) \\
&\leqslant \mathbb{P}\left(\{-\sqrt{n}A_n \leqslant -a\} \cap \{\sqrt{n}A_n \leqslant b\} \cap \left\{|\sqrt{n}\mathsf{R}_n|_\infty \leqslant t\right\}\right) + \mathbb{P}\left(|\sqrt{n}\mathsf{R}_n|_\infty > t\right) \\
&\leqslant \mathbb{P}(\{-r\sqrt{n}U_n^{(1)}(\pi_1 h) \leqslant -a + t\} \cap \{r\sqrt{n}U_n^{(1)}(\pi_1 h) \leqslant b + t\}) + Ct^{-1}n^{-1/2}D_n\log^2 d \\
&\leqslant \gamma_A(\{y \in \mathbb{R}^d : -y \leqslant -a + t, y \leqslant b + t\}) + C\varpi_n + Ct^{-1}n^{-1/2}D_n\log^2 d \\
&\leqslant \gamma_A(R) + Ct\sqrt{\log d} + C\varpi_n + Ct^{-1}n^{-1/2}D_n\log^2 d
\end{aligned}$$

for every $t > 0$, where the last inequality follows from Nazarov's inequality stated in Lemma A.1 in [12]. Choosing $t = (n^{-1}D_n^2\log^3 d)^{1/4}$, we conclude that

$$\mathbb{P}(\sqrt{n}A_n \in R) - \gamma_A(R) \leqslant C\left(\frac{D_n^2\log^5 d}{n}\right)^{1/4} + C\varpi_n \leqslant C\varpi_n$$

because of the assumption (16). Likewise, we have $\mathbb{P}(\sqrt{n}A_n \in R) \geqslant \gamma_A(R) - C\varpi_n$. Therefore, we obtain the conclusion (17).

Step 4: Gaussian approximation to W_n. Pick any hyperrectangle $R \in \mathcal{R}$. Recall that $\alpha_n = n/N$, and observe that

$$\mathbb{P}(\sqrt{n}W_n \in R) = \mathbb{E}\left[\mathbb{P}_{|X_1^n}\left(\sqrt{N}B_n \in \left[\frac{1}{\sqrt{\alpha_n(1-p_n)}}R - \sqrt{\frac{N}{1-p_n}}A_n\right]\right)\right].$$

Now, we freeze the random variables X_1^n. From Steps 1 and 2, the conditional probability inside the expectation is bounded from above by $\gamma_B\left(\left[\frac{1}{\sqrt{\alpha_n(1-p_n)}}R - \sqrt{\frac{N}{1-p_n}}A_n\right]\right) + C\varpi_n$ with probability at least $1 - Cn^{-1}$. Since the probability is bounded by 1 and $n^{-1} \lesssim \varpi_n$, we have

$$\begin{aligned}\mathbb{P}(\sqrt{n}W_n \in R) &\leqslant \mathbb{E}\left[\gamma_B\left(\left[\frac{1}{\sqrt{\alpha_n(1-p_n)}}R - \sqrt{\frac{N}{1-p_n}}A_n\right]\right)\right] + C\varpi_n \\ &= \mathbb{P}\left(\sqrt{1-p_n}Y_B \in [\alpha_n^{-1/2}R - \sqrt{N}A_n]\right) + C\varpi_n = \mathbb{P}\left(\sqrt{n}A_n \in [R - \sqrt{\alpha_n(1-p_n)}Y_B]\right) + C\varpi_n,\end{aligned}$$

where $Y_B \sim N(0, \Gamma_h) = \gamma_B$ independent of X_1^n. Next, we freeze the random variable Y_B. Since Y_B is independent of X_1^n, Step 3 yields that

$$\mathbb{P}_{|Y_B}\left(\sqrt{n}A_n \in [R - \sqrt{\alpha_n(1-p_n)}Y_B]\right) \leqslant \gamma_A\left([R - \sqrt{\alpha_n(1-p_n)}Y_B]\right) + C\varpi_n.$$

By Fubini, we conclude that

$$\begin{aligned}\mathbb{P}(\sqrt{n}W_n \in R) &\leqslant \mathbb{E}\left[\gamma_A\left([R - \sqrt{\alpha_n(1-p_n)}Y_B]\right)\right] + C\varpi_n \\ &= \mathbb{P}\left(Y_A \in [R - \sqrt{\alpha_n(1-p_n)}Y_B]\right) + C\varpi_n = \mathbb{P}(Y_A + \sqrt{\alpha_n(1-p_n)}Y_B \in R) + C\varpi_n,\end{aligned}$$

where $Y_A \sim N(0, r^2\Gamma_g) = \gamma_A$ is independent of Y_B. Since $\alpha_n p_n = n/|I_{n,r}| \lesssim n^{-r+1} \leqslant n^{-1}$ and $|\Gamma_h|_\infty \lesssim D_n$, using the Gaussian comparison inequality (Lemma A.4), we have

$$\mathbb{P}(Y_A + \sqrt{\alpha_n(1-p_n)}Y_B \in R) \leqslant \mathbb{P}(Y_A + \alpha_n^{1/2}Y_B \in R) + C\left(\frac{D_n \log^2 d}{n}\right)^{1/3},$$

and the second term on the right hand side is bounded from above by $C\varpi_n$. Likewise, we have $\mathbb{P}(\sqrt{n}W_n \in R) \geqslant \mathbb{P}(Y_A + \alpha_n^{1/2}Y_B \in R) - C\varpi_n$. Hence, for $Y = Y_A + \alpha_n^{1/2}Y_B \sim N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$, we have

$$\sup_{R\in\mathcal{R}}\left|\mathbb{P}(\sqrt{n}W_n \in R) - \mathbb{P}(Y \in R)\right| \leqslant C\varpi_n. \qquad (18)$$

Step 5: Gaussian approximation to $U'_{n,N}$. We shall verify that the inequality (18) holds with $\sqrt{n}W_n$ replaced by $\sqrt{n}U'_{n,N}$. Since Y is centered Gaussian and $\max_{1\leqslant j\leqslant d}\operatorname{Var}(Y_j) \lesssim D_n(1+\alpha_n)$, we have $\mathbb{E}[|Y|_\infty] \lesssim \sqrt{D_n(1+\alpha_n)\log d}$. By the Borell-Sudakov-Tsirel'son inequality (cf. Theorem 2.5.8 in [18]), we have

$$\mathbb{P}\left(|\alpha_n^{-1/2}Y|_\infty > C\sqrt{D_n(1+\alpha_n^{-1})\log(dn)}\right) \leqslant 2n^{-1}.$$

Combining this estimate with (18), we have

$$\mathbb{P}\left(|\sqrt{N}W_n|_\infty > C\sqrt{D_n(1+\alpha_n^{-1})\log(dn)}\right) \leqslant C\varpi_n.$$

Next, since $\widehat{N} = \sum_{\iota \in I_{n,r}} Z_\iota$ and $Z_\iota, \iota \in I_{n,r}$ are i.i.d. $\mathsf{Ber}(p_n)$ with $p_n = N/|I_{n,r}|$, by Bernstein's inequality (cf. Lemma 2.2.9 in [39]), we have

$$\mathbb{P}\left(|\widehat{N} - N| > \sqrt{2Nt} + 2t/3\right) \leqslant 2e^{-t}$$

for every $t > 0$. Choosing $t = \log n$ and choosing c_1 sufficiently small in (16) such that $\sqrt{(\log n)/N} \leqslant 1/4$, we have

$$\mathbb{P}\left(|\widehat{N}/N - 1| > 2\sqrt{(\log n)/N}\right) \leqslant 2n^{-1}.$$

Since $|z^{-1} - 1| \leqslant 2|z-1|$ for $|z-1| \leqslant 1/2$, we have that

$$|N/\widehat{N} - 1| \leqslant 2|\widehat{N}/N - 1| \leqslant 4\sqrt{(\log n)/N}$$

with probability at least $1 - 2n^{-1}$.

Now, observe that $\sqrt{N}U'_{n,N} = \sqrt{N}W_n + (N/\widehat{N} - 1)\sqrt{N}W_n$, and with probability at least $1 - C\varpi_n$,

$$|(N/\widehat{N} - 1)\sqrt{N}W_n|_\infty \leqslant C\sqrt{\frac{D_n(\log n)\log(dn)}{n \wedge N}}.$$

Arguing as in Step 3 and noting that $\min_{1 \leqslant j \leqslant d} \operatorname{Var}(\alpha_n^{-1/2}Y_j) \geqslant \min_{1 \leqslant j \leqslant d} P^r h_j^2 \gtrsim 1$, we conclude that for every $R \in \mathcal{R}$,

$$\begin{aligned}\mathbb{P}(\sqrt{N}U'_{n,N} \in R) &\leqslant \mathbb{P}(\alpha_n^{-1/2}Y \in R) + C\varpi_n + C\sqrt{\frac{D_n(\log n)\log^2(dn)}{n \wedge N}} \\ &\leqslant \mathbb{P}(\alpha_n^{-1/2}Y \in R) + C\varpi_n.\end{aligned}$$

Likewise, we have $\mathbb{P}(\sqrt{N}U'_{n,N} \in R) \geqslant \mathbb{P}(\alpha_n^{-1/2}Y \in R) - C\varpi_n$. This leads to the conclusion of the theorem in the Bernoulli sampling case.

Sampling with replacement case. Next, consider sampling with replacement. The proof is similar to the Bernoulli sampling case, so we only point out the differences. Recall that we assume $\theta = 0$. Observe that $U'_{n,N} = U_n + N^{-1}\sum_{j=1}^N \{h(X^*_{\iota_j}) - U_n\} =: A_n + B_n$. Since $X^*_{\iota_1}, \ldots, X^*_{\iota_N}$ are i.i.d. draws from the empirical distribution $|I_{n,r}|^{-1}\sum_{\iota \in I_{n,r}} \delta_{X_\iota}$ conditionally on X_1^n, $\sqrt{N}B_n$ is $\sqrt{N}$ times the average of i.i.d. random vectors with mean zero and covariance matrix $\widehat{\Gamma}_h - U_n U_n^T$ conditionally on X_1^n, where $\widehat{\Gamma}_h = |I_{n,r}|^{-1}\sum_{\iota \in I_{n,r}} h(X_\iota)h(X_\iota)^T$. Let $\widehat{Y}$ be a random vector in $\mathbb{R}^d$ such that $\widehat{Y} \mid X_1^n \sim N(0, \widehat{\Gamma}_h - U_n U_n^T)$. We first verify that

$$\rho^{\mathcal{R}}_{|X_1^n}(\sqrt{N}B_n, \widehat{Y}) \leqslant C\varpi_n$$

with probability at least $1 - Cn^{-1}$. Define

$$\widehat{L}_n := \max_{1 \leqslant j \leqslant d} \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} |h_j(X_\iota) - U_{n,j}|^3.$$

By Jensen's inequality, $\widehat{L}_n \leqslant 8\mathsf{Z}_1$, where Z_1 is defined in Step 1.2 for the Bernoulli sampling case. By Step 1.2, we have $\mathbb{P}(\widehat{L}_n \geqslant CD_n) \leqslant 3n^{-1}$ under the assumption (16). So we can take $\overline{L}_n = CD_n$

and $\phi_n = C_2(N^{-1}\overline{L}_n^2 \log^4 d)^{-1/6} \geqslant 1$ by taking the constant C_2 large enough. For $\phi \geqslant 1$, define

$$\widehat{M}_{n,X}(\phi) = \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} \left[\max_{1 \leqslant j \leqslant d} |h_j(X_\iota) - U_{n,j}|^3 \, \mathbf{1}\left(\max_{1 \leqslant j \leqslant d} |h_j(X_\iota) - U_{n,j}| > \frac{\sqrt{N}}{4\phi \log d} \right) \right],$$

$$\widehat{M}_{n,Y}(\phi) = \mathbb{E}_{|X_1^n} \left[\max_{1 \leqslant j \leqslant d} |\widehat{Y}_j|^3 \mathbf{1}\left(\max_{1 \leqslant j \leqslant d} |\widehat{Y}_j| > \frac{\sqrt{N}}{4\phi \log d} \right) \right],$$

and $\widehat{M}_n(\phi) = \widehat{M}_{n,X}(\phi) + \widehat{M}_{n,Y}(\phi)$. Observe that

$$\begin{aligned}\left\| \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota) - U_{n,j}| \right\|_{\psi_1} &\lesssim \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} \|h_j(X_\iota) - U_{n,j}\|_{\psi_1} \log(dn) \\ &\lesssim \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} \|h_j(X_\iota)\|_{\psi_1} \log(dn) \leqslant D_n \log(dn),\end{aligned}$$

and hence

$$\max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota) - U_{n,j}| \leqslant C D_n \log^2(dn)$$

with probability at least $1 - 2n^{-1}$. Using similar calculations to those in Step 1.3, we have that $\widehat{M}_{n,X}(\phi_n) = 0$ with probability at least $1 - 2n^{-1}$. Step 1.4 needs a modification. Since $\widehat{Y}_j \mid X_1^n \sim N(0, |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} (h_j(X_\iota) - U_{n,j})^2)$, we have $\| \max_{1 \leqslant j \leqslant d} |\widehat{Y}_j| \|_{\psi_1} \lesssim \| \max_{1 \leqslant j \leqslant d} |\widehat{Y}_j| \|_{\psi_2} \lesssim \sqrt{V_n \log d}$ conditionally on X_1^n where $V_n = \max_{1 \leqslant j \leqslant d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_j^2(X_\iota)$, from which we have

$$\mathbb{P}_{|X_1^n} \left(\max_{1 \leqslant j \leqslant d} |\widehat{Y}_j| \geqslant t \right) \leqslant 2 \exp\left(- \frac{t}{C\sqrt{V_n \log d}} \right).$$

Let $m = \lfloor n/r \rfloor$ and $\breve{V}_n = \max_{1 \leqslant j \leqslant d} \sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir})$. Then, Lemma E.3 in [8] yields that

$$\mathbb{P}\left(m V_n \geqslant 2\mathbb{E}[\breve{V}_n] + C \|\mathsf{M}_1^2\|_{\psi_{1/2}} t^2 \right) \leqslant 3e^{-t}$$

for every $t > 0$, where $\mathsf{M}_1 = \max_{1 \leqslant i \leqslant m} \max_{1 \leqslant j \leqslant d} |h_j(X_{(i-1)r+1}^{ir})|$. Further, Lemma 9 in [11] yields that

$$\mathbb{E}[\breve{V}_n] \lesssim \max_{1 \leqslant j \leqslant d} \sum_{i=1}^m \mathbb{E}\left[h_j^2(X_{(i-1)r+1}^{ir}) \right] + \mathbb{E}[\mathsf{M}_1^2] \log d \lesssim m D_n + \mathbb{E}[\mathsf{M}_1^2] \log d.$$

Since $\mathbb{E}[\mathsf{M}_1^2] \lesssim \|\mathsf{M}_1^2\|_{\psi_{1/2}} = \|\mathsf{M}_1\|_{\psi_1}^2 \lesssim D_n^2 \log^2(dn)$, we have

$$\mathbb{P}\left(V_n \geqslant C D_n \{1 + n^{-1} D_n \log^3(dn) + t^2 n^{-1} D_n \log^2(dn)\} \right) \leqslant 3e^{-t}.$$

Since $D_n \geqslant 1$ and $D_n^2 \log^7(dn) \leqslant c_1 n$, by choosing $t = \log n$, we conclude that $\mathbb{P}(V_n \geqslant C D_n) \leqslant 3n^{-1}$. Now, suppose that $V_n \leqslant C D_n$, which holds with probability at least $1 - 3n^{-1}$. Then, since

$$\mathbb{P}_{|X_1^n} \left(\max_{1 \leqslant j \leqslant d} |\widehat{Y}_j| \geqslant t \right) \leqslant 2 \exp\left(- \frac{t}{C\sqrt{D_n \log d}} \right),$$

it follows from Lemma C.1 in [12] that

$$\begin{aligned}
&\mathbb{E}_{|X_1^n}\left[\max_{1\leqslant j\leqslant d}|\widehat{Y}_j|^3\mathbf{1}\left(\max_{1\leqslant j\leqslant d}|\widehat{Y}_j|>\frac{\sqrt{N}}{4\phi_n\log d}\right)\right]\\
&\lesssim\left[\frac{\sqrt{N}}{4\phi_n\log d}+\sqrt{D_n\log d}\right]^3\exp\left(-\frac{\sqrt{N}}{C\phi_n D_n^{1/2}\log^{3/2}d}\right)\\
&\lesssim N^{3/2}\exp\left(-\frac{N^{1/3}}{CD_n^{1/6}\log^{5/6}d}\right)\leqslant N^{3/2}\exp(-N^{1/6}/C),
\end{aligned}$$

where we have used the assumption (16). Therefore, we conclude that $\widehat{M}_{n,Y}(\phi_n)\leqslant CN^{3/2}\exp(-N^{1/6}/C)$ with probability at least $1-2n^{-1}$.

Step 1.5 also needs a modification. Note that $|\widehat{\Gamma}_h-U_nU_n^T-\Gamma_h|_\infty\leqslant|\widehat{\Gamma}_h-\Gamma_h|_\infty+|U_n|_\infty^2$. In the Bernoulli sampling case, we have shown in Step 1.5 that $|\widehat{\Gamma}_h-\Gamma_h|_\infty\leqslant\underline{\sigma}^2/4$ with probability at least $1-Cn^{-1}$ (changing the constant from 1/2 to 1/4 does not affect the proof). So we only need to show that $|U_n|_\infty^2\leqslant\underline{\sigma}^2/4$ with probability at least $1-Cn^{-1}$. By Lemma A.3,

$$\mathbb{P}\left(m|U_n|_\infty\geqslant 2\mathbb{E}[\breve{\mathsf{Z}}_3]+t\right)\leqslant e^{-t^2/(3mD_n)}+3\exp\left\{-\frac{t}{C\|\mathsf{M}_1\|_{\psi_1}}\right\},$$

where $\breve{\mathsf{Z}}_3=\max_{1\leqslant j\leqslant d}|\sum_{i=1}^m h_j(X_{(i-1)r+1}^{ir})|$. Observe that $\|\mathsf{M}_1\|_{\psi_1}\lesssim D_n\log(dn)$. In addition, Lemma 8 in [11] yields that

$$\mathbb{E}[\breve{\mathsf{Z}}_3]\lesssim\sqrt{mD_n\log d}+\sqrt{\mathbb{E}[\mathsf{M}_1^2]}\log d\lesssim\sqrt{nD_n\log d}+D_n\log^2(dn).$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(|U_n|_\infty\geqslant C\{n^{-1/2}D_n^{1/2}\log^{1/2}d+n^{-1}D_n\log^2(dn)\}+t\right)\\
&\leqslant\exp\left(-\frac{nt^2}{3rD_n}\right)+3\exp\left(-\frac{nt}{CD_n\log(dn)}\right).
\end{aligned}$$

Choosing $t=Cn^{-1/2}D_n^{1/2}(\log n)^{1/2}\bigvee Cn^{-1}D_n(\log n)\log(dn)$ for large enough C leads to

$$\mathbb{P}\left(|U_n|_\infty\geqslant C\{n^{-1/2}D_n^{1/2}\log^{1/2}(dn)+n^{-1}D_n\log^2(dn)\}\right)\leqslant Cn^{-1}.$$

Choosing c_1 in (16) small enough, we conclude that $|U_n|_\infty^2\leqslant\underline{\sigma}^2/4$ and hence $\min_{1\leqslant j\leqslant d}\{\widehat{\Gamma}_{h,jj}-U_{n,j}^2\}\geqslant\underline{\sigma}^2/2$ with probability at least $1-Cn^{-1}$. Therefore, the overall bound in Step 1 for the sampling with replacement case is given by

$$\rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n,\widehat{Y})\leqslant C\left\{\left(\frac{D_n^2\log^7 d}{N}\right)^{1/6}+\frac{N^{3/2}e^{-N^{1/6}/C}}{D_n}\right\}\lesssim\varpi_n$$

with probability at least $1-Cn^{-1}$.

Step 2 in the Bernoulli sampling case goes through under the assumption (16). Step 3 remains exactly the same as the Bernoulli sampling case. Step 4 follows similarly as the Bernoulli sampling case with $p_n=0$. Step 5 is not needed in the sampling with replacement case. This completes the proof. □

Proof of Corollary 3.2. In view of Theorem 3.1, the corollary follows from the Gaussian comparison inequality (Lemma A.4) and the fact that $|\Gamma_g|_\infty \leqslant |\Gamma_h|_\infty \leqslant CD_n$. □

Proof of Theorem 3.3. We shall follow the notation used in the proof of Theorem 3.1. In this proof, without loss generality, we may assume that

$$ND_n^2 \log^{k+3} d \leqslant c_2 n^k, \quad D_n^2(\log n)\log^5(dn) \leqslant c_2 n, \quad \text{and} \quad D_n^2 \log^7(dn) \leqslant c_2 N \tag{19}$$

for some sufficiently small constant c_2 depending only on $\underline{\sigma}$ and r, since otherwise the conclusion of the theorem is trivial by taking C sufficiently large.

Bernoulli sampling case. We first verify that

$$\rho^{\mathcal{R}}_{|X_1^n}(\sqrt{N}B_n, \widehat{Y}) \leqslant C\left\{\left(\frac{D_n^2 \log^7 d}{N}\right)^{1/6} + \frac{p_n^{1/2}}{D_n} n^{3r/2} e^{-n^{-1/10}/C}\right\} \tag{20}$$

with probability at least $1 - Cn^{-1}$.

It is not difficult to verify from Step 1.2 in the proof of Theorem 3.1 that $\mathbb{P}(\widehat{L}_n \geqslant Cp_n^{-1/2}D_n) \leqslant 3n^{-1}$ under the assumption that $D_n^2(\log n)\log^5(dn) \leqslant c_2 n$, and so take $\overline{L} = Cp_n^{-1/2}D_n$. Step 1.3 goes through as it is. Step 1.4 needs a modification. From Step 1.4, we have that on the event $\max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \leqslant CD_n \log^2(dn)$,

$$\mathbb{E}_{|X_1^n}\left[\max_{1\leqslant j \leqslant d} |\widehat{Y}_{\iota,j}|^3 \mathbf{1}\left(\max_{1 \leqslant j \leqslant d} |\widehat{Y}_{\iota,j}| > \frac{\sqrt{|I_{n,r}|}}{4\phi_n \log d}\right)\right] \leqslant Cn^{3r/2} \exp\left(-\frac{n^{2/3}}{CD_n^{2/3}\log^{17/6}(dn)}\right),$$

and the assumption that $D_n^2(\log n)\log^5(dn) \leqslant c_2 n$ yields that the right hand side is bounded from above by

$$n^{3r/2} \exp\left(-\frac{(n\log n)^{1/3}}{C\log^{7/6}(dn)}\right) \leqslant n^{3r/2} e^{-n^{1/10}/C}.$$

Since $\max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)| \leqslant CD_n \log^2(dn)$ with probability at least $1 - 2n^{-1}$, we have that $\widehat{M}_{n,Y}(\phi_n) \leqslant Cn^{3r/2}e^{-n^{1/10}/C}$ with probability at least $1 - 2n^{-1}$. Step 1.5 holds under the present assumption. Hence, the inequality (20) holds with probability at least $1 - Cn^{-1}$. In addition, Step 2 in the proof of Theorem 3.1 goes through under the present assumption (19), so that

$$\begin{aligned}\sup_{R \in \mathcal{R}} \left|\mathbb{P}_{|X_1^n}(\widehat{Y} \in R) - \gamma_B(R)\right| &\leqslant C\left\{\left(\frac{D_n^2 \log^5(dn)}{n}\right)^{1/6} + \left(\frac{D_n^2(\log n)\log^5(dn)}{n}\right)^{1/3}\right\} \\ &\leqslant C\left(\frac{D_n^2(\log n)\log^5(dn)}{n}\right)^{1/6}\end{aligned}$$

with probability at least $1 - Cn^{-1}$. Therefore, we have that

$$\sup_{R \in \mathcal{R}} \left|\mathbb{P}_{|X_1^n}(\sqrt{N}B_n \in R) - \gamma_B(R)\right| \leqslant C\left\{\left(\frac{D_n^2(\log n)\log^5(nd)}{n}\right)^{1/6} + \left(\frac{D_n^2\log^7(dn)}{N}\right)^{1/6}\right\} =: C\breve{\varpi}_n$$

with probability at least $1 - Cn^{-1}$, and in view of the fact that $n^{-1} \lesssim \breve{\varpi}_n$, we conclude that

$$\sup_{R \in \mathcal{R}} \left|\mathbb{P}(\sqrt{N}B_n \in R) - \gamma_B(R)\right| \leqslant C\breve{\varpi}_n.$$

Since h is degenerate of order $k-1$, we have

$$A_n = \sum_{\ell=k}^{r} \binom{r}{\ell} U_n^{(\ell)}(\pi_\ell h),$$

and Step 3 in the proof of Theorem 3.1 yields that

$$\mathbb{E}\left[|A_n|_\infty\right] \leqslant CD_n \sum_{\ell=k}^{r} n^{-\ell/2} \log^{\ell/2+1} d \leqslant CD_n n^{-k/2} \log^{k/2+1} d. \tag{21}$$

Hence, for $R = \prod_{j=1}^{d}[a_j, b_j], a = (a_1, \dots, a_d)^T, b = (b_1, \dots, b_d)^T$, and $t > 0$, we have

$$\begin{aligned}
&\mathbb{P}(\sqrt{N}W_n \in R) = \mathbb{P}(\{-\sqrt{N}W_n \leqslant -a\} \cap \{\sqrt{N}W_n \leqslant b\}) \\
&\quad \leqslant \mathbb{P}\left(\{-\sqrt{N}W_n \leqslant -a\} \cap \{\sqrt{N}W_n \leqslant b\} \cap \left\{|\sqrt{N}A_n|_\infty \leqslant t\right\}\right) + \mathbb{P}\left(|\sqrt{N}A_n|_\infty > t\right) \\
&\quad \leqslant \mathbb{P}\left(\{-\sqrt{N(1-p_n)}B_n \leqslant -a+t\} \cap \{\sqrt{N(1-p_n)}B_n \leqslant b+t\}\right) + Ct^{-1}\sqrt{N}n^{-k/2}D_n \log^{k/2+1} d \\
&\quad \leqslant \gamma_B(\{y \in \mathbb{R}^d : -\sqrt{1-p_n}y \leqslant -a+t, \sqrt{1-p_n}y \leqslant b+t\}) + C\breve{\varpi}_n + Ct^{-1}\sqrt{N}n^{-k/2}D_n \log^{k/2+1} d \\
&\quad \leqslant \gamma_B([(1-p_n)^{-1/2}R]) + Ct\sqrt{\log d} + C\breve{\varpi}_n + Ct^{-1}\sqrt{N}n^{-k/2}D_n \log^{k/2+1} d,
\end{aligned}$$

where the last inequality follows from Nazarov's inequality ([12], Lemma A.1). Choosing $t = (Nn^{-k}D_n^2 \log^{k+1} d)^{1/4}$, we conclude that

$$\mathbb{P}(\sqrt{N}W_n \in R) \leqslant \gamma_B([(1-p_n)^{-1/2}R]) + C\left(\frac{ND_n^2 \log^{k+3} d}{n^k}\right)^{1/4} + C\breve{\varpi}_n.$$

Finally, since $p_n \lesssim N/n^r$, the Gaussian comparison inequality (Lemma A.4) yields that

$$\gamma_B([(1-p_n)^{-1/2}R]) \leqslant \gamma_B(R) + C\left(\frac{ND_n \log^2 d}{n^r}\right)^{1/3},$$

and the second term on the right hand side is bounded from above by $C\left(\frac{ND_n^2 \log^{k+3} d}{n^k}\right)^{1/4}$. Hence,

$$\mathbb{P}(\sqrt{N}W_n \in R) \leqslant \gamma_B(R) + C\left(\frac{ND_n^2 \log^{k+3} d}{n^k}\right)^{1/4} + C\breve{\varpi}_n.$$

Likewise, we have

$$\mathbb{P}(\sqrt{N}W_n \in R) \geqslant \gamma_B(R) - C\left(\frac{ND_n^2 \log^{k+3} d}{n^k}\right)^{1/4} - C\breve{\varpi}_n.$$

Finally, arguing as in Step 5 in the proof of Theorem 3.1, we obtain the conclusion of Theorem 3.3 for the Bernoulli sampling case.

Sampling with replacement case. This case is similar to but easier than the Bernoulli sampling case under degeneracy. Recall that $U'_{n,N} = A_n + B_n$, where $A_n = U_n$ and $B_n = N^{-1}\sum_{j=1}^{N}\{h(X^*_{\iota_j}) - U_n\}$. Under the assumptions that $D_n^2(\log n)\log^5(dn) \leqslant c_2 n$ and $D_n^2 \log^7(dn) \leqslant c_2 N$, all the substeps of Step 1 in the proof of Theorem 3.1 carry over to the degenerate case, i.e., we have that

$$\rho^{\mathcal{R}}_{|X_1^n}(\sqrt{N}B_n, \widehat{Y}) \leqslant C\left\{\left(\frac{D_n^2 \log^7 d}{N}\right)^{1/6} + \frac{N^{3/2}e^{-N^{1/6}/C}}{D_n}\right\}$$

with probability at least $1 - Cn^{-1}$. In addition, the error bound in Step 2 remains the same as the Bernoulli sampling case under degeneracy. Hence, we have that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|X_1^n}(\sqrt{N} B_n \in R) - \gamma_B(R) \right| \leqslant C \breve{\varpi}_n$$

with probability at least $1 - Cn^{-1}$. Now, using the estimate (21), for $R = \prod_{j=1}^d [a_j, b_j], a = (a_1, \dots, a_d)^T, b = (b_1, \dots, b_d)^T$, and $t > 0$, we have

$$\begin{aligned}
\mathbb{P}(\sqrt{N} U'_{n,N} \in R) &= \mathbb{P}(\{-\sqrt{N} U'_{n,N} \leqslant -a\} \cap \{\sqrt{N} U'_{n,N} \leqslant b\}) \\
&\leqslant \mathbb{P}\left(\{-\sqrt{N} U'_{n,N} \leqslant -a\} \cap \{\sqrt{N} U'_{n,N} \leqslant b\} \cap \left\{|\sqrt{N} A_n|_\infty \leqslant t\right\}\right) + \mathbb{P}\left(|\sqrt{N} A_n|_\infty > t\right) \\
&\leqslant \mathbb{P}\left(\{-\sqrt{N} B_n \leqslant -a + t\} \cap \{\sqrt{N} B_n \leqslant b + t\}\right) + Ct^{-1}\sqrt{N} n^{-k/2} D_n \log^{k/2+1} d \\
&\leqslant \gamma_B(\{y \in \mathbb{R}^d : -y \leqslant -a + t, y \leqslant b + t\}) + C\breve{\varpi}_n + Ct^{-1}\sqrt{N} n^{-k/2} D_n \log^{k/2+1} d \\
&\leqslant \gamma_B(R) + Ct\sqrt{\log d} + C\breve{\varpi}_n + Ct^{-1}\sqrt{N} n^{-k/2} D_n \log^{k/2+1} d,
\end{aligned}$$

where the last inequality follows from Nazarov's inequality ([12], Lemma A.1). Choosing $t = (Nn^{-k} D_n^2 \log^{k+1} d)^{1/4}$, we conclude that

$$\mathbb{P}(\sqrt{N} U'_{n,N} \in R) \leqslant \gamma_B(R) + C\left(\frac{N D_n^2 \log^{k+3} d}{n^k}\right)^{1/4} + C\breve{\varpi}_n.$$

Likewise, we have the reverse inequality and the conclusion of Theorem 3.3 for the sampling with replacement case follows. □

A.3. **Proofs of Theorems 4.1 and 4.2.** As before, we will assume that $\theta = P^r h = 0$. Throughout 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 $\underline{\sigma}, r$, and C_1. Let C denote a generic constant that depends only on $\underline{\sigma}, r$, and C_1; its value may change from place to place. Recall that $Y_A \sim N(0, r^2 \Gamma_r) = \gamma_A$ and $Y_B \sim N(0, \Gamma_h) = \gamma_B$, and Y_A and Y_B are independent. Define

$$\rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^\sharp_{n,\star}, Y_\star) := \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(U^\sharp_{n,\star} \in R) - \mathbb{P}(Y_\star \in R) \right|, \ \star = A, B.$$

Proof of Theorem 4.1. Bernoulli sampling case. Conditionally on $\mathcal{D}_n$, the vector $U^\sharp_{n,B}$ is Gaussian with mean zero and covariance matrix

$$\frac{1}{\widehat{N}} \sum_{\iota \in I_{n,r}} Z_\iota \{h(X_\iota) - U'_{n,N}\}\{h(X_\iota) - U'_{n,N}\}^T.$$

On the other hand, $Y_B \sim N(0, \Gamma_h)$ and $\min_{1 \leqslant j \leqslant d} P^r h_j^2 \geqslant \underline{\sigma}^2$. Hence, the Gaussian comparison inequality (Lemma A.4) yields that

$$\rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^\sharp_{n,B}, Y_B) \lesssim (\widehat{\Delta}_B \log^2 d)^{1/3}, \tag{22}$$

where $\widehat{\Delta}_B$ is defined by

$$\widehat{\Delta}_B = \left| \widehat{N}^{-1} \textstyle\sum_{\iota \in I_{n,r}} Z_\iota \{h(X_\iota) - U'_{n,N}\}\{h(X_\iota) - U'_{n,N}\}^T - \Gamma_h \right|_\infty.$$

Observe that

$$\begin{aligned}\widehat{\Delta}_B &\leqslant |N/\widehat{N}| \cdot \left(\left| N^{-1} \textstyle\sum_{\iota \in I_{n,r}} (Z_\iota - p_n) h(X_\iota) h(X_\iota)^T \right|_\infty + |\widehat{\Gamma}_h - \Gamma_h|_\infty \right) \\ &\quad + |N/\widehat{N} - 1| \cdot |\Gamma_h|_\infty + |U'_{n,N}|_\infty^2 \\ &=: |N/\widehat{N}| (\widehat{\Delta}_{B,1} + \widehat{\Delta}_{B,2}) + \widehat{\Delta}_{B,3} + \widehat{\Delta}_{B,4},\end{aligned}$$

where $\widehat{\Gamma}_h = |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h(X_\iota) h(X_\iota)^T$.

From Step 5 in the proof of Theorem 3.3, $|\widehat{N}/N - 1| \leqslant C(N^{-1/2} \log^{1/2} n + N^{-1} \log n) \leqslant CN^{-1/2} \log^{1/2} n \leqslant Cn^{-\zeta/2}$ with probability at least $1 - 2n^{-1}$. Choose the smallest n_0 such that $Cn^{-\zeta/2} \leqslant 1/2$ for all $n \geqslant n_0$. Clearly, n_0 depends only on $\underline{\sigma}, C_1$, and ζ, and since for $n < n_0$, the conclusion of the theorem is trivial by taking the constant C in (11) sufficiently large (the constant C in (11) can be taken independent of ζ), we may assume in what follows that $n \geqslant n_0$. Then, $|\widehat{N}/N - 1| \leqslant CN^{-1/2} \log^{1/2} n \leqslant 1/2$ with probability at least $1 - 2n^{-1}$, and hence using the inequality $|z^{-1} - 1| \leqslant 2|z - 1|$ for $|z - 1| \leqslant 1/2$, we have that $|N/\widehat{N} - 1| \leqslant CN^{-1/2} \log^{-1/2} n$ with probability at least $1 - 2n^{-1}$. In particular, $|N/\widehat{N}| \leqslant C$ with probability at least $1 - 2n^{-1}$. In addition, since $|\Gamma_h|_\infty \lesssim D_n$, we have that

$$\widehat{\Delta}_{B,3} \log^2 d \leqslant CD_n N^{-1/2} (\log^{1/2} d) \log^2 d \leqslant Cn^{-\zeta/2} \leqslant Cn^{-3\zeta/8}$$

with probability at least $1 - 2n^{-1}$.

For $\widehat{\Delta}_{B,2}$, Hoeffding's averaging (cf. [36], Section 5.1.6) together with the computation in Step 1.5 in the proof of Theorem 3.1 yield that

$$\begin{aligned}\mathbb{E}[\widehat{\Delta}_{B,2}] &\leqslant \mathbb{E}\left[\max_{1 \leqslant j,\ell \leqslant d} \left| m^{-1} \sum_{i=1}^m \{ h_j(X_{(i-1)r+1}^{ir}) h_\ell(X_{(i-1)r+1}^{ir}) - P^r h_j h_\ell \} \right| \right] \\ &\lesssim n^{-1/2} D_n \log^{1/2} d + n^{-1} D_n^2 \log^3(dn),\end{aligned}$$

where $m = \lfloor n/r \rfloor$. For $\widehat{\Delta}_{B,1}$, Lemma 8 in [11] (applied conditionally on X_1^n) yields that

$$\mathbb{E}_{|X_1^n}[N\widehat{\Delta}_{B,1}] \lesssim \sqrt{N(\log d) \max_{1 \leqslant j,\ell \leqslant d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_j^2(X_\iota) h_k^2(X_\iota)} + \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} h_j^2(X_\iota) \log d,$$

and $\mathbb{E}[\max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} h_j^2(X_\iota)] \lesssim D_n^2 \log^2(dn)$. In addition, Hoeffding's averaging together with Lemma 9 in [11] yield that

$$\begin{aligned}&\mathbb{E}\left[\max_{1 \leqslant j,\ell \leqslant d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_j^2(X_\iota) h_\ell^2(X_\iota) \right] \leqslant \mathbb{E}\left[\max_{1 \leqslant j,\ell \leqslant d} m^{-1} \sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir}) h_\ell^2(X_{(i-1)r+1}^{ir}) \right] \\ &\lesssim \max_{1 \leqslant j,\ell \leqslant d} m^{-1} \sum_{i=1}^m \mathbb{E}[h_j^2(X_{(i-1)r+1}^{ir}) h_\ell^2(X_{(i-1)r+1}^{ir})] + m^{-1} \mathbb{E}\left[\max_{1 \leqslant i \leqslant m} \max_{1 \leqslant j \leqslant d} h_j^4(X_{(i-1)r+1}^{ir}) \right] \log d \\ &\lesssim D_n^2 + n^{-1} D_n^4 \log^5(dn),\end{aligned}$$

so that by Fubini,

$$\begin{aligned}\mathbb{E}[\widehat{\Delta}_{B,1}] &\lesssim N^{-1/2} \{ D_n \log^{1/2} d + n^{-1/2} D_n^2 \log^3(dn) \} + N^{-1} D_n^2 \log^3(dn) \\ &\lesssim N^{-1/2} D_n \log^{1/2} d + (n \wedge N)^{-1} D_n^2 \log^3(dn).\end{aligned}$$

Hence,

$$(\mathbb{E}[\widehat{\Delta}_{B,1}+\widehat{\Delta}_{B,2}])\log^2 d \lesssim (n\wedge N)^{-1/2}D_n\log^{5/2} d+(n\wedge N)^{-1}D_n^2\log^5(dn)\lesssim n^{-\zeta/2}$$

by Condition (9), so that by Markov's inequality,

$$(\widehat{\Delta}_{B,1}+\widehat{\Delta}_{B,2})\log^2 d\leqslant n^{-3\zeta/8}$$

with probability at least $1-Cn^{-\zeta/8}$.

Finally, for $\widehat{\Delta}_{B,4}$, observe that

$$\widehat{\Delta}_{B,4}=|N/\widehat{N}|^2|W_n|_\infty^2\leqslant 2|N/\widehat{N}|^2(|A_n|_\infty^2+|B_n|_\infty^2),$$

where $A_n=U_n=|I_{n,r}|^{-1}\sum_{\iota\in I_{n,r}}h(X_\iota)$ and $B_n=|I_{n,r}|^{-1}\sum_{\iota\in I_{n,r}}p_n^{-1}(Z_\iota-p_n)h(X_\iota)$. Conditionally on X_1^n, the Hoffmann-Jørgensen inequality yields that

$$\mathbb{E}_{|X_1^n}[|B_n|_\infty^2]\lesssim(\mathbb{E}_{|X_1^n}[|B_n|_\infty])^2+N^{-2}\max_{\iota\in I_{n,r}}\max_{1\leqslant j\leqslant d}h_j^2(X_\iota),$$

and Lemma 8 in [11] yields that

$$\mathbb{E}_{|X_1^n}[|B_n|_\infty]\lesssim\sqrt{N^{-1}(\log d)\max_{1\leqslant j\leqslant d}|I_{n,r}|^{-1}\sum_{\iota\in I_{n,r}}h_j^2(X_\iota)}+N^{-1}(\log d)\max_{\iota\in I_{n,r}}\max_{1\leqslant j\leqslant d}|h_j(X_\iota)|.$$

Hence,

$$\mathbb{E}_{|X_1^n}[|B_n|_\infty^2]\lesssim N^{-1}(\log d)\max_{1\leqslant j\leqslant d}|I_{n,r}|^{-1}\sum_{\iota\in I_{n,r}}h_j^2(X_\iota)+N^{-2}(\log d)^2\max_{\iota\in I_{n,r}}\max_{1\leqslant j\leqslant d}h_j^2(X_\iota),$$

and $\mathbb{E}[\max_{\iota\in I_{n,r}}\max_{1\leqslant j\leqslant d}h_j^2(X_\iota)]\lesssim D_n^2\log^2(dn)$. In addition, Hoeffding's averaging together with Lemma 9 in [11] yield that

$$\begin{aligned}\mathbb{E}\left[\max_{1\leqslant j\leqslant d}|I_{n,r}|^{-1}\sum_{\iota\in I_{n,r}}h_j^2(X_\iota)\right]&\leqslant\mathbb{E}\left[\max_{1\leqslant j\leqslant d}m^{-1}\sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir})\right]\\&\lesssim\max_{1\leqslant j\leqslant d}m^{-1}\sum_{i=1}^m\mathbb{E}[h_j^2(X_{(i-1)r+1}^{ir})]+m^{-1}\mathbb{E}\left[\max_{1\leqslant i\leqslant m}\max_{1\leqslant j\leqslant d}h_j^2(X_{(i-1)r+1}^{ir})\right]\log d\\&\lesssim D_n+n^{-1}D_n^2\log^3(dn),\end{aligned}$$

so that by Fubini,

$$\mathbb{E}[|B_n|_\infty^2]\lesssim N^{-1}D_n\log d+(nN)^{-1}D_n^2\log^4(dn)+N^{-2}D_n^2\log^4(dn).$$

Next, applying Hoeffding's averaging and Jensen's inequality, we have

$$\mathbb{E}[|A_n|_\infty^2]\leqslant\mathbb{E}\left[\max_{1\leqslant j\leqslant d}\left|m^{-1}\sum_{i=1}^m h_j(X_{(i-1)r+1}^{ir})\right|^2\right]$$

which is bounded from above by

$$\lesssim n^{-1}(\log d)\mathbb{E}\left[\max_{1\leqslant j\leqslant d}h_j^2(X_1^r)\right]\lesssim n^{-1}D_n^2\log^2 d$$

by Theorem 2.14.1 in [39]. Hence,

$$
\begin{aligned}
(\mathbb{E}[|A_n|_\infty^2 + |B_n|_\infty^2]) \log^2 d &\lesssim n^{-1} D_n^2 \log^4 d + N^{-1} D_n \log^3 d \\
&\qquad + (nN)^{-1} D_n^2 \log^6(dn) + N^{-2} D_n^2 \log^6(dn) \lesssim n^{-\zeta} \leqslant n^{-\zeta/2},
\end{aligned}
$$

and by Markov's inequality,

$$
(|A_n|_\infty^2 + |B_n|_\infty^2) \log^2 d \leqslant n^{-3\zeta/8}
$$

with probability at least $1 - Cn^{-\zeta/8}$.

In conclusion, we have that $\widehat{\Delta}_B \log^2 d \leqslant Cn^{-3\zeta/8}$ with probability at least $1 - Cn^{-\zeta/8}$, and in view of (22), this leads to the desired conclusion.

Sampling with replacement case. Conditionally on $\mathcal{D}_n$, the vector $U_{n,B}^\sharp$ is Gaussian with mean zero and covariance matrix

$$
\frac{1}{N} \sum_{j=1}^N \{h(X_{\iota_j}^*) - U'_{n,N}\}\{h(X_{\iota_j}^*) - U'_{n,N}\}^T.
$$

In view of the previous proof, it suffices to prove that $\widehat{\Delta}_B \log^2 \leqslant Cn^{-3\zeta/8}$ with probability at least $1 - Cn^{-\zeta/8}$, where $\widehat{\Delta}_B$ is now defined by

$$
\widehat{\Delta}_B = \left| N^{-1} \textstyle\sum_{j=1}^N \{h(X_{\iota_j}^*) - U'_{n,N}\}\{h(X_{\iota_j}^*) - U'_{n,N}\}^T - \Gamma_h \right|_\infty .
$$

To this end, by Markov's inequality, it suffices to prove that $\mathbb{E}[\widehat{\Delta}_B] \log^2 d \lesssim n^{-\zeta/2}$. Observe that

$$
\begin{aligned}
\widehat{\Delta}_B &\leqslant \left| N^{-1} \textstyle\sum_{j=1}^N h(X_{\iota_j}^*) h(X_{\iota_j}^*)^T - \widehat{\Gamma}_h \right|_\infty + |\widehat{\Gamma}_h - \Gamma_h|_\infty + |U'_{n,N}|_\infty^2 \\
&=: \widehat{\Delta}_{B,1} + \widehat{\Delta}_{B,2} + \widehat{\Delta}_{B,3}.
\end{aligned}
$$

We have shown that $\mathbb{E}[\widehat{\Delta}_{B,2}] \lesssim n^{-1/2} D_n \log^{1/2} d + n^{-1} D_n^2 \log^3(dn)$ and so $\mathbb{E}[\widehat{\Delta}_{B,2}] \log^2 d \lesssim n^{-\zeta/2}$. In addition, $\widehat{\Delta}_{B,3} \leqslant 2(|A_n|_\infty^2 + |B_n|_\infty^2)$ where $A_n = U_n$ and $B_n = N^{-1} \sum_{j=1}^N \{h(X_{\iota_j}^*) - U_n\}$. We have shown that $\mathbb{E}[|A_n|_\infty^2] \lesssim n^{-1} D_n^2 \log^2 d$. Next, since $h(X_{\iota_j}^*), j = 1, \dots, N$ are i.i.d. with mean U_n conditionally on X_1^n, by the Hoffmann-Jørgensen inequality and Lemma 8 in [11], we have

$$
\begin{aligned}
&\mathbb{E}_{|X_1^n}[|B_n|^2] \lesssim (\mathbb{E}_{|X_1^n}[|B_n|_\infty])^2 + N^{-2} \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} h_j^2(X_\iota), \quad \text{and} \\
&\mathbb{E}_{|X_1^n}[|B_n|_\infty] \lesssim \sqrt{N^{-1} (\log d) \max_{1 \leqslant j \leqslant d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_j^2(X_\iota)} + N^{-1} (\log d) \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)|.
\end{aligned}
$$

Hence, using the calculations in the previous proof, we have

$$
\mathbb{E}[|B_n|_\infty^2] \lesssim N^{-1} D_n \log d + (nN)^{-1} D_n^2 \log^4(dn) + N^{-2} D_n^2 \log^4(dn),
$$

and hence $\mathbb{E}[\widehat{\Delta}_{B,3}] \log^2 d \lesssim n^{-\zeta/2}$. Finally, by Lemma 8 in [11], we have

$$
\mathbb{E}_{|X_1^n}[\widehat{\Delta}_{B,1}] \lesssim \sqrt{N^{-1} (\log d) \max_{1 \leqslant j,k \leqslant d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_j^2(X_\iota) h_k^2(X_\iota)} + N^{-1} \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} h_j^2(X_\iota) \log d,
$$

and by the calculations in the previous proof, we have

$$
\mathbb{E}[\widehat{\Delta}_{B,1}] \lesssim N^{-1/2} D_n \log^{1/2} d + (n \wedge N)^{-1} D_n^2 \log^3(dn).
$$

Hence, $\mathbb{E}[\widehat{\Delta}_{B,1}] \log^2 d \lesssim n^{-\zeta/2}$. This completes the proof. □

Proof of Theorem 4.2. The proof is divided into three steps.

Step 1: Bounding $\rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^{\sharp}_{n,B}, Y_B)$. Since Condition (C3-ND) implies Condition (C3-D), and $n_1 \leqslant n$ by definition, by Theorem 4.1, we have that

$$\rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^{\sharp}_{n,B}, Y_B) \leqslant Cn^{-\zeta/8}$$

with probability at least $1 - Cn^{-\zeta/8}$.

Step 2: Bounding $\rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^{\sharp}_{n,A}, Y_A)$. In this step, we shall show that

$$\rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^{\sharp}_{n,A}, Y_A) \leqslant Cn^{-\zeta/8}$$

with probability at least $1 - Cn^{-\zeta/8}$.

Without loss of generality, we may assume $S_1 = \{1, \ldots, n_1\}$. Conditionally on $\mathcal{D}_n$, the vector $U^{\sharp}_{n,A}$ is Gaussian with mean zero and covariance matrix

$$\frac{r^2}{n_1} \sum_{i_1=1}^{n_1} \{\widehat{g}^{(i_1)}(X_{i_1}) - \breve{g}\}\{\widehat{g}^{(i_1)}(X_{i_1}) - \breve{g}\}^T.$$

On the other hand, $Y_A \sim N(0, r^2\Gamma_g)$ and $\min_{1\leqslant j\leqslant d} Pg_j^2 \geqslant \underline{\sigma}^2$. Hence, the Gaussian comparison inequality (Lemma A.4) yields that

$$\rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^{\sharp}_{n,A}, Y_A) \lesssim (\widehat{\Delta}_A \log^2 d)^{1/3},$$

where

$$\widehat{\Delta}_A = \max_{1\leqslant j,\ell\leqslant d} \left| n_1^{-1} \sum_{i_1=1}^{n_1} \{\widehat{g}^{(i_1)}_j(X_{i_1}) - \breve{g}_j\}\{\widehat{g}^{(i_1)}_\ell(X_{i_1}) - \breve{g}_\ell\} - Pg_j g_\ell \right|.$$

Observe that for every $1 \leqslant j, \ell \leqslant d$,

$$\begin{aligned}
& n_1^{-1} \sum_{i_1=1}^{n_1} \{\widehat{g}^{(i_1)}_j(X_{i_1}) - \breve{g}_j\}\{\widehat{g}^{(i_1)}_\ell(X_{i_1}) - \breve{g}_\ell\} \\
&= n_1^{-1} \sum_{i_1=1}^{n_1} \widehat{g}^{(i_1)}_j(X_{i_1})\widehat{g}^{(i_1)}_\ell(X_{i_1}) - \breve{g}_j\breve{g}_\ell \\
&= n_1^{-1} \sum_{i_1=1}^{n_1} \{\widehat{g}^{(i_1)}_j(X_{i_1}) - g_j(X_{i_1})\}\{\widehat{g}^{(i_1)}_\ell(X_{i_1}) - g_\ell(X_{i_1})\} \\
&\quad + n_1^{-1} \sum_{i_1=1}^{n_1} \{\widehat{g}^{(i_1)}_j(X_{i_1}) - g_j(X_{i_1})\}g_\ell(X_{i_1}) + n_1^{-1} \sum_{i_1=1}^{n_1} \{\widehat{g}^{(i_1)}_\ell(X_{i_1}) - g_\ell(X_{i_1})\}g_j(X_{i_1}) \\
&\quad + n_1^{-1} \sum_{i_1=1}^{n_1} g_j(X_{i_1})g_\ell(X_{i_1}) - \breve{g}_j\breve{g}_\ell,
\end{aligned}$$

so that by the Cauchy-Schwarz inequality,

$$\widehat{\Delta}_A \leqslant \underbrace{\max_{1\leqslant j\leqslant d} n_1^{-1}\sum_{i_1=1}^{n_1}\{\widehat{g}_j^{(i_1)}(X_{i_1})-g_j(X_{i_1})\}^2}_{=\widehat{\Delta}_{A,1}} + 2\widehat{\Delta}_{A,1}^{1/2}\sqrt{\max_{1\leqslant j\leqslant d} n_1^{-1}\sum_{i_1=1}^{n_1} g_j^2(X_{i_1})}$$
$$+ \max_{1\leqslant j,\ell\leqslant d}\left| n_1^{-1}\sum_{i_1=1}^{n_1}\{g_j(X_{i_1})g_\ell(X_{i_1})-Pg_jg_\ell\}\right| + \max_{1\leqslant j\leqslant d}|\breve{g}_j|^2.$$

For the notational convenience, define

$$\widehat{\Delta}_{A,2} := \max_{1\leqslant j,\ell\leqslant d}\left| n_1^{-1}\sum_{i_1=1}^{n_1}\{g_j(X_{i_1})g_\ell(X_{i_1})-Pg_jg_\ell\}\right|.$$

Then, since $n_1^{-1}\sum_{i_1=1}^{n_1} g_j^2(X_{i_1}) \leqslant Pg_j^2 + |n_1^{-1}\sum_{i_1=1}^{n_1}\{g_j^2(X_{i_1})-Pg_j^2\}| \leqslant \overline{\sigma}_g^2 + \widehat{\Delta}_{A,2}$, and $\breve{g}_j = n_1^{-1}\sum_{i_1=1}^{n}\{g_j^{(i_1)}(X_{i_1})-g_j(X_{i_1})\} + n_1^{-1}\sum_{i_1=1}^{n_1} g_j(X_{i_1})$, so that

$$\max_{1\leqslant j\leqslant d}|\breve{g}_j|^2 \lesssim \widehat{\Delta}_{A,1} + \widehat{\Delta}_{A,3}^2, \quad \text{with } \widehat{\Delta}_{A,3} := \max_{1\leqslant j\leqslant d}\left| n_1^{-1}\sum_{i_1=1}^{n_1} g_j(X_{i_1})\right|,$$

we have

$$\widehat{\Delta}_A \lesssim \overline{\sigma}_g\widehat{\Delta}_{A,1}^{1/2} + \widehat{\Delta}_{A,1} + \widehat{\Delta}_{A,2} + \widehat{\Delta}_{A,3}^2,$$

where we have used the inequality $2ab \leqslant a^2+b^2$ for $a,b\in\mathbb{R}$.

Now, by assumption, $\overline{\sigma}_g\widehat{\Delta}_{A,1}^{1/2}\log^2 d \leqslant Cn^{-3\zeta/8}$ with probability at least $1-Cn^{-\zeta/8}$. For $\widehat{\Delta}_{A,2}$, Lemma 8 in [11] yields that

$$\begin{aligned}\mathbb{E}[\widehat{\Delta}_{A,2}] &\lesssim n_1^{-1}\sqrt{(\log d)\max_{1\leqslant j,\ell\leqslant d}\sum_{i_1=1}^{n_1}\mathbb{E}[g_j^2(X_{i_1})g_\ell^2(X_{i_1})]} + n_1^{-1}\sqrt{\mathbb{E}\left[\max_{1\leqslant i_1\leqslant n_1}\max_{1\leqslant j\leqslant d} g_j^4(X_{i_1})\right]}\log d \\ &\lesssim n_1^{-1/2}D_n\log^{1/2} d + n_1^{-1}D_n^2\log^3(dn).\end{aligned} \tag{23}$$

For $\widehat{\Delta}_{A,3}^2$, the Hoffmann-Jørgensen inequality [39, Proposition A.1.6] yields that

$$\mathbb{E}[\widehat{\Delta}_{A,3}^2] \lesssim (\mathbb{E}[\widehat{\Delta}_{A,3}])^2 + n_1^{-2}\mathbb{E}\left[\max_{1\leqslant i_1\leqslant n_1}\max_{1\leqslant j\leqslant d} g_j^2(X_{i_1})\right] \lesssim (\mathbb{E}[\widehat{\Delta}_{A,3}])^2 + n_1^{-2}D_n^2\log^2(dn),$$

where Lemma 8 in [11] yields that

$$\begin{aligned}\mathbb{E}[\widehat{\Delta}_{A,3}] &\lesssim n_1^{-1}\sqrt{(\log d)\max_{1\leqslant j\leqslant d}\sum_{i_1=1}^{n_1}\mathbb{E}[g_j^2(X_{i_1})]} + n_1^{-1}\sqrt{\mathbb{E}\left[\max_{1\leqslant i_1\leqslant n_1}\max_{1\leqslant j\leqslant d} g_j^2(X_{i_1})\right]}\log d \\ &\lesssim n_1^{-1/2}\overline{\sigma}_g\log^{1/2} d + n_1^{-1}D_n\log^2(dn).\end{aligned}$$

Hence, using $\overline{\sigma}_g^2 \leqslant 1+\max_j P|g_j|^3 \lesssim D_n$, we have

$$\mathbb{E}[\widehat{\Delta}_{A,3}^2] \lesssim n_1^{-1}D_n\log d + n_1^{-2}D_n^2\log^4(dn). \tag{24}$$

Combining (23) and (24), and using Condition (10), we conclude that

$$(\mathbb{E}[\widehat{\Delta}_{A,2}+\widehat{\Delta}_{A,3}^2])\log^2 d \lesssim n_1^{-1/2}D_n\log^{5/2}(dn) \lesssim n^{-\zeta/2},$$

so that by Markov's inequality, $(\widehat{\Delta}_{A,2}+\widehat{\Delta}_{A,3}^2)\log^2 d \leqslant Cn^{-3\zeta/8}$ with probability at least $1-Cn^{-\zeta/8}$, which leads to the conclusion of this step.

Step 3: Conclusion. Let $\Xi = \{\xi_{i_1} : i_1 \in S_1\}$ and $\Xi' = \{\xi'_\iota : \iota \in I_{n,r}\}$. Recall that Ξ, Ξ', and $\mathcal{D}_n$ are mutually independent. Suppose that

$$\rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^\sharp_{n,A}, Y_A) \bigvee \rho^{\mathcal{R}}_{|\mathcal{D}_n}(U^\sharp_{n,B}, Y_B) \leqslant Cn^{-\zeta/8},$$

which holds with probability at least $1 - Cn^{-\zeta/8}$. Pick any hyperrectangle $R \in \mathcal{R}$. Observe that

$$\mathbb{P}_{|\mathcal{D}_n}(U^\sharp_n \in R) = \mathbb{E}_{|\mathcal{D}_n}\left[\mathbb{P}_{|(\mathcal{D}_n,\Xi)}\left(U^\sharp_{n,B} \in [\alpha_n^{-1/2}R - \alpha_n^{-1/2}U^\sharp_{n,A}]\right)\right].$$

The conditional probability on the right hand side is bounded from above by $\gamma_B([\alpha_n^{-1/2}R - \alpha_n^{-1/2}U^\sharp_{n,A}]) + Cn^{-\zeta/8}$, and hence

$$\begin{aligned}\mathbb{P}_{|\mathcal{D}_n}(U^\sharp_n \in R) &\leqslant \mathbb{E}_{|\mathcal{D}_n}\left[\gamma_B\left([\alpha_n^{-1/2}R - \alpha_n^{-1/2}U^\sharp_{n,A}]\right)\right] + Cn^{-\zeta/8}\\ &= \mathbb{P}_{|\mathcal{D}_n}\left(\breve{Y}_B \in [\alpha_n^{-1/2}R - \alpha_n^{-1/2}U^\sharp_{n,A}]\right) + Cn^{-\zeta/8} = \mathbb{P}_{|\mathcal{D}_n}\left(U^\sharp_{n,A} \in [R - \alpha_n^{1/2}\breve{Y}_B]\right) + Cn^{-\zeta/8},\end{aligned}$$

where $\breve{Y}_B \sim N(0,(1-p_n)\Gamma_h)$ independent of $\mathcal{D}_n$ and Ξ. The first term on the far right hand side can be written as $\mathbb{E}_{|\mathcal{D}_n}[\mathbb{P}_{|(\mathcal{D}_n,\breve{Y}_B)}(U^\sharp_{n,A} \in [R - \alpha_n^{1/2}\breve{Y}_B])]$, and the inner conditional probability is bounded from above by $\gamma_A([R-\alpha_n^{1/2}\breve{Y}_B]) + Cn^{-\zeta/8}$. Hence,

$$\mathbb{P}_{|\mathcal{D}_n}(U^\sharp_n \in R) \leqslant \mathbb{E}_{|\mathcal{D}_n}\left[\gamma_A\left([R - \alpha_n^{1/2}\breve{Y}_B]\right)\right] + Cn^{-\zeta/8} = \mathbb{P}\left(Y \in R\right) + Cn^{-\zeta/8}.$$

Likewise, we have $\mathbb{P}_{|\mathcal{D}_n}(U^\sharp_n \in R) \geqslant \mathbb{P}(Y \in R) - Cn^{-\zeta/8}$.

Finally, the last statement of the theorem is trivial since the bootstrap distribution is taken only with respect to $\{\xi_{i_1} : i_1 \in S_1\}$ and $\{\xi'_\iota : \iota \in I_{n,r}\}$. This completes the proof. □

Proof of Corollary 4.3. This follows from Step 2 in the proof of Theorem 4.2. □

A.4. **Proofs of Proposition 4.4 and 4.5.** For the notational convenience, let $H = \max_{1\leqslant j\leqslant d}|h_j|$. For each fixed $x \in S$, denote by $\delta_x h$ the function of $(r-1)$ variables, $(\delta_x h)(x_2,\dots,x_r) = h(x,x_2,\dots,x_r)$.

Proof of Proposition 4.4. 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 and q. For each $i_1 \in S_1$ and $k = 1,\dots,K$, let

$$g^{(i_1,k)}(x) = \frac{1}{|I_{L,r-1}|}\sum_{\substack{i_2,\dots,i_r \in S^{(i_1)}_{2,k}\\ i_2<\dots<i_r}} (\delta_x h)(X_{i_2},\dots,X_{i_r}),$$

which is the U-statistic with kernel $\delta_x h$ for the sample $\{X_i : i \in S^{(i_1)}_{2,k}\}$. Recall that the size of each block $S^{(i_1)}_{2,k}$ is L, $|S^{(i_1)}_{2,k}| = L$. Then, $\widehat{g}^{(i_1)}(x)$ is the average of $g^{(i_1,k)}(x), k = 1,\dots,K$,

$$\widehat{g}^{(i_1)}(x) = \frac{1}{K}\sum_{k=1}^K g^{(i_1,k)}(x).$$

For each $i_1 \in S_1$, since the blocks $S_{2,k}^{(i_1)}, k = 1, \dots, K$ are disjoint and do not contain i_1, the vectors $g^{(i_1,k)}(X_{i_1}), k = 1, \dots, K$ are independent with mean $g(X_{i_1})$ conditionally on X_{i_1}. Hence, applying first the Hoffmann-Jørgensen inequality [39, Proposition A.1.6] conditionally on X_{i_1}, we have

$$\begin{aligned}
&\mathbb{E}_{|X_{i_1}}\left[\left\{\max_{1\leqslant j\leqslant d}\left|\widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1})\right|\right\}^2\right] \\
&\quad\lesssim \left(\mathbb{E}_{|X_{i_1}}\left[\max_{1\leqslant j\leqslant d}\left|\widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1})\right|\right]\right)^2 + K^{-2}\mathbb{E}_{|X_{i_1}}\left[\max_{1\leqslant k\leqslant K}\max_{1\leqslant j\leqslant d}|g_j^{(i_1,k)}(X_{i_1}) - g_j(X_{i_1})|^2\right].
\end{aligned}$$

Further, applying Lemma 8 in [11] conditionally on X_{i_1}, we have

$$\begin{aligned}
&\mathbb{E}_{|X_{i_1}}\left[\max_{1\leqslant j\leqslant d}\left|\widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1})\right|\right] \\
&\quad\lesssim K^{-1}\sqrt{(\log d)\max_{1\leqslant j\leqslant d}\sum_{k=1}^{K}\mathbb{E}_{|X_{i_1}}\left[\left\{\widehat{g}_j^{(i_1,k)}(X_{i_1}) - g_j(X_{i_1})\right\}^2\right]} \\
&\qquad + K^{-1}\sqrt{\mathbb{E}_{|X_{i_1}}\left[\max_{1\leqslant k\leqslant K}\max_{1\leqslant j\leqslant d}|g_j^{(i_1,k)}(X_{i_1}) - g_j(X_{i_1})|^2\right]}\log d.
\end{aligned}$$

From the variance formula for U-statistics (cf. [28], Theorem 3), we have

$$\begin{aligned}
\mathbb{E}_{|X_{i_1}}\left[\left\{\widehat{g}_j^{(i_1,k)}(X_{i_1}) - g_j(X_{i_1})\right\}^2\right] &\leqslant \binom{L}{r-1}^{-1}\sum_{\ell=1}^{r-1}\binom{r-1}{\ell}\binom{L-r+1}{r-1-\ell}P^{r-1}(\delta_{X_{i_1}}h_j)^2 \\
&\lesssim L^{-1}P^{r-1}(\delta_{X_{i_1}}h_j)^2.
\end{aligned}$$

It remains to bound

$$\mathbb{E}_{|X_{i_1}}\left[\max_{1\leqslant k\leqslant K}\max_{1\leqslant j\leqslant d}|g_j^{(i_1,k)}(X_{i_1}) - g_j(X_{i_1})|^2\right]. \tag{25}$$

Observe that the term (25) is bounded from above by

$$\left(\sum_{k=1}^{K}\mathbb{E}_{|X_{i_1}}\left[\max_{1\leqslant j\leqslant d}|g_j^{(i_1,k)}(X_{i_1}) - g_j(X_{i_1})|^{2q}\right]\right)^{1/q}.$$

Applying Hoeffiding's averaging and Theorem 2.14.1 in [39], we have

$$\mathbb{E}_{|X_{i_1}}\left[\max_{1\leqslant j\leqslant d}|g_j^{(i_1,k)}(X_{i_1}) - g_j(X_{i_1})|^{2q}\right] \lesssim L^{-q}(\log d)^q P^{r-1}|\delta_{X_{i_1}}H|^{2q}.$$

Hence, the term (25) is bounded from above by $K^{1/q}L^{-1}(\log d)(P^{r-1}|\delta_{X_{i_1}}H|^{2q})^{1/q}$ up to a constant that depends only on r and q. By Fubini and Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\{\max_{1\leqslant j\leqslant d}\left|\widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1})\right|\right\}^2\right] \\
&\lesssim (KL)^{-1}(\log d)\mathbb{E}\left[P^{r-1}|\delta_{X_{i_1}}H|^2\right] + K^{-2+1/q}L^{-1}(\log^3 d)\mathbb{E}\left[\left(P^{r-1}|\delta_{X_{i_1}}H|^{2q}\right)^{1/q}\right] \\
&\lesssim (KL)^{-1}(\log d)P^rH^2 + K^{-2+1/q}L^{-1}(\log^3 d)(P^rH^{2q})^{1/q} \\
&\lesssim (KL)^{-1}D_n^2\log^3 d + K^{-2+1/q}L^{-1}D_n^2\log^5 d,
\end{aligned}$$

from which we conclude that

$$\overline{\sigma}_g^2 \mathbb{E}[\widehat{\Delta}_{A,1}] \log^4 d \lesssim \overline{\sigma}_g^2 (KL)^{-1} D_n^2 (\log^7 d)(1 + K^{-1+1/q} \log^2 d) \lesssim n^{-7\zeta/8},$$

so that by Markov's inequality,

$$\mathbb{P}\left(\overline{\sigma}_g^2 \widehat{\Delta}_{A,1} \log^4 d > n^{-3\zeta/4}\right) \lesssim n^{-\zeta/8}.$$

In view of Theorem 4.2 and Corollary 4.3, this leads to the conclusion of the proposition. □

Proof of Proposition 4.5. 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 and C_1. Observe that

$$\begin{aligned}\widehat{g}^{(i_1)}(X_{i_1}) - g(X_{i_1}) = M^{-1} &\sum_{\iota' \in I_{n-1,r-1}} (Z'_{\iota'} - \vartheta_n)(\delta_{X_{i_1}} h)(X_{\sigma_{i_1}(\iota')}) \\ &+ |I_{n-1,r-1}|^{-1} \sum_{\iota' \in I_{n-1,r-1}} \{(\delta_{X_{i_1}} h)(X_{\sigma_{i_1}(\iota')}) - g(X_{i_1})\}.\end{aligned}$$

Conditionally on X_1^n, the first term is the sum of centered independent random vectors, and hence the Hoffmann-Jørgensen inequality yields that

$$\begin{aligned}&\mathbb{E}_{|X_1^n}\left[\max_{1 \leqslant j \leqslant d}\left|\sum_{\iota' \in I_{n-1,r-1}} (Z'_{\iota'} - \vartheta_n)(\delta_{X_{i_1}} h)(X_{\sigma_{i_1}(\iota')})\right|^2\right] \\ &\lesssim \left(\mathbb{E}_{|X_1^n}\left[\max_{1 \leqslant j \leqslant d}\left|\sum_{\iota' \in I_{n-1,r-1}} (Z'_{\iota'} - \vartheta_n)(\delta_{X_{i_1}} h)(X_{\sigma_{i_1}(\iota')})\right|\right]\right)^2 + \max_{\iota' \in I_{n-1,r-1}} \max_{1 \leqslant j \leqslant d} |(\delta_{X_{i_1}} h_j)(X_{\sigma_{i_1}(\iota')})|^2.\end{aligned}$$

By Lemma 8 in [11],

$$\begin{aligned}&\mathbb{E}_{|X_1^n}\left[\max_{1 \leqslant j \leqslant d}\left|\sum_{\iota' \in I_{n-1,r-1}} (Z'_{\iota'} - \vartheta_n)(\delta_{X_{i_1}} h)(X_{\sigma_{i_1}(\iota')})\right|\right] \\ &\lesssim \sqrt{M(\log d) \max_{1 \leqslant j \leqslant d} |I_{n-1,r-1}|^{-1} \sum_{\iota' \in I_{n-1,r-1}} (\delta_{X_{i_1}} h_j)^2(X_{\sigma_{i_1}(\iota')})} \\ &\quad + (\log d) \max_{\iota' \in I_{n-1,r-1}} \max_{1 \leqslant j \leqslant d} |(\delta_{X_{i_1}} h_j)(X_{\sigma_{i_1}(\iota')})|.\end{aligned}$$

Observe that

$$\max_{\iota' \in I_{n-1,r-1}} \max_{1 \leqslant j \leqslant d} |(\delta_{X_{i_1}} h_j)(X_{\sigma_{i_1}(\iota')})| \leqslant \max_{\iota \in I_{n,r}} \max_{1 \leqslant j \leqslant d} |h_j(X_\iota)|.$$

In addition, applying the Hoeffding averaging and Lemma 9 in [11] conditionally on X_{i_1}, we have

$$\begin{aligned}&\mathbb{E}_{|X_{i_1}}\left[\max_{1 \leqslant j \leqslant d} |I_{n-1,r-1}|^{-1} \sum_{\iota' \in I_{n-1,r-1}} (\delta_{X_{i_1}} h_j)^2(X_{\sigma_{i_1}(\iota')})\right] \\ &\lesssim \max_{1 \leqslant j \leqslant d} P^{r-1}(\delta_{X_{i_1}} h_j)^2 + n^{-1}(\log d)\mathbb{E}_{|X_{i_1}}\left[\max_{\iota' \in I_{n-1,r-1}} \max_{1 \leqslant j \leqslant d} |(\delta_{X_{i_1}} h_j)(X_{\sigma_{i_1}(\iota')})|^2\right].\end{aligned}$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\left[\max_{1\leqslant j\leqslant d}\left|M^{-1}\sum_{\iota'\in I_{n-1,r-1}}(Z'_{\iota'}-\vartheta_n)(\delta_{X_{i_1}}h)(X_{\sigma_{i_1}(\iota')})\right|^2\right]\\
&\lesssim M^{-1}(\log d)\left\{P^rH^2+n^{-1}(\log d)\mathbb{E}\left[\max_{\iota\in I_{n,r}}H^2(X_\iota)\right]\right\}+M^{-2}(\log d)^2\mathbb{E}\left[\max_{\iota\in I_{n,r}}H^2(X_\iota)\right]\\
&\lesssim M^{-1}(\log d)\{D_n^2\log^2 d+n^{-1}D_n^2\log^3(dn)\}+M^{-2}D_n^2\log^4(dn).
\end{aligned}
$$

On the other hand, applying Hoeffding averaging and Theorem 2.14.1 in [39] conditionally on X_{i_1}, we have

$$
\begin{aligned}
&\mathbb{E}_{|X_{i_1}}\left[\left|\max_{1\leqslant j\leqslant d}|I_{n-1,r-1}|^{-1}\sum_{\iota'\in I_{n-1,r-1}}\{(\delta_{X_{i_1}}h)(X_{\sigma_{i_1}(\iota')})-g(X_{i_1})\}\right|^2\right]\\
&\lesssim n^{-1}(\log d)P^{r-1}|\delta_{X_{i_1}}H|^2.
\end{aligned}
$$

The expectation of the left hand side is bounded by $\lesssim n^{-1}(\log d)P^rH^2\lesssim n^{-1}D_n^2\log^3 d$.

Therefore, using Condition (15), we conclude that

$$
\begin{aligned}
\overline{\sigma}_g^2\mathbb{E}[\widehat{\Delta}_{A,1}]\log^4 d &\lesssim \overline{\sigma}_g^2\left\{M^{-1}D_n^2\log^7 d+(nM)^{-1}D_n^2\log^8(dn)+M^{-2}D_n^2\log^8(dn)+n^{-1}D_n^2\log^7 d\right\}\\
&\lesssim n^{-7\zeta/8}.
\end{aligned}
$$

In view of Theorem 4.2 and Corollary 4.3, this leads to the conclusion of the proposition. □

A.5. **Proofs for Section 6.**

Proof of Lemma 6.1. Recall the notation used in the proofs of Theorems 4.1 and 4.2. Then we have $\max_{1\leqslant j\leqslant d}|\widehat{\sigma}_{A,j}^2-\sigma_{A,j}^2|\leqslant r^2\widehat{\Delta}_A$ and $\max_{1\leqslant j\leqslant d}|\widehat{\sigma}_{B,j}^2-\sigma_{B,j}^2|\leqslant\widehat{\Delta}_B$. Since $\min_{1\leqslant j\leqslant d}\sigma_{A,j}^2\geqslant r^2\underline{\sigma}^2$ in Case (i) and $\min_{1\leqslant j\leqslant d}\sigma_{B,j}^2\geqslant\underline{\sigma}^2$ in Case (ii), the conclusion of the lemma follows from the bounds on $\widehat{\Delta}_B$ and $\widehat{\Delta}_A$ established in the proofs of Theorems 4.1 and 4.2, respectively. □

Proof of Corollary 6.2. We only prove Case (i) since the proof for Case (ii) is analogous. As before, we will assume that $\theta=P^rh=0$. In this proof, let C denote a generic constant that depends only on $\underline{\sigma}, r$, and C_1; its value may change from place to place. In addition, without loss of generality, we may assume that $n^{-\zeta/8}\leqslant c_1$ for some sufficiently small constant c_1 depending only on $\underline{\sigma}, r$ and C_1, since otherwise the conclusion of Case (i) is trivial by taking C in the bounds sufficiently large (say, $C\geqslant 1/c_1$). We begin with noting that

$$
\left|\frac{\widehat{\sigma}_j^2}{\sigma_j^2}-1\right|=\left|\frac{\widehat{\sigma}_{A,j}^2-\sigma_{A,j}^2+\alpha_n(\widehat{\sigma}_{B,j}^2-\sigma_{B,j}^2)}{\sigma_{A,j}^2+\alpha_n\sigma_{B,j}^2}\right|\leqslant\left|\frac{\widehat{\sigma}_{A,j}^2}{\sigma_{A,j}^2}-1\right|+\left|\frac{\widehat{\sigma}_{B,j}^2}{\sigma_{B,j}^2}-1\right|,
$$

so that by Lemma 6.1, we have that $\max_{1\leqslant j\leqslant d}|\widehat{\sigma}_j^2/\sigma_j^2-1|\leqslant Cn^{-3\zeta/8}/\log^2 d$ with probability at least $1-Cn^{-\zeta/8}$. Choosing c_1 sufficiently small so that $Cn^{-3\zeta/8}/\log^2 d\leqslant 1/2$, and using the inequalities that $|z-1|\leqslant|z^2-1|$ for $z\geqslant 0$ and $|z^{-1}-1|\leqslant 2|z-1|$ for $|z-1|\leqslant 1/2$, we have that

$$
\max_{1\leqslant j\leqslant d}|\sigma_j/\widehat{\sigma}_j-1|\leqslant Cn^{-3\zeta/8}/\log^2 d
$$

with probability at least $1 - Cn^{-\zeta/8}$. Now, by Theorem 3.1, we have

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\Lambda^{-1/2}U'_{n,N} \in R) - \mathbb{P}(\Lambda^{-1/2}Y \in R) \right| \leqslant Cn^{-\zeta/8}.$$

Since $\Lambda^{-1/2}Y$ is Gaussian with mean zero and covariance matrix whose diagonal elements are 1, by the Borell-Sudakov-Tsirel'son inequality together with the bound $\mathbb{E}[|\Lambda^{-1/2}Y|_\infty] \leqslant C\sqrt{\log d}$, we have $\mathbb{P}(|\Lambda^{-1/2}Y|_\infty > C\sqrt{\log(dn)}) \leqslant 2n^{-1}$. Hence, $\mathbb{P}(|\sqrt{n}\Lambda^{-1/2}U'_{n,N}|_\infty > C\sqrt{\log(dn)}) \leqslant Cn^{-\zeta/8}$. Since $\frac{n^{-3\zeta/8}}{\log^2 d} \times \sqrt{\log(dn)} \leqslant \frac{Cn^{-\zeta/8}}{\log^{3/2} d}$, we have

$$\mathbb{P}(|\sqrt{n}(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2})U'_{n,N}|_\infty > t_n) \leqslant Cn^{-\zeta/8},$$

where $t_n = Cn^{-\zeta/8}/\log^{3/2} d$.

Now, for $R = \prod_{j=1}^d [a_j, b_j], a = (a_1, \dots, a_d)^T$, and $b = (b_1, \dots, b_d)^T$, we have

$$\begin{aligned}
\mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}U'_{n,N} \in R) &\leqslant \mathbb{P}(\{-\sqrt{n}\Lambda^{-1/2}U'_{n,N} \leqslant -a + t_n\} \cap \{\sqrt{n}\Lambda^{-1/2}U'_{n,N} \leqslant b + t_n\}) \\
&\quad + \mathbb{P}(|\sqrt{n}(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2})U_{n,N}|_\infty > t_n) \\
&\leqslant \mathbb{P}(\{-\Lambda^{-1/2}Y \leqslant -a + t_n\} \cap \{\Lambda^{-1/2}Y \leqslant b + t_n\}) + Cn^{-\zeta/8} \\
&\leqslant \mathbb{P}(\Lambda^{-1/2}Y \in R) + Ct_n\sqrt{\log d} + Cn^{-\zeta/8},
\end{aligned}$$

where the last inequality follows from Nazarov's inequality. Since $t_n\sqrt{\log d} \leqslant Cn^{-\zeta/8}/\log d \leqslant Cn^{-\zeta/8}$, we conclude that $\mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}U'_{n,N} \in R) \leqslant \mathbb{P}(\Lambda^{-1/2}Y \in R) + Cn^{-\zeta/8}$. Likewise, we have $\mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}U'_{n,N} \in R) \geqslant \mathbb{P}(\Lambda^{-1/2}Y \in R) - Cn^{-\zeta/8}$. Hence we have shown that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}U'_{n,N} \in R) - \mathbb{P}(\Lambda^{-1/2}Y \in R) \right| \leqslant Cn^{-\zeta/8}.$$

Similarly, using Theorem 4.2, we have that

$$\mathbb{P}_{|\mathcal{D}_n}(|(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2})U_n^\sharp|_\infty > t_n) \leqslant Cn^{-\zeta/8}$$

with probability at least $1 - Cn^{-\zeta}$. Following arguments similar to those above, we conclude that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(\widehat{\Lambda}^{-1/2}U_n^\sharp \in R) - \mathbb{P}(\Lambda^{-1/2}Y \in R) \right| \leqslant Cn^{-\zeta/8}$$

with probability at least $1 - Cn^{-\zeta/8}$. This completes the proof. □

Appendix B. Additional simulation results

In this section, we provide additional results of the partial bootstrap $U_{n,A}^\sharp$ for the non-degenerate Spearman's ρ statistic. As in Section 5, we test the performance of MB-NDG-DC and MB-NDG-RS. The computational budget parameter value is set as $N = 4n^{3/2}$ and other parameter values remain the same as the simulation examples in Section 5. The exponent of $n^{3/2}$ in N is chosen by minimizing the error bound in the Gaussian approximation (cf. Corollary 3.2). We empirically observe that the bootstrap approximation is sensitive to small constant values in $N = Cn^{3/2}$ and we find that $C \geqslant 4$ can produce reasonably accurate bootstrap approximation quality (cf. Figure 5 and 6).

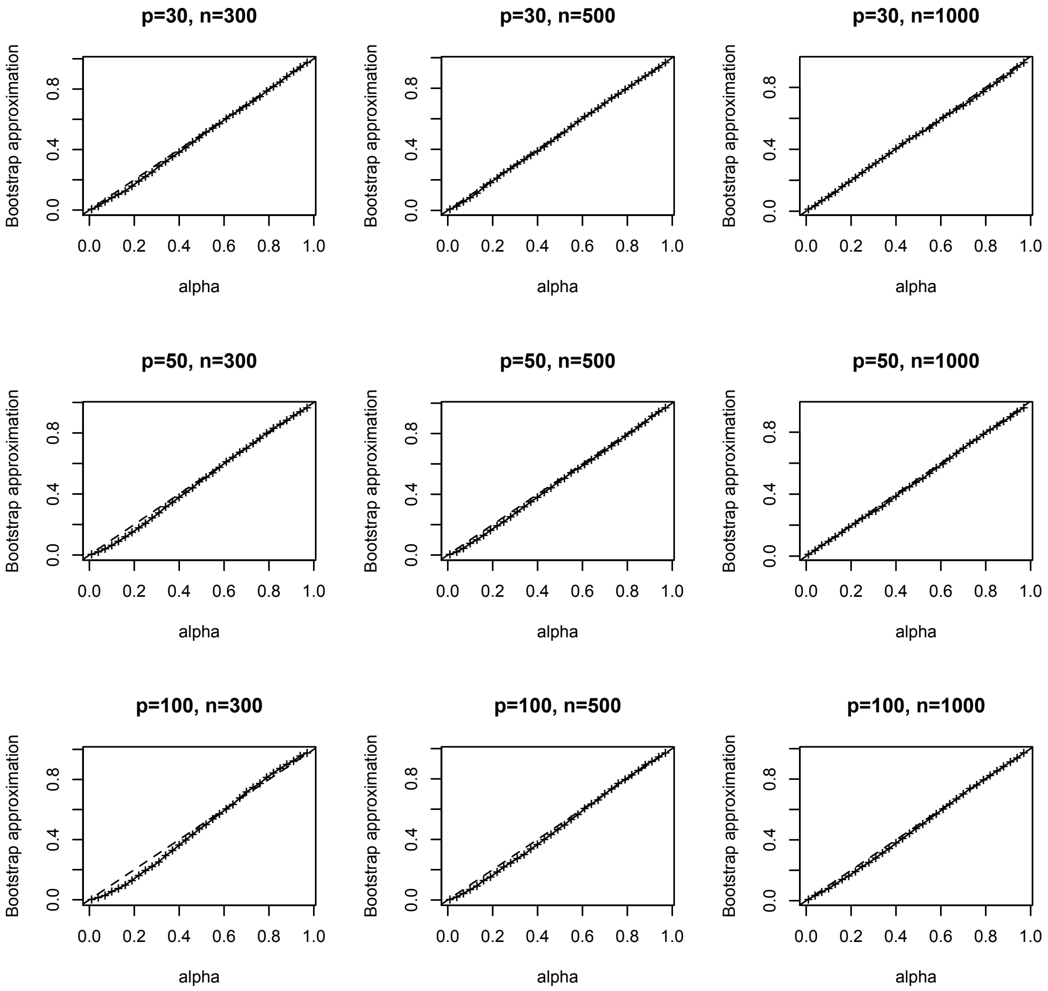

FIGURE 5. Bootstrap approximation $U^{\sharp}_{n,A}$ for Spearman's ρ test statistic with the divide and conquer estimation (MB-NDG-DC). Plot of the nominal size α versus the empirical rejection probability $\widehat{R}(\alpha)$.

Fitting a linear model with the (log-)running time for the bootstrap methods as the response variable and the (log-)sample size as the covariate (with the intercept term), we find that the slope coefficient for $p = (30, 50, 100)$ is $(1.830, 1.829, 1.810)$ in the case MB-NDG-DC, and $(1.955, 1.961, 1.950)$ in the case MB-NDG-RS. In either case, the slope coefficient again matches very well to the theoretic value 2.

References

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

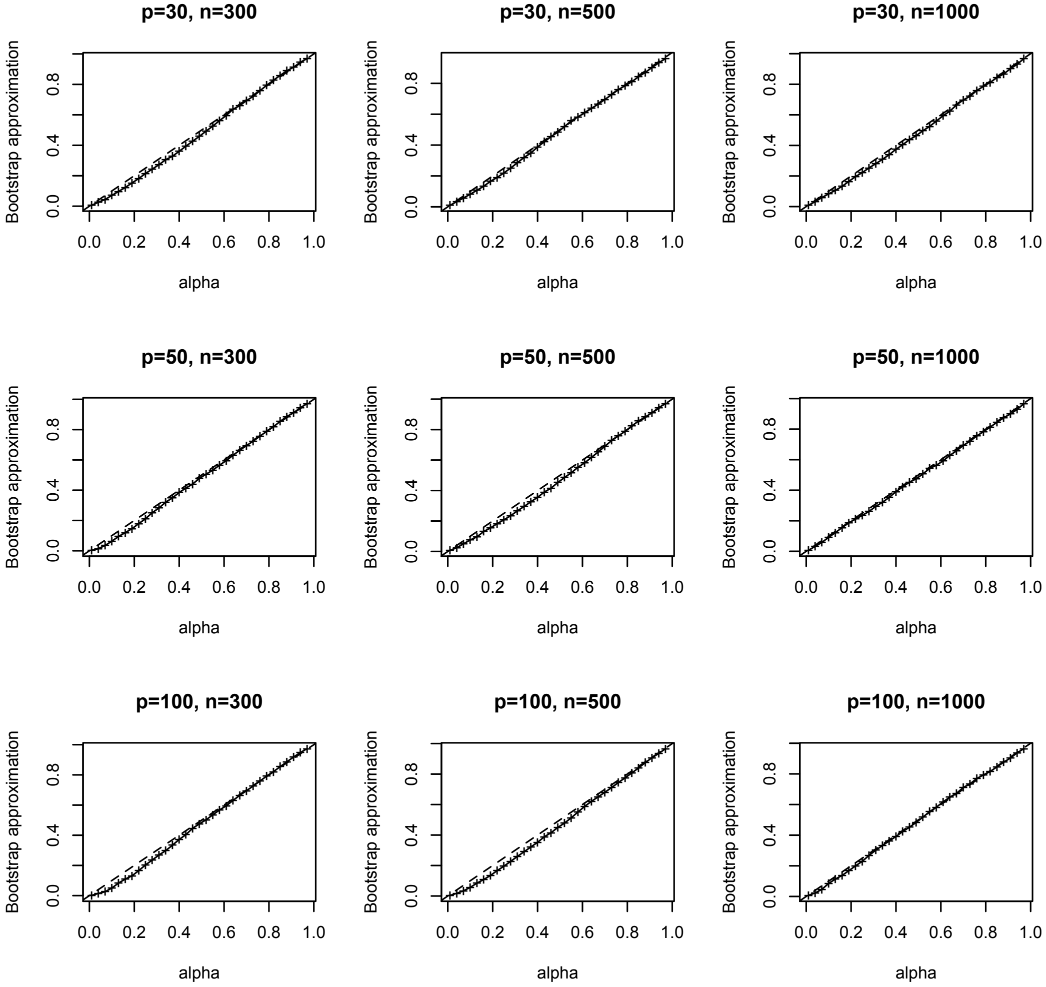

FIGURE 6. Bootstrap approximation $U^{\sharp}_{n,A}$ for Spearman's ρ test statistic with the random sampling estimation (MB-NDG-RS). Plot of the nominal size α versus the empirical rejection probability $\widehat{R}(\alpha)$.

[2] W. Bergsma and A. Dassios. A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20(2):1006–1028, 2014.

[3] Patrice Bertail and Jessica Tressou. Incomplete generalized U-statistics for food risk assessment. *Biometrics*, 62:66–74, 2006.

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

[5] Gunnar Blom. Some properties of incomplete U-statistics. *Biometrika*, 63(3):573–580, 1976.

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

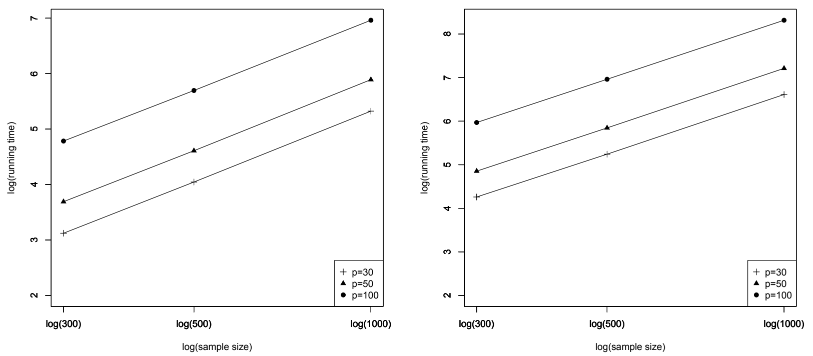

Figure 7. Computer running time of the bootstrap versus the sample size on the log-scale. Left: bootstrap $U^{\sharp}_{n,A}$ for Spearman's ρ with the divide and conquer estimation (MB-NDG-DC). Right: bootstrap $U^{\sharp}_{n,A}$ for Spearman's ρ with the random sampling estimation (MB-NDG-RS).

[7] B.M. Brown and D.G. Kildea. Reduced U-statistics and the Hodges-Lehmann estimator. *Annals of Statistics*, 6:828–835, 1978.

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

[9] Xiaohui Chen and Kengo Kato. Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. 2017. arXiv:1708.02705.

[10] 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.

[11] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162:47–70, 2015.

[12] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4):2309–2352, 2017.

[13] Stephan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up empirical risk minimization: optimization of incomplete U-statistics. *Journal of Machine Learning Research*, 17:1–36, 2016.

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

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

[16] Edward W. Frees. Infinite order U-statistics. *Scandinavian Journal of Statistics*, 16(1), 1989.

[17] Seymour Geisser and Nathan Mantel. Pairwise independence of jointly dependent random variables. *Annals of Mathematical Statistics*, 33(1):290–291, 1962.

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

[19] Fang Han, Shizhe Chen, and Han Liu. Distribution-free tests of independence in high dimensions. *Biometrika*, 2017, to appear.

[20] Fang Han and Tianchen Qian. Asymptotics for asymmetric weighted U-statistics: Central limit theorem and bootstrap under data heterogeneity. *Preprint*, 2016.

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

[22] Wassily Hoeffding. A nonparametric test of independence. *Annals of Mathematical Statistics*, 19:546–557, 1948.

[23] Tailen Hsing and Wei Biao Wu. On weighted U-statistics for stationary processes. *Annals of Probability*, 32(2):1600–1631, 2004.

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

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

[26] Svante Janson. The asymptotic distributions of incomplete U-statistics. *Z, Wahrscheinlichkeitstheorie verw. Gebiete*, 66:495–505, 1984.

[27] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(4):795–816, 2014.

[28] Alan J. Lee. *U-Statistics: Theory and Practice.* Statistics: A Series of Textbooks and Monographs (Book 110). CRC Press, 1990.

[29] Eric L. Lehmann. *Elements of Large-Sample Theory.* Springer Texts in Statistics, 1998.

[30] Dennis Leung and Mathias Drton. Testing independence in high dimensions with sums of rank correlations. *Annals of Statistics*, 2017, to appear.

[31] P. Major. Asymptotic distributions for weighted U-statistics. *Annals of Probability*, 21(2):1514–1535, 1994.

[32] Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17:1–41, 2016.

[33] K.A. O'Neil and R.A. Redner. Asymptotic distributions of weighted U-statistics of degree 2. *Annals of Probability*, 21(2):1159–1169, 1993.

[34] M. Rifi and F. Utzet. On the asymptotic behavior of weighted U-statistics. *Journal of Theoretical Probability*, 13(1):141–167, 2000.

[35] H. Rubin and R.A. Vitale. Asymptotic distribution of symmetric statistics. *Annals of Statistics*, 8(1):165–170, 1980.

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

[37] C.P. Shapiro and L. Hubert. Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *Annals of Statistics*, 7(4):788–794, 1979.

[38] Aad van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998.

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

[40] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of American Statistical Association*, 2017, to appear. arXiv:1510:04342.

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

[42] Sun Yao, Xianyang Zhang, and Xiaofeng Shao. Testing mutual independence in high dimension via distance covariance. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 2017, to appear.

[43] Yuchen Zhang, John Duchi, and Martin J. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.

(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