

# RANDOMIZED INCOMPLETE $U$ -STATISTICS IN HIGH DIMENSIONS<sup>||</sup>

BY XIAOHUI CHEN<sup>†,\*\*</sup>, AND KENGO KATO<sup>††</sup>

*University of Illinois at Urbana-Champaign<sup>\*\*</sup> and Cornell University<sup>††</sup>*

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 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. Our methods are illustrated on the application to nonparametric testing for the pair-wise 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  $r \geq 2$  and  $d \geq 1$  be given positive integers, and let  $h = (h_1, \dots, h_d)^T : S^r \rightarrow \mathbb{R}^d$  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  $\mathbb{E}[|h_j(X_1, \dots, X_r)|] <$

<sup>||</sup>First arXiv version: December 3, 2017. This version: January 29, 2019.

<sup>†</sup>Supported in part by NSF DMS-1404891, NSF CAREER Award DMS-1752614, and UIUC Research Board Awards (RB17092, RB18099).

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

*MSC 2010 subject classifications:* Primary 62E17, 62F40; secondary 62H15

*Keywords and phrases:* Incomplete  $U$ -statistics, randomized inference, Gaussian approximation, bootstrap, divide and conquer, Bernoulli sampling, sampling with replacement

$\infty$  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  $\{1, \dots, n\}$

$$(1.1) \quad U_n := U_n^{(r)}(h) := \frac{1}{|I_{n,r}|} \sum_{(i_1, \dots, i_r) \in I_{n,r}} h(X_{i_1}, \dots, X_{i_r}),$$

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

$U$ -statistics are an important and general class of statistics, and applied in a wide variety of statistical problems; we refer to [30] as an excellent monograph on  $U$ -statistics. For univariate  $U$ -statistics ( $d = 1$ ), the asymptotic distributions are derived in the seminal paper [23] for the non-degenerate case and in [38] for the degenerate case. There is also a large literature on bootstrap methods for univariate  $U$ -statistics [4, 6, 1, 26, 27, 15, 44]. 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; see also [20].

However, a major obstacle of inference using the complete  $U$ -statistic (1.1) is its computational intractability. Namely, the computation of the complete  $U$ -statistic (1.1) 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  $r \geq 3$ . 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 (1.1) 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*  $N \leq |I_{n,r}|$ , 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  $h(X_{i_1}, \dots, X_{i_r})$  taken over the subset of chosen indices  $(i_1, \dots, i_r)$ . 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. The Gaussian approximation results are, however, 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 the non-degenerate case. 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  $U$ -statistic 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  $X = (X^{(1)}, \dots, X^{(p)})^T$ , i.e., testing for the hypothesis that

$$(1.2) \quad H_0 : X^{(1)}, \dots, X^{(p)} \text{ are pairwise independent.}$$

Let  $X_1, \dots, X_n$  be i.i.d. copies of  $X$ . Several dependence measures are proposed in the literature, including: Kendall's  $\tau$ , Spearman's  $\rho$ , Hoeffding's  $D$  [24], Bergsma and Dassios'  $t^*$  [2], and the distance covariance [40, 45], all

of which can be estimated by  $U$ -statistics. So various nonparametric tests for  $H_0$  can be constructed based on those  $U$ -statistics. To compute the test statistics, we have to compute  $U$ -statistics with dimension  $d = p(p-1)/2$ , which corresponds to the number of upper triangular entries in the  $p \times p$  dependence matrix and can be quite large. In addition, the orders of the  $U$ -statistics are at least 3 (except for Kendall's  $\tau$  which is of order 2). 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. 32, 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 applicable to calibrating critical values for test statistics based upon incomplete versions of those  $U$ -statistics. Detailed comparisons and discussions of nonparametric pairwise independence test statistics are presented in Section 5. In addition to pairwise independence testing, values of the dependence measures are also interesting *per se* in some applications. For instance, Spearman's  $\rho$  is related to the copula correlation if the marginal distributions are continuous [16, Chapter 8], and our bootstrap methods can be used to construct simultaneous confidence intervals for the copula correlations uniformly over many pairs of variables.

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 (1.2). Specifically, we consider to approximate the null distributions of the incomplete versions of the (leading term of) Spearman  $\rho$  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 Gaussian approximation of the test statistics is quite accurate and 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 are first considered in [5], and the asymptotic distributions of incomplete  $U$ -statistics (for fixed  $d$ ) are derived in [7] and [28]; see also Section 4.3 in [30] for a review on incomplete  $U$ -statistics. Closely related to the present paper is [28], 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 [28]. Incomplete  $U$ -statistics can be viewed as a special case of weighted  $U$ -statistics, and there is a large literature on limit theorems for weighted  $U$ -statistics; see [39, 36, 33, 37, 25, 22] 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, 34], 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 extends 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 in the Supplementary Material (SM). All the technical proofs are gathered in Appendix C in the SM. We conclude the paper in Section 6 with a brief discussion on some extensions.

**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  $y \leq z$  means that  $y_j \leq z_j$  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  $\psi_\beta$  be the function on  $[0, \infty)$  defined by  $\psi_\beta(x) = e^{x^\beta} - 1$ , and for a real-valued random variable  $\xi$ , define  $\|\xi\|_{\psi_\beta} = \inf\{C > 0 : \mathbb{E}[\psi_\beta(|\xi|/C)] \leq 1\}$ . 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  $C_\beta$  depending only on  $\beta$  such that  $\|\xi_1 + \xi_2\|_{\psi_\beta} \leq C_\beta(\|\xi_1\|_{\psi_\beta} + \|\xi_2\|_{\psi_\beta})$ . (Indeed, there is a norm equivalent to  $\|\cdot\|_{\psi_\beta}$  obtained by linearizing  $\psi_\beta$  in a neighborhood of the origin; cf. Lemma C.2 in the SM.)

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  $1 \leq k \leq \ell \leq n$ , we use the notation  $X_k^\ell = (X_k, \dots, X_\ell)$ . Throughout the paper, we assume that  $n \geq 4 \vee r$  and  $d \geq 3$ .

**2. Randomized incomplete  $U$ -statistics.** In this paper, to construct sparsely weighted  $U$ -statistics, we shall use random sparse weights. For  $\iota = (i_1, \dots, i_r) \in I_{n,r}$ , let us write  $X_\iota = (X_{i_1}, \dots, X_{i_r})$ , and observe that the complete  $U$ -statistic (1.1) 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  $0 < N \leq |I_{n,r}|$ , and let  $p_n = N/|I_{n,r}|$ . Instead of taking the average over all possible  $\iota$  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.  $\text{Ber}(p_n)$  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

$$(2.1) \quad U'_{n,N} = \frac{1}{\widehat{N}} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota),$$

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 on the Bernoulli sampling. The variable  $\widehat{N}$  follows  $\text{Bin}(|I_{n,r}|, p_n)$ , 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 (2.1) only requires  $O(Nd)$  operations on average. In addition, by Bernstein's inequality (cf. Lemma 2.2.9 in [42]),

$$(2.2) \quad \mathbb{P}\left(|\widehat{N}/N - 1| > \sqrt{2t/N} + 2t/(3N)\right) \leq 2e^{-t}$$

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 (2.1) as follows.

1. Generate  $\widehat{N} \sim \text{Bin}(|I_{n,r}|, p_n)$ .
2. Choose indices  $\iota_1, \dots, \iota_{\widehat{N}}$  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  $\iota$  is one of  $\iota_1, \dots, \iota_{\widehat{N}}$ , and  $Z_\iota = 0$  otherwise; then, it is not difficult to see that  $\{Z_\iota : \iota \in I_{n,r}\}$  are i.i.d.  $\text{Ber}(p_n)$  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 (2.1) *does* affect approximating distributions

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

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

$$(2.3) \quad U'_{n,N} = \frac{1}{N} \sum_{j=1}^N h(X_{\iota_j}^*)$$

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. Observe that  $U'_{n,N}$  in (2.3) can be efficiently computed by sampling  $r$  distinct terms from  $\{X_1, \dots, X_n\}$  independently for  $N$  times. The statistic  $U'_{n,N}$  can be written as a weighted  $U$ -statistic. Indeed, for each  $\iota \in I_{n,r}$ , let  $Z_\iota$  denote the number of times that  $X_\iota$  is redrawn in the sample  $\{X_{\iota_1}^*, \dots, X_{\iota_N}^*\}$ . 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  $1/|I_{n,r}|, \dots, 1/|I_{n,r}|$  independent of  $X_1^n$ , and  $U'_{n,N}$  can be written as

$$(2.4) \quad U'_{n,N} = \frac{1}{N} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota).$$

Hence we can think of  $U'_{n,N}$  as a statistic of  $X_1, \dots, X_n$  and  $Z_\iota, \iota \in I_{n,r}$ , but we will use both representations (2.3) and (2.4) 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 (2.1) and (2.3) 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  $-\infty \leq a_j \leq b_j \leq \infty$  for  $j = 1, \dots, d$  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  $\min_{1 \leq j \leq d} \text{Var}(\mathbb{E}[h_j(X_1, \dots, X_r) \mid X_1])$  is bounded away from zero in the following discussion. However, our Gaussian approximation results also cover the degenerate case (cf. Theorem 3.3). The intuition behind and the proof sketch for the Gaussian approximation results are given in Section C.2 in the SM.

To state the formal Gaussian approximation results, we assume the following conditions. Let  $\underline{\sigma} > 0$  and  $D_n \geq 1$  be given constants, and define  $g := (g_1, \dots, g_d)^T := P^{r-1}h$ . Suppose that

- (C1)  $P^r |h_j|^{2+k} \leq D_n^k$  for all  $j = 1, \dots, d$  and  $k = 1, 2$ .
- (C2)  $\|h_j(X_1^r)\|_{\psi_1} \leq D_n$  for all  $j = 1, \dots, d$ .

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

- (C3-ND)  $P(g_j - \theta_j)^2 \geq \underline{\sigma}^2$  for all  $j = 1, \dots, d$ .
- (C3-D)  $P^r(h_j - \theta_j)^2 \geq \underline{\sigma}^2$  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 bounded kernels. 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 mainly work with Condition (C2) and point out the differences when the kernel satisfies a polynomial moment condition in Remark 3.5. By Jensen's inequality, Conditions (C1) and (C2) imply that  $P|g_j|^{2+k} \leq D_n^k$  for all  $j$  and for  $k = 1, 2$ , and  $\|g_j(X_1)\|_{\psi_1} \leq D_n$  for all  $j$ . Here we allow the exponential moment bound  $D_n$  to depend on  $n$  since the distribution  $P$  may depend on  $n$  in the high-dimensional setting. In addition, Condition (C1) implies that  $P^r h_j^2 \leq 1 + P|h_j|^3 \leq 1 + D_n$  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

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

without further mentioning. The value  $1/2$  has no special meaning; we can allow  $p_n \leq c$  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*

$$(3.1) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P} \{ \sqrt{n}(U'_{n,N} - \theta) \in R \} - \mathbb{P}(Y \in R) \right| \\ = \sup_{R \in \mathcal{R}} \left| \mathbb{P} \{ \sqrt{N}(U'_{n,N} - \theta) \in R \} - \mathbb{P}(\alpha_n^{-1/2} Y \in R) \right| \leq C \left( \frac{D_n^2 \log^7(dn)}{n \wedge N} \right)^{1/6},$$

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  $d \gg n$ . Asymptotically, if e.g.  $D_n$  is bounded in  $n$  and  $N \geq n$ , then as  $n \rightarrow \infty$ ,

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P} \{ \sqrt{n}(U'_{n,N} - \theta) \in R \} - \mathbb{P}(Y \in R) \right| \rightarrow 0$$

whenever  $d = d_n$  satisfies that  $\log d = o(n^{1/7})$ , so that the high-dimensional CLT on the hyperrectangles holds for the incomplete  $U$ -statistics even in ultra-high dimensional cases where  $d$  is much larger than  $n$ . Similar comments apply to all the other results we will derive.

For complete and non-degenerate  $U$ -statistics (a special case of incomplete  $U$ -statistics with the complete design and  $N = |I_{n,r}|$ ), it has been argued in [12] ( $r = 1$ ) and [8] ( $r = 2$ ) that the rate of convergence in Theorem 3.1 is nearly optimal in the regime where  $d$  grows sub-exponentially fast in  $n$ . On the other hand, the rate of convergence can be improved to  $n^{-1/4}$  (up to logarithmic factors) if  $d = O(n^{1/7})$ , namely if the dimension increases at most polynomially fast with the sample size.

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

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

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

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

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 of the same order as  $n$ , then the scaling factor of  $U'_{n,N}$  remains the same as for  $U_n$ , namely,  $\sqrt{n}$ . However, the approximating Gaussian distribution for  $\sqrt{n}(U'_{n,N} - \theta)$  has covariance matrix  $r^2 \Gamma_g + \alpha_n \Gamma_h$ , which is larger than the corresponding covariance matrix  $r^2 \Gamma_g$  for  $\sqrt{n}(U_n - \theta)$  in the sense that their difference  $\alpha_n \Gamma_h$  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  $j = 1, \dots, d$ . We consider the case where the kernel  $h$  is degenerate of order  $k - 1$  for some  $k = 2, \dots, r$ , i.e.,  $P^{r-k+1}h(x_1, \dots, x_{k-1}) = P^r h$

for all  $(x_1, \dots, x_{k-1}) \in S^{k-1}$ . 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 logarithmic 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  $k = 2, \dots, r$ . 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*

$$(3.2) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P} \left\{ \sqrt{N}(U'_{n,N} - \theta) \in R \right\} - \gamma_B(R) \right| \\ \leq C \left\{ \left( \frac{ND_n^2 \log^{k+3} d}{n^k} \right)^{1/4} + \left( \frac{D_n^2 (\log n) \log^5(dn)}{n} \right)^{1/6} + \left( \frac{D_n^2 \log^7(dn)}{N} \right)^{1/6} \right\},$$

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 observe that  $\sqrt{N}(U'_{n,N} - \theta) = (N/\hat{N}) \cdot \sqrt{N}W_n = (N/\hat{N})(\sqrt{N}A_n + \sqrt{N}(1 - p_n)B_n)$ , where  $A_n = U_n - \theta$  and  $B_n = U'_{n,N} - U_n$ . According to Theorem 12.10 in [41],  $n^{k/2}A_n$  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 logarithmic 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  $\gamma_B$  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  $\theta$  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.,  $\check{U}'_{n,N} = N^{-1} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota)$ , instead of the random one, i.e.,  $U'_{n,N} = \hat{N}^{-1} \sum_{\iota \in I_{n,r}} Z_\iota h(X_\iota)$ . Then, in the non-degenerate case, the distribution of  $\sqrt{n}(\check{U}'_{n,N} - \theta)$  can be approximated by  $N(0, r^2 \Gamma_g + \alpha_n P^r h h^T)$ , and in the degenerate case,  $\sqrt{N}(\check{U}'_{n,N} - \theta)$  can be approximated by  $N(0, P^r h h^T)$  (provided that  $N \ll n^k$  for the degenerate case). To see this, observe that  $\check{U}'_{n,N} - \theta = (U_n - \theta) + N^{-1} \sum_{\iota \in I_{n,r}} (Z_\iota - p_n) h(X_\iota)$ , and the distribution of  $N^{-1} \sum_{\iota \in I_{n,r}} (Z_\iota - p_n) h(X_\iota)$  can be approximated by  $N(0, (1 - p_n) P^r h h^T)$ . Since  $P^r h h^T$  is larger than  $\Gamma_h$  unless  $\theta = 0$  (in the sense that  $P^r h h^T - \Gamma_h = \theta \theta^T$  is positive semi-definite), the approximating Gaussian distributions have larger covariance matrices for  $\check{U}'_{n,N}$  than those for  $U'_{n,N}$ , and hence it is in general recommended to use the random normalization rather than the deterministic one. A numerical comparison between these normalizations can be found in Section E of the SM.

REMARK 3.4 (Comparisons with [28] 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 [28] to high dimensions. [28] 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 ([28] considers the deterministic normalization for the Bernoulli sampling case). For the illustrative purpose, consider sampling with replacement. Suppose that  $p_n \rightarrow p \in [0, 1]$  and the kernel  $h$  is degenerate of order  $k - 1$  for some  $k = 1, \dots, r$  (the  $k = 1$  case corre-

sponds to a non-degenerate kernel). Then Theorem 1 in [28] 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  $n^k/N \rightarrow \alpha \in [0, \infty]$ ,  $n^{k/2}(U'_{n,N} - \theta) \xrightarrow{d} V + \alpha W$  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  $\alpha$  is finite, or  $k \geq 2$  and  $\alpha = \infty$ ), but covers high-dimensional kernels and derives explicit and non-asymptotic Gaussian approximation error bounds that are not obtained in [28]. In addition, the proof strategy of our Gaussian approximation results differs substantially from that of [28]. [28] 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.

Finally, we expect that the results of the present paper can be extended to the case where  $k \geq 2$  and  $\alpha$  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.

REMARK 3.5 (Relaxation of sub-exponential moment Condition (C2)). It is possible to relax the sub-exponential moment Condition (C2) to a polynomial moment condition. Suppose that

$$(C2') \quad (P^r |h|_\infty^q)^{1/q} \leq D_n \text{ for some } q \in [4, \infty).$$

Condition (C2') covers a kernel with bounded polynomial moment of a finite degree  $q$ .

THEOREM 3.4 (Gaussian approximation under polynomial moment condition). *(i) If Conditions (C1), (C2'), and (C3-ND) hold, then there exists a constant  $C$  depending only on  $\underline{\sigma}, r$ , and  $q$  such that*

$$(3.3) \quad \sup_{R \in \mathcal{R}} |\mathbb{P} \{ \sqrt{n}(U'_{n,N} - \theta) \in R \} - \mathbb{P}(Y \in R)| \leq C \left\{ \left( \frac{D_n^2 \log^7(dn)}{n \wedge N} \right)^{1/6} + \left( \frac{D_n^2 n^{2r/q} \log^3(dn)}{(n \wedge N)^{1-2/q}} \right)^{1/3} \right\},$$

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

*(ii) Suppose the kernel  $h$  is degenerate of order  $k - 1$  for some  $k = 2, \dots, r$ .*

If Conditions (C1), (C2'), and (C3-D) hold, then there exists a constant  $C$  depending only on  $\underline{\sigma}, r$ , and  $q$  such that

$$\begin{aligned}
 (3.4) \quad & \sup_{R \in \mathcal{R}} \left| \mathbb{P} \left\{ \sqrt{N}(U'_{n,N} - \theta) \in R \right\} - \gamma_B(R) \right| \\
 & \leq C \left\{ \left( \frac{ND_n^2 \log^{k+3} d}{n^k} \right)^{1/4} + \left( \frac{D_n^2 \log^5(dn)}{n} \right)^{1/6} \right. \\
 & \quad \left. + \left( \frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \left( \frac{D_n^2 n^{2r/q} \log^3 d}{(n \wedge N)^{1-2/q}} \right)^{1/3} \right\},
 \end{aligned}$$

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

Comparing Theorem 3.4 with Theorems 3.1 and 3.3, we see that the same approximating Gaussian distributions under the sub-exponential moment condition (C2) are valid under the polynomial moment condition (C2') as well. The rates of convergence to the Gaussian distributions under (C2') involve an extra Nagaev-type term similarly to the sample average and complete  $U$ -statistic cases (cf. [12, 8]), and so the rates may be slower than those obtained under the sub-exponential moment condition (C2). In particular, the rates in (3.3) and (3.4) now depend on the order  $r$  through the term  $n^{2r/q}$ . Still, the leading orders in (3.3) and (3.4) coincide with those under the sub-exponential moment condition (C2) as long as  $q$  is sufficiently large compared with  $r$ . For example, if  $D_n$  is bounded in  $n$ ,  $N \geq n$ , and  $q \geq 4(r+1)$ , then the leading order of (3.3) is  $(n^{-1} \log^7(dn))^{1/6}$ , which coincides with that in the sub-exponential case.

**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  $\Gamma_h$  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  $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$  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  $Y \stackrel{d}{=} Y_A + \alpha_n^{1/2} Y_B$  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  $U_{n,A}^\#$  and  $U_{n,B}^\#$  such that, conditionally on  $\mathcal{D}_n$ , (i)  $U_{n,A}^\#$  and  $U_{n,B}^\#$  are independent, and (ii) the conditional distributions of  $U_{n,A}^\#$  and  $U_{n,B}^\#$  are computable and “close” enough to  $N(0, r^2 \Gamma_g)$  and  $N(0, \Gamma_h)$ , respectively. Then, the conditional distribution of  $U_n^\# = U_{n,A}^\# + \alpha_n^{1/2} U_{n,B}^\#$  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  $U_{n,A}^\#$  or  $U_{n,B}^\#$  alone, respectively.

Construction of  $U_{n,B}^\#$  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  $U_{n,B}^\#$ .

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

$$U_{n,B}^\# = \frac{1}{\sqrt{\widehat{N}}} \sum_{\iota \in I_{n,r}} \xi'_\iota \sqrt{Z_\iota} \{h(X_\iota) - U'_{n,N}\},$$

where  $\widehat{N}$  is replaced by  $N$  for the sampling with replacement case.

In the Bernoulli sampling case,  $U_{n,B}^\#$  reduces to  $U_{n,B}^\# = \widehat{N}^{-1/2} \sum_{j=1}^{\widehat{N}} \xi'_{\iota_j} \{h(X_{\iota_j}) - U'_{n,N}\}$ , while in the sampling with replacement case, simulating  $U_{n,B}^\#$  can be equivalently implemented by simulating  $U_{n,B}^\# = N^{-1/2} \sum_{j=1}^N \eta_j \{h(X_{\iota_j}^*) - U'_{n,N}\}$  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  $U_{n,B}^\#$  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  $U_{n,B}^\#$  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  $U_{n,B}^\#$  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  $U_{n,B}^\#$ ). *Suppose that (C1), (C2), and (C3-D)*



hold. If

$$(4.1) \quad \frac{D_n^2(\log^2 n) \log^5(dn)}{n \wedge N} \leq C_1 n^{-\zeta}$$

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

$$(4.2) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(U_{n,B}^\# \in R) - \gamma_B(R) \right| \leq C n^{-\zeta/6}$$

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

REMARK 4.1 (Bootstrap validity under the polynomial moment condition). Analogous bootstrap validity results for  $U_{n,B}^\#$  in Theorem 4.1, as well as those for  $U_n^\#$  and  $U_{n,A}^\#$  in Theorem 4.2, 4.3 and Proposition 4.4, 4.5 ahead, can be obtained under the polynomial moment Condition (C2'). Due to the space concern, detailed results can be found in Section B of the SM.

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  $U_{n,B}^\#$ , which can be simulated by drawing multiplier variables many times. We call the simulation of  $U_{n,B}^\#$  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  $U_{n,B}^\#$ , construction of  $U_{n,A}^\#$  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  $U_{n,A}^\#$  is as follows.

Generic construction of  $U_{n,A}^\#$ .

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  $\hat{g}^{(i_1)}$  of  $g$  based on  $X_1^n$ .

3. Construct

$$U_{n,A}^\# = \frac{r}{\sqrt{n_1}} \sum_{i_1 \in S_1} \xi_{i_1} \{\hat{g}^{(i_1)}(X_{i_1}) - \check{g}\},$$

$$\text{where } \check{g} = n_1^{-1} \sum_{i_1 \in S_1} \hat{g}^{(i_1)}(X_{i_1}).$$

Step 1 chooses a subset  $S_1$  to reduce the computational cost of the resulting bootstrap. Construction of estimates  $\hat{g}^{(i_1)}, i_1 \in S_1$  can be flexible. For instance, the estimates  $\hat{g}^{(i_1)}, i_1 \in S_1$  may depend on another randomization independent of everything else. In Sections 4.2 and 4.3, we will consider deterministic and random constructions of  $\hat{g}^{(i_1)}, i_1 \in S_1$ , 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  $U_{n,A}^\#$  where  $S_1 = \{1, \dots, n\}$  and  $\hat{g}^{(i_1)}(X_{i_1})$  is realized by its jackknife estimate, i.e., by the  $U$ -statistic with kernel  $(x_2, \dots, x_r) \mapsto h(X_{i_1}, x_2, \dots, x_r)$  for the sample without the  $i_1$ -th observation. Nevertheless, the bottleneck is that the computation of the jackknife estimates of  $g(X_{i_1}), i_1 = 1, \dots, n$  requires  $O(n^r d)$  operations and hence implementing the JMB can be computationally demanding.

Now, consider  $U_n^\# = U_{n,A}^\# + \alpha_n^{1/2} U_{n,B}^\#$ . We call the simulation of  $U_n^\#$  the *multiplier bootstrap under non-degeneracy* (MB-NDG). The following theorem establishes conditions under which the conditional distribution of  $U_n^\#$  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

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

which quantifies the errors of the estimates  $\hat{g}^{(i_1)}, i_1 \in S_1$ . In addition, let  $\bar{\sigma}_g := \max_{1 \leq j \leq d} \sqrt{P(g_j - \theta_j)^2}$ .

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

$$(4.3) \quad \begin{aligned} & \frac{D_n^2(\log^2 n) \log^5(dn)}{n_1 \wedge N} \leq C_1 n^{-\zeta_1} \quad \text{and} \\ & \mathbb{P}\left(\bar{\sigma}_g^2 \hat{\Delta}_{A,1} \log^4 d > C_1 n^{-\zeta_2}\right) \leq C_1 n^{-1} \end{aligned}$$

for some constants  $0 < C_1 < \infty$  and  $\zeta_1, \zeta_2 \in (0, 1)$ . Then there exists a constant  $C$  depending only on  $\underline{\sigma}, r$ , and  $C_1$  such that

$$(4.4) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(U_n^\# \in R) - \mathbb{P}(Y \in R) \right| \leq C n^{-(\zeta_1 \wedge \zeta_2)/6}$$

with probability at least  $1 - Cn^{-1}$ , 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 (4.4), with  $\mathcal{D}_n$  replaced by the augmentation of  $\mathcal{D}_n$  with variables used in the additional randomization, holds with probability at least  $1 - Cn^{-1}$ .

The second part of Condition (4.3) is a high-level condition on the estimation accuracy of  $\hat{g}^{(i_1)}, i_1 \in S_1$ . In Sections 4.2 and 4.3, we will verify the second part of Condition (4.3) for deterministic and random constructions of  $\hat{g}^{(i_1)}, i_1 \in S_1$ . 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.

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

**COROLLARY 4.3** (Validity of  $U_{n,A}^\sharp$ ). *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*

$$(4.5) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(U_{n,A}^\sharp \in R) - \gamma_A(R) \right| \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}$$

with probability at least  $1 - Cn^{-1}$ . 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 (4.5), with  $\mathcal{D}_n$  replaced by the augmentation of  $\mathcal{D}_n$  with variables used in the additional randomization, holds with probability at least  $1 - Cn^{-1}$ .

**REMARK 4.2** (Comments on the partial bootstrap simplification under non-degeneracy). When the approximating 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  $U_{n,B}^\sharp$  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  $U_{n,B}^\sharp$ .

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

**4.2. Divide and conquer estimation.** We first propose a deterministic construction of  $\hat{g}^{(i_1)}, i_1 \in S_1$  via the divide and conquer (DC) algorithm (cf. [46]).

1. For each  $i_1 \in S_1$ , choose  $K$  *disjoint* subsets  $S_{2,k}^{(i_1)}, k = 1, \dots, K$  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  $\{X_i : i \in S_{2,k}^{(i_1)}, k = 1, \dots, K\}$ , and taking the average of those  $U$ -statistics of order  $r - 1$ , i.e.,

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

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 the simulation of  $U_n^\sharp$  with the DC algorithm 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  $\hat{g}^{(i_1)}$  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  $U_{n,A}^\sharp$ , then the computational cost is  $O(n^2d + Bnd)$ , since the  $O(BNd)$  computations come from simulating  $U_{n,B}^\sharp$ . 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  $U_{n,A}^\sharp$ ). 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 substantially differs from the the Bag of Little Bootstraps (BLB) proposed in [29], which is another generically scalable bootstrap method for large datasets based on the DC algorithm. Due to the space concern, we defer the comparison of our MB-NDG-DC with the BLB in Section A.1 of the SM.

The following proposition provides conditions for the 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*

$$(4.6) \quad \frac{D_n^2(\log^2 n) \log^5(dn)}{n_1 \wedge N} \bigvee \left\{ \frac{\bar{\sigma}_g^2 D_n^2 \log^7 d}{KL} \left( 1 + \frac{\log^2 d}{K^{1-1/\nu}} \right) \right\} \leq C_1 n^{-\zeta}$$

for some constants  $0 < C_1 < \infty$ ,  $\zeta \in (0, 1)$ , and  $\nu \in (1/\zeta, \infty)$ . Then, there exists a constant  $C$  depending only on  $\underline{\sigma}, r, \nu$ , and  $C_1$  such that each of the results (4.4) and (4.5) with  $(\zeta_1, \zeta_2) = (\zeta, \zeta - 1/\nu)$  holds with probability at least  $1 - Cn^{-1}$ .

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  $D_n^2 \log^7(dn) \leq n^{1-\zeta}$  for some  $\zeta \in (0, 1)$ . Then, by Theorem 3.1 and Proposition 4.4, for arbitrarily large  $\nu \in (1/\zeta, \infty)$ , there exists a constant  $C$  depending only on  $\bar{\sigma}_g, \underline{\sigma}, r$ , and  $\nu$  such that

$$(4.7) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}(U'_{n,N} - \theta) \in R) - \mathbb{P}_{|\mathcal{D}_n}(U_n^\# \in R) \right| \leq Cn^{-(\zeta-1/\nu)/6}$$

with probability at least  $1 - Cn^{-1}$ . 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  $i_1 = 1, \dots, n$ , let  $I_{n-1,r-1}(i_1) = \{(i_2, \dots, i_r) : 1 \leq i_2 < \dots < i_r \leq n, i_j \neq i_1 \forall j \neq 1\}$ . In addition, define  $\sigma_{i_1} : \{1, \dots, n-1\} \rightarrow \{1, \dots, n\} \setminus \{i_1\}$  as follows: if  $\{1, \dots, n\} \setminus \{i_1\} = \{j_1, \dots, j_{n-1}\}$  with  $j_1 < \dots < j_{n-1}$ , then  $\sigma_{i_1}(\ell) = j_\ell$  for  $\ell = 1, \dots, n-1$ . For the notational convenience, for  $\iota' = (i_2, \dots, i_r) \in I_{n-1,r-1}$ , we write  $\sigma_{i_1}(\iota') = (\sigma_{i_1}(i_2), \dots, \sigma_{i_1}(i_r)) \in I_{n-1,r-1}(i_1)$ .

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

1. Let  $0 < M = M_n \leq |I_{n-1, r-1}|$  be a positive integer, and generate i.i.d.  $\text{Ber}(\vartheta_n)$  random variables  $\{Z'_{\iota'} : \iota' = (i_2, \dots, i_r) \in I_{n-1, r-1}\}$  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  $\hat{g}^{(i_1)}(x) = M^{-1} \sum_{\iota' \in I_{n-1, r-1}} Z'_{\iota'} h(x, X_{\sigma_{i_1}(\iota')})$ .

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

So, on average, the computational cost to construct  $\hat{g}^{(i_1)}, i_1 \in S_1$  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  $U_{n, A}^\#$ ). As a practical guidance to implement the random sampling estimation, we suggest to choose  $S_1 = \{1, \dots, n\}$  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  $O(n^2 d + B n d)$  if we only simulate  $U_{n, A}^\#$ ), 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  $O(B n d)$  if we only simulate  $U_{n, A}^\#$ ), 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*

$$(4.8) \quad \frac{D_n^2(\log^2 n) \log^5(dn)}{n_1 \wedge N} \bigvee \frac{\bar{\sigma}_g^2 D_n^2 \log^7(dn)}{n \wedge M} \leq C_1 n^{-\zeta}$$

for some constants  $0 < C_1 < \infty$  and  $\zeta \in (0, 1)$ . Then, for arbitrarily large  $\nu \in (1/\zeta, \infty)$ , there exists a constant  $C$  depending only on  $\underline{\sigma}, r, \nu$ , and  $C_1$  such that each of the results (4.4) and (4.5), with  $\mathcal{D}_n$  replaced by  $\mathcal{D}'_n = \mathcal{D}_n \cup \{Z'_{\iota'} : \iota' \in I_{n-1, r-1}\}$  and with  $(\zeta_1, \zeta_2) = (\zeta, \zeta - 1/\nu)$ , holds with probability at least  $1 - C n^{-1}$ .

For instance, consider to take  $N = n$ ,  $S_1 = \{1, \dots, n\}$ , and  $M$  proportional to  $n - 1$ , and suppose that  $D_n^2 \log^7(dn) \leq n^{1-\zeta}$  for some  $\zeta \in (0, 1)$ . Then, by Theorem 3.1 and Proposition 4.5, for arbitrarily large  $\nu \in (1/\zeta, \infty)$ , there exists a constant  $C$  depending only on  $\bar{\sigma}_g, \underline{\sigma}, r$ , and  $\nu$  such that the result (4.7) holds with probability at least  $1 - Cn^{-1}$ .

REMARK 4.3 (Alternative options for random sampling estimation). In construction of  $\hat{g}^{(i_1)}$ , 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  $\hat{g}^{(i_1)}(x)$ . In view of the concentration inequality for  $\widehat{M}$  (cf. equation (2.2)), it is not difficult to see that the same conclusion of Proposition 4.5 holds for  $\hat{g}^{(i_1)}(x) = \widehat{M}^{-1} \sum_{j=1}^{\widehat{M}} h(x, X_{\sigma_{i_1}(\iota'_j)})$ .

Next, alternatively to the Bernoulli sampling, we may use sampling with replacement to construct  $\hat{g}^{(i_1)}$ , 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  $\hat{g}^{(i_1)}(x) = M^{-1} \sum_{j=1}^M h(x, X_{\sigma_{i_1}(\iota'_j)})$  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 (4.8) (here  $Z_{\iota'}$  is the number of times that  $\iota'$  is redrawn in the sample  $\{\iota'_1, \dots, \iota'_M\}$ , for which  $\hat{g}^{(i_1)}(x)$  can be expressed as  $\hat{g}^{(i_1)}(x) = M^{-1} \sum_{\iota' \in I_{n-1, r-1}} Z_{\iota'} h(x, X_{\sigma_{i_1}(\iota')})$ ).

**5. Numerical examples.** In this section, we provide some numerical examples to verify the validity of our Gaussian approximation results and the 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 Gaussian approximation and bootstrap algorithms in the leading example of testing for the pairwise independence of a high-dimensional vector.

5.1. *Test statistics.* In this section, we discuss several nonparametric statistics in the literature for the testing problem of the pairwise independence.

EXAMPLE 5.1 (Spearman's  $\rho$ ). Let  $\Pi_r$  be the collection of all possible permutations on  $\{1, \dots, r\}$ . [23] shows that Spearman's rank correlation coefficient matrix  $\rho$  can be written as

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

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

$$\begin{aligned} h^S(X_1, X_2, X_3) &= (h_{j,k}^S(X_1, X_2, X_3))_{1 \leq j, k \leq p} \\ &= \frac{1}{2} \sum_{\pi \in \Pi_3} \text{sign} \{ (X_{\pi(1)} - X_{\pi(2)})(X_{\pi(1)} - X_{\pi(3)})^T \}, \end{aligned}$$

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

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

Here, for a matrix  $A = (a_{j,k})_{1 \leq j, k \leq p}$ ,  $\text{sign}\{A\}$  is the matrix of the same size as  $A$  whose  $(j, k)$ -th element is  $\text{sign}(a_{j,k}) = \mathbf{1}(a_{j,k} > 0) - \mathbf{1}(a_{j,k} < 0)$ . It is seen that the leading term in Spearman's  $\rho$  is  $\hat{\rho}$ , and so it is reasonable to reject the null hypothesis (1.2) if  $\max_{1 \leq j < k \leq p} |\hat{\rho}_{j,k}|$  is large. Precisely speaking, this test is testing for a weaker hypothesis that

$$H'_0 : \mathbb{E}[\text{sign}(X_1^{(j)} - X_2^{(j)})\text{sign}(X_1^{(k)} - X_3^{(k)})] = 0 \text{ for all } 1 \leq j < k \leq p.$$

EXAMPLE 5.2 (Bergsma and Dassios'  $t^*$ ). [2] propose a  $U$ -statistic  $t^* = (t_{j,k}^*)_{1 \leq j, k \leq p} = U_n^{(4)}(h^{BD})$  of order 4 with the kernel

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

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

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

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 (1.2) if  $\max_{1 \leq j < k \leq p} |t_{j,k}^*|$  is large (or  $\max_{1 \leq j < k \leq p} t_{j,k}^*$  is large, since in general  $\mathbb{E}[t_{j,k}^*] \geq 0$ ).

EXAMPLE 5.3 (Hoeffding's  $D$ ). [24] proposes a  $U$ -statistic  $D = (D_{j,k})_{1 \leq j, k \leq p} = U_n^{(5)}(h^D)$  of order 5 with the kernel

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



where  $\phi(X_1, \dots, X_5) = (\phi_j(X_1, \dots, X_5))_{j=1}^p$  and  $\phi_j(X_1, \dots, X_5) = [\mathbf{1}(X_1^{(j)} \geq X_2^{(j)}) - \mathbf{1}(X_1^{(j)} \geq X_3^{(j)})][\mathbf{1}(X_1^{(j)} \geq X_4^{(j)}) - \mathbf{1}(X_1^{(j)} \geq X_5^{(j)})]/4$ . Under the assumption that the joint distribution of  $(X^{(j)}, X^{(k)})$  has continuous joint and marginal densities, [24] 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 (1.2) if  $\max_{1 \leq j < k \leq p} |D_{j,k}|$  is large (or  $\max_{1 \leq j < k \leq p} D_{j,k}$  is large, since in general  $\mathbb{E}[D_{j,k}] \geq 0$ ). 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.

Here  $h^S$  is non-degenerate, while  $h^{BD}$  and  $h^D$  are degenerate of order 1 under  $H_0$ . The above testing problem is motivated from recent papers by [32] and [21], which study testing for the null hypothesis

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

and develop tests based on functions of the  $U$ -statistics appearing in Examples 5.1–5.3. Note that  $H_0''$  is a stronger hypothesis than  $H_0$ . Specifically, [32] consider tests statistics such as, e.g.,  $S_{\hat{\rho}} = \sum_{1 \leq j < k \leq p} \hat{\rho}_{j,k}^2 - 3\mu_{\hat{\rho}}$  with  $\mu_{\hat{\rho}} = \mathbb{E}[\hat{\rho}_{1,2}^2]$  under  $H_0''$  and show that  $nS_{\hat{\rho}}/(9p\zeta_1^{\hat{\rho}}) \xrightarrow{d} N(0, 1)$  under  $H_0''$  as  $(n, p) \rightarrow \infty$  where  $\zeta_1^{\hat{\rho}} = \text{Var}(\mathbb{E}[h_{1,2}^S(X_1, X_2, X_3) \mid X_1])$ . On the other hand, [21] consider test statistics such as, e.g.,  $L_n = \max_{1 \leq j < k \leq p} |\hat{\rho}_{j,k}|$  and show that  $L_n^2/\text{Var}(\hat{\rho}_{1,2}) - 4 \log p + \log \log p$  converges in distribution to a Gumbel distribution as  $n \rightarrow \infty$  and  $p = p_n \rightarrow \infty$  under  $H_0''$  provided that  $\log p = o(n^{1/3})$  (precisely speaking, [21] rule out degenerate kernels). Importantly, compared with the tests developed in [32] and [21] based on analytical critical values, our bootstrap-based tests can directly detect the pairwise dependence for some pair of coordinates (or  $\mathbb{E}[\text{sign}(X_1^{(j)} - X_2^{(j)})\text{sign}(X_1^{(k)} - X_3^{(k)})] \neq 0$  for some  $1 \leq j < k \leq p$  for Spearman's  $\rho$ ) rather than the non-mutual-independence and also work for non-continuous random vectors (see, e.g., [18] 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 [32] and [21] 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 5.1–5.3 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  $X$  [35].

In our simulation studies, we consider two test statistics: Spearman's  $\rho$  and Bergsma-Dassios'  $t^*$ . Under  $H_0$  in (1.2), the leading term  $\hat{\rho}$  of Spearman's  $\rho$  is non-degenerate while Bergsma-Dassios'  $t^*$  is degenerate of order 1, both having zero mean. Slightly abusing notation, we will use  $\hat{\rho}$  as Spearman's  $\rho$  statistic throughout this section. We consider tests of the forms

$$\max_{1 \leq j < k \leq p} |\hat{\rho}_{j,k}'| > c \Rightarrow \text{reject } H_0 \quad \text{and} \quad \max_{1 \leq j < k \leq p} |t_{j,k}^{*'}| > c \Rightarrow \text{reject } H_0,$$

where  $\hat{\rho}_{j,k}'$  and  $t_{j,k}^{*'}$  are incomplete versions of  $\hat{\rho}_{j,k}$  and  $t_{j,k}^*$ , respectively, and their critical values are calibrated by the bootstrap methods. In particular, for any nominal size  $\alpha \in (0, 1)$ , the value of  $c := c(\alpha)$  can be chosen as the  $(1 - \alpha)$ -th quantile of an appropriate bootstrap conditional distribution given  $\mathcal{D}_n$ . For Spearman's  $\rho$ , 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  $\rho$  statistic when its distribution can be approximated by  $\gamma_A = N(0, r^2 \Gamma_g)$  (cf. Corollary 3.2).

**5.2. Simulation setup.** We simulate i.i.d. data from the non-central  $t$ -distribution with 3 degrees of freedom and non-centrality parameter 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  $\rho$ , 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  $K = \lfloor (n - 1)/L \rfloor$ ), and the MB-NDG-RS with the parameter values suggested in Section 4.3 (i.e.,  $S_1 = \{1, \dots, n\}$  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  $N = 4n^{3/2}$ . These computational budget parameter values are chosen to minimize the rate in the error bounds of 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.

5.3. *Simulation results.* We first examine the statistical accuracy of the bootstrap tests in terms of size for  $U_n^\sharp$  for the incomplete versions of Spearman's  $\rho$  and  $U_{n,B}^\sharp$  for Bergsma-Dassios'  $t^*$ . For 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. The uniform errors-in-size on  $\alpha \in [0.01, 0.10]$  of our bootstrap tests are summarized in Table 1. We observe that the bootstrap approximations become more accurate as  $n$  increases, and they work quite well for small values of  $\alpha$ , which are relevant in the testing application. Due to the space concern, we defer the empirical size graph  $\{(\alpha, \widehat{R}(\alpha)) : \alpha \in (0, 1)\}$  of the bootstrap tests for MB-NDG-DC (Spearman's  $\rho$ ), MB-NDG-RS (Spearman's  $\rho$ ), and MB-DG (Bergsma-Dassios'  $t^*$ ) to Appendix D in the SM. In addition, we also report the simulation results of the partial bootstrap  $U_{n,A}^\sharp$  for Spearman's  $\rho$  in Appendix D in the SM.

Setup	Spearman's $\rho$ (MB-NDG-DC)	Spearman's $\rho$ (MB-NDG-RS)	Bergsma-Dassios' $t^*$ (MB-DG)
$p = 30, n = 300$	0.0080	0.0110	0.0280
$p = 30, n = 500$	0.0065	0.0130	0.0225
$p = 30, n = 1000$	0.0060	0.0055	0.0095
$p = 50, n = 300$	0.0250	0.0135	0.0385
$p = 50, n = 500$	0.0105	0.0035	0.0260
$p = 50, n = 1000$	0.0145	0.0095	0.0235
$p = 100, n = 300$	0.0180	0.0125	0.0660
$p = 100, n = 500$	0.0135	0.0100	0.0290
$p = 100, n = 1000$	0.0075	0.0020	0.0170

TABLE 1  
Uniform error-in-size  $\sup_{\alpha \in [0.01, 0.10]} |\widehat{R}(\alpha) - \alpha|$  of the bootstrap tests, where  $\alpha$  is the nominal size.

We also report the empirical performance of the Gaussian approximation for the test statistics. The P-P plots for Spearman's  $\rho$  (i.e.,  $\sqrt{n}U'_{n,N}$  versus  $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$ ) and Bergsma-Dassios'  $t^*$  (i.e.,  $\sqrt{N}U'_{n,N}$  versus  $N(0, \Gamma_h)$ ) are shown in Figure 1 and 2, respectively. Similarly as the bootstrap approximations, Gaussian approximations become more accurate as  $n$  increases.

Next, we report the computer running time of the bootstrap tests. Figure 3 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  $p$ . For Spearman's  $\rho$ , 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)

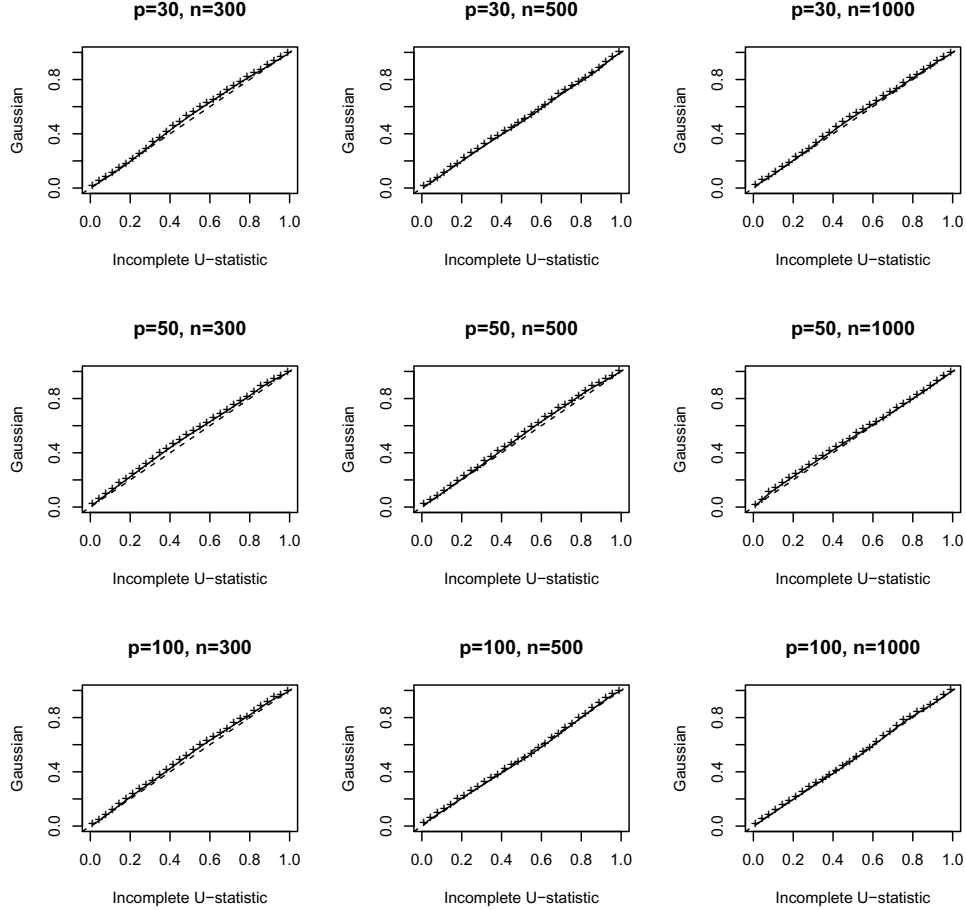


FIG 1. *P-P plots for the Gaussian approximation  $N(0, r^2\Gamma_g + \alpha_n\Gamma_h)$  of  $\sqrt{n}U'_{n,N}$  for Spearman's  $\rho$  test statistic with the Bernoulli sampling.*

in the case MB-NDG-RS. In both cases, the slope coefficients are 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 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

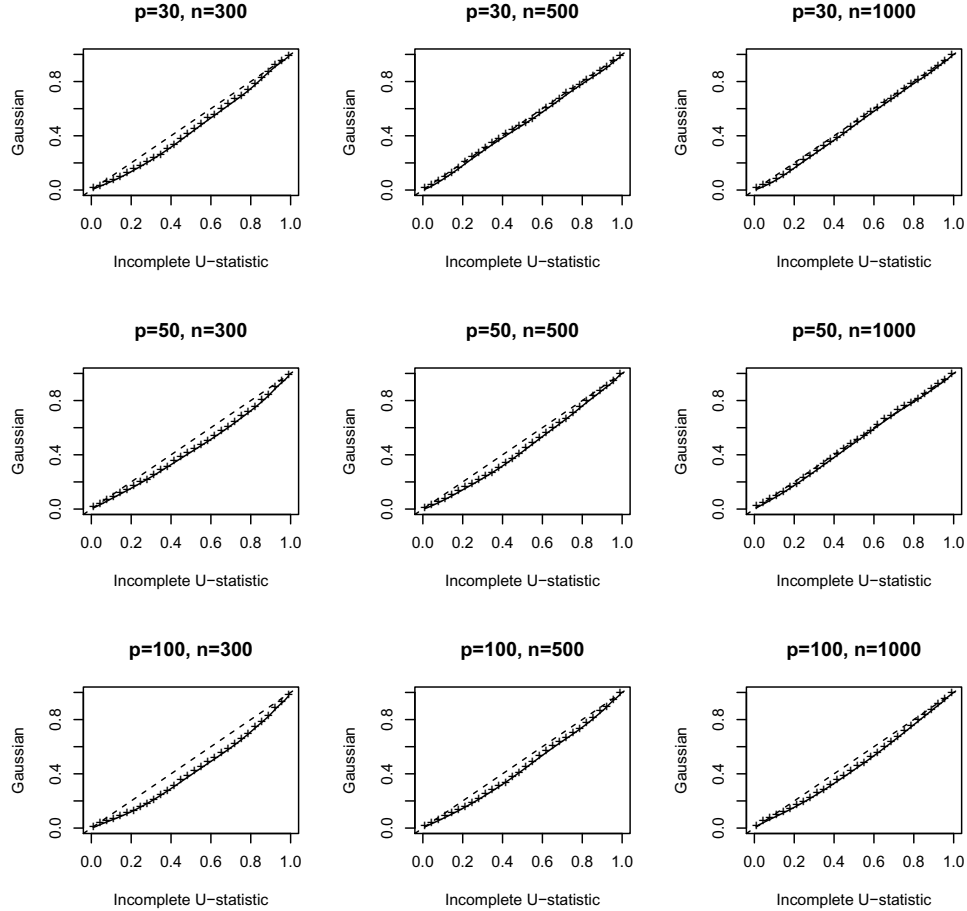


FIG 2.  $P$ - $P$  plots for the Gaussian approximation  $N(0, \Gamma_h)$  of  $\sqrt{N}U'_{n,N}$  for Bergsma-Dassios'  $t^*$  test statistic with the Bernoulli sampling.

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.

**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.

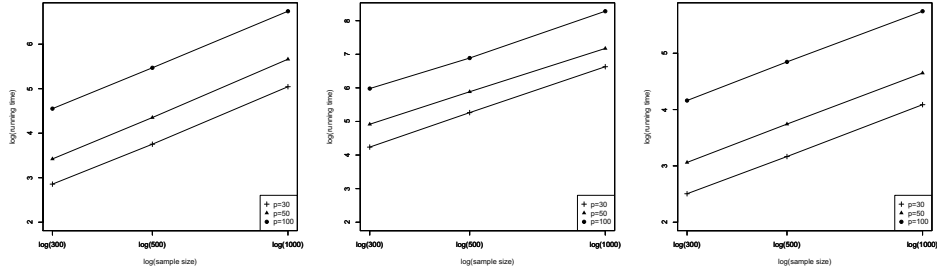


FIG 3. *Computer running time of the bootstrap versus the sample size on the log-scale. Left: bootstrap  $U_n^\sharp$  for Spearman's  $\rho$  with the divide and conquer estimation (MB-NDG-DC). Middle: bootstrap  $U_n^\sharp$  for Spearman's  $\rho$  with the random sampling estimation (MB-NDG-RS). Right: bootstrap  $U_{n,B}^\sharp$  for Bergsma-Dassios'  $t^*$  (MB-DG).*

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. Additional discussions on two extensions (extensions to normalized  $U$ -statistics and incomplete  $U$ -statistics with increasing orders) can be found in Section A of the SM.

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

## SUPPLEMENTARY MATERIAL

### Supplementary Material:

( ). The Supplementary Material contains the proofs and additional discussions, simulation results, and applications of the main paper.

### References.

- [1] ARCONES, M. and GINÉ, E. (1992). On the bootstrap of  $U$ - and  $V$ -statistics. *Annals of Statistics* **20** 655-674.
- [2] BERGSMAS, W. and DASSIOS, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli* **20** 1006-1028.

- [3] BERTAIL, P. and TRESSOU, J. (2006). Incomplete generalized  $U$ -statistics for food risk assessment. *Biometrics* **62** 66-74.
- [4] BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9** 1196-1217.
- [5] BLOM, G. (1976). Some properties of incomplete  $U$ -statistics. *Biometrika* **63** 573-580.
- [6] BRETAGNOLLE, J. (1983). Lois limites du Bootstrap de certaines fonctionnelles. *Annales de l'Institut Henri Poincaré Section B* **XIX** 281-296.
- [7] BROWN, B. M. and KILDEA, D. G. (1978). Reduced  $U$ -statistics and the Hodges-Lehmann estimator. *Annals of Statistics* **6** 828-835.
- [8] CHEN, X. (2018). Gaussian and bootstrap approximations for high-dimensional  $U$ -statistics and their applications. *Annals of Statistics* **46** 642-678.
- [9] CHEN, X. and KATO, K. (2017). Jackknife multiplier bootstrap: finite sample approximations to the  $U$ -process supremum with applications. arXiv:1708.02705.
- [10] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics* **41** 2786-2819.
- [11] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields* **162** 47-70.
- [12] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability* **45** 2309-2352.
- [13] CLÉMENÇON, S., COLIN, I. and BELLET, A. (2016). Scaling-up empirical risk minimization: optimization of incomplete  $U$ -statistics. *Journal of Machine Learning Research* **17** 1-36.
- [14] DE LA PEÑA, V. and GINÉ, E. (1999). *Decoupling: From Dependence to Independence*. Springer.
- [15] DEHLING, H. and MIKOSCH, T. (1994). Random quadratic forms and the bootstrap for  $U$ -statistics. *Journal of Multivariate Analysis* **51** 392-413.
- [16] EMBRECHTS, P., LINDSKOG, F. and MCNEIL, A. (2003). Modelling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance* (S. T. Rachev, ed.) 8 North Holland.
- [17] FREES, E. W. (1989). Infinite order  $U$ -statistics. *Scandinavian Journal of Statistics* **16**.
- [18] GEISSER, S. and MANTEL, N. (1962). Pairwise independence of jointly dependent random variables. *Annals of Mathematical Statistics* **33** 290-291.
- [19] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.
- [20] GU, Q., CAO, Y., NING, Y. and LIU, H. (2015). Local and global inference for high dimensional nonparanormal graphical models. arXiv:1502.02347.
- [21] HAN, F., CHEN, S. and LIU, H. (2017, to appear). Distribution-free tests of independence in high dimensions. *Biometrika*.
- [22] HAN, F. and QIAN, T. (2016). Asymptotics for asymmetric weighted  $U$ -statistics: Central limit theorem and bootstrap under data heterogeneity. *Preprint*.
- [23] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics* **19** 293-325.
- [24] HOEFFDING, W. (1948). A nonparametric test of independence. *Annals of Mathematical Statistics* **19** 546-557.
- [25] HSING, T. and WU, W. B. (2004). On weighted  $U$ -statistics for stationary processes. *Annals of Probability* **32** 1600-1631.
- [26] HUŠKOVÁ, M. and JANSSEN, P. (1993). Consistency of the generalized bootstrap for

- degenerate  $U$ -statistics. *Annals of Statistics* **21** 1811-1823.
- [27] HUŠKOVÁ, M. and JANSSEN, P. (1993). Generalized bootstrap for studentized  $U$ -statistics: a rank statistic approach. *Statistics and Probability Letters* **16** 225-233.
  - [28] JANSON, S. (1984). The asymptotic distributions of incomplete  $U$ -statistics. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **66** 495-505.
  - [29] KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **76** 795-816.
  - [30] LEE, A. J. (1990). *U-Statistics: Theory and Practice. Statistics: A Series of Textbooks and Monographs (Book 110)*. CRC Press.
  - [31] LEHMANN, E. L. (1998). *Elements of Large-Sample Theory*. Springer Texts in Statistics.
  - [32] LEUNG, D. and DRTON, M. (2017, to appear). Testing independence in high dimensions with sums of rank correlations. *Annals of Statistics*.
  - [33] MAJOR, P. (1994). Asymptotic distributions for weighted  $U$ -statistics. *Annals of Probability* **21** 1514-1535.
  - [34] MENTCH, L. and HOOKER, G. (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *Journal of Machine Learning Research* **17** 1-41.
  - [35] NANDY, P., WEIHS, L. and DRTON, M. (2016). Large-sample theory for the Bergsma-Dassios sign covariance. *Electronic Journal of Statistics* **10** 2287-2311.
  - [36] O'NEIL, K. A. and REDNER, R. A. (1993). Asymptotic distributions of weighted  $U$ -statistics of degree 2. *Annals of Probability* **21** 1159-1169.
  - [37] RIFI, M. and UTZET, F. (2000). On the asymptotic behavior of weighted  $U$ -statistics. *Journal of Theoretical Probability* **13** 141-167.
  - [38] RUBIN, H. and VITALE, R. A. (1980). Asymptotic distribution of symmetric statistics. *Annals of Statistics* **8** 165-170.
  - [39] SHAPIRO, C. P. and HUBERT, L. (1979). Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *Annals of Statistics* **7** 788-794.
  - [40] SZÉKELY, G., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics* **35** 2769-2794.
  - [41] VAN DER VAART, A. (1998). *Asymptotic Statistics*. Cambridge University Press.
  - [42] VAN DER VAART, A. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
  - [43] WAGER, S. and ATHEY, S. (2017, to appear). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of American Statistical Association*. arXiv:1510:04342.
  - [44] WANG, Q. and JING, B.-Y. (2004). Weighted bootstrap for  $U$ -statistics. *Journal of Multivariate Analysis* **91** 177-198.
  - [45] YAO, S., ZHANG, X. and SHAO, X. (2017, to appear). Testing mutual independence in high dimension via distance covariance. *J. R. Stat. Soc. Ser. B Stat. Methodol.*
  - [46] ZHANG, Y., DUCHI, J. and WAINWRIGHT, M. J. (2015). Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research* **16** 3299-3340.



## SUPPLEMENTARY MATERIAL TO “RANDOMIZED INCOMPLETE $U$ -STATISTICS IN HIGH DIMENSIONS”<sup>||</sup>

BY XIAOHUI CHEN<sup>\*\*</sup> AND KENGO KATO<sup>††</sup>

*University of Illinois at Urbana-Champaign<sup>\*\*</sup> and Cornell University<sup>††</sup>*

This Supplementary Material contains the additional discussions, bootstrap validity under the polynomial moment condition, proofs omitted from the main text, and additional simulation results of the paper “Randomized incomplete  $U$ -statistics in high dimensions.”

### A. Additional discussions.

A.1. *Comparison of MB-NDG-DC with BLB.* Our MB-NDG-DC in Section 4.2 differs from the Bag of Little Bootstraps (BLB) proposed in [29], which is another generically scalable bootstrap method for large datasets based on the divide and conquer (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 [31]). For instance,  $\lambda(Q_n(P))$  can be the 95%-quantile of the distribution of  $\max_{1 \leq j \leq d} \sqrt{n}(U_{n,j} - \theta_j)$ . 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 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  $\mathbb{P}_{n,L}^{(k)} = L^{-1} \sum_{i \in \mathcal{I}_k} \delta_{X_i}$  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  $\mathbb{P}_{n,L}^{(k)}$ , 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  $\lambda(Q_n(\mathbb{P}_{n,L}^{(k)}))$  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  $O(BKL^r d) = O(BnL^{r-1}d)$ . Note that the asymptotic validity of the BLB requires that  $L \rightarrow \infty$  (cf. Theorem 1 of [29]), so that  $\mathbb{P}_{n,L}^{(k)}$  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

---

<sup>||</sup>This version: January 29, 2019.

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 of the same order as  $n$ .

*A.2. 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

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

where  $\widehat{N}$  is replaced by  $N$  in the definition of  $\hat{\sigma}_{B,j}^2$  for the sampling with replacement case. These estimators are the  $(j, j)$ -elements of the conditional covariance matrices of  $U_{n,A}^\#$  and  $U_{n,B}^\#$ , respectively. Note that the computational cost to construct  $\hat{\sigma}_{B,j}^2, j = 1, \dots, d$  is (on average)  $O(Nd)$ , while that of  $\hat{\sigma}_{A,j}^2, j = 1, \dots, d$  is  $O(n^2d)$  if the DC estimation with the parameter values suggested in Section 4.2 is used for estimation of  $g$ . Now, let  $\Lambda_A = \text{diag}\{\sigma_{A,1}^2, \dots, \sigma_{A,d}^2\}$ ,  $\Lambda_B = \text{diag}\{\sigma_{B,1}^2, \dots, \sigma_{B,d}^2\}$ ,  $\widehat{\Lambda}_A = \text{diag}\{\hat{\sigma}_{A,1}^2, \dots, \hat{\sigma}_{A,d}^2\}$ ,  $\widehat{\Lambda}_B = \text{diag}\{\hat{\sigma}_{B,1}^2, \dots, \hat{\sigma}_{B,d}^2\}$ ,  $\Lambda = \text{diag}\{\sigma_1^2, \dots, \sigma_d^2\} = \Lambda_A + \alpha_n \Lambda_B$ , and  $\widehat{\Lambda} = \text{diag}\{\hat{\sigma}_1^2, \dots, \hat{\sigma}_d^2\} = \widehat{\Lambda}_A + \alpha_n \widehat{\Lambda}_B$ . 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 A.1** (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 (4.3) together with  $D_n^2(\log^7(dn))/(n \wedge N) \leq C_1 n^{-(\zeta_1 \wedge \zeta_2)}$  hold for some constants  $0 < C_1 < \infty$  and  $\zeta_1, \zeta_2 \in (0, 1)$ . Then there exists a constant  $C$  depending only on  $\underline{\sigma}, r$ ,*

and  $C_1$  such that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n} \hat{\Lambda}^{-1/2} (U'_{n,N} - \theta) \in R) - \mathbb{P}(\Lambda^{-1/2} Y \in R) \right| \leq C n^{-(\zeta_1 \wedge \zeta_2)/6} \quad \text{and} \\ \mathbb{P} \left\{ \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(\hat{\Lambda}^{-1/2} U_n^\# \in R) - \mathbb{P}(\Lambda^{-1/2} Y \in R) \right| > C n^{-(\zeta_1 \wedge \zeta_2)/6} \right\} \leq C n^{-1},$$

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 (4.1) 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

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(\hat{\Lambda}_B^{-1/2} U_{n,B}^\# \in R) - \gamma_B^\dagger(R) \right| \leq C n^{-\zeta/6}$$

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

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{N} \hat{\Lambda}_B^{-1/2} (U'_{n,N} - \theta) \in R) - \gamma_B^\dagger(R) \right| \leq C' n^{-\zeta/6}.$$

**A.3. 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 [34] studies asymptotic theory for one-dimensional incomplete  $U$ -statistics with increasing orders (i.e.,  $r = r_n \rightarrow \infty$ ). 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 (subsampling) random forests, which, from a mathematical point of view, are defined as *infinite order*  $U$ -statistics [17] where the order of the  $U$ -statistics corresponds to the subsample size for a single tree and so  $r = r_n \rightarrow \infty$ . 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, [34] develop pointwise confidence intervals for subbagging and random forests; see also [43] for related results. Extending the results of [34] 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.

**B. Bootstrap validity under the polynomial moment condition.**

In this section, we present the bootstrap validity under the polynomial moment condition (C2'). Recall the assumption

$$(C2') \quad (P^r |h|_\infty^q)^{1/q} \leq D_n \text{ for some } q \in [4, \infty),$$

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

**THEOREM B.1** (Validity of  $U_{n,B}^\sharp$ ). *Suppose that (C1), (C2'), and (C3-D) hold. If*

$$(B.1) \quad \frac{D_n^2 \log^5(dn)}{n \wedge N} \leq C_1 n^{-\zeta} \quad \text{and} \quad \frac{D_n^2 n^{2 \max(r+1,4)/q} \log^3(dn)}{n \wedge N} \leq C_1 n^{-\zeta/2}$$

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

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(U_{n,B}^\sharp \in R) - \gamma_B(R) \right| \leq C n^{-\zeta/6}$$

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

**THEOREM B.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*

$$(B.2) \quad \begin{aligned} \frac{D_n^2 \log^5(dn)}{n_1 \wedge N} &\leq C_1 n^{-\zeta_1}, \quad \frac{D_n^2 n^{2 \max(r+1,4)/q} \log^3(dn)}{n_1 \wedge N} \leq C_1 n^{-\zeta_1/2}, \quad \text{and} \\ \mathbb{P} \left( \bar{\sigma}_g^2 \widehat{\Delta}_{A,1} \log^4 d > C_1 n^{-\zeta_2} \right) &\leq C_1 n^{-1} \end{aligned}$$

for some constants  $0 < C_1 < \infty$  and  $\zeta_1, \zeta_2 \in (0, 1)$ , where  $\bar{\sigma}_g := \max_{1 \leq j \leq d} \sqrt{P(g_j - \theta_j)^2}$ . Then there exists a constant  $C$  depending only on  $\underline{\sigma}, r, q$ , and  $C_1$  such that

$$(B.3) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(U_n^\sharp \in R) - \mathbb{P}(Y \in R) \right| \leq C n^{-(\zeta_1 \wedge \zeta_2)/6}$$

with probability at least  $1 - C n^{-1}$ , 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 (B.3), with  $\mathcal{D}_n$  replaced by the augmentation of  $\mathcal{D}_n$  with variables used in the additional randomization, holds with probability at least  $1 - C n^{-1}$ .

The following two propositions are concerned with extensions of Propositions 4.4 and 4.5 under the polynomial moment condition (C2').

**PROPOSITION B.3** (Validity of bootstrap with DC estimation). *Consider the MB-NDG-DC defined in Section 4.2. Suppose that Conditions (C1), (C2'), and (C3-ND) hold. In addition, suppose that*

$$\frac{D_n^2(\log^2 n) \log^5(dn)}{n_1 \wedge N} \bigvee \left\{ \frac{\bar{\sigma}_g^2 D_n^2 \log^5 d}{KL} \left( 1 + \frac{\log^2 d}{K^{1-2/q}} \right) \right\} \leq C_1 n^{-\zeta} \quad \text{and}$$

$$\frac{D_n^2 n^{2 \max(r+1, 4)/q} \log^3(dn)}{n_1 \wedge N} \leq C_1 n^{-\zeta/2}$$

for some constants  $0 < C_1 < \infty$  and  $\zeta \in (2/q, 1)$ . Then there exists a constant  $C$  depending only on  $\underline{\sigma}, r, q$ , and  $C_1$  such that the result (B.3) with  $(\zeta_1, \zeta_2) = (\zeta, \zeta - 2/q)$  holds with probability at least  $1 - Cn^{-1}$ .

**PROPOSITION B.4** (Validity of bootstrap with Bernoulli sampling estimation). *Consider the MB-NDG-RS defined in Section 4.3. Suppose that Conditions (C1), (C2'), and (C3-ND) hold. In addition, suppose that*

$$\frac{D_n^2(\log^2 n) \log^5(dn)}{n_1 \wedge N} \bigvee \frac{\bar{\sigma}_g^2 D_n^2 \log^5 d}{n} \bigvee \frac{\bar{\sigma}_g^2 D_n^2 n^{2(r-1)/q} (\log d)^{5+2/q}}{M} \bigvee \frac{\bar{\sigma}_g^2 D_n^2 n^{2r/q} \log^6 d}{M^2} \leq C_1 n^{-\zeta} \quad \text{and}$$

$$\frac{D_n^2 n^{2 \max(r+1, 4)/q} \log^3(dn)}{n_1 \wedge N} \leq C_1 n^{-\zeta/2}$$

for some constants  $0 < C_1 < \infty$  and  $\zeta \in (2/q, 1)$ . Then there exists a constant  $C$  depending only on  $\underline{\sigma}, r, q$ , and  $C_1$  such that result (B.3), with  $\mathcal{D}_n$  replaced by  $\mathcal{D}'_n = \mathcal{D}_n \cup \{Z'_{\iota'} : \iota' \in I_{n-1, r-1}\}$  and with  $(\zeta_1, \zeta_2) = (\zeta, \zeta - 2/q)$ , holds with probability at least  $1 - Cn^{-1}$ .

### C. Proofs.

**C.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  $\psi_\beta$ -norms.

**LEMMA C.1** (Maximal inequalities for the  $\psi_\beta$ -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  $k \geq 2$ . Then*

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

where  $C_\beta$  is a constant that depends only on  $\beta$ .

PROOF OF LEMMA C.1. For  $\beta \geq 1$ , the lemma follows from Lemma 2.2.2 in [42]. For  $\beta \in (0, 1)$ ,  $\psi_\beta$  is not convex and so we can not directly apply Lemma 2.2.2 in [42], but apply the lemma for the norm equivalent to  $\|\cdot\|_{\psi_\beta}$  obtained by linearizing  $\psi_\beta$  in a neighborhood of the origin; see Lemma C.2 below.  $\square$

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

$$\tilde{\psi}_\beta(x) = \begin{cases} \psi_\beta(x) & \text{if } x \geq x_\beta \\ \frac{\psi_\beta(x_\beta)}{x_\beta}x & \text{if } 0 \leq x \leq x_\beta \end{cases}$$

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

$$C_\beta^{-1} \|\xi\|_{\tilde{\psi}_\beta} \leq \|\xi\|_{\psi_\beta} \leq C_\beta \|\xi\|_{\tilde{\psi}_\beta}$$

*for every real-valued random variable  $\xi$ .*

PROOF OF LEMMA C.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  $\beta$ . We first show that  $\|\xi\|_{\tilde{\psi}_\beta} \lesssim \|\xi\|_{\psi_\beta}$ . To this end, we may assume that  $\|\xi\|_{\psi_\beta} = 1$ , i.e.,  $\mathbb{E}[\psi_\beta(|\xi|)] = 1$ , and show that  $\|\xi\|_{\tilde{\psi}_\beta} \lesssim 1$ . By Taylor's theorem, we have  $\psi_\beta(x) \gtrsim x$  and  $\psi_\beta(x/C) \leq C^{-\beta} \psi_\beta(x)$  for  $C > 1$ , so that

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

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

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

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

LEMMA C.3 (Useful maximal inequalities for  $U$ -statistics: sub-exponential moment). *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  $h = (h_1, \dots, h_d)^T : S^r \rightarrow \mathbb{R}^d$  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  $U_n(h) = |I_{n,r}|^{-1} \sum_{i \in I_{n,r}} h(X_i)$  with kernel  $h$ , and let  $Z = \max_{1 \leq j \leq d} |U_n(h_j) - P^r h_j|$ . In addition, let

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

$$M = \max_{1 \leq i \leq m} \max_{1 \leq j \leq d} |h_j(X_{(i-1)r+1}^{ir}) - P^r h_j|,$$

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  $\beta, \eta$ , and  $\delta$  such that

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

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

**PROOF OF LEMMA C.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],  $\check{Z}$  is defined as  $\max_{1 \leq j \leq d} |\sum_{i=1}^m \{\bar{h}_j(X_{(i-1)r+1}^{ir}) - P^r \bar{h}_j\}|$  where  $\bar{h}$  is to be defined below. Without loss of generality, we may assume that  $t \geq C_1 \|M\|_{\psi_\beta}$  for some sufficiently large constant  $C_1$  that depends only on  $\beta, \eta$  and  $\delta$ . For  $\tau = 8\mathbb{E}[M]$ , let  $\bar{h}(x_1, \dots, x_r) = h(x_1, \dots, x_r) \mathbf{1}(|h(x_1, \dots, x_r)|_\infty \leq \tau)$  and  $\underline{h} = h - \bar{h}$ . 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,  $Z \leq T_1 + T_2$ , and since  $h = \bar{h} + \underline{h}$  and hence  $\bar{h} = h + (-\underline{h})$ , we have  $\mathbb{E}[|W_1(X_1^n)|_\infty] \leq \mathbb{E}[\check{Z}] + \mathbb{E}[|W_2(X_1^n)|_\infty]$ , so that  $\mathbb{E}[\check{Z}] \geq \mathbb{E}[|W_1(X_1^n)|_\infty] - \mathbb{E}[|W_2(X_1^n)|_\infty]$ . Hence, for every  $\eta > 0$  and  $\varepsilon \in (0, 1)$ ,

$$\begin{aligned} \mathbb{P} \left( Z \geq (1 + \eta) \mathbb{E}[\check{Z}] + t \right) \\ \leq \mathbb{P} (T_1 \geq (1 + \eta)(\mathbb{E}[|W_1(X_1^n)|_\infty] - \mathbb{E}[|W_2(X_1^n)|_\infty]) + (1 - \varepsilon)t) + \mathbb{P}(T_2 \geq \varepsilon t). \end{aligned}$$

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

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

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

LEMMA C.4 (Useful maximal inequalities for  $U$ -statistics: polynomial moment). *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  $h = (h_1, \dots, h_d)^T : S^r \rightarrow \mathbb{R}^d$  be a symmetric and jointly measurable function such that  $P^r|h_j|^q < \infty$  for all  $j = 1, \dots, d$  for some  $q \in [1, \infty)$ . Consider the associated  $U$ -statistic  $U_n(h) = |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h(X_\iota)$  with kernel  $h$ . Let  $Z$ ,  $\check{Z}$ , and  $M$  have the same definition as in Lemma C.3. Then, for every  $\eta \in (0, 1]$  and  $\delta > 0$ , there exists a constant  $C$  depending only on  $q, \eta$ , and  $\delta$  such that*

$$\mathbb{P}\left(mZ \geq (1 + \eta)\mathbb{E}[\check{Z}] + t\right) \leq \exp\left(-\frac{t^2}{2(1 + \delta)m\sigma^2}\right) + \frac{C\mathbb{E}[M^q]}{t^q}$$

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

PROOF OF LEMMA C.4. The proof follows from that of Lemma E.2 in [8] and a similar modification argument of Lemma C.3. Details are omitted.  $\square$

LEMMA C.5 (Gaussian comparison on hyperrectangles). *Let  $Y, W$  be centered Gaussian random vectors in  $\mathbb{R}^d$  with covariance matrices  $\Sigma^Y = (\Sigma_{j,k}^Y)_{1 \leq j,k \leq d}$ ,  $\Sigma^W = (\Sigma_{j,k}^W)_{1 \leq j,k \leq d}$ , respectively, and let  $\Delta = \|\Sigma^Y - \Sigma^W\|_\infty$ . Suppose that  $\min_{1 \leq j \leq d} \Sigma_{j,j}^Y \vee \min_{1 \leq j \leq d} \Sigma_{j,j}^W \geq \underline{\sigma}^2$  for some constant  $\underline{\sigma} > 0$ . Then*

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

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

PROOF OF LEMMA C.5. The proof is implicit in the proof of Theorem 4.1 in [12].  $\square$

C.2. *Proof sketch of Theorem 3.1.* Before we formally prove Theorem 3.1, we first sketch the proof for the Bernoulli sampling case. Similar arguments carry through the sampling with replacement case as well. The rigorous proof of Theorem 3.1 is given in Section C.3. Let  $C$  denote a generic constant that depends only on  $\underline{\sigma}$  and  $r$ . 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 given by

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

Subject to the moment conditions (C1) and (C2), Lemma C.3 in the SM and Lemma 8 in [11] yield that with probability at least  $1 - Cn^{-1}$ ,

$$(C.1) \quad |\hat{\Gamma}_h - \Gamma_h|_\infty \leq C\{n^{-1/2}D_n \log^{1/2}(dn) + n^{-1}D_n(\log n) \log^3(dn)\},$$

where  $\Gamma_h = P^r(h - \theta)(h - \theta)^T$ . Under the non-degeneracy condition (C3-ND), by the Gaussian approximation result (cf. Theorem 2.1 in [12]) applied conditionally on  $X_1^n$ , we can show that with probability at least  $1 - Cn^{-1}$ ,

$$(C.2) \quad \rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \hat{Y}) := \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|X_1^n}(\sqrt{N}B_n \in R) - \mathbb{P}_{|X_1^n}(\hat{Y} \in R) \right| \leq C\varpi_n,$$

where  $\varpi_n = \{D_n^2 \log^7(dn)/(n \wedge N)\}^{1/6}$  and  $\hat{Y}$  is a random vector in  $\mathbb{R}^d$  such that  $\hat{Y} \mid X_1^n \sim N(0, \hat{\Gamma}_h)$ . This step is involved and we defer the details of deriving (C.2) to Appendix C in the SM. Then, in view of (C.1), the Gaussian comparison inequality (Lemma C.5 in the SM) yields that

$$(C.3) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|X_1^n}(\hat{Y} \in R) - \gamma_B(R) \right| \leq C\varpi_n$$

with probability at least  $1 - Cn^{-1}$ , where  $\gamma_B = N(0, \Gamma_h)$ . Since  $A_n$  is a mean-zero complete  $U$ -statistic, the high-dimensional CLT for the Hájek projection term (cf. Proposition 2.1 in [12]), the maximal inequality for the higher-order degenerate terms (cf. Corollary 5.6 in [9]), and Nazarov's inequality (cf. Lemma A.1 in [12]) together yield that

$$(C.4) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}A_n \in R) - \gamma_A(R) \right| \leq C\varpi_n,$$

where  $\gamma_A = N(0, r^2 \Gamma_g)$  and  $\Gamma_g = P(g - \theta)(g - \theta)^T$  with  $g = P^{r-1}h$ . For any hyperrectangle  $R \in \mathcal{R}$ , observe that

$$\begin{aligned} \mathbb{P}(\sqrt{n}W_n \in R) &= \mathbb{E} \left[ \mathbb{P}_{|X_1^n} \left( \sqrt{N}B_n \in \left[ \frac{1}{\sqrt{\alpha_n(1-p_n)}}R - \sqrt{\frac{N}{1-p_n}}A_n \right] \right) \right] \\ &\leq_{(*)} \mathbb{E} \left[ \gamma_B \left( \left[ \frac{1}{\sqrt{\alpha_n(1-p_n)}}R - \sqrt{\frac{N}{1-p_n}}A_n \right] \right) \right] + C\varpi_n \\ &= \mathbb{P} \left( \sqrt{1-p_n}Y_B \in [\alpha_n^{-1/2}R - \sqrt{N}A_n] \right) + C\varpi_n \\ &= \mathbb{P} \left( \sqrt{n}A_n \in [R - \sqrt{\alpha_n(1-p_n)}Y_B] \right) + C\varpi_n, \end{aligned}$$

where  $Y_B \sim N(0, \Gamma_h) = \gamma_B$  independent of  $X_1^n$ , and the inequality  $(*)$  follows from (C.2) and (C.3). Now we freeze the random variable  $Y_B$ . Since  $Y_B$  is independent of  $X_1^n$ , (C.4) yields that

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

By Fubini, we conclude that

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

where  $Y_A \sim N(0, r^2 \Gamma_g) = \gamma_A$  is independent of  $Y_B$ . Since  $\alpha_n p_n = n/|I_{n,r}| \leq Cn^{-r+1} \leq Cn^{-1}$  and  $|\Gamma_h|_\infty \leq CD_n$ , a second application of the Gaussian comparison inequality (Lemma C.5) yields that

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

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

$$(C.5) \quad \sup_{R \in \mathcal{R}} |\mathbb{P}(\sqrt{n}W_n \in R) - \mathbb{P}(Y \in R)| \leq C\varpi_n.$$

In the last step of the proof, using Bernstein's inequality (cf. Lemma 2.2.9 in [42]), we can verify that the inequality (C.5) holds with  $\sqrt{n}W_n$  replaced by  $\sqrt{n}(U'_{n,N} - \theta)$ . This leads to the conclusion of Theorem 3.1 in the Bernoulli sampling case.  $\square$

C.3. *Proofs for Section 3.* Observe that  $P^r|h_j - \theta_j|^{2+k} \leq 2^{2+k}D_n^k$  by Jensen's inequality for all  $j$  and  $k = 1, 2$ , and  $\|h_j(X_1^r) - \theta_j\|_{\psi_1} \leq (1 + \log 2)D_n$  for all  $j$ . So, in view of the identity  $U'_{n,N} - \theta = \hat{N}^{-1} \sum_{\iota \in I_{n,r}} Z_\iota \{h(X_\iota) - \theta\}$  where  $\hat{N}$  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  $\underline{\sigma}, r$  under (C2) and only on  $\underline{\sigma}, r, q$  under (C2'). In addition, let  $C$  denote a generic constant that depends only on  $\underline{\sigma}, r$  under (C2) and only on  $\underline{\sigma}, r, q$  under (C2'); 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 (3.1) holds since  $n = N\alpha_n$ . So it suffices to prove the second line in (3.1). In this proof, without loss of generality, we may assume that

$$(C.6) \quad D_n^2 \log^7(dn) \leq c_1(n \wedge N)$$

for some sufficiently small constant  $c_1 \in (0, 1)$  depending only on  $\underline{\sigma}$  and  $r$ , since otherwise the conclusion of the theorem is trivial by taking  $C$  in (3.1) sufficiently large. In addition, for the notational convenience, let

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

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

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

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

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

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

$$(C.7) \quad \hat{L}_n = \max_{1 \leq j \leq d} \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} (p_n(1-p_n))^{-3/2} |h_j(X_\iota)|^3 \mathbb{E}[|Z_\iota - p_n|^3].$$

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

$$(C.8) \quad \begin{aligned} \widehat{M}_{n,X}(\phi) &= \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} \mathbb{E}_{|X_1^n} \left[ \max_{1 \leq j \leq d} \left| \frac{(Z_\iota - p_n)h_j(X_\iota)}{\sqrt{p_n(1-p_n)}} \right|^3 \mathbf{1} \left( \max_{1 \leq j \leq d} \left| \frac{(Z_\iota - p_n)h_j(X_\iota)}{\sqrt{p_n(1-p_n)}} \right| > \frac{\sqrt{|I_{n,r}|}}{4\phi \log d} \right) \right], \\ \widehat{M}_{n,Y}(\phi) &= \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} \mathbb{E}_{|X_1^n} \left[ \max_{1 \leq j \leq d} |\hat{Y}_{\iota,j}|^3 \mathbf{1} \left( \max_{1 \leq j \leq d} |\hat{Y}_{\iota,j}| > \frac{\sqrt{|I_{n,r}|}}{4\phi \log d} \right) \right], \end{aligned}$$

and  $\widehat{M}_n(\phi) = \widehat{M}_{n,X}(\phi) + \widehat{M}_{n,Y}(\phi)$ . Let  $\bar{L}_n$  and  $\bar{M}_n$  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

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

we have that

$$(C.9) \quad \rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \hat{Y}) \leq C \left\{ \left( \frac{\bar{L}_n^2 \log^7 d}{|I_{n,r}|} \right)^{1/6} + \frac{\bar{M}_n}{\bar{L}_n} \right\}$$

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

Step 1.2: Bounding  $\hat{L}_n$ . Since  $p_n \leq 1/2$  and  $\mathbb{E}[|Z_\iota - p_n|^3] = p_n(1-p_n)\{p_n^2 + (1-p_n)^2\} \leq Cp_n$ ,  $\hat{L}_n$  is bounded from above by  $Cp_n^{-1/2}$  times  $\max_{1 \leq j \leq d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} |h_j(X_\iota)|^3 =: Z_1$ . Let  $m = \lfloor n/r \rfloor$ ,  $\check{Z}_1 = \max_{1 \leq j \leq d} \sum_{i=1}^m |h_j(X_{(i-1)r+1}^{ir})|^3$ ,

and  $M_1 = \max_{1 \leq i \leq m} \max_{1 \leq j \leq d} |h_j(X_{(i-1)r+1}^{ir})|$ . Then, Lemma E.3 in [8] yields that

$$\mathbb{P}\left(mZ_1 \geq 2\mathbb{E}[\check{Z}_1] + C\|M_1^3\|_{\psi_{1/3}}t^3\right) \leq 3e^{-t}$$

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

$$\mathbb{E}[\check{Z}_1] \lesssim \max_{1 \leq j \leq d} \sum_{i=1}^m \mathbb{E}\left[|h_j(X_{(i-1)r+1}^{ir})|^3\right] + \mathbb{E}[M_1^3] \log d \lesssim mD_n + \mathbb{E}[M_1^3] \log d.$$

Since  $\mathbb{E}[M_1^3] \lesssim \|M_1^3\|_{\psi_{1/3}} = \|M_1\|_{\psi_1}^3 \lesssim D_n^3 \log^3(dn)$ , we have

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

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

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

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

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

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

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

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

by choosing  $c_1$  in (C.6) sufficiently small. In addition,  $\phi_n^{-1} = C_2^{-1}C^{1/3}(N^{-1}D_n^2 \log^4 d)^{1/6} \leq C_2^{-1}C^{1/3}c_1^{1/6} \leq 1$  for a sufficiently small  $c_1$ . Hence 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

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

which holds with probability at least  $1 - 2n^{-1}$  by Step 1.3. Recall that  $\|\xi\|_{\psi_1} \leq (1+e)\|\xi\|_{\psi_2}$  for every real-valued random variable. (For completeness, we provide its proof: assume  $\|\xi\|_{\psi_2} = 1$ , and observe that  $\mathbb{E}[e^{|\xi|}] \leq$

$e + \mathbb{E}[e^{\xi^2}] \leq e + 2$ , so that  $\mathbb{E}[\psi_1(|\xi|)] \leq 1 + e$ . The desired result follows from the observation that  $\psi_1(x/C) \leq C^{-1}\psi_1(x)$  for  $C > 1$  and  $x \geq 0$ .) Conditionally on  $X_1^n$ , since  $\hat{Y}_{\iota,j} \mid X_1^n \sim N(0, h_j^2(X_\iota))$  for every  $\iota \in I_{n,r}$ , we have  $\|\max_{1 \leq j \leq d} |\hat{Y}_{\iota,j}|\|_{\psi_1} \leq (1+e) \|\max_{1 \leq j \leq d} |\hat{Y}_{\iota,j}|\|_{\psi_2} \lesssim \max_{1 \leq j \leq d} |h_j(X_\iota)| \log^{1/2} d$ , so that

$$(C.10) \quad \mathbb{P}_{|X_1^n} \left( \max_{1 \leq j \leq d} |\hat{Y}_{\iota,j}| \geq t \right) \leq 2 \exp \left( - \frac{t}{C \max_{1 \leq j \leq d} |h_j(X_\iota)| \log^{1/2} d} \right)$$

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

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

where we have used the assumption (C.6). Therefore, we conclude that  $\widehat{M}_{n,Y}(\phi_n) \leq C n^{3r/2} e^{-n^{1/4}/C}$  with probability at least  $1 - 2n^{-1}$ .

Step 1.5: Bounding  $|\widehat{\Gamma}_h - \Gamma_h|_\infty$ . Let  $Z_2 = |\widehat{\Gamma}_h - \Gamma_h|_\infty$ , and observe that

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

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

$$\mathbb{P} \left( m Z_2 \geq 2\mathbb{E}[\check{Z}_2] + t \right) \leq e^{-t^2/(3mD_n^2)} + 3 \exp\{- (t/(C\|\mathbf{M}_2\|_{\psi_{1/2}}))^{1/2}\},$$

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

$$\check{Z}_2 = \max_{1 \leq j, k \leq d} \left| \sum_{i=1}^m \{h_j(X_{(i-1)r+1}^{ir})h_k(X_{(i-1)r+1}^{ir}) - P^r h_j h_k\} \right| \quad \text{and}$$

$$M_2 = \max_{1 \leq j, k \leq d} \max_{1 \leq i \leq m} \left| h_j(X_{(i-1)r+1}^{ir})h_k(X_{(i-1)r+1}^{ir}) - P^r h_j h_k \right|.$$

Observe that  $\|M_2\|_{\psi_{1/2}} \lesssim D_n^2 \log^2(dn)$ . In addition, Lemma 8 in [11] yields that

$$\mathbb{E}[\check{Z}_2] \lesssim \sqrt{m D_n^2 \log d} + \sqrt{\mathbb{E}[M_2^2] \log d} \lesssim D_n \sqrt{n \log d} + D_n^2 \log^3(dn).$$

Hence,

$$\begin{aligned} \mathbb{P}\left(Z_2 \geq C\{n^{-1/2} D_n \log^{1/2} d + n^{-1} D_n^2 \log^3(dn)\} + t\right) \\ \leq \exp\left(-\frac{nt^2}{3r D_n^2}\right) + 3 \exp\left(-\frac{(nt)^{1/2}}{C D_n \log(dn)}\right). \end{aligned}$$

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

(C.11)

$$\mathbb{P}\left(Z_2 \geq C\{n^{-1/2} D_n \log^{1/2}(dn) + n^{-1} D_n^2 (\log n) \log^3(dn)\}\right) \leq Cn^{-1}.$$

Choosing  $c_1$  in (C.6) small enough, we conclude that  $|\hat{\Gamma}_h - \Gamma_h|_\infty \leq \underline{\sigma}^2/2$  and hence  $\min_{1 \leq j \leq d} \hat{\Gamma}_{h,jj} \geq \underline{\sigma}^2/2$  with probability at least  $1 - Cn^{-1}$ .

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

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

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

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

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

with probability at least  $1 - Cn^{-1}$ , where  $\gamma_B = N(0, \Gamma_h)$ . First, the Gaussian comparison inequality (Lemma C.5) yields that the left hand side is bounded by  $C\bar{\Delta}^{1/3} \log^{2/3} d$  on the event  $\{|\hat{\Gamma}_h - \Gamma_h|_\infty \leq \bar{\Delta}\}$ . From Step 1.5,  $|\hat{\Gamma}_h -$

$\Gamma_h|_\infty \leq C\{n^{-1/2}D_n \log^{1/2}(dn) + n^{-1}D_n^2(\log n) \log^3(dn)\}$  with probability at least  $1 - Cn^{-1}$ , so that

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

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

$$(C.12) \quad \sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}A_n \in R) - \gamma_A(R) \right| \leq C\varpi_n,$$

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

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

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

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

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

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

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

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

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

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

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

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

because of the assumption (C.6). Likewise, we have  $\mathbb{P}(\sqrt{n}A_n \in R) \geq \gamma_A(R) - C\varpi_n$ . Therefore, we obtain the conclusion (C.12).

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

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

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

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

By Fubini, we conclude that

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

where  $Y_A \sim N(0, r^2 \Gamma_g) = \gamma_A$  is independent of  $Y_B$ . Since  $\alpha_n p_n = n/|I_{n,r}| \lesssim n^{-r+1} \leq n^{-1}$  and  $|\Gamma_h|_\infty \lesssim D_n$ , using the Gaussian comparison inequality (Lemma C.5), we have

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

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

$$(C.13) \quad \sup_{R \in \mathcal{R}} |\mathbb{P}(\sqrt{n}W_n \in R) - \mathbb{P}(Y \in R)| \leq C\varpi_n.$$

Step 5: Gaussian approximation to  $U'_{n,N}$ . We shall verify that the inequality (C.13) holds with  $\sqrt{n}W_n$  replaced by  $\sqrt{n}U'_{n,N}$ . Since  $Y$  is centered Gaussian and  $\max_{1 \leq j \leq d} \text{Var}(Y_j) \lesssim D_n(1 + \alpha_n)$ , 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 [19]), we have

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

Combining this estimate with (C.13), we have

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

Next, since  $\hat{N} = \sum_{\iota \in I_{n,r}} Z_\iota$  and  $Z_\iota, \iota \in I_{n,r}$  are i.i.d.  $\text{Ber}(p_n)$  with  $p_n = N/|I_{n,r}|$ , by Bernstein's inequality (cf. Lemma 2.2.9 in [42]), we have

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

for every  $t > 0$ . Choosing  $t = \log n$  and choosing  $c_1$  sufficiently small in (C.6) such that  $\sqrt{(\log n)/N} \leq 1/4$ , we have

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

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

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

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

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

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

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

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

Likewise, we have  $\mathbb{P}(\sqrt{N}U'_{n,N} \in R) \geq \mathbb{P}(\alpha_n^{-1/2}Y \in R) - C\varpi_n$ . 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  $U'_{n,N} = U_n + N^{-1} \sum_{j=1}^N \{h(X_{l_j}^*) - U_n\} =: A_n + B_n$ . Since  $X_{l_1}^*, \dots, X_{l_N}^*$  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  $\hat{\Gamma}_h - U_n U_n^T$  conditionally on  $X_1^n$ , where  $\hat{\Gamma}_h = |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h(X_\iota)h(X_\iota)^T$ . Let  $\hat{Y}$  be a random vector in  $\mathbb{R}^d$  such that  $\hat{Y} \mid X_1^n \sim N(0, \hat{\Gamma}_h - U_n U_n^T)$ . We first verify that

$$\rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \hat{Y}) \leq C\varpi_n$$

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

$$(C.14) \quad \hat{L}_n := \max_{1 \leq j \leq d} \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} |h_j(X_\iota) - U_{n,j}|^3.$$

By Jensen's inequality,  $\hat{L}_n \leq 8Z_1$ , where  $Z_1$  is defined in Step 1.2 for the Bernoulli sampling case. By Step 1.2, we have  $\mathbb{P}(\hat{L}_n \geq CD_n) \leq 3n^{-1}$  under the assumption (C.6). So we can take  $\bar{L}_n = CD_n$  and  $\phi_n = C_2(N^{-1}\bar{L}_n^2 \log^4 d)^{-1/6} \geq 1$  by choosing  $c_1$  in (C.6) sufficiently small. For  $\phi \geq 1$ , define

$$(C.15) \quad \begin{aligned} \widehat{M}_{n,X}(\phi) &= \frac{1}{|I_{n,r}|} \sum_{\iota \in I_{n,r}} \left[ \max_{1 \leq j \leq d} |h_j(X_\iota) - U_{n,j}|^3 \mathbf{1} \left( \max_{1 \leq j \leq d} |h_j(X_\iota) - U_{n,j}| > \frac{\sqrt{N}}{4\phi \log d} \right) \right], \\ \widehat{M}_{n,Y}(\phi) &= \mathbb{E}_{|X_1^n} \left[ \max_{1 \leq j \leq d} |\hat{Y}_j|^3 \mathbf{1} \left( \max_{1 \leq j \leq d} |\hat{Y}_j| > \frac{\sqrt{N}}{4\phi \log d} \right) \right], \end{aligned}$$

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

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

and hence

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

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  $\|\max_{1 \leq j \leq d} |\widehat{Y}_j|\|_{\psi_1} \lesssim \|\max_{1 \leq j \leq d} |\widehat{Y}_j|\|_{\psi_2} \lesssim \sqrt{V_n \log d}$  conditionally on  $X_1^n$  where  $V_n = \max_{1 \leq j \leq d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_j^2(X_\iota)$ , from which we have

$$(C.16) \quad \mathbb{P}_{|X_1^n} \left( \max_{1 \leq j \leq d} |\widehat{Y}_j| \geq t \right) \leq 2 \exp \left( -\frac{t}{C\sqrt{V_n \log d}} \right).$$

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

$$\mathbb{P} \left( mV_n \geq 2\mathbb{E}[\check{V}_n] + C\|\mathbf{M}_1^2\|_{\psi_{1/2}} t^2 \right) \leq 3e^{-t}$$

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

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

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

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

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

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

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

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

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

Step 1.5 also needs a modification. Note that  $|\widehat{\Gamma}_h - U_n U_n^T - \Gamma_h|_\infty \leq |\widehat{\Gamma}_h - \Gamma_h|_\infty + |U_n|_\infty^2$ . In the Bernoulli sampling case, we have shown in Step 1.5 that  $|\widehat{\Gamma}_h - \Gamma_h|_\infty \leq \underline{\sigma}^2/4$  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  $|U_n|_\infty^2 \leq \underline{\sigma}^2/4$  with probability at least  $1 - Cn^{-1}$ . By Lemma C.3,

$$\mathbb{P} \left( m|U_n|_\infty \geq 2\mathbb{E}[\check{Z}_3] + t \right) \leq e^{-t^2/(3mD_n)} + 3 \exp\{-t/(C\|\mathbf{M}_1\|_{\psi_1})\},$$

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

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

Hence,

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

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

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

Choosing  $c_1$  in (C.6) small enough, we conclude that  $|U_n|_\infty^2 \leq \underline{\sigma}^2/4$  and hence  $\min_{1 \leq j \leq d} \{\widehat{\Gamma}_{h,jj} - U_{n,j}^2\} \geq \underline{\sigma}^2/2$  with probability at least  $1 - Cn^{-1}$ .

Therefore, the overall bound in Step 1 for the sampling with replacement case is given by

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

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

Step 2 in the Bernoulli sampling case goes through under the assumption (C.6). 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.  $\square$

PROOF OF COROLLARY 3.2. In view of Theorem 3.1, the corollary follows from the Gaussian comparison inequality (Lemma C.5) and the fact that  $|\Gamma_g|_\infty \leq |\Gamma_h|_\infty \leq CD_n$ .  $\square$

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 (C.17)

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

for some sufficiently small constant  $c_2$  depending only on  $\sigma$  and  $r$ , since otherwise the conclusion of the theorem is trivial by taking  $C$  in (3.2) sufficiently large.

Bernoulli sampling case. We first verify that

$$(C.18) \quad \rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \hat{Y}) \leq C \left\{ \left( \frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \frac{p_n^{1/2}}{D_n} n^{3r/2} e^{-n^{-1/10}/C} \right\}$$

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  $\mathbb{P}(\hat{L}_n \geq Cp_n^{-1/2} D_n) \leq 3n^{-1}$  under the assumption that  $D_n^2 (\log n) \log^5(dn) \leq c_2 n$ , and so take  $\bar{L}_n = Cp_n^{-1/2} D_n$ . Step 1.3 goes through as it is. Step 1.4 needs a modification. From Step 1.4, we have that on the event  $\max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} |h_j(X_\iota)| \leq CD_n \log^2(dn)$ ,

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

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

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

Since  $\max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} |h_j(X_\iota)| \leq CD_n \log^2(dn)$  with probability at least  $1 - 2n^{-1}$ , we have that  $\widehat{M}_{n,Y}(\phi_n) \leq Cn^{3r/2}e^{-n^{1/10}/C}$  with probability at least  $1 - 2n^{-1}$ .

Step 1.5 holds under the present assumption. Hence, the inequality (C.18) 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 (C.17), so that

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

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

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

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

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{N}B_n \in R) - \gamma_B(R) \right| \leq C\check{\omega}_n.$$

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

$$(C.19) \quad \mathbb{E}[|A_n|_\infty] \leq CD_n \sum_{\ell=k}^r n^{-\ell/2} \log^{\ell/2+1} d \leq CD_n n^{-k/2} \log^{k/2+1} d.$$

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

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

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

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

Finally, since  $p_n \lesssim N/n^r$ , the Gaussian comparison inequality (Lemma C.5) yields that

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

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,

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

Likewise, we have

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

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  $B_n = N^{-1} \sum_{j=1}^N \{h(X_{t_j}^*) - U_n\}$ . Under the assumptions that  $D_n^2(\log n) \log^5(dn) \leq c_2 n$  and  $D_n^2 \log^7(dn) \leq c_2 N$ , all the sub-steps of Step 1 in the proof of Theorem 3.1 carry over to the degenerate case, i.e., we have that

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

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

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



with probability at least  $1 - Cn^{-1}$ . Now, using the estimate (C.19), 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

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

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

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

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

**PROOF OF THEOREM 3.4.** We first prove Part (i). We shall follow the notation used in the proof of Theorem 3.1. In addition to assuming (C.6), we may also assume that

$$(C.20) \quad D_n^2 n^{2r/q} \log^3(dn) \leq c_1(n \wedge N)^{1-2/q},$$

for some sufficiently small constant  $c_1 \in (0, 1)$  depending only on  $\underline{\sigma}, r$ , and  $q$ . In this proof, the constant  $c_1$  in (C.6) may also depend on  $\underline{\sigma}, r$ , and  $q$ . Let

$$\varphi_n := \left( \frac{D_n^2 n^{2r/q} \log^3(dn)}{(n \wedge N)^{1-2/q}} \right)^{1/3} \leq 1.$$

Without loss of generality, we may assume  $q > 2(r+1)$  since otherwise  $\varphi_n > 1$  and the bound (3.3) holds trivially.

**Bernoulli sampling case.** Recall that  $\hat{Y} \mid X_1^n \sim N(0, \hat{\Gamma}_h)$ , where  $\hat{\Gamma}_h = |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h(X_\iota)h(X_\iota)^T$ . We first verify that

$$\rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \hat{Y}) \leq C(\varpi_n + \varphi_n)$$

with probability at least  $1 - C(n \wedge N)^{-1}$ . Step 1.1 in the proof of Theorem 3.1 goes through, namely that there exists a constant  $C_2$  depending only on

$\underline{\sigma}$  and  $q$  such that the bound (C.9) holds for

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

on the event  $\mathcal{E}_n := \{\widehat{M}_n(\phi_n) \leq \bar{M}_n\} \cap \{\widehat{L}_n \leq \bar{L}_n\} \cap \{\min_{1 \leq j \leq d} \widehat{\Gamma}_{h,jj} \geq \underline{\sigma}^2/2\}$ , where  $\widehat{L}_n$  and  $\widehat{M}_n(\phi)$  are defined in (C.7) and (C.8), respectively. Step 1.2–Step 1.5 need modifications due to the polynomial moment condition (C2'). Observe that  $(\mathbb{E}[M_1^q])^{3/q} \leq m^{3/q} D_n^3$ , and by Lemma 9 in [11], we have

$$\mathbb{E}[\check{Z}_1] \lesssim \max_{1 \leq j \leq d} \sum_{i=1}^m \mathbb{E} \left[ |h_j(X_{(i-1)r+1}^{ir})|^3 \right] + \mathbb{E}[M_1^3] \log d \lesssim m D_n + m^{3/q} D_n^3 \log d.$$

Applying Lemma E.4 in [8] and recalling that  $\widehat{L}_n \leq C p_n^{-1/2} Z_1$ , we have

$$\mathbb{P} \left( \widehat{L}_n \geq C p_n^{-1/2} D_n \{1 + D_n^2 n^{-1+6/q} \log d\} \right) \leq n^{-1}.$$

Choose

$$\bar{L}_n = C p_n^{-1/2} D_n \left\{ 1 + \frac{D_n^2 n^{3r/q} \log d}{(n \wedge N)^{1-3/q}} \right\}.$$

Then  $\mathbb{P}(\widehat{L}_n \geq \bar{L}_n) \leq n^{-1}$  since  $r \geq 2$ . Now for our choice of  $\bar{L}_n$  and  $\phi_n$ , we observe that

$$\begin{aligned} \frac{\sqrt{N}}{4 D_n \phi_n \log d} &= \frac{C^{1/3}}{4 C_2} \left( \frac{N}{D_n^2 \log d} \right)^{1/3} \left( 1 + \frac{D_n^2 n^{3r/q} \log d}{(n \wedge N)^{1-3/q}} \right)^{1/3} \\ &\geq \frac{C^{1/3}}{4 C_2} \frac{N^{1/3} n^{r/q}}{(n \wedge N)^{1/3-1/q}} \geq \frac{C^{1/3}}{4 C_2} n^{r/q} N^{1/q}. \end{aligned}$$

Then Markov's inequality and Condition (C2') yield that

$$\mathbb{P} \left( \max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} |h_j(X_\iota)| > \frac{\sqrt{N}}{4 \phi_n \log d} \right) \leq \frac{n^r (4 D_n \phi_n \log d)^q}{N^{q/2}} \leq \frac{(4 C_2)^q}{C^{q/3} N}.$$

In addition, similar calculations under the assumptions (C.6) and (C.20) show that  $\phi_n^{-1} \leq 1$  for a sufficiently small  $c_1$ . Indeed, since  $\varpi_n \leq c_1^{1/6} \leq 1$

and  $\varphi_n \leq c_1^{1/3} \leq 1$  under (C.20), we have

$$\begin{aligned} \phi_n^{-1} &= \frac{C^{1/3}}{C_2} \left( \frac{D_n^2 \log^4 d}{N} \right)^{1/6} \left( 1 + \frac{D_n^2 n^{3r/q} \log d}{(n \wedge N)^{1-3/q}} \right)^{1/3} \\ &\leq \frac{C^{1/3}}{C_2} \left\{ \left( \frac{D_n^2 \log^4 d}{N} \right)^{1/6} + \left( \frac{D_n^2 n^{2r/q} \log^2 d}{(n \wedge N)^{1-2/q}} \right)^{1/2} \right\} \\ &\leq C(\varpi_n + \varphi_n^{3/2}) \leq C(\varpi_n + \varphi_n) \leq Cc_1^{1/6} \leq 1. \end{aligned}$$

Hence we have  $\mathbb{P}(\widehat{M}_{n,X}(\phi_n) = 0) \geq 1 - CN^{-1}$ . For  $\widehat{M}_{n,Y}(\phi)$ , suppose that

$$\max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} |h_j(X_\iota)| \leq D_n n^{(r+1)/q},$$

which holds with probability at least  $1 - n^{-1}$  under (C2'). Observe that

$$\phi_n^{-1} \geq \frac{C^{1/3}}{C_2} \frac{D_n n^{r/q} \log d}{|I_{n,r}|^{1/6} (n \wedge N)^{1/3-1/q}}.$$

Then in view of (C.10) and Lemma C.1 in [12], we have

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

where we have used the assumption (C.6) and  $q > 2(r+1)$ . Therefore, we conclude that  $\widehat{M}_{n,Y}(\phi_n) \leq Cn^{3r/2} e^{-n^{1/4}/C}$  with probability at least  $1 - n^{-1}$ . Now we can choose  $\overline{M}_n = Cn^{3r/2} e^{-n^{1/4}/C}$  such that  $\mathbb{P}(\widehat{M}_n(\phi_n) > \overline{M}_n) \leq C(n \wedge N)^{-1}$ . For Step 1.5, we note that

$$\max_{1 \leq i \leq m} \max_{1 \leq j, k \leq d} |h_j(X_{(i-1)r+1}^{ir}) h_k(X_{(i-1)r+1}^{ir})| \leq \max_{1 \leq i \leq m} \max_{1 \leq j \leq d} h_j^2(X_{(i-1)r+1}^{ir}),$$

which implies that

$$\mathbb{E}[\mathbf{M}_2^{q/2}] \lesssim \mathbb{E}[\max_{1 \leq i \leq m} \max_{1 \leq j \leq d} |h_j(X_{(i-1)r+1}^{ir})|^q] \leq mD_n^q;$$

i.e.,  $(\mathbb{E}[M_2^{q/2}])^{2/q} \lesssim m^{2/q} D_n^2$ . Lemma 9 in [11] yields that

$$\mathbb{E}[\check{Z}_2] \lesssim \sqrt{m D_n^2 \log d} + \sqrt{\mathbb{E}[M_2]} \log d \lesssim D_n \sqrt{m \log d} + D_n^2 m^{2/q} \log d.$$

Now Lemma C.4 yields that

$$(C.21) \quad \mathbb{P}\left(Z_2 \geq C\{n^{-1/2} D_n \log^{1/2}(dn) + n^{-1+4/q} D_n^2 \log d\}\right) \leq C n^{-1}.$$

Then under the assumptions (C.6) and (C.20), we conclude that  $|\hat{\Gamma}_h - \Gamma_h|_\infty \leq \underline{\sigma}^2/2$  and hence  $\min_{1 \leq j \leq d} \hat{\Gamma}_{h,jj} \geq \underline{\sigma}^2/2$  with probability at least  $1 - C n^{-1}$ . Step 1.6 is similar to that step in Theorem 3.1, namely we have with probability at least  $1 - C(n \wedge N)^{-1}$  that

$$\begin{aligned} \rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N} B_n, \hat{Y}) &\leq C \left\{ \left( \frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \left( \frac{D_n^6 n^{6r/q} \log^9 d}{N(n \wedge N)^{2-6/q}} \right)^{1/6} + \frac{p_n^{1/2}}{D_n} n^{3r/2} e^{-n^{1/4}/C} \right\} \\ &\leq C \left\{ \varpi_n + \varphi_n^{3/2} + \frac{p_n^{1/2}}{D_n} n^{3r/2} e^{-n^{1/4}/C} \right\} \leq C(\varpi_n + \varphi_n). \end{aligned}$$

Step 2 is also similar to that step in Theorem 3.1. Indeed, by the Gaussian comparison inequality (Lemma C.5) and (C.21), we have

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|X_1^n}(\hat{Y} \in R) - \gamma_B(R) \right| \leq C \left\{ \left( \frac{D_n^2 \log^5(dn)}{n} \right)^{1/6} + \left( \frac{D_n^2 \log^3 d}{n^{1-4/q}} \right)^{1/3} \right\} \lesssim \varpi_n + \varphi_n$$

with probability at least  $1 - C n^{-1}$ . Next, Proposition 2.1 in [12] yields that

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

under our assumptions. Following the argument in Step 3 of Theorem 3.1, we have

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

Likewise, we have the reverse inequality and therefore  $\sup_{R \in \mathcal{R}} |\mathbb{P}(\sqrt{n} A_n \in R) - \gamma_A(R)| \leq C(\varpi_n + \varphi_n)$ . Since  $(n \wedge N)^{-1} \lesssim \varpi_n$ , Steps 4 and 5 follow similar lines with  $\varpi_n$  being replaced by  $\varpi_n + \varphi_n$  verbatim. 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. In this case, recall that  $\hat{Y} \mid X_1^n \sim N(0, \hat{\Gamma}_h - U_n U_n^T)$ . We first verify that

$$\rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \hat{Y}) \leq C(\varpi_n + \varphi_n)$$

with probability at least  $1 - C(n \wedge N)^{-1}$ . Let  $\hat{L}_n$  and  $\widehat{M}_n(\phi)$  be defined in (C.14) and (C.15), respectively. By Jensen's inequality,  $\hat{L}_n \leq 8Z_1$  and by the calculations of Step 1 in the Bernoulli sampling case, we have  $\mathbb{P}(\hat{L}_n \geq \bar{L}_n) \leq n^{-1}$ , where

$$\bar{L}_n = CD_n \left\{ 1 + \frac{D_n^2 n^{3r/q} \log d}{(n \wedge N)^{1-3/q}} \right\}.$$

Let  $\phi_n = C_2(N^{-1}\bar{L}_n \log^4 d)^{-1/6}$ . Then  $\phi_n \geq 1$  by choosing  $c_1$  in (C.6) and (C.20) sufficiently small. By Markov's and Jensen's inequalities together with Condition (C2'), we have

$$\begin{aligned} \mathbb{P} \left( \max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} |h_j(X_\iota) - U_{n,j}| > \frac{\sqrt{N}}{4\phi_n \log d} \right) &\leq \mathbb{E} \left[ \max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} |h_j(X_\iota) - U_{n,j}|^q \right] \frac{(4\phi_n \log d)^q}{N^{q/2}} \\ &\leq \frac{(4\phi_n \log d)^q}{N^{q/2}} 2^{q-1} (n^r D_n^q + \mathbb{E}[|U_n|_\infty^q]) \leq \frac{n^r (8\phi_n \log d)^q}{N^{q/2}}. \end{aligned}$$

By the calculations in Step 1.3, we have

$$\mathbb{P} \left( \max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} |h_j(X_\iota) - U_{n,j}| > \frac{\sqrt{N}}{4\phi_n \log d} \right) \leq \frac{(8C_2)^q}{C^{q/3} N}.$$

For Step 1.4, recall that  $V_n = \max_{1 \leq j \leq d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_j^2(X_\iota)$ . Then Lemma E.4 in [8] and Lemma 9 in [11] yield that

$$\mathbb{P} \left( V_n \geq C\{D_n + n^{-1+4/q} D_n^2 \log d\} \right) \leq Cn^{-1}.$$

Under (C.20) and using  $D_n \geq 1$ , we have that  $V_n \leq CD_n$  with probability at least  $1 - Cn^{-1}$ . Now in view of (C.16) and following the lines of Step 1.4 in the proof of Theorem 3.1, we conclude that  $\widehat{M}_{n,Y}(\phi_n) \leq CN^{3/2} \exp(-N^{1/6}/C)$  with probability at least  $1 - Cn^{-1}$ . Hence we have  $\mathbb{P}(\widehat{M}_n(\phi_n) > \bar{M}_n) \leq C(n \wedge N)^{-1}$ , where  $\bar{M}_n = CN^{3/2} e^{-N^{1/6}/C}$ . Since  $\mathbb{E}[M_1^q] \leq mD_n^q \leq nD_n^q$ , Lemma 8 in [11] yields that

$$\mathbb{E}[\check{Z}_3] \lesssim \sqrt{mD_n \log d} + \sqrt{\mathbb{E}[M_1^2] \log d} \lesssim \sqrt{nD_n \log d} + n^{1/q} D_n \log d.$$

Then using Lemma C.4, we have

$$\mathbb{P} \left( |U_n|_\infty \geq C \left\{ \sqrt{\frac{D_n \log(dn)}{n}} + \frac{D_n \log d}{n^{1-2/q}} \right\} \right) \leq Cn^{-1},$$

so that  $|U_n|_\infty^2 \leq \underline{\sigma}^2/4$  with probability at least  $1 - Cn^{-1}$  by choosing  $c_1$  sufficiently small in (C.6) and (C.20). As in Step 1.5 in the Bernoulli sampling case, we have shown that  $|\hat{\Gamma}_h - \Gamma_h|_\infty \leq \underline{\sigma}^2/4$  with probability at least  $1 - Cn^{-1}$ . Hence  $|\hat{\Gamma}_h - U_n U_n^T - \Gamma_h|_\infty \leq \underline{\sigma}^2/2$  and  $\min_{1 \leq j \leq d} \{\hat{\Gamma}_{h,jj} - U_{n,j}^2\} \geq \underline{\sigma}^2/2$  with probability at least  $1 - Cn^{-1}$ . Then we have with probability at least  $1 - C(n \wedge N)^{-1}$  that

$$\begin{aligned} \rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \hat{Y}) &\leq C \left\{ \left( \frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \left( \frac{D_n^6 n^{6r/q} \log^9 d}{N(n \wedge N)^{2-6/q}} \right)^{1/6} + \frac{N^{3/2} e^{-N^{1/6}/C}}{D_n} \right\} \\ &\leq C \left\{ \varpi_n + \varphi_n^{3/2} + \frac{N^{3/2} e^{-N^{1/6}/C}}{D_n} \right\} \leq C(\varpi_n + \varphi_n). \end{aligned}$$

Step 2 goes through in view of the current Step 1.5 and Step 3 is exactly the same as in the Bernoulli sampling case. Step 4 follows similarly as the Bernoulli sampling case with  $p_n = 0$  and Step 5 is not needed in the sampling with replacement case. This completes the proof of Part (i).

We next prove Part (ii). We may assume that

$$(C.22) \quad \begin{aligned} ND_n^2 \log^{k+3} d &\leq c_2 n^k, \quad D_n^2 \log^5(dn) \leq c_2 n, \\ D_n^2 \log^7 d &\leq c_2 N, \quad D_n^2 n^{2r/q} \log^3(dn) \leq c_2 (N \wedge n)^{1-2/q}, \end{aligned}$$

for some sufficiently small constant  $c_2$  depending only on  $\underline{\sigma}, r$ , and  $q$ .

Bernoulli sampling case. We first verify that

$$(C.23) \quad \rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \hat{Y}) \leq C \left\{ \left( \frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \left( \frac{D_n^6 n^{6r/q} \log^9 d}{N(n \wedge N)^{2-6/q}} \right)^{1/6} + \frac{p_n^{1/2}}{D_n} n^{3r/2} e^{-n^{-1/10}/C} \right\}$$

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 under the present assumptions  $\mathbb{P}(\hat{L}_n \geq \bar{L}_n) \leq n^{-1}$ , where

$$\bar{L}_n = Cp_n^{-1/2} D_n \left\{ 1 + \frac{D_n^2 n^{3r/q} \log d}{(n \wedge N)^{1-3/q}} \right\}.$$

Step 1.3 goes through as in Part (i), namely  $\mathbb{P}(\widehat{M}_{n,X}(\phi_n) = 0) \geq 1 - CN^{-1}$ . Step 1.4 needs a modification. From Step 1.4, we have on the event  $\max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} |h_j(X_\iota)| \leq CD_n n^{(r+1)/q}$  that,

$$\begin{aligned} \mathbb{E}_{|X_1^n} \left[ \max_{1 \leq j \leq d} |\widehat{Y}_{\iota,j}|^3 \mathbf{1} \left( \max_{1 \leq j \leq d} |\widehat{Y}_{\iota,j}| > \frac{\sqrt{|I_{n,r}|}}{4\phi_n \log d} \right) \right] &\leq Cn^{3r/2} \exp \left( -\frac{n^{1/3}}{C \log^{1/2} d} \right) \\ &\leq n^{3r/2} \exp(-n^{7/30}/C), \end{aligned}$$

under the assumption that  $D_n^2 \log^5(dn) \leq c_2 n$ . Thus  $\widehat{M}_{n,Y}(\phi_n) \leq Cn^{3r/2} e^{-n^{7/30}/C}$  with probability at least  $1 - n^{-1}$ , so that

$$\begin{aligned} \rho_{|X_1^n}^{\mathcal{R}}(\sqrt{N}B_n, \widehat{Y}) &\leq C \left\{ \left( \frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \left( \frac{D_n^6 n^{6r/q} \log^9 d}{N(n \wedge N)^{2-6/q}} \right)^{1/6} + \frac{p_n^{1/2} n^{3r/2} e^{-n^{7/30}/C}}{D_n} \right\} \\ &\leq C \left\{ \left( \frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \frac{D_n n^{r/q} \log^{3/2} d}{(n \wedge N)^{1/2-1/q}} + \frac{p_n^{1/2} n^{3r/2} e^{-n^{7/30}/C}}{D_n} \right\} \end{aligned}$$

with probability at least  $1 - C(n \wedge N)^{-1}$ . Step 2 remains exactly the same as in Part (i). Under (C.22), we have

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{N}B_n \in R) - \gamma_B(R) \right| \leq C \left\{ \left( \frac{D_n^2 \log^7 d}{N} \right)^{1/6} + \left( \frac{D_n^2 \log^5(dn)}{n} \right)^{1/6} + \left( \frac{D_n^2 n^{2r/q} \log^3 d}{(n \wedge N)^{1-2/q}} \right)^{1/3} \right\}.$$

The rest of the proof follows similar lines as in the proof of Theorem 3.3. This leads to the conclusion (3.4) in the Bernoulli sampling case.

Sampling with replacement case. The proof is similar to 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 in (C.22), the overall bound for Step 1 in the sampling with replacement case becomes

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

with probability at least  $1 - C(n \wedge N)^{-1}$ . The Gaussian comparison inequality of Step 2 in the Bernoulli sampling case remains exactly the same under degeneracy. The rest of the proof goes through as in the proof of Theorem 3.3.  $\square$

C.4. *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

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

PROOF OF THEOREM 4.1. Bernoulli sampling case. Conditionally on  $\mathcal{D}_n$ , the vector  $U_{n,B}^\sharp$  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  $\min_{1 \leq j \leq d} P^r h_j^2 \geq \underline{\sigma}^2$ . Hence, the Gaussian comparison inequality (Lemma C.5) yields that

$$(C.24) \quad \rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,B}^\sharp, Y_B) \lesssim (\widehat{\Delta}_B \log^2 d)^{1/3},$$

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

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

Observe that

$$(C.25) \quad \begin{aligned} \widehat{\Delta}_B &\leq |N/\widehat{N}| \cdot \left( \left| N^{-1} \sum_{\iota \in I_{n,r}} (Z_\iota - p_n) h(X_\iota) h(X_\iota)^T \right|_\infty + |\widehat{\Gamma}_h - \Gamma_h|_\infty \right) \\ &\quad + |N/\widehat{N} - 1| \cdot |\Gamma_h|_\infty + |U'_{n,N}|_\infty^2 \\ &=: |N/\widehat{N}| (\widehat{\Delta}_{B,1} + \widehat{\Delta}_{B,2}) + \widehat{\Delta}_{B,3} + \widehat{\Delta}_{B,4}, \end{aligned}$$

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 3.1,  $|\widehat{N}/N - 1| \leq C(N^{-1/2} \log^{1/2} n + N^{-1} \log n) \leq CN^{-1/2} \log^{1/2} n \leq Cn^{-\zeta/2}$  with probability at least  $1 - 2n^{-1}$ . Choose the smallest  $n_0$  such that  $Cn^{-\zeta/2} \leq 1/2$  for all  $n \geq n_0$ . Clearly,  $n_0$  depends only on  $\underline{\sigma}, C_1$ , and  $\zeta$ , and since for  $n < n_0$ , the conclusion of the theorem is trivial by taking the constant  $C$  in (4.1) sufficiently large (the constant  $C$  in (4.1) can be taken independent of  $\zeta$ ), we may assume in



what follows that  $n \geq n_0$ . Then,  $|\hat{N}/N - 1| \leq CN^{-1/2} \log^{1/2} n \leq 1/2$  with probability at least  $1 - 2n^{-1}$ , and hence using the inequality  $|z^{-1} - 1| \leq 2|z - 1|$  for  $|z - 1| \leq 1/2$ , we have that  $|N/\hat{N} - 1| \leq CN^{-1/2} \log^{1/2} n$  with probability at least  $1 - 2n^{-1}$ . In particular,  $|N/\hat{N}| \leq C$  with probability at least  $1 - 2n^{-1}$ . In addition, since  $|\Gamma_h|_\infty \lesssim D_n$ , we have that

$$\hat{\Delta}_{B,3} \log^2 d \leq CD_n N^{-1/2} (\log n)^{1/2} \log^2 d \leq Cn^{-\zeta/2}$$

with probability at least  $1 - 2n^{-1}$ . For  $\hat{\Delta}_{B,2}$ , (C.11) in Step 1.5 in the proof of Theorem 3.1 yields that

$$\mathbb{P} \left( \hat{\Delta}_{B,2} \geq C \{n^{-1/2} D_n \log^{1/2}(dn) + n^{-1} D_n^2 (\log n) \log^3(dn)\} \right) \leq Cn^{-1}.$$

Then under Condition (4.1), we have that

$$\hat{\Delta}_{B,2} \log^2 d \leq Cn^{-1/2} D_n \log^{5/2}(dn) \leq Cn^{-\zeta/2}$$

with probability at least  $1 - Cn^{-1}$ . Next we deal with  $\hat{\Delta}_{B,1}$ . Let

$$V_n = \max_{1 \leq j, k \leq d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_j^2(X_\iota) h_k^2(X_\iota) \quad \text{and} \quad M_1 = \max_{\iota \in I_{n,r}} \max_{1 \leq j \leq d} h_j^2(X_\iota).$$

Since  $|Z_\iota - p_n| \leq 1$  and  $\text{Var}(Z_\iota - p_n) \leq p_n = N/|I_{n,r}|$ , by Lemma E.2 in [12] (applied conditionally on  $X_1^n$ ), we have

$$\mathbb{P}_{|X_1^n} \left( N \hat{\Delta}_{B,1} \geq 2\mathbb{E}_{|X_1^n} [N \hat{\Delta}_{B,1}] + t \right) \leq \exp \left( -\frac{t^2}{3NV_n} \right) + 3 \exp \left( -\frac{t}{CM_1} \right).$$

By Lemma 8 in [11] (applied again conditionally on  $X_1^n$ ), we have

$$\mathbb{E}_{|X_1^n} [N \hat{\Delta}_{B,1}] \lesssim \sqrt{NV_n \log d} + M_1 \log d.$$

Then we have

$$(C.27) \quad \mathbb{P}_{|X_1^n} \left( N \hat{\Delta}_{B,1} \geq C \{ \sqrt{NV_n \log(dn)} + M_1 \log(dn) \} \right) \leq Cn^{-1}.$$

Next we shall find an upper bound for  $V_n$  and  $M_1$  with probability at least  $1 - Cn^{-1}$ . Since  $\|M_1\|_{\psi_{1/2}} \lesssim D_n^2 \log^2(dn)$ , we have  $M_1 \leq CD_n^2 (\log^2 n) \log^2(dn)$  with probability at least  $1 - 2n^{-1}$ . Let  $m = \lfloor n/r \rfloor$ . By Lemma E.3 in [8], we have for all  $t > 0$  that

$$\mathbb{P} \left( mV_n \geq 2\mathbb{E} \left[ \max_{1 \leq j, k \leq d} \sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir}) h_k^2(X_{(i-1)r+1}^{ir}) \right] + t \right) \leq 3 \exp \left( -\frac{t^{1/4}}{C \|M_1'\|_{\psi_1}} \right),$$

where  $M'_1 = \max_{1 \leq i \leq m} \max_{1 \leq j \leq d} |h_j(X_{(i-1)r+1}^{ir})|$ . Clearly,  $\|M'_1\|_{\psi_1} \lesssim D_n \log(dn)$ . In addition, Lemma 9 in [11] yields that

$$\begin{aligned} & \mathbb{E} \left[ \max_{1 \leq j, k \leq d} \sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir}) h_k^2(X_{(i-1)r+1}^{ir}) \right] \\ & \lesssim \max_{1 \leq j, k \leq d} \sum_{i=1}^m \mathbb{E}[h_j^2(X_{(i-1)r+1}^{ir}) h_k^2(X_{(i-1)r+1}^{ir})] + \mathbb{E} \left[ \max_{1 \leq i \leq m} \max_{1 \leq j \leq d} h_j^4(X_{(i-1)r+1}^{ir}) \right] \log d \\ & \lesssim nD_n^2 + D_n^4 \log^5(dn). \end{aligned}$$

Hence we have  $\mathbb{P}(V_n \geq CD_n^2 \{1 + n^{-1} D_n^2 (\log^3 n) \log^5(dn)\}) \leq 3n^{-1}$ , so that  $V_n \leq CD_n^2$  with probability at least  $1 - 3n^{-1}$  under Condition (4.1). Then it follows from (C.27) that

$$\mathbb{P}_{|X_1^n} \left( N \hat{\Delta}_{B,1} \geq C \{ \sqrt{ND_n^2 \log(dn)} + D_n^2 (\log^2 n) \log^3(dn) \} \right) \leq Cn^{-1}$$

with probability at least  $1 - 5n^{-1}$ . Then Fubini yields that

$$\mathbb{P} \left( \hat{\Delta}_{B,1} \geq C \{ N^{-1/2} D_n \log^{1/2}(dn) + N^{-1} D_n^2 (\log^2 n) \log^3(dn) \} \right) \leq Cn^{-1}.$$

Under Condition (4.1), we have that

$$\hat{\Delta}_{B,1} \log^2 d \leq C \{ N^{-1} D_n^2 (\log^2 n) \log^5(dn) \}^{1/2} \leq Cn^{-\zeta/2}$$

with probability at least  $1 - Cn^{-1}$ . Finally, for  $\hat{\Delta}_{B,4}$ , observe that

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

where  $A_n = U_n = |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h(X_\iota)$  and  $B_n = |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} p_n^{-1} (Z_\iota - p_n) h(X_\iota) = N^{-1} \sum_{\iota \in I_{n,r}} (Z_\iota - p_n) h(X_\iota)$ . Note that bounding  $|B_n|_\infty$  is similar to bounding  $\hat{\Delta}_{B,1}$ . Here, for completeness and later usage of similar argument, we give the proof. Let  $\tilde{V}_n = \max_{1 \leq k \leq d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_k^2(X_\iota)$  and  $\tilde{M}_1 = \max_{\iota \in I_{n,r}} \max_{1 \leq k \leq d} |h_k(X_\iota)|$ . By Lemma E.2 in [12] (applied conditionally on  $X_1^n$ ), we have

(C.28)

$$\mathbb{P}_{|X_1^n} \left( |NB_n|_\infty \geq 2\mathbb{E}_{|X_1^n}[|NB_n|_\infty] + t \right) \leq \exp \left( -\frac{t^2}{3N\tilde{V}_n} \right) + 3 \exp \left( -\frac{t}{C\tilde{M}_1} \right).$$

By Lemma 8 in [11] (applied again conditionally on  $X_1^n$ ), we have  $\mathbb{E}_{|X_1^n}[|NB_n|_\infty] \lesssim \sqrt{N\tilde{V}_n \log d} + \tilde{M}_1 \log d$ , which together with (C.28) implies that

$$(C.29) \quad \mathbb{P}_{|X_1^n} \left( |NB_n|_\infty \geq C \{ \sqrt{N\tilde{V}_n \log(dn)} + \tilde{M}_1 \log(dn) \} \right) \leq Cn^{-1}.$$

Since  $\|\tilde{\mathbf{M}}_1\|_{\psi_1} \lesssim D_n \log(dn)$  so that  $\tilde{\mathbf{M}}_1 \lesssim D_n(\log n) \log(dn)$  with probability at least  $1 - 2n^{-1}$ . By Lemma E.3 in [8], we have for all  $t > 0$  that

$$\mathbb{P} \left( m\tilde{\mathbf{V}}_n \geq 2\mathbb{E} \left[ \max_{1 \leq j \leq d} \sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir}) \right] + t \right) \leq 3 \exp \left( -\frac{t^{1/2}}{C\|\mathbf{M}'_1\|_{\psi_1}} \right),$$

where we recall that  $\mathbf{M}'_1 = \max_{1 \leq i \leq m} \max_{1 \leq j \leq d} |h_j(X_{(i-1)r+1}^{ir})|$ . Clearly,  $\|\mathbf{M}'_1\|_{\psi_1} \lesssim D_n \log(dn)$ . In addition, Lemma 9 in [11] yields that

$$\begin{aligned} \mathbb{E} \left[ \max_{1 \leq j \leq d} \sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir}) \right] &\lesssim \max_{1 \leq j \leq d} \sum_{i=1}^m \mathbb{E}[h_j^2(X_{(i-1)r+1}^{ir})] + \mathbb{E} \left[ \max_{1 \leq i \leq m} \max_{1 \leq j \leq d} h_j^2(X_{(i-1)r+1}^{ir}) \right] \log d \\ &\lesssim nD_n + D_n^2 \log^3(dn). \end{aligned}$$

Hence we have  $\mathbb{P}(\tilde{\mathbf{V}}_n \geq CD_n\{1 + n^{-1}D_n \log^3(dn)\}) \leq 3n^{-1}$ , so that  $\tilde{\mathbf{V}}_n \leq CD_n$  with probability at least  $1 - 3n^{-1}$  under Condition (4.1). Then it follows from (C.29) that

$$\mathbb{P}_{|X_1^n} \left( |NB_n|_\infty \geq C\{\sqrt{ND_n \log(dn)} + D_n(\log n) \log^2(dn)\} \right) \leq Cn^{-1}$$

with probability at least  $1 - 5n^{-1}$ . Then Fubini yields that

$$\mathbb{P} \left( |B_n|_\infty \geq C\{(N^{-1}D_n \log(dn))^{1/2} + N^{-1}D_n(\log n) \log^2(dn)\} \right) \leq Cn^{-1},$$

which implies that  $|B_n|_\infty \leq C\{N^{-1}D_n \log(dn)\}^{1/2}$  with probability at least  $1 - Cn^{-1}$  under Condition (4.1). In addition, Lemma C.3 and Lemma 8 in [11] yield that  $|A_n|_\infty \leq C\{n^{-1}D_n \log(dn)\}^{1/2}$  with probability at least  $1 - Cn^{-1}$ . Hence we have that

$$(|A_n|_\infty^2 + |B_n|_\infty^2) \log^2 d \leq Cn^{-\zeta/2}$$

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

In conclusion, we have shown that  $\hat{\Delta}_B \log^2 d \leq Cn^{-\zeta/2}$  with probability at least  $1 - Cn^{-1}$ , and in view of (C.24), 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  $\hat{\Delta}_B \log^2 d \leq Cn^{-\zeta/2}$  with probability at least  $1 - Cn^{-1}$ , where  $\hat{\Delta}_B$  is now defined by

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

Observe that

$$(C.30) \quad \begin{aligned} \widehat{\Delta}_B &\leq \left| N^{-1} \sum_{j=1}^N h(X_{\iota_j}^*) h(X_{\iota_j}^*)^T - \widehat{\Gamma}_h \right|_\infty + |\widehat{\Gamma}_h - \Gamma_h|_\infty + |U'_{n,N}|_\infty^2 \\ &=: \widehat{\Delta}_{B,1} + \widehat{\Delta}_{B,2} + \widehat{\Delta}_{B,3}. \end{aligned}$$

We have shown that  $\widehat{\Delta}_{B,2} \log^2 d \leq Cn^{-1/2} D_n \log^{5/2}(dn) \leq Cn^{-\zeta/2}$  with probability at least  $1 - Cn^{-1}$ . Since  $X_{\iota_j}^*, j = 1, \dots, N$  are (conditionally on  $X_1^n$ ) i.i.d. draws from the empirical distribution  $|I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} \delta_{X_\iota}$ , we have  $\mathbb{E}_{|X_1^n} [h(X_{\iota_j}^*) h(X_{\iota_j}^*)^T] = \widehat{\Gamma}_h$ . In addition, by Jensen's inequality, we have

$$\begin{aligned} \max_{1 \leq k, \ell \leq d} \sum_{j=1}^N \mathbb{E}_{|X_1^n} \left[ \left( h_k(X_{\iota_j}^*) h_\ell(X_{\iota_j}^*) - \widehat{\Gamma}_{h,k\ell} \right)^2 \right] &\leq \max_{1 \leq k, \ell \leq d} \sum_{j=1}^N \mathbb{E}_{|X_1^n} \left[ h_k^2(X_{\iota_j}^*) h_\ell^2(X_{\iota_j}^*) \right] \\ &= N \max_{1 \leq k, \ell \leq d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_k^2(X_\iota) h_\ell^2(X_\iota) = NV_n, \end{aligned}$$

$$\max_{1 \leq j \leq N} \max_{1 \leq k, \ell \leq d} |h_k(X_{\iota_j}^*) h_\ell(X_{\iota_j}^*)| \leq \max_{\iota \in I_{n,r}} \max_{1 \leq k, \ell \leq d} |h_k(X_\iota) h_\ell(X_\iota)| \leq \max_{\iota \in I_{n,r}} \max_{1 \leq k \leq d} h_k^2(X_\iota) = M_1,$$

and  $\max_{1 \leq k, \ell \leq d} |\widehat{\Gamma}_{h,k\ell}| \leq \max_{\iota \in I_{n,r}} \max_{1 \leq k, \ell \leq d} |h_k(X_\iota) h_\ell(X_\iota)| \leq M_1$ . By Lemma E.2 in [12] (applied conditionally on  $X_1^n$ ), we deduce that (C.26) continues to hold for every  $t > 0$  for the sampling with replacement case. Then following the argument in the previous proof, we have

$$\widehat{\Delta}_{B,1} \log^2 d \leq C \{N^{-1} D_n^2 (\log^2 n) \log^5(dn)\}^{1/2} \leq Cn^{-\zeta/2}$$

with probability at least  $1 - Cn^{-1}$ . Finally,  $\widehat{\Delta}_{B,3} \leq 2(|A_n|_\infty^2 + |B_n|_\infty^2)$  where  $A_n = U_n$  and  $B_n = N^{-1} \sum_{j=1}^N \{h(X_{\iota_j}^*) - U_n\}$ . We have shown that  $|A_n|_\infty \leq Cn^{-1} D_n \log(dn)$  with probability at least  $1 - Cn^{-1}$ . Next, since  $h(X_{\iota_j}^*), j = 1, \dots, N$  are i.i.d. with mean  $U_n$  conditionally on  $X_1^n$ , we have

$$\begin{aligned} \max_{1 \leq k \leq d} \sum_{j=1}^N \mathbb{E}_{|X_1^n} \left[ \left( h_k(X_{\iota_j}^*) - U_{n,k} \right)^2 \right] &\leq \max_{1 \leq k \leq d} \sum_{j=1}^N \mathbb{E}_{|X_1^n} \left[ h_k^2(X_{\iota_j}^*) \right] \\ &= N \max_{1 \leq k \leq d} |I_{n,r}|^{-1} \sum_{\iota \in I_{n,r}} h_k^2(X_\iota) = N\widetilde{V}_n. \end{aligned}$$

In addition,  $\max_{1 \leq j \leq N} \max_{1 \leq k \leq d} |h_k(X_{\iota_j}^*)| \leq \max_{\iota \in I_{n,r}} \max_{1 \leq k \leq d} |h_k(X_\iota)| = \widetilde{M}_1$  and  $\max_{1 \leq k \leq d} |U_{n,k}| \leq \widetilde{M}_1$ . By Lemma E.2 in [12] (applied conditionally on  $X_1^n$ ), we deduce that (C.28) continues to hold for every  $t > 0$  for the sampling with replacement case. Then following the argument in

the previous proof, we have  $|B_n|_\infty \leq C\{N^{-1}D_n \log(dn)\}^{1/2}$  with probability at least  $1 - Cn^{-1}$  under Condition (4.1). Now we conclude that  $(|A_n|_\infty^2 + |B_n|_\infty^2) \log^2 d \leq Cn^{-\zeta/2}$  and hence  $\hat{\Delta}_B \log^2 d \leq Cn^{-\zeta/2}$  with probability at least  $1 - Cn^{-1}$ .  $\square$

PROOF OF THEOREM 4.2. The proof is divided into three steps.

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

$$\rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,B}^\#, Y_B) \leq Cn^{-\zeta_1/6}$$

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

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

$$\rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,A}^\#, Y_A) \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}$$

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

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

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

On the other hand,  $Y_A \sim N(0, r^2 \Gamma_g)$  and  $\min_{1 \leq j \leq d} P g_j^2 \geq \underline{\sigma}^2$ . Hence, the Gaussian comparison inequality (Lemma C.5) yields that

$$\rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,A}^\#, Y_A) \lesssim (\hat{\Delta}_A \log^2 d)^{1/3},$$

where

$$\hat{\Delta}_A = \max_{1 \leq j, \ell \leq d} \left| n_1^{-1} \sum_{i_1=1}^{n_1} \{\hat{g}_j^{(i_1)}(X_{i_1}) - \check{g}_j\} \{\hat{g}_\ell^{(i_1)}(X_{i_1}) - \check{g}_\ell\} - P g_j g_\ell \right|.$$

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

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

so that by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\widehat{\Delta}_A &\leq \underbrace{\max_{1 \leq j \leq d} n_1^{-1} \sum_{i_1=1}^{n_1} \{\widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1})\}^2}_{=\widehat{\Delta}_{A,1}} + 2\widehat{\Delta}_{A,1}^{1/2} \sqrt{\max_{1 \leq j \leq d} n_1^{-1} \sum_{i_1=1}^{n_1} g_j^2(X_{i_1})} \\
&\quad + \max_{1 \leq j, \ell \leq d} \left| n_1^{-1} \sum_{i_1=1}^{n_1} \{g_j(X_{i_1}) g_\ell(X_{i_1}) - P g_j g_\ell\} \right| + \max_{1 \leq j \leq d} |\check{g}_j|^2.
\end{aligned}$$

For the notational convenience, define

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

Then, since  $n_1^{-1} \sum_{i_1=1}^{n_1} g_j^2(X_{i_1}) \leq P g_j^2 + |n_1^{-1} \sum_{i_1=1}^{n_1} \{g_j^2(X_{i_1}) - P g_j^2\}| \leq \bar{\sigma}_g^2 + \widehat{\Delta}_{A,2}$ , and  $\check{g}_j = n_1^{-1} \sum_{i_1=1}^{n_1} \{\widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1})\} + n_1^{-1} \sum_{i_1=1}^{n_1} g_j(X_{i_1})$ , so that

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

we have

$$\text{(C.31)} \quad \widehat{\Delta}_A \lesssim \bar{\sigma}_g \widehat{\Delta}_{A,1}^{1/2} + \widehat{\Delta}_{A,1} + \widehat{\Delta}_{A,2} + \widehat{\Delta}_{A,3}^2,$$

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

Now, by assumption,  $\bar{\sigma}_g \hat{\Delta}_{A,1}^{1/2} \log^2 d \leq Cn^{-\zeta_2/2}$  with probability at least  $1 - Cn^{-1}$ . For  $\hat{\Delta}_{A,2}$ , Lemma 8 in [11] yields that

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

By Lemma E.2 in [12], we have for every  $t > 0$

$$\mathbb{P}(\hat{\Delta}_{A,2} \geq 2\mathbb{E}[\hat{\Delta}_{A,2}] + t) \leq \exp\left(-\frac{n_1 t^2}{3D_n^2}\right) + 3 \exp\left\{-\left(\frac{n_1 t}{CD_n^2 \log^2(dn)}\right)^{1/2}\right\}.$$

Choosing  $t = \{Cn_1^{-1} D_n^2 \log n\}^{1/2} \vee \{Cn_1^{-1} D_n^2 (\log^2 n) \log^3(dn)\}$ , we have

$$\mathbb{P}\left(\hat{\Delta}_{A,2} \geq C\{(n_1^{-1} D_n^2 \log(dn))^{1/2} + n_1^{-1} D_n^2 (\log^2 n) \log^3(dn)\}\right) \leq Cn^{-1}.$$

Now, using the first part of Condition (4.3), we deduce that

$$\hat{\Delta}_{A,2} \log^2 d \leq C\{n_1^{-1} D_n^2 (\log n)^2 \log^5(dn)\}^{1/2} \leq Cn^{-\zeta_1/2}$$

with probability at least  $1 - Cn^{-1}$ . The term  $\hat{\Delta}_{A,3}^2$  can be similarly dealt with. In particular, using  $\bar{\sigma}_g^2 \leq 1 + \max_j P|g_j|^3 \lesssim D_n$ , we have

$$(C.32) \quad \mathbb{E}[\hat{\Delta}_{A,3}] \lesssim (n_1^{-1} D_n \log d)^{1/2} + n_1^{-1} D_n \log^2(dn).$$

Then using Lemma 8 in [11] and Lemma E.2 in [12], we can show that

$$\hat{\Delta}_{A,3}^2 \log^2 d \leq C\{n_1^{-1} D_n \log^3(dn) + n_1^{-2} D_n^2 \log^6(dn)\} \leq Cn^{-\zeta_1} \leq Cn^{-\zeta_1/2}$$

with probability at least  $1 - Cn^{-1}$ . Hence  $\hat{\Delta}_A \log^2 d \leq Cn^{-(\zeta_1 \wedge \zeta_2)/2}$  with probability at least  $1 - Cn^{-1}$ , 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  $\Xi, \Xi'$ , and  $\mathcal{D}_n$  are mutually independent. Suppose that

$$\rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,A}^\#, Y_A) \bigvee \rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,B}^\#, Y_B) \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6},$$

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

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

The conditional probability on the right hand side is bounded from above by  $\gamma_B([\alpha_n^{-1/2}R - \alpha_n^{-1/2}U_{n,A}^\#]) + Cn^{-(\zeta_1 \wedge \zeta_2)/6}$ , and hence

$$\begin{aligned} \mathbb{P}_{|\mathcal{D}_n}(U_n^\# \in R) &\leq \mathbb{E}_{|\mathcal{D}_n} \left[ \gamma_B([\alpha_n^{-1/2}R - \alpha_n^{-1/2}U_{n,A}^\#]) \right] + Cn^{-(\zeta_1 \wedge \zeta_2)/6} \\ &= \mathbb{P}_{|\mathcal{D}_n} \left( \check{Y}_B \in [\alpha_n^{-1/2}R - \alpha_n^{-1/2}U_{n,A}^\#] \right) + Cn^{-(\zeta_1 \wedge \zeta_2)/6} = \mathbb{P}_{|\mathcal{D}_n} \left( U_{n,A}^\# \in [R - \alpha_n^{1/2}\check{Y}_B] \right) + Cn^{-(\zeta_1 \wedge \zeta_2)/6}, \end{aligned}$$

where  $\check{Y}_B \sim N(0, \Gamma_h)$  independent of  $\mathcal{D}_n$  and  $\Xi$ . The first term on the far right hand side can be written as  $\mathbb{E}_{|\mathcal{D}_n}[\mathbb{P}_{|(\mathcal{D}_n, \check{Y}_B)}(U_{n,A}^\# \in [R - \alpha_n^{1/2}\check{Y}_B])]$ , and the inner conditional probability is bounded from above by  $\gamma_A([R - \alpha_n^{1/2}\check{Y}_B]) + Cn^{-(\zeta_1 \wedge \zeta_2)/6}$ . Hence,

$$\mathbb{P}_{|\mathcal{D}_n}(U_n^\# \in R) \leq \mathbb{E}_{|\mathcal{D}_n} \left[ \gamma_A([R - \alpha_n^{1/2}\check{Y}_B]) \right] + Cn^{-(\zeta_1 \wedge \zeta_2)/6} = \mathbb{P}(Y \in R) + Cn^{-(\zeta_1 \wedge \zeta_2)/6}.$$

Likewise, we have  $\mathbb{P}_{|\mathcal{D}_n}(U_n^\# \in R) \geq \mathbb{P}(Y \in R) - Cn^{-(\zeta_1 \wedge \zeta_2)/6}$ .

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.  $\square$

PROOF OF COROLLARY 4.3. This follows from Step 2 in the proof of Theorem 4.2.  $\square$

C.5. *Proofs of Theorem B.1 and B.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  $\underline{\sigma}, r, q$ , and  $C_1$ . Let  $C$  denote a generic constant that depends only on  $\underline{\sigma}, r, q$ , and  $C_1$ ; its value may change from place to place. Recall that  $\rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,\star}^\#, Y_\star) := \sup_{R \in \mathcal{R}} |\mathbb{P}_{|\mathcal{D}_n}(U_{n,\star}^\# \in R) - \mathbb{P}(Y_\star \in R)|$  for  $\star = A, B$ .

PROOF OF THEOREM B.1. We shall follow the notation used in the proof of Theorem 4.1.

Bernoulli sampling case. We consider the decomposition of  $\widehat{\Delta}_B$  in (C.25). First note that  $\widehat{\Delta}_{B,3} \log^2 d \leq Cn^{-\zeta/2}$  with probability at least  $1 - 2n^{-1}$  since the randomness of  $\widehat{\Delta}_{B,3}$  only comes from  $\widehat{N}$  which is independent of  $X_1^n$ . For  $\widehat{\Delta}_{B,2}$ , by (C.21), we have

$$\mathbb{P} \left( \widehat{\Delta}_{B,2} \geq C \{ n^{-1/2} D_n \log^{1/2}(dn) + n^{-1+4/q} D_n^2 \log d \} \right) \leq Cn^{-1}.$$

Since  $r \geq 2$ , under Condition (B.1), we have  $\widehat{\Delta}_{B,2} \log^2 d \leq Cn^{-\zeta/2}$  with probability at least  $1 - Cn^{-1}$ . For  $\widehat{\Delta}_{B,1}$ , note that, under (C2'), (C.27)



continues to hold and we only need to find an upper bound for  $V_n$  and  $M_1$  with probability at least  $1 - Cn^{-1}$ . Note that  $(\mathbb{E}[M_1^{q/2}])^{2/q} \leq n^{2r/q} D_n^2$ , so that by Markov's inequality we have  $\mathbb{P}(M_1 \geq D_n^2 n^{2(r+1)/q}) \leq n^{-1}$ . Let  $m = \lfloor n/r \rfloor$ . By Lemma E.4 in [8], we have for all  $t > 0$  that

$$\mathbb{P} \left( mV_n \geq 2\mathbb{E} \left[ \max_{1 \leq j, k \leq d} \sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir}) h_k^2(X_{(i-1)r+1}^{ir}) \right] + t \right) \leq \frac{C\mathbb{E}[M_1'^q]}{t^{q/4}}.$$

Now  $(\mathbb{E}[M_1'^q])^{1/q} \lesssim n^{1/q} D_n$  and Lemma 9 in [11] yields that

$$\begin{aligned} & \mathbb{E} \left[ \max_{1 \leq j, k \leq d} \sum_{i=1}^m h_j^2(X_{(i-1)r+1}^{ir}) h_k^2(X_{(i-1)r+1}^{ir}) \right] \\ & \lesssim \max_{1 \leq j, k \leq d} \sum_{i=1}^m \mathbb{E}[h_j^2(X_{(i-1)r+1}^{ir}) h_k^2(X_{(i-1)r+1}^{ir})] + \mathbb{E} \left[ \max_{1 \leq i \leq m} \max_{1 \leq j \leq d} h_j^4(X_{(i-1)r+1}^{ir}) \right] \log d \\ & \lesssim nD_n^2 + n^{4/q} D_n^4 \log d. \end{aligned}$$

Hence we have  $\mathbb{P}(V_n \geq CD_n^2 \{1 + n^{-1+4/q} D_n^2 \log d + n^{-1+8/q} D_n^2\}) \leq Cn^{-1}$ , so that  $V_n \leq CD_n^2$  with probability at least  $1 - Cn^{-1}$  under Condition (B.1) because  $r \geq 2$ . Then it follows from (C.27) and Fubini that

$$\mathbb{P} \left( \widehat{\Delta}_{B,1} \geq C \{N^{-1/2} D_n \log^{1/2}(dn) + N^{-1} D_n^2 n^{2(r+1)/q} \log(dn)\} \right) \leq Cn^{-1}, \quad (\text{C.33})$$

which in turn implies that

$$\widehat{\Delta}_{B,1} \log^2 d \leq C \{N^{-1/2} D_n \log^{5/2}(dn) + N^{-1} D_n^2 n^{2(r+1)/q} \log^3(dn)\} \leq Cn^{-\zeta/2}$$

with probability at least  $1 - Cn^{-1}$ . For  $\widehat{\Delta}_{B,4}$ , by similar calculations to those in bounding  $\widehat{\Delta}_{B,1}$  (cf. the details for bounding the term  $\widehat{\Delta}_{B,4}$  in the proof of Theorem 4.1), we can show that  $|B_n|_\infty \leq C \{(N^{-1} D_n \log(dn))^{1/2} + N^{-1} D_n n^{(r+1)/q} \log(dn)\}$  with probability at least  $1 - Cn^{-1}$ . In addition, Lemma C.4 and Lemma 8 in [11] yield that  $|A_n|_\infty \leq C \{(n^{-1} D_n \log(dn))^{1/2} + n^{-1+2/q} D_n \log d\}$  with probability at least  $1 - Cn^{-1}$ . Hence we have that

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

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

Combining the above bounds, we have  $\widehat{\Delta}_B \log^2 d \leq Cn^{-\zeta/2}$  with probability at least  $1 - Cn^{-1}$ , and in view of (C.24), this leads to the desired conclusion.

Sampling with replacement case. We consider the decomposition of  $\hat{\Delta}_B$  in (C.30). We have shown that  $\hat{\Delta}_{B,2} \log^2 d \leq C\{n^{-1/2} D_n \log^{5/2}(dn) + n^{-1+4/q} D_n^2 \log^3 d\} \leq Cn^{-\zeta/2}$  with probability at least  $1 - Cn^{-1}$ . Note that the bound (C.33) on  $\hat{\Delta}_{B,1}$  holds exactly as in the Bernoulli sampling case in view of (C.27) and the proof of sampling with replacement case in Theorem 4.1; that is, we have  $\hat{\Delta}_{B,1} \log^2 d \leq C\{N^{-1/2} D_n \log^{5/2}(dn) + N^{-1} D_n^2 n^{2(r+1)/q} \log^3(dn)\} \leq Cn^{-\zeta/2}$  with probability at least  $1 - Cn^{-1}$ . Finally,  $\hat{\Delta}_{B,3} \leq 2(|A_n|_\infty^2 + |B_n|_\infty^2)$  where  $A_n = U_n$  and  $B_n = N^{-1} \sum_{j=1}^N \{h(X_{i_j}^*) - U_n\}$ . We have shown that  $|A_n|_\infty \leq C\{(n^{-1} D_n \log(dn))^{1/2} + n^{-1+2/q} D_n \log d\}$  with probability at least  $1 - Cn^{-1}$ . Similarly,  $|B_n|_\infty \leq C\{(N^{-1} D_n \log(dn))^{1/2} + N^{-1} D_n n^{(r+1)/q} \log(dn)\}$  with probability at least  $1 - Cn^{-1}$  in view of (C.29) and the proof of sampling with replacement case in Theorem 4.1.  $\square$

**PROOF OF THEOREM B.2.** We shall follow the notation and the proof structure in the proof of Theorem 4.2. In Step 1, it follows from Theorem B.1 that  $\rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,B}^\sharp, Y_B) \leq Cn^{-\zeta_1/6}$  with probability at least  $1 - Cn^{-1}$  in view that (C3-ND) implies Condition (C3-D) and  $n_1 \leq n$ . Step 2 in Theorem 4.2 needs modifications. In particular, we shall show in this step that  $\rho_{|\mathcal{D}_n}^{\mathcal{R}}(U_{n,A}^\sharp, Y_A) \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}$  holds with probability at least  $1 - Cn^{-1}$  under (C2') and (B.2). Consider the decomposition of  $\hat{\Delta}_A$  in (C.31). For  $\hat{\Delta}_{A,2}$ , Lemma 8 in [11] yields that

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

By Lemma E.2 in [12], we have for every  $t > 0$

$$\mathbb{P}(\hat{\Delta}_{A,2} \geq 2\mathbb{E}[\hat{\Delta}_{A,2}] + t) \leq \exp\left(-\frac{n_1 t^2}{3D_n^2}\right) + \frac{CD_n^q}{n_1^{q/2-1} t^{q/2}}.$$

Choosing  $t = \{Cn_1^{-1} D_n^2 \log n\}^{1/2} \vee \{Cn_1^{-1+2/q} n^{2/q} D_n^2\}$ , we have

$$\mathbb{P}\left(\hat{\Delta}_{A,2} \geq C\{(n_1^{-1} D_n^2 \log(dn))^{1/2} + n_1^{-1+2/q} n^{2/q} D_n^2 \log d\}\right) \leq Cn^{-1}.$$

Now, using Condition (B.2), we deduce that

$$\hat{\Delta}_{A,2} \log^2 d \leq C\{(n_1^{-1} D_n^2 \log^5(dn))^{1/2} + n_1^{-1+2/q} n^{2/q} D_n^2 \log^3 d\} \leq Cn^{-\zeta_1/2}$$

with probability at least  $1 - Cn^{-1}$ . The term  $\widehat{\Delta}_{A,3}^2$  can be similarly dealt with. In particular, using  $\bar{\sigma}_g^2 \leq 1 + \max_j P|g_j|^3 \lesssim D_n$ , we have

$$(C.34) \quad \mathbb{E}[\widehat{\Delta}_{A,3}] \lesssim (n_1^{-1} D_n \log d)^{1/2} + n_1^{-1+1/q} D_n \log d.$$

Then using Lemma 8 in [11] and Lemma E.2 in [12], we can show that

$$\widehat{\Delta}_{A,3}^2 \log^2 d \leq C \{n_1^{-1} D_n \log^3(dn) + n_1^{-2+2/q} n^{2/q} D_n^2 \log^4 d\} \leq Cn^{-\zeta_1} \leq Cn^{-\zeta_1/2}$$

with probability at least  $1 - Cn^{-1}$ . Hence  $\widehat{\Delta}_A \log^2 d \leq Cn^{-(\zeta_1 \wedge \zeta_2)/2}$  with probability at least  $1 - Cn^{-1}$ , which leads to the conclusion of this step.  $\square$

C.6. *Proofs of Proposition 4.4 and 4.5.* For the notational convenience, let  $H = \max_{1 \leq j \leq d} |h_j|$ . 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  $r, \nu$ , and  $C_1$ . We will bound  $\mathbb{E}[\widehat{\Delta}_{A,1}^\nu]$ . To this end, we begin with observing that

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

and by Jensen's inequality, we have

$$\mathbb{E}[\widehat{\Delta}_{A,1}^\nu] \leq \frac{1}{n_1} \sum_{i_1 \in S_1} \mathbb{E} \left[ \max_{1 \leq j \leq d} \left| \widehat{g}_j^{(i_1)}(X_{i_1}) - g_j(X_{i_1}) \right|^{2\nu} \right].$$

For each  $i_1 \in S_1$  and  $k = 1, \dots, K$ , let

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

which is the  $U$ -statistic with kernel  $\delta_x h$  for the sample  $\{X_i : i \in S_{2,k}^{(i_1)}\}$ . Recall that the size of each block  $S_{2,k}^{(i_1)}$  is  $L$ ,  $|S_{2,k}^{(i_1)}| = L$ . 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 [42, Proposition A.1.6] conditionally on  $X_{i_1}$ , we have

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

Applying Hoeffding's averaging and Theorem 2.14.1 in [42], we have

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

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

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

From the variance formula for  $U$ -statistics (cf. [30], Theorem 3), we have

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

It remains to bound

$$(C.35) \quad \mathbb{E}_{|X_{i_1}} \left[ \max_{1 \leq k \leq K} \max_{1 \leq j \leq d} |g_j^{(i_1,k)}(X_{i_1}) - g_j(X_{i_1})|^2 \right].$$

Observe that the term (C.35) is bounded from above by

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

which is bounded from above by  $K^{1/\nu} L^{-1} (\log d) (P^{r-1} |\delta_{X_{i_1}} H|^{2\nu})^{1/\nu}$  up to a constant that depends only on  $r$  and  $\nu$ .

By Fubini and Jensen's inequality, we have

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

from which we conclude that  $\mathbb{E}[(\bar{\sigma}_g^2 \hat{\Delta}_{A,1} \log^4 d)^\nu] \lesssim n^{-\zeta\nu}$ , so that by Markov's inequality,

$$\mathbb{P} \left( \bar{\sigma}_g^2 \hat{\Delta}_{A,1} \log^4 d > n^{-\zeta+1/\nu} \right) \lesssim n^{-1}.$$

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

**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  $r, \nu$ , and  $C_1$ . Observe that

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

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

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

By Lemma 8 in [11],

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

Observe that

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

Hence, using Jensen's inequality, we have

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

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

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

Hence,

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

where we have used the fact that  $(n \wedge M)^{-1} \log(dn) \lesssim 1$ .

On the other hand, applying Hoeffding averaging and Theorem 2.14.1 in [42] conditionally on  $X_{i_1}$ , we have

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

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

Therefore, using Condition (4.8), we conclude that

$$\left( \mathbb{E}[(\bar{\sigma}_g^2 \hat{\Delta}_{A,1} \log^4 d)^\nu] \right)^{1/\nu} \lesssim \bar{\sigma}_g^2 \{M^{-1} D_n^2 \log^7(dn) + n^{-1} D_n^2 \log^7 d\} \lesssim n^{-\zeta}.$$

By Markov's inequality, we conclude that

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

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

**C.7. Proofs of Propositions B.3 and B.4.** The proofs are almost the same as those of Propositions 4.4 and 4.5. The required modifications are to take  $\nu = q/2$ , and to bound  $P^r H^q$  and  $\mathbb{E}[\max_{\iota \in I_{n,r}} H^q(X_\iota)]$  by  $D_n^q$  and  $n^r D_n^q$ , respectively. We omit the detail for brevity.  $\square$

**C.8. Proofs for Section A.2.**

**LEMMA C.6 (Variance estimation).** (i) Suppose that Conditions (C1), (C2), and (C3-ND) hold, and in addition suppose that Condition (4.3) holds for some constants  $0 < C_1 < \infty$  and  $\zeta_1, \zeta_2 \in (0, 1)$ . Then there exists a constant  $C$  depending only on  $\underline{\sigma}, r$ , and  $C_1$  such that  $\max_{1 \leq j \leq d} |\hat{\sigma}_{A,j}^2 / \sigma_{A,j}^2 - 1| \leq C n^{-(\zeta_1 \wedge \zeta_2)/2} \log^2 d$  with probability at least  $1 - C n^{-1}$ .

(ii) Suppose that Conditions (C1), (C2), and (C3-D) hold, and in addition suppose that Condition (4.1) 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  $\max_{1 \leq j \leq d} |\hat{\sigma}_{B,j}^2 / \sigma_{B,j}^2 - 1| \leq C n^{-\zeta/2} / \log^2 d$  with probability at least  $1 - C n^{-1}$ .

PROOF OF LEMMA C.6. The proofs of Theorems 4.1 and 4.2 immediately imply the lemma. Recall the notation used in the proofs of Theorems 4.1 and 4.2. Then we have  $\max_{1 \leq j \leq d} |\hat{\sigma}_{A,j}^2 - \sigma_{A,j}^2| \leq r^2 \hat{\Delta}_A$  and  $\max_{1 \leq j \leq d} |\hat{\sigma}_{B,j}^2 - \sigma_{B,j}^2| \leq \hat{\Delta}_B$ . Since  $\min_{1 \leq j \leq d} \sigma_{A,j}^2 \geq r^2 \underline{\sigma}^2$  in Case (i) and  $\min_{1 \leq j \leq d} \sigma_{B,j}^2 \geq \underline{\sigma}^2$  in Case (ii), the conclusion of the lemma follows from the bounds on  $\hat{\Delta}_B$  and  $\hat{\Delta}_A$  established in the proofs of Theorems 4.1 and 4.2, respectively.  $\square$

PROOF OF COROLLARY A.1. We only prove Case (i) since the proof for Case (ii) is analogous. As before, we will assume that  $\theta = P^r h = 0$ . 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  $n^{-(\zeta_1 \wedge \zeta_2)/6} \leq c_1$  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,  $C \geq 1/c_1$ ). We begin with noting that

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

so that by Lemma C.6, we have that  $\max_{1 \leq j \leq d} |\hat{\sigma}_j^2/\sigma_j^2 - 1| \leq Cn^{-(\zeta_1 \wedge \zeta_2)/2}/\log^2 d$  with probability at least  $1 - Cn^{-1}$ . Choosing  $c_1$  sufficiently small so that  $Cn^{-(\zeta_1 \wedge \zeta_2)/2}/\log^2 d \leq 1/2$ , and using the inequalities that  $|z - 1| \leq |z^2 - 1|$  for  $z \geq 0$  and  $|z^{-1} - 1| \leq 2|z - 1|$  for  $|z - 1| \leq 1/2$ , we have that

$$\max_{1 \leq j \leq d} |\sigma_j/\hat{\sigma}_j - 1| \leq Cn^{-(\zeta_1 \wedge \zeta_2)/2}/\log^2 d$$

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

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\Lambda^{-1/2}U'_{n,N} \in R) - \mathbb{P}(\Lambda^{-1/2}Y \in R) \right| \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}.$$

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  $\mathbb{E}[|\Lambda^{-1/2}Y|_\infty] \leq C\sqrt{\log d}$ , we have  $\mathbb{P}(|\Lambda^{-1/2}Y|_\infty > C\sqrt{\log(dn)}) \leq 2n^{-1}$ . Hence,  $\mathbb{P}(|\sqrt{n}\Lambda^{-1/2}U'_{n,N}|_\infty > C\sqrt{\log(dn)}) \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}$ . Since  $\frac{n^{-(\zeta_1 \wedge \zeta_2)/2}}{\log^2 d} \times \sqrt{\log(dn)} \leq \frac{Cn^{-(\zeta_1 \wedge \zeta_2)/6}}{\log^{3/2} d}$ , we have

$$\mathbb{P}(|\sqrt{n}(\hat{\Lambda}^{-1/2} - \Lambda^{-1/2})U'_{n,N}|_\infty > t_n) \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6},$$

where  $t_n = Cn^{-(\zeta_1 \wedge \zeta_2)/6}/\log^{3/2} d$ .



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

$$\begin{aligned} \mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}U'_{n,N} \in R) &\leq \mathbb{P}(\{-\sqrt{n}\Lambda^{-1/2}U'_{n,N} \leq -a + t_n\} \cap \{\sqrt{n}\Lambda^{-1/2}U'_{n,N} \leq b + t_n\}) \\ &\quad + \mathbb{P}(|\sqrt{n}(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2})U'_{n,N}|_\infty > t_n) \\ &\leq \mathbb{P}(\{-\Lambda^{-1/2}Y \leq -a + t_n\} \cap \{\Lambda^{-1/2}Y \leq b + t_n\}) + Cn^{-(\zeta_1 \wedge \zeta_2)/8} \\ &\leq \mathbb{P}(\Lambda^{-1/2}Y \in R) + Ct_n\sqrt{\log d} + Cn^{-(\zeta_1 \wedge \zeta_2)/6}, \end{aligned}$$

where the last inequality follows from Nazarov's inequality. Since  $t_n\sqrt{\log d} \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}/\log d \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}$ , we conclude that  $\mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}U'_{n,N} \in R) \leq \mathbb{P}(\Lambda^{-1/2}Y \in R) + Cn^{-(\zeta_1 \wedge \zeta_2)/6}$ . Likewise, we have  $\mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}U'_{n,N} \in R) \geq \mathbb{P}(\Lambda^{-1/2}Y \in R) - Cn^{-(\zeta_1 \wedge \zeta_2)/6}$ . Hence we have shown that

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}(\sqrt{n}\widehat{\Lambda}^{-1/2}U'_{n,N} \in R) - \mathbb{P}(\Lambda^{-1/2}Y \in R) \right| \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}.$$

Similarly, using Theorem 4.2, we have that

$$\mathbb{P}_{|\mathcal{D}_n}(|(\widehat{\Lambda}^{-1/2} - \Lambda^{-1/2})U_n^\#|_\infty > t_n) \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}$$

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

$$\sup_{R \in \mathcal{R}} \left| \mathbb{P}_{|\mathcal{D}_n}(\widehat{\Lambda}^{-1/2}U_n^\# \in R) - \mathbb{P}(\Lambda^{-1/2}Y \in R) \right| \leq Cn^{-(\zeta_1 \wedge \zeta_2)/6}$$

with probability at least  $1 - Cn^{-1}$ . This completes the proof.  $\square$

**D. Additional simulation results.** In Section D.1, we first examine the statistical accuracy of the bootstrap test statistic  $U_n^\#$  in terms of the size for MB-NDG-DC (Spearman's  $\rho$ ), MB-NDG-RS (Spearman's  $\rho$ ), and the bootstrap test statistic  $U_{n,B}^\#$  for MB-DG (Bergsma-Dassios'  $t^*$ ). In Section D.2, we report the simulation results of the partial bootstrap  $U_{n,A}^\#$  for Spearman's  $\rho$ .

*D.1. Empirical performance of the bootstrap tests.* Recall that, for each nominal size  $\alpha \in (0, 1)$ ,  $\widehat{R}(\alpha)$  is the empirical rejection probability of the null hypothesis, where the critical values are calibrated by our bootstrap methods. Figures 4, 5, and 6 display the plots of the empirical size graphs  $\{(\alpha, \widehat{R}(\alpha)) : \alpha \in (0, 1)\}$  for MB-NDG-DC (Spearman's  $\rho$ ), MB-NDG-RS (Spearman's  $\rho$ ), and MB-DG (Bergsma-Dassios'  $t^*$ ), respectively. Clearly,

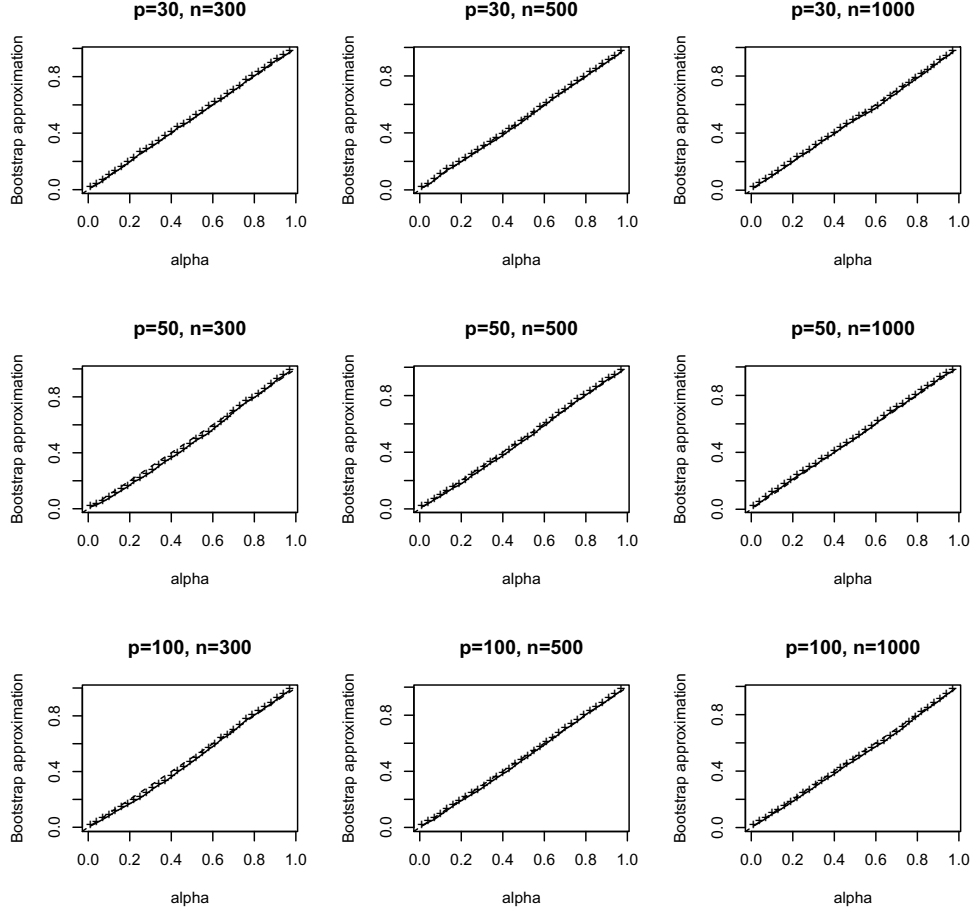


FIG 4. Bootstrap approximation  $U_n^{\sharp}$  for Spearman's  $\rho$  test statistic with the divide and conquer estimation (MB-NDG-DC). Plot of the nominal size  $\alpha$  versus the empirical rejection probability  $\hat{R}(\alpha)$ .

the bootstrap approximations becomes more accurate as  $n$  increases. Qualitatively, it is worth noting that the bootstrap approximations work quite well on the lower (left) tail, which is relevant in the testing application for small values of  $\alpha$  (say  $\alpha \leq 0.10$ ). Quantitatively, the uniform errors-in-size on  $\alpha \in [0.01, 0.10]$  of our bootstrap tests are summarized in Table 1 in the main paper.

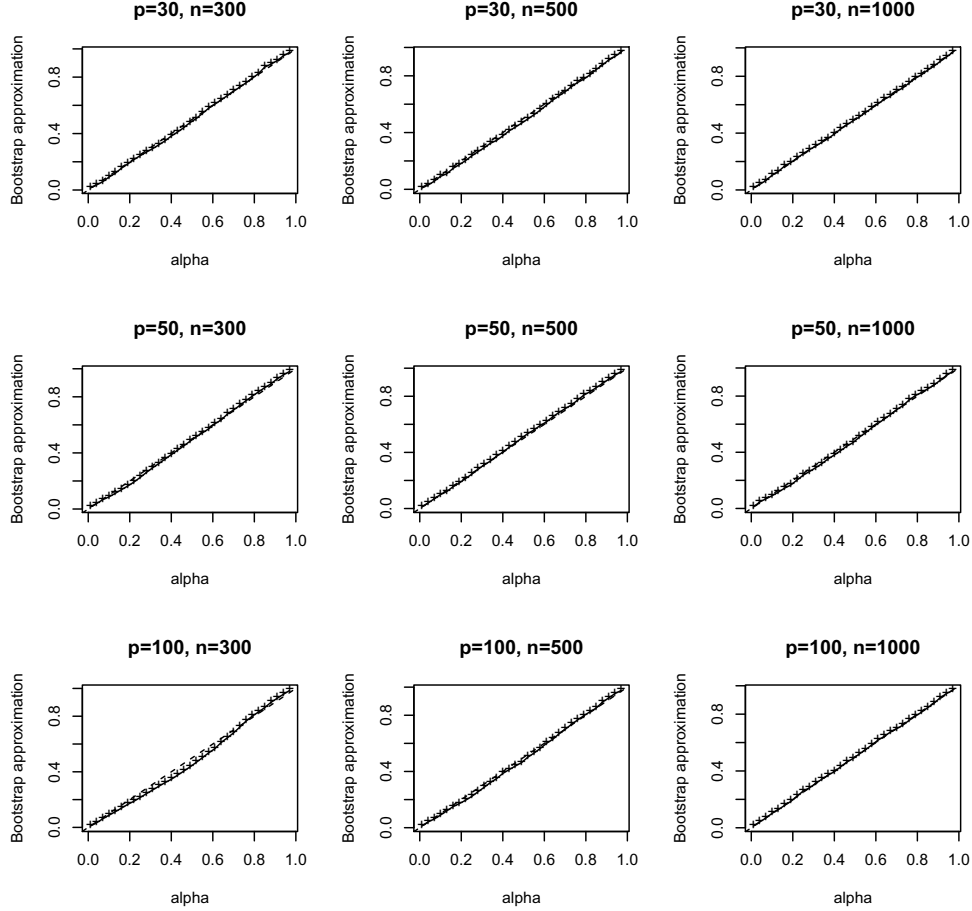


FIG 5. Bootstrap approximation  $U_n^\sharp$  for Spearman's  $\rho$  test statistic with the random sampling estimation (MB-NDG-RS). Plot of the nominal size  $\alpha$  versus the empirical rejection probability  $\hat{R}(\alpha)$ .

D.2. *Partial bootstrap*  $U_{n,A}^\sharp$ . We also provide additional results of the partial bootstrap  $U_{n,A}^\sharp$  for the non-degenerate Spearman's  $\rho$  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 rate in the error bound of the Gaussian approximation (cf. Corollary 3.2). We empirically observe

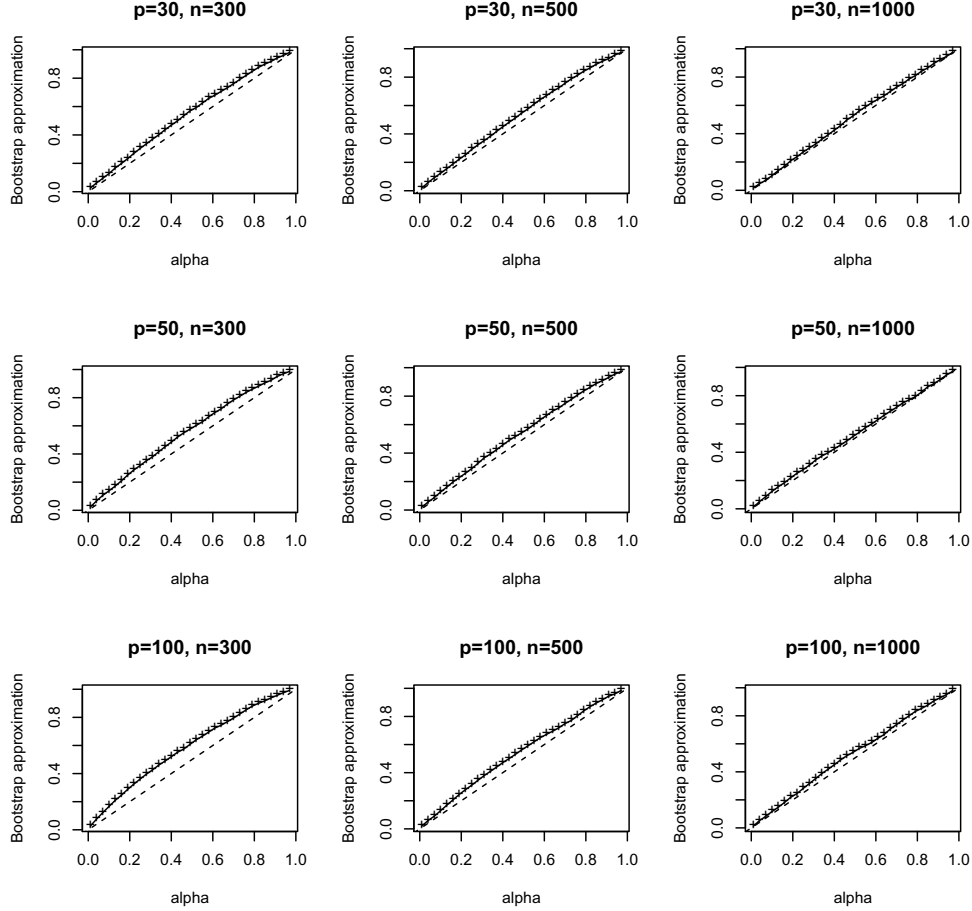


FIG 6. Bootstrap approximation  $U_{n,B}^{\#}$  for Bergsma-Dassios'  $t^*$  test statistic (MB-DG). Plot of the nominal size  $\alpha$  versus the empirical rejection probability  $\hat{R}(\alpha)$ .

that the bootstrap approximation is sensitive to small constant values in  $N = Cn^{3/2}$  and we find that  $C \geq 4$  can produce reasonably accurate bootstrap approximation quality (cf. Figure 8 and 9). Table 2 shows the uniform errors-in-size on  $\alpha \in [0.01, 0.10]$  of the partial bootstrap tests. We also show the P-P plots of the (simplified) Gaussian approximation for Spearman's  $\rho$  (i.e.,  $\sqrt{n}U'_{n,N}$  versus  $N(0, r^2\Gamma_g)$ ) in Figure 7.

Plots of the computer running time of the partial bootstraps are shown in Figure 10. Fitting a linear model with the (log-)running time for the

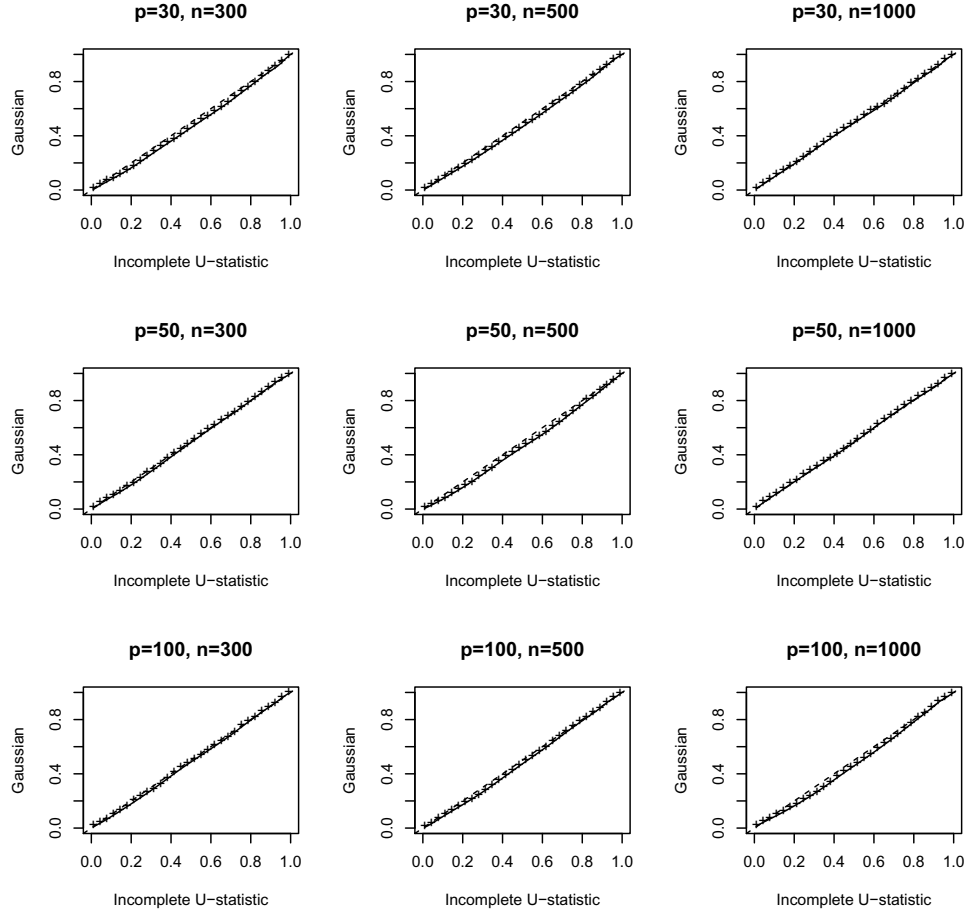


FIG 7.  $P$ - $P$  plots for the Gaussian approximation  $N(0, r^2 \Gamma_g)$  of  $\sqrt{n}U'_{n,N}$  for Spearman's  $\rho$  test statistic with the Bernoulli sampling.

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 both cases, the slope coefficients again match very well to the theoretic value 2.

**E. Random versus deterministic normalization in the Bernoulli sampling case.** In this section, we report the empirical effect of the random and deterministic normalization in the Bernoulli sample case on the

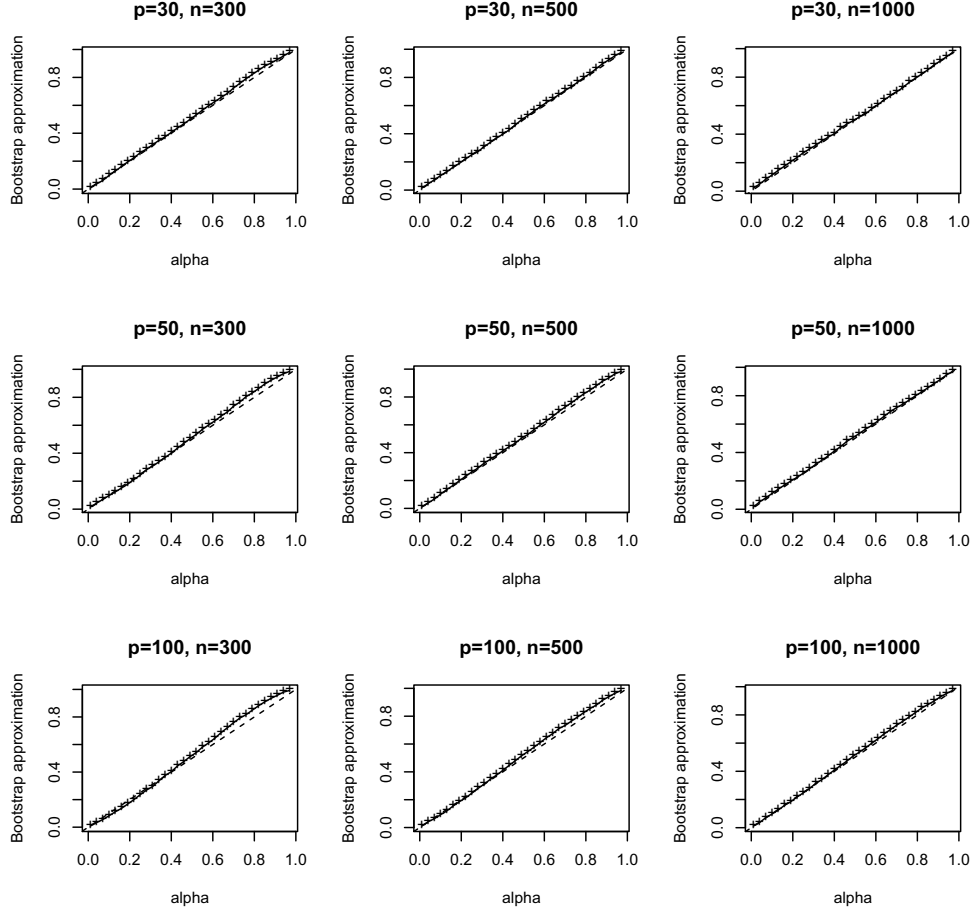


FIG 8. Bootstrap approximation  $U_{n,A}^\#$  for Spearman's  $\rho$  test statistic with the divide and conquer estimation (MB-NDG-DC). Plot of the nominal size  $\alpha$  versus the empirical rejection probability  $\hat{R}(\alpha)$ .

inference of the copula dependence structures [16, Chapter 8]. Let  $F$  be the joint distribution of  $X = (X^{(1)}, \dots, X^{(p)})$  and  $F_j$  be the marginal distribution of  $X^{(j)}$ ,  $j = 1, \dots, p$ . Then it is easy to see that

$$\theta := \mathbb{E}[\hat{\rho}_{jk}] = 12\text{Cov}(F_j(X^{(j)}), F_k(X^{(k)})) = \text{Corr}(F_j(X^{(j)}), F_k(X^{(k)})),$$

where  $\hat{\rho}_{jk} = U_n^{(3)}(h_{jk}^S)$ ,  $j, k = 1, \dots, p$ , is defined in Example 5.1. Thus Spearman's rank correlation coefficient  $\rho$  is equivalent to the copula correlation,

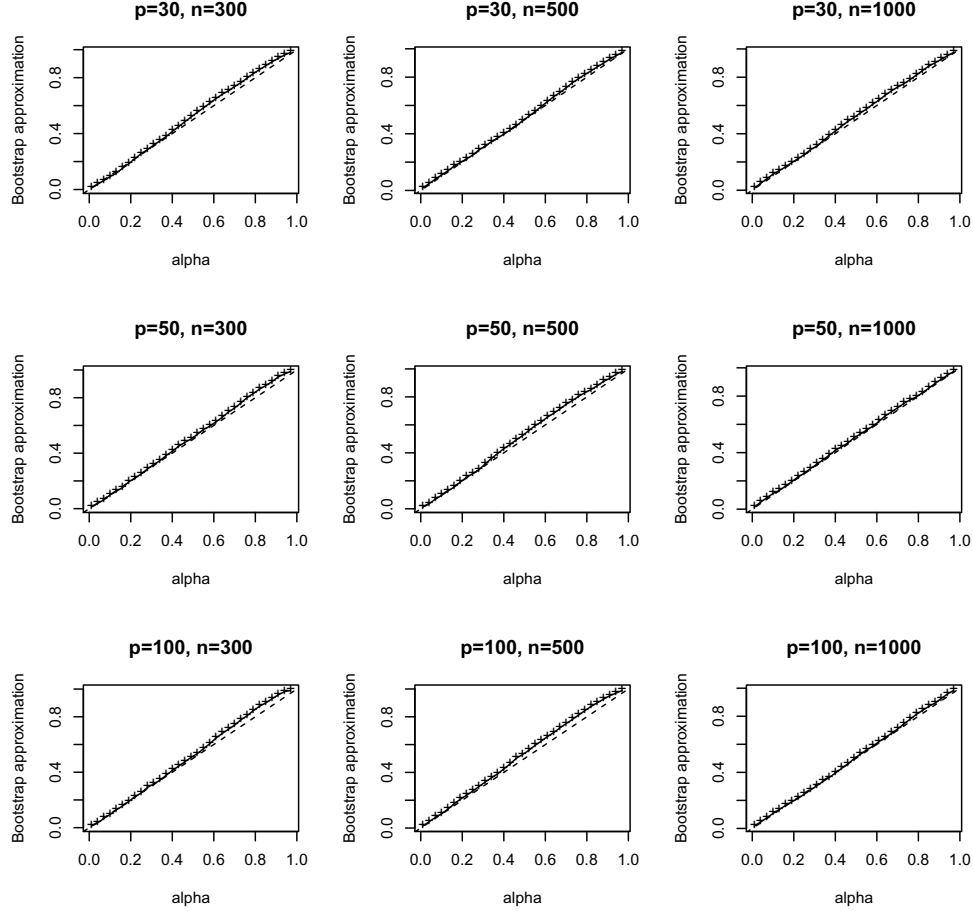


FIG 9. Bootstrap approximation  $U_{n,A}^\sharp$  for Spearman's  $\rho$  test statistic with the random sampling estimation (MB-NDG-RS). Plot of the nominal size  $\alpha$  versus the empirical rejection probability  $\hat{R}(\alpha)$ .

and a natural statistic to estimate the copula correlation is the  $U$ -statistic associated with the kernel  $h^S$ . To compare the empirical effect of normalization of the incomplete  $U$ -statistic in the Bernoulli sampling, we simulate  $n = 300$  i.i.d. copies  $X_1, \dots, X_n$  of  $X$  from the bivariate (i.e.,  $p = 2$ ) Gaussian distribution with the following parameters:

$$\begin{pmatrix} X_i^{(1)} \\ X_i^{(2)} \end{pmatrix} \stackrel{\text{i.i.d.}}{\sim} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} \right), \quad i = 1, \dots, n,$$

Setup	Spearman's $\rho$ (MB-NDG-DC)	Spearman's $\rho$ (MB-NDG-RS)
$p = 30, n = 300$	0.0085	0.0080
$p = 30, n = 500$	0.0045	0.0095
$p = 30, n = 1000$	0.0160	0.0135
$p = 50, n = 300$	0.0070	0.0070
$p = 50, n = 500$	0.0050	0.0045
$p = 50, n = 1000$	0.0120	0.0110
$p = 100, n = 300$	0.0175	0.0085
$p = 100, n = 500$	0.0080	0.0065
$p = 100, n = 1000$	0.0040	0.0070

TABLE 2  
Uniform error-in-size  $\sup_{\alpha \in [0.01, 0.10]} |\hat{R}(\alpha) - \alpha|$  of the partial bootstrap tests, where  $\alpha$  is the nominal size.

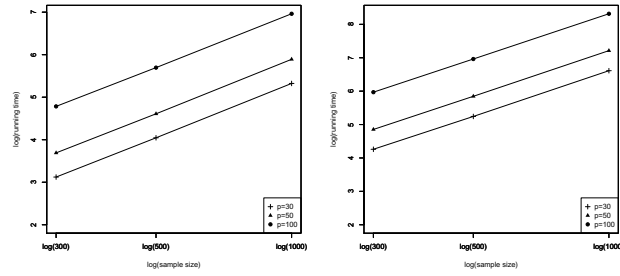


FIG 10. Computer running time of the bootstrap versus the sample size on the log-scale. Left: bootstrap  $U_{n,A}^\#$  for Spearman's  $\rho$  with the divide and conquer estimation (MB-NDG-DC). Right: bootstrap  $U_{n,A}^\#$  for Spearman's  $\rho$  with the random sampling estimation (MB-NDG-RS).

and  $a \in (-1, 1)$ . In our simulation example, we choose  $a = 0.9$  so that the effect of the centering term in the Gaussian approximations can be clearly seen. Histograms of  $\sqrt{n}(U'_{n,N} - \theta)$  for the above Gaussian copula model under the Bernoulli sampling are shown in Figure 11. Two empirical observations can be drawn from Figure 11. First, the Gaussian approximation is quite accurate for  $\sqrt{n}(U'_{n,N} - \theta)$  under both random and deterministic normalizations. Second, the variance of the approximating Gaussian distribution in the random normalization case is smaller than that in the deterministic case. Both observations are consistent with our theory; cf. Theorem 3.1, Remark 2.1, and Remark 3.3 in the main paper.



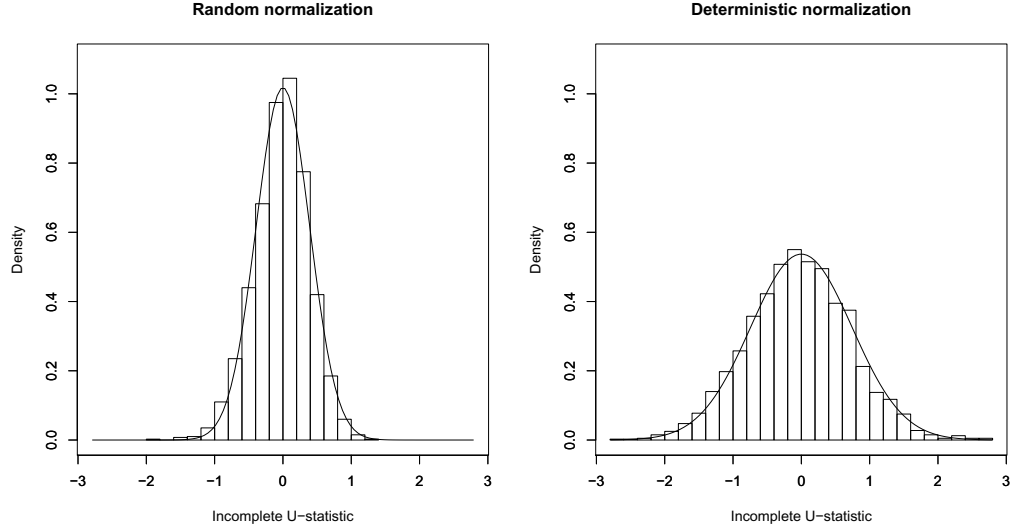


FIG 11. Histogram of  $\sqrt{n}(U'_{n,N} - \theta)$  for the copula model in the Bernoulli sampling, where  $\theta$  is the copula correlation. Data are generate from a bivariate Gaussian distribution with mean zero, unit variance, and covariance 0.9. The curve is the density of the approximating Gaussian distribution for  $\sqrt{n}(U'_{n,N} - \theta)$ .

DEPARTMENT OF STATISTICS  
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN  
725 S. WRIGHT STREET  
CHAMPAIGN, ILLINOIS 61874  
USA  
E-MAIL: [xhchen@illinois.edu](mailto:xhchen@illinois.edu)

DEPARTMENT OF STATISTICAL SCIENCE  
CORNELL UNIVERSITY  
1194 COMSTOCK HALL  
ITHACA, NEW YORK 14853  
USA  
E-MAIL: [kk976@cornell.edu](mailto:kk976@cornell.edu)