Statistica Sinica **31** (2021), 1593-1618
doi:https://doi.org/10.5705/ss.202019.0098

DESIGN BASED INCOMPLETE U-STATISTICS

Xiangshun Kong and Wei Zheng

Beijing Institute of Technology and University of Tennessee

Abstract: U-statistics are widely used in fields such as economics, machine learning, and statistics. However, while they enjoy desirable statistical properties, they have an obvious drawback in that the computation becomes impractical as the data size n increases. Specifically, the number of combinations, say m, that a U-statistic of order d has to evaluate is $O(n^d)$. Many efforts have been made to approximate the original U-statistic using a small subset of combinations since Blom (1976), who referred to such an approximation as an incomplete U-statistic. To the best of our knowledge, all existing methods require m to grow at least faster than n, albeit more slowly than n^d, in order for the corresponding incomplete U-statistic to be asymptotically efficient in terms of the mean squared error. In this paper, we introduce a new type of incomplete U-statistic that can be asymptotically efficient, even when m grows more slowly than n. In some cases, m is only required to grow faster than $\sqrt{n}$. Our theoretical and empirical results both show significant improvements in the statistical efficiency of the new incomplete U-statistic.

Key words and phrases: Asymptotically efficient, BIBD, big data, design of experiment, subsampling.

1. Introduction

The U-statistic has been a fundamental statistical estimator since the work of Hoeffding (1948), who studied its theoretical properties and established central limit theorems for non-degenerate U-statistics. Eagleson (1979) derived asymptotic distributions of some degenerate U-statistics of order two, which were then extended to all degenerate U-statistics by Lee (1979). Other extensions include a variant of U-statistics called V-statistics by von Mises (1948), U-statistics for stationary processes by Enqvist (1985), and multi-sample U-statistics by Lehmann (1951) and Sen (1974, 1977).

The theory of U-statistics admits a minimum variance unbiased estimator of an estimable parameter for a large class of probability distributions, hence its popularity in applications. However, U-statistics can also be time consuming to compute. For a U-statistic of order d, the number of combinations, say m, to be

Corresponding author: Wei Zheng, Department of Business Analytics and Statistics, University of Tennessee, Knoxville, TN 37996, USA. E-mail: wzheng9@utk.edu.

evaluated is $\binom{n}{d}$, that is $O(n^d)$, where n is the data size. Suppose $n = 10^4$ and $d = 3$. Then, listing the $\binom{10^4}{3}$ combinations requires 667 GB of memory and a computing time of approximately 100 hours on a Macbook Pro with Intel Core i7 2.9 GHz CPU. With $n = 10^5$ and $d = 4$, the required memory is roughly 16.7 EB and the computing time is projected to be 285,000 years. To provide context, Hilbert and López (2011) estimated that humankind was able to store 295 EB of optimally compressed data in 2007. The issue of computational difficulty becomes even more severe in the bootstrap approximation of the asymptotic distribution of a U-statistic; see, for instance, Bickel and Freedman (1981), Bretagnolle (1983), Dehling and Mikosch (1994), and Huskova and Janssen (1993a,b), among others.

For certain U-statistics, the computational complexity can be reduced to $O(n)$ by exploiting the structure of the kernel function, especially when the data are univariate and consist of one sample. However, in practice, such a computational reduction is often not feasible. Note that we do not focus here on which U-statistics are candidates for a reduction in the original computational complexity of $O(n^d)$ because our goal is to study a generic scheme for the fast approximation of U-statistics. A natural remedy is to take a sample of size $m \ll \binom{n}{d}$ from all possible combinations. Blom (1976) referred to the resulting estimator as an *incomplete U-statistic*. The problem of identifying a good incomplete U-statistic is related to the design of the sampling scheme. Of the various options, the vanilla scheme of simple random sampling by Blom (1976) has received much attention in the literature. Janson (1984) established the asymptotic distributions of incomplete U-statistics based on random sampling (ICUR), Herrndorf (1986) established the invariance principle for the statsitics, and Chen and Kengo (2019) studied the vector- and matrix-valued ICUR. For a more detailed discussion on incomplete U-statistics, refer to Wang (2012) and Wang and Lindsay (2014).

First, we introduce some required notation. For $\alpha > 0$, we use $m \prec n^\alpha$, $m \asymp n^\alpha$, and $m \succ n^\alpha$ to mean $m/n^\alpha \to 0$, $0 < \underline{\lim}\, m/n^\alpha \leq \overline{\lim}\, m/n^\alpha < \infty$, and $m/n^\alpha \to \infty$, respectively. For a given incomplete U-statistic, say U, its efficiency is defined in terms of the mean squared error (MSE): $\mathrm{Eff}(U) = \mathrm{MSE}(U_0)/\mathrm{MSE}(U)$, where U_0 is the complete U-statistic. An incomplete U-statistic is said to be *asymptotically efficient* if $\mathrm{Eff}(U) \to 1$ as $n \to \infty$. Note that the ICUR is asymptotically efficient for the non-degenerate case when $m \succ n$; see (2.5) for a theoretical verification, and Table 1 for empirical evidence.

Blom (1976) also proposed sampling schemes based on the design of an experiment. In particular, balanced incomplete block designs (BIBDs) have been examined by Brown and Kildea (1978) and Lee (1982). The latter also proved that incomplete U-statistics based on BIBDs achieve the minimum variance among all unbiased estimators for a given m. By Raghavarao (1971), a BIBD exists whenever $n = 6a + 3$ for any positive integer a. Unfortunately, the optimality of the BIBD does not make it practically attractive because its construction requires $m \asymp n^2$; see Table 1. The same issue exists for the permanent design of Rempala and Wesolowski (2003) and the rectangular design of Rempala and Srivastav (2004). For the case of $m/n \to 1$, Blom (1976) proposed using a Latin square and a Graeco-Latin square to guide the sampling scheme. However, the efficiency of the estimator derived in this way is essentially asymptotically the same as that of the ICUR. Moreover, the limit of the efficiency does not exceed $d/(1 + d)$ as $n \to \infty$; see (2.5) and the follow-up discussion.

Another method recently proposed in the literature is the divide and conquer (DC) strategy of Lin and Xi (2010), which randomly divides the data into many groups, calculates the complete U-statistic within each group, and then takes the average of these complete U-statistics. Unfortunately, the DC is even less efficient than the ICUR. Moreover, it is not available when $m \le n$; see Table 1.

We conclude that the ICUR is still the most viable of the existing choices of incomplete U-statistics. It performs as well as a design-based method when a design exists. It also possesses several advantages, such as a flexible choice of m, the availability of asymptotic properties, and being extendable to multi-sample cases.

In this paper, we introduce a new type of incomplete U-statistic that is substantially more efficient than the ICUR, while maintaining the latter's aforementioned advantages. It has three main steps: (*i*) Divide the data into $L(\ll n)$ groups of homogeneous units. (*ii*) Judiciously select a collection of the combinations of the groups based on a design structure called an orthogonal array (OA). (*iii*) Randomly select a combination of inputs from each selected group combination. We call the derived estimator the incomplete U-statistic based on division and an orthogonal array (ICUDO). Our first example provides a snapshot of the performance of the major incomplete U-statistics mentioned so far.

Example 1. (The symmetry of distribution). The kernel function $g(x_1, x_2, x_3) = \text{sign}(2x_1 - x_2 - x_3) + \text{sign}(2x_2 - x_1 - x_3) + \text{sign}(2x_3 - x_1 - x_2)$ has mean zero when the distribution of the data is symmetric. The data consists of $n = 10^3$ independent and identically distributed (i.i.d.) observations generated iid from the standard normal distribution. The performance of the ICUR, BIBD, DC, and ICUDO is measured by their efficiency at different values of m.

Note that the DC is unavailable when $m \le n$, and the BIBD does not exist in most cases, except for $m = 166{,}167$. For $m \le 166{,}167$, the sample size is

Table 1. Comparison of efficiencies in Example 1.

m	m/n	$m/\binom{n}{3}$	ICUR	BIBD	DC	ICUDO
1.0×10^3	1.0	6.018×10^{-6}	21.62%	2.706%	–	36.31%
1.2×10^4	12.0	7.222×10^{-5}	74.97%	9.155%	55.60%	100%
5.7×10^4	57.0	3.430×10^{-4}	97.40%	21.81%	76.70%	100%
1.66×10^5	116.0	1.000×10^{-3}	100%	100%	84.22%	100%
3.92×10^5	392.0	2.359×10^{-3}	100%	100%	90.71%	100%
1.617×10^6	1617.0	9.731×10^{-3}	100%	100%	95.64%	100%

separately reduced for the BIBD in order to make it available. The ICUR has the same efficiency as the BIBD method at 100% when the BIBD exists. It is more efficient than the DC method whenever the DC is available. However, the ICUDO methods outperforms the ICUR for all m.

Here, we briefly explain why our ICUDO performs so well. Note that existing design-based methods focus on the arrangement of indices of units, without referring to their actual values. The ICUDO method exploits the fact that replacing a unit by another one with a similar value does not change the value of the kernel function g too much. For example, suppose the first six numbers of the data are $(1, 2, 3, 1, 2, 3)$. Then, a kernel function of order three yields the same value by evaluating the first three and the next three units. Beyond the grouping idea, we use the OA to achieve the projective uniformity of the group combinations in the dominating lower-dimensional spaces. This allows us to recover information on the lower dimension's variability in the U-statistics, which is the dominating part of Hoeffding's decomposition of the U-statistics. As shown later, in the non-degenerate case, whereas the ICUR needs $m \succ n$ to be asymptotically efficient, the ICUDO requires a substantially smaller m; sometimes even $m \succ \sqrt{n}$ will suffice. See Theorem 2 for the latter case. When the U-statistic is degenerate, both methods require larger m, but the ICUDO still requires a substantially smaller m than that of the ICUR.

The rest of the paper is organized as follows. Section 2 introduces the construction of the ICUDO for univariate data and derives its asymptotic properties. Section 3 discusses the debiasing issues of the ICUDO for the degenerate case. Section 4 constructs a debiased ICUDO for multi-dimensional data. Simulations are presented in each section to support the theoretical results. Section 5 concludes the paper and points out some future research topics. All proofs are postponed to the Appendix. Additional theorems are given in the online Supplementary Material.

2. ICUDO Based on Univariate Data

Let $X_1, \ldots, X_n$ be a random sample of size n from a univariate distribution, say F. For a given symmetric kernel function, say $g : R^d \to R$, of order d, the uniformly minimum variance unbiased estimator (UMVUE) of the parameter $\Theta = \int g(x_1, \ldots, x_d)\, dF(x_1) \ldots dF(x_d)$ is given by the U-statistic

$$U_0 = \binom{n}{d}^{-1} \sum_{\boldsymbol{\eta} \in S_{n,d}} g(\mathcal{X}_{\boldsymbol{\eta}}), \tag{2.1}$$

where $S_{n,d} = \{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d) : 1 \le \eta_1 < \eta_2 < \cdots < \eta_d \le n\}$ and $\mathcal{X}_{\boldsymbol{\eta}} = (X_{\eta_1}, \ldots, X_{\eta_d})$. When $S_{n,d}$ is replaced with the set of all n^d ordered combinations, the corresponding average in (2.1) is called a *V-Statistic* (von Mises (1948)). The main difference is that V-statistics include combinations with duplicated units, such as $(1, 1, 2)$. Throughout this paper, we adopt the mild assumption $Eg^2\,(X_1, \ldots, X_d) < \infty$.

Unless there is some special structure of g that can be exploited to reduce the computational burden, in general, (2.1) becomes impractical to compute as n increases. To address this problem, Blom (1976) proposed using the following incomplete U-statistic as a fast approximation:

$$U = \frac{1}{m} \sum_{\boldsymbol{\eta} \in S} g(\mathcal{X}_{\boldsymbol{\eta}}), \tag{2.2}$$

where $S \subset S_{n,d}$, with its cardinality $m = |S|$ being only a fraction of $\binom{n}{d}$. The statistic in (2.2) becomes an ICUR when S is a simple random sample, which we denote as U_{RND}.

Here, we briefly review the properties of U_0 and U_{RND}. For arbitrary positive integers N and p, define $\mathcal{Z}_N = \{1, \ldots, N\}$ and $\mathcal{Z}_N^p = \{(z_1 \ldots, z_p) : z_j \in \mathcal{Z}_N, 1 \le j \le p\}$. Following Hoeffding (1948), for $\boldsymbol{u} \subseteq \mathcal{Z}_d$ and $\boldsymbol{x} = (x_1, \ldots, x_d)$, denote $g_{\boldsymbol{u}}(\boldsymbol{x}) = \int g(\boldsymbol{x}) dF_{\boldsymbol{u}^c}$, with $\boldsymbol{u}^c = \mathcal{Z}_d \setminus \boldsymbol{u}$ and $dF_{\boldsymbol{u}} = \prod_{j \in \boldsymbol{u}} dF(x_j)$. With the conventions $g_{\emptyset}(\boldsymbol{x}) = \Theta$ and $h_{\emptyset}(\boldsymbol{x}) = 0$, we recursively define the projection

$$h_{\boldsymbol{u}}(\boldsymbol{x}) = g_{\boldsymbol{u}}(\boldsymbol{x}) - \sum_{\boldsymbol{v} \subseteq \mathcal{Z}_d : \boldsymbol{v} \subset \boldsymbol{u}} h_{\boldsymbol{v}}(x).$$

Because g is symmetric, we have $Eg_{\boldsymbol{v}}^2 = Eg_{\boldsymbol{u}}^2$ and $Eh_{\boldsymbol{v}}^2 = Eh_{\boldsymbol{u}}^2$ for any pair $\boldsymbol{u}, \boldsymbol{v} \subseteq \mathcal{Z}_d$, with $|\boldsymbol{v}| = |\boldsymbol{u}|$. Hence, we can now define

$$\sigma_j^2 = \text{Var}(g_{\boldsymbol{u}}) \text{ and } \delta_j^2 = \text{Var}(h_{\boldsymbol{u}}), \quad \text{with } |\boldsymbol{u}| = j.$$

Following Hoeffding (1948) and Blom (1976), we have

$$\mathrm{MSE}(U_0) = \binom{n}{d}^{-1} \sum_{j=1}^{d} \binom{d}{j}\binom{n-d}{d-j}\sigma_j^2 \;=\; \sum_{j=1}^{d} \binom{d}{j}^2 \binom{n}{j}^{-1} \delta_j^2, \quad (2.3)$$

$$\begin{aligned}\mathrm{MSE}(U_{\mathrm{RND}}) &= \mathrm{MSE}(U_0) + \frac{\sigma_d^2}{m} + O\left(\frac{1}{nm}\right) \\ &= \mathrm{MSE}(U_0) + \frac{1}{m}\sum_{j=1}^{d}\binom{d}{j}\delta_j^2 + O\left(\frac{1}{nm}\right). \quad (2.4)\end{aligned}$$

In (2.3) and (2.4), the MSEs are expressed in terms of both σ_j^2 and δ_j^2. The equivalences are established by $\sigma_j^2 = \sum_{j'=1}^{j}\binom{j}{j'}\delta_{j'}^2$, for $1 \le j \le d$. The U-statistic and the kernel function g are called *non-degenerate* if $\delta_1^2 = \sigma_1^2 > 0$, and are called *order-$q$ degenerate* if $\sigma_q^2 = 0$ and $\sigma_{q+1}^2 > 0$, or equivalently $\delta_1^2 = \cdots = \delta_q^2 = 0$ and $\delta_{q+1}^2 > 0$. For the non-degenerate case, we have $\mathrm{Var}(U_0) \asymp n^{-1}$, which together with (2.4) yields

$$\mathrm{Eff}(U_{\mathrm{RND}}) = \begin{cases} 1 - O\left(\dfrac{n}{m}\right), & m \succ n \\ \dfrac{1}{1 + (n/m)(\sigma_d^2/(d^2\delta_1^2))} + O\left(\dfrac{1}{n}\right), & m \asymp n \\ O\left(\dfrac{m}{n}\right), & m \prec n. \end{cases} \quad (2.5)$$

As a result, we have $\mathrm{Eff}(U_{\mathrm{RND}}) \to 1$ when $m \succ n$, $\mathrm{Eff}(U_{\mathrm{RND}}) \to 0$ when $m \prec n$, and $\mathrm{Eff}(U_{\mathrm{RND}}) \to \left(1 + \sigma_d^2/(cd^2\sigma_1^2)\right)^{-1}$ when $m/n \to c$, for a constant $c > 0$. With $c = 1$, Blom (1976) proposed using Latin squares and Graeco-Latin squares to construct the incomplete U-statistics. In such a case, we can verify that its efficiency is asymptotically the same as that of U_{RND}, and $\lim_{n\to\infty} \mathrm{Eff}(U_{\mathrm{RND}}) \le d/(1+d)$, from (2.5) and $\sigma_d^2 \le d\sigma_1^2$. In contrast, Theorem 1 shows that the ICUDO is asymptotically efficient when $m \asymp n$. Stronger results are stated in Theorem 2 in Section 2.1 and in similar theorems in the Supplementary Material under various conditions on g and F.

2.1. One-sample U-statistics

Recall that $\delta_j^2 = \mathrm{Var}(h_{\boldsymbol{u}})$, for $|\boldsymbol{u}| = j$, $1 \le j \le d$, and note that the coefficient of δ_j^2 in (2.3) is $O(n^{-j})$. Hence, it is more important to capture the variability of g in its lower-dimensional projected space. This idea matches perfectly with the projective property of the OA. An OA denoted by $OA(m, d, L, t)$, is an m by d array with entries from $\{1, \dots, L\}$, arranged in such a way that for any m by

t subarray, all ordered t-tuples of the entries from $\{1,\dots,L\}$ appear $\lambda = m/L^t$ times in the rows. The number t is called the strength of the OA; see the matrix A defined in (2.7) as an example of $OA(9,4,3,2)$. In this case, the ordered 2-arrays are $\{(i_1,i_2) : 1 \le i_1, i_2 \le 3\}$. Consider any two columns of A, we can see that all these ordered 2-tuples appear once, that is, $\lambda = 1$. For sets $\mathcal{S}_1,\dots,\mathcal{S}_q$, define $\prod_{i=1}^q \mathcal{S}_i = \{(\boldsymbol{s}_1,\dots,\boldsymbol{s}_q) : \boldsymbol{s}_i \in \mathcal{S}_i\}$. The ICUDO is constructed as follows. For ease of illustration, we assume n is a multiple of L. Actually, throughout the manuscript, we assume that $L \ll n$. Thus, we may randomly draw an $n' = \lfloor n/L \rfloor \cdot L$ subsample as the new data set. The information loss in this process is negligible compared with the original size n.

Step 1. Let A_0 be an $OA(m,d,L,t)$. Apply random level permutations $\{\pi_1,\dots,\pi_d\}$ to columns of A_0 independently. Specifically, for $l \in \mathcal{Z}_L$, change all elements l in the jth column of A_0 to $\pi_j(l)$. The new OA is denoted by $A = (a_{ij})_{m\times d}$.

Step 2. Create the partition $\mathcal{Z}_n = \bigcup_{l=1}^L G_l$ such that $|G_l| = n/L$ for $l \in \mathcal{Z}_L$, and $X_{i_1} \le X_{i_2}$ for any $i_1 \in G_{l_1}, i_2 \in G_{l_2}$, with $l_1 < l_2$.

Step 3. For $i = 1,\dots,m$, independently draw an element, say $\boldsymbol{\eta}^i$, uniformly from $\prod_{j=1}^d G_{a_{ij}}$. the ICUDO based on the OA A is defined as

$$U_{oa} = \frac{1}{m}\sum_{i=1}^m g(\mathcal{X}_{\boldsymbol{\eta}^i}). \tag{2.6}$$

The level permutation in step 1 ensures that each row of A takes each d-tuple with equal probability. At the same time, the projective uniformity of the beginning OA, A_0, carries over to A. Here, we ensure that A is free of a coincidence defect, which means no two rows are the same in any $m \times (t+1)$ subarray. This property is necessary for the relevant theorems to hold. Step 2 divides the data into homogeneous groups. Step 3 is built on the first two steps. It chooses representative elements from selected groups, and the selection of groups is guided by the structure of A. Note (2.6) is in the form of (2.2) by taking S as $S_{oa} = \{\boldsymbol{\eta}^1,\dots,\boldsymbol{\eta}^m\}$. We now give a toy example of choosing η^i, for $i = 1,\dots,m$. Suppose $d = 4$, $n = 9$, and

$$X_6 \le X_8 \le X_2 \le X_4 \le X_7 \le X_5 \le X_3 \le X_9 \le X_1.$$

Then, we have $L = 3$ groups listed as $G_1 = \{6,8,2\}, G_2 = \{4,7,5\}$, and $G_3 = \{3,9,1\}$. An example of $OA(m = 9, d = 4, L = 3, t = 2)$ in step 1 is given as follows in transpose:

$$A^T = \begin{pmatrix} 1\,1\,1\,\mathbf{2}\,2\,2\,3\,3\,3 \\ 1\,2\,3\,\mathbf{1}\,2\,3\,1\,2\,3 \\ 1\,2\,3\,\mathbf{2}\,3\,1\,3\,1\,2 \\ 1\,2\,3\,\mathbf{3}\,1\,2\,2\,3\,1 \end{pmatrix}. \tag{2.7}$$

The fourth row of A, namely $(2,1,2,3)$, means we are sampling $\boldsymbol{\eta}^4$ from $G_2 \times G_1 \times G_2 \times G_3$. One possible outcome for $\boldsymbol{\eta}^4$ could be $(4,8,7,3)$. Repeating this for each row of A, we could possibly have the $\mathcal{X}_{\boldsymbol{\eta}^i}$, for $i = 1,\dots,9$, used in the construction as follows:

$$\{\mathcal{X}_{\boldsymbol{\eta}^1},\dots,\mathcal{X}_{\boldsymbol{\eta}^9}\} = \begin{Bmatrix} X_6\ X_8\ X_2\ X_{\mathbf{4}}\ X_4\ X_5\ X_9\ X_3\ X_1 \\ X_2\ X_4\ X_3\ X_{\mathbf{8}}\ X_7\ X_9\ X_8\ X_5\ X_9 \\ X_8\ X_6\ X_9\ X_{\mathbf{7}}\ X_1\ X_2\ X_1\ X_2\ X_4 \\ X_6\ X_5\ X_1\ X_{\mathbf{3}}\ X_6\ X_4\ X_7\ X_9\ X_6 \end{Bmatrix}. \tag{2.8}$$

To proceed with the asymptotic properties of U_{oa}, we define

$$R(t) = \sum_{j>t} \binom{d}{j} \delta_j^2. \tag{2.9}$$

Theorem 1. *For any (g,F), using $OA(m,d,L,t)$ in step 1 of the ICUDO algorithm, we have*

$$\mathrm{MSE}(U_{oa}) = \mathrm{MSE}(U_0) + \frac{R(t)}{m} + o\left(\frac{1}{m}\right) + O\left(\frac{1}{n^2}\right). \tag{2.10}$$

We now explain the meanings of the three terms in (2.10) generated in the process of approximating the complete U-statistic U_0 using U_{oa}. The term $O(n^{-2})$ is the bias square of U_{oa} due to the inclusion of combinations with duplicate units, such as the first column of (2.8). Essentially, U_{oa} is approximating the V-statistic, which is biased for Θ itself. The term $o(m^{-1})$ is due to the sampling variability when we draw one point from each selected group, that is, step 3 of the algorithm. The term $R(t)/m$ is due to the usage of the OA structure in place of a complete enumeration of all group combinations. Compared with the second term in (2.4) for the ICUR, $R(0)/m$, we are able to eliminate all δ_j^2 with $j \le t$ owing to the projective uniformity of the OA in all t-dimensional projected spaces. If $\delta_j^2 = 0$ for $d' \le j \le d$, an OA with strength $t \ge d'$ yields $R(t) = 0$. We discuss the hidden benefit of using a lower strength OA in Example 2.

In the non-degenerate case, recall the $\mathrm{MSE}(U_0) \asymp n^{-1}$ and $\lim_{n\to\infty} \mathrm{Eff}(U_{\mathrm{RND}}) \le d/(1+d)$ for the ICUR when $m \asymp n$. Under the same situation, Theorem 1 implies that U_{oa} is asymptotically efficient by simply taking $t = d$. In fact, stronger results can be derived for the ICUDO so that m is allowed to grow more slowly than n under various conditions. We give Theorem 2 here as one example; additional results can be found in the Supplementary Material.

Theorem 2. *Suppose (i) the kernel function g is Lipschitz continuous, and (ii) F has density function $f(x) > c$ for some fixed $c > 0$ and $x \in [a, b]$, and $f(x) = 0$ otherwise. For U_{oa} based on $OA(m, d, L, t)$ with $L^2 \leq n(\log n)^{-1}$, we have*

$$\mathrm{MSE}(U_{oa}) = \mathrm{MSE}(U_0) + \frac{R(t)}{m} + O\left(\frac{1}{mL^2}\right) + O\left(\frac{1}{n^2}\right). \tag{2.11}$$

For $t = d = 2$, we automatically have $R(t) = 0$. If the conditions in Theorem 2 hold, we only need $m \succ \sqrt{n}$ to achieve $\mathrm{Eff}(U_{oa}) \to 1$, while the ICUR requires $m \succ n$. In general, $R(t)$ decreases in t and could vanish if we take t large enough so that $\delta_j^2 = 0$, for all $j > t$. Without knowledge of δ_j^2, simply taking $t = d$ will eliminate $R(t)$ too. On the other hand, the term $O(1/(mL^2))$ in (2.11) is decreasing in L, meaning the more groups we use to divide the data, the more homogeneous the units we could have in each group. However, L and t are subject to the constraint $m = \lambda L^t$, where λ is the number of replicates of each t-tuple in OA and is equal to one in all examples presented here. As a result, L and t cannot be increased simultaneously. To gain insight to the trade-off between L and t, we need to determine the constant term for $O(1/(mL^2))$. For this, we derive the following theorem. A more detailed discussion on how to choose L and t, given m, is provided in the Supplementary Material. Denote by $U(0, 1)$ the uniform distribution on $[0, 1]$.

Theorem 3. *Suppose g has a continuous first-order derivative on $[0, 1]^d$, $X \sim U(0, 1)$, and there exists some $c \in (0, 1/2)$, such that $L \preceq n^c$. For U_{oa} based on $OA(m, d, L, t)$,*

$$\mathrm{MSE}(U_{oa}) = \mathrm{MSE}(U_0) + \frac{R(t)}{m} + \frac{d}{12mL^2} E\gamma^2(X_1, \ldots, X_d) + o\left(\frac{1}{mL^2}\right), \tag{2.12}$$

where $\gamma(x_1, \ldots, x_d) = \partial g / \partial x_1 (x_1, \ldots, x_d)$.

The assumption of a uniform distribution for X is not as strict as it seems. To see this, for $X \sim F$, let $Z = F(X) \sim U(0, 1)$. Applying Theorem 3 to $g_F(Z_1, \ldots, Z_d) := g(F^{-1}(Z_1), \ldots, F^{-1}(Z_d)) = g(X_1, \ldots, X_d)$, we have the following corollary.

Corollary 1. *Suppose g_F has a continuous first-order derivative on $[0, 1]^d$, and*

Table 2. Result of Example 2.

m/n	$\mu = 0.5$			$\mu = 2$		
	Eff(U_{RND})	Eff(U_{oa_2})	Eff(U_{oa_3})	Eff(U_{RND})	Eff(U_{oa_2})	Eff(U_{oa_3})
0.005	0.133%	0.171%	**0.218**%	1.110%	**9.908**%	2.323%
0.01	0.290%	0.464%	**0.579**%	2.485%	**26.84**%	8.455%
0.05	1.291%	2.448%	**6.096**%	10.31%	**75.12**%	51.71%
0.1	2.936%	4.527%	**16.62**%	20.13%	**91.87**%	76.80%
0.5	12.58%	21.89%	**71.78**%	50.78%	**100.0**%	98.53%
1.0	21.05%	33.26%	**99.94**%	67.51%	**100.0**%	99.64%

there exists some $c \in (0, 1/2)$*, such that* $L \preceq n^c$*. Then,* (2.12) *still holds.*

The term $E\gamma^2$ in (2.12) provides a nice interpretation of the trade-off between t and L. When the kernel function g has a large variability (large $E\gamma^2$), it is more challenging to make each group as homogeneous as possible, which enforces larger values of L. On the other hand, if g is quite flat on the domain (small $E\gamma^2$), we prefer fewer groups to improve the strength of the OA.

Example 2. The kernel function $g(x_1, x_2, x_3) = x_1x_2x_3$ estimates μ^3, where $\mu = E(X)$. We compare the performance of three methods: U_{RND}; U_{oa_2} based on $OA(m, 3, \sqrt{m}, 2)$, with strength $t = 2$; and U_{oa_3} based on $OA(m, 3, m^{1/3}, 3)$, with strength $t = 3$. The data consist of $n = 10^4$ i.i.d. observations simulated from $N(\mu, 1)$, where μ takes the values of 0.5 and 2; see Table 2 for the simulation results.

In Table 2, both U_{oa_2} and U_{oa_3} outperform U_{RND} significantly. The advantage of the ICUDO over the ICUR is discussed below in additional examples. Furthermore, we find that the winning strategy changes from U_{oa_3} to U_{oa_2} as we increase the mean μ of the distribution. This observation well illustrates the comments after Theorem 3 on the relevance of $E\gamma^2$ in determining the optimal value of the strength t. That is, for larger $E\gamma^2$, we are more inclined to choose a smaller strength. This is validated by our second observation together with $E\gamma^2 = (\mu^2 + 1)^2$, which increases in $\mu(> 0)$.

Note that the applicability of Theorem 2 and its variants, Theorems 7–9 in the Supplementary Material is broader than it appears. To see this, let $\phi : R \to R$ be a one-to-one mapping. Denote by F_ϕ the distribution of the transformed random variable $Z = \phi(X)$, which leads to the following representation:

$$g_\phi(z_1, \ldots, z_d) := g(\phi^{-1}(z_1), \ldots, \phi^{-1}(z_d)) = g(x_1, \ldots, x_d).$$

If (g_ϕ, F_ϕ) satisfies the conditions in these theorems, corresponding results also hold for the pair (g, F). For example, suppose $g(x_1, x_2) = x_1^{-a} x_2^{-a}$ and F is a Pareto distribution with shape and scale parameters a and b, respectively. The Pareto distribution is neither light-tailed nor bounded, and hence violates the conditions in Theorem 2. By taking $\phi(x) = 1 - (b/x)^a$, we have $\phi(X) \sim U(0, 1)$. It can be verified that the conditions in Theorem 2 are satisfied by (g_ϕ, F_ϕ).

2.2. Multi-sample U-statistics

For $k = 1, \ldots, K$, let $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ be a random sample of size n_k from the distribution F_k. The UMVUE of

$$\Theta = \int g(x_1^{(1)}, \ldots, x_{d_1}^{(1)}, \ldots, x_1^{(K)}, \ldots, x_{d_K}^{(K)}) dF_1(x_1^{(1)}) \cdots dF_K(x_{d_K}^{(K)})$$

is given by the generalized U-statistic

$$U_0 = \prod_{k=1}^{K} \binom{n_k}{d_k}^{-1} \sum_{\boldsymbol{\eta} \in \prod_{k=1}^{K} S_{n_k, d_k}} g(\mathcal{X}_{\boldsymbol{\eta}}),$$

$$S_{n_k, d_k} = \{\boldsymbol{\eta}_k = (\eta_{k,1}, \ldots, \eta_{k,d_k}) : 1 \leq \eta_{k,1} < \eta_{k,2} < \cdots < \eta_{k,d_k} \leq n_k\},$$
$$\mathcal{X}_{\boldsymbol{\eta}} = (\mathcal{X}_{\boldsymbol{\eta}_1}, \ldots, \mathcal{X}_{\boldsymbol{\eta}_K}) = (X_{\eta_{1,1}}^{(1)}, \ldots, X_{\eta_{1,d_1}}^{(1)}, \ldots, X_{\eta_{K,1}}^{(K)}, \ldots, X_{\eta_{K,d_K}}^{(K)}).$$

The $d (= \sum_{k=1}^{K} d_k)$-dimensional kernel function g is symmetric about any d_k-dimensional sub-input $\{x_1^{(k)}, \ldots, x_{d_k}^{(k)}\}$. The generalized U-statistic reduces to the traditional U-statistic when $K = 1$. An incomplete generalized U-statistic is given by

$$U = \frac{1}{m} \sum_{\boldsymbol{\eta} \in S} g(\mathcal{X}_{\boldsymbol{\eta}}), \tag{2.13}$$

where $S \subset \prod_{k=1}^{K} S_{n_k, d_k}$ and $m = |S|$. We construct the multi-sample ICUDO as follows. For ease of illustration, we assume n_k' is a multiple of L.

Step 1. Let A_0 be an $OA(m, d, L, t)$. Adopt random level permutations $\{\pi_1, \ldots, \pi_d\}$ of columns of A_0 independently. Specifically, for each $l \in \mathcal{Z}_L$, change all elements l in the jth column of A_0 to $\pi_j(l)$. The m rows of the resulting array A are denoted by $\{\boldsymbol{a}^i = (\boldsymbol{a}_1^i, \ldots, \boldsymbol{a}_K^i) : i = 1, \ldots, m; \boldsymbol{a}_k^i \in \mathcal{Z}_L^{d_k}, k = 1, \ldots, K\}$.

Step 2. For each $k = 1, \ldots, K$, create the partition $\mathcal{Z}_{n_k} = \bigcup_{l=1}^{L} G_l^{(k)}$, such that $|G_l^{(k)}| = n_k L^{-1}$ for $l \in \mathcal{Z}_L$, and $X_{i_1}^{(k)} \leq X_{i_2}^{(k)}$ for any $i_1 \in G_{l_1}^{(k)}$, $i_2 \in G_{l_2}^{(k)}$, with $l_1 < l_2$. For any $\boldsymbol{a} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K)$ with $\boldsymbol{a}_k = (a_{k,1}, \ldots, a_{k,d_k}) \in \mathcal{Z}_L^{d_k}$, define

$$\mathcal{G}_{\boldsymbol{a}} = \prod_{k=1}^{K} \prod_{j=1}^{d_k} G^{(k)}_{a_{k,j}}. \tag{2.14}$$

Step 3. For $i = 1, \ldots, m$, independently draw an element $\boldsymbol{\eta}^i$ uniformly from $\mathcal{G}_{\boldsymbol{a}^i}$, where $\boldsymbol{a}^i$ is the ith row of A:

$$U_{oa} = \frac{1}{m} \sum_{i=1}^{m} g(\mathcal{X}_{\boldsymbol{\eta}^i}). \tag{2.15}$$

An example is given in the Supplementary Material. For any $j_{k,1}, \ldots, j_{k,d_k} \in \mathcal{Z}_{d_k}$ and $k \in \mathcal{Z}_K$, assume

$$Eg^2\left(X^{(1)}_{j_{1,1}}, \ldots, X^{(1)}_{j_{1,d_1}}, \ldots, X^{(K)}_{j_{K,1}}, \ldots, X^{(K)}_{j_{K,d_K}}\right) < \infty.$$

Let $n_{\min} = \min\{n_1, \ldots, n_K\}$ and $n_{\max} = \max\{n_1, \ldots, n_K\}$. Here, we assume $n_{\min} \asymp n_{\max}$ and $L \prec n_{\min}$. Let $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K)$, where $\boldsymbol{u}_k \subseteq \mathcal{Z}_{d_k}$. Define $dF_{\boldsymbol{u}} = \prod_{k=1}^{K} \prod_{j \in \boldsymbol{u}_k} dF_k(x^{(k)}_j)$. For any $\boldsymbol{u}$ and $\boldsymbol{x} = (x^{(1)}_1, \ldots, x^{(1)}_{d_1}, \ldots, x^{(K)}_1, \ldots, x^{(K)}_{d_K})$, we recursively define

$$g_{\boldsymbol{u}}(\boldsymbol{x}) = \int g(\boldsymbol{x}) dF_{\boldsymbol{u}^c} \qquad h_{\boldsymbol{u}}(\boldsymbol{x}) = g(\boldsymbol{x}) - \sum_{\boldsymbol{v} \subset \boldsymbol{u}} h_{\boldsymbol{v}}(x),$$

where $\boldsymbol{u}^c = (\boldsymbol{u}^c_1, \ldots, \boldsymbol{u}^c_K) = (\mathcal{Z}_{d_1} \setminus \boldsymbol{u}_1, \ldots, \mathcal{Z}_{d_K} \setminus \boldsymbol{u}_K)$, $g_{\emptyset}(\boldsymbol{x}) = \Theta$ and $h_{\emptyset}(\boldsymbol{x}) = 0$, $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)$, and $\boldsymbol{v} \subset \boldsymbol{u}$ means $\boldsymbol{v}_k \subseteq \boldsymbol{u}_k$ ($\boldsymbol{v} \neq \boldsymbol{u}$).

For $\boldsymbol{u}$, we can define $\sigma^2_{\boldsymbol{u}} = \mathrm{Var}(g_{\boldsymbol{u}})$ and $\delta^2_{\boldsymbol{u}} = \mathrm{Var}(h_{\boldsymbol{u}})$. The MSE of the complete generalized U-statistic is given by Sen (1974) as

$$\mathrm{MSE}(U_0) = \prod_{k=1}^{K} \binom{n_k}{d_k}^{-1} \sum_{\boldsymbol{u}=(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_K)} \left\{ \prod_{k=1}^{K} \binom{d_k}{|\boldsymbol{u}_k|} \binom{n_k - d_k}{d_k - |\boldsymbol{u}_k|} \right\} \sigma^2_{\boldsymbol{u}}.$$

Let $|\boldsymbol{u}| = \sum_{k=1}^{K} |\boldsymbol{u}_k|$. The generalized U-statistic and the kernel function are called *order-q degenerate* if $\sigma^2_{\boldsymbol{u}} = \sum_{\boldsymbol{v} \in \boldsymbol{u}} \delta^2_{\boldsymbol{v}} = 0$, for all $|\boldsymbol{u}| \leq q$, and there exists $\boldsymbol{u}'$ such that $\sigma^2_{\boldsymbol{u}'} > 0$ and $|\boldsymbol{u}'| = q + 1$. We have $\mathrm{MSE}(U_0) = O(n^{-(q+1)})$ in this case. For the non-degenerate case $q = 0$, we have $\mathrm{MSE}(U_0) \asymp n^{-1}$. With a slight abuse of notation, let $\sigma_{(j_1,\ldots,j_K)} = \sigma_{\boldsymbol{u}}$ and $\delta_{(j_1,\ldots,j_K)} = \delta_{\boldsymbol{u}}$, for $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K)$, with $|\boldsymbol{u}_k| = j_k$, $k = 1, \ldots, K$. For the ICUR, we have

$$\mathrm{MSE}(U_{\mathrm{RND}}) = \mathrm{MSE}(U_0) + \frac{R(0)}{m} + O\left(\frac{1}{m n_{\min}}\right),$$

$$R(t) = \boldsymbol{u} : |\boldsymbol{u}| > t\delta_{\boldsymbol{u}}^2 = \sum_{j_1=0}^{d_1} \cdots \sum_{j_K=0}^{d_K} I(j_1 + \cdots + j_K > t) \prod_{k=1}^{K} \binom{d_k}{j_k} \delta^2_{(j_1,\ldots,j_K)}.$$

The last term above reduces to the form of $R(t)$ for the one-sample case, but the second term yields a parsimonious presentation for the multi-sample case. The corresponding properties of U_{oa} are given as follows.

Theorem 4. *For U_{oa} based on $OA(m, d, L, t)$, for any pair of (g, F), we have*

$$\mathrm{MSE}(U_{oa}) = \mathrm{MSE}(U_0) + \frac{R(t)}{m} + o\left(\frac{1}{m}\right) + O\left(\frac{1}{n_{\min}^2}\right). \qquad (2.16)$$

Theorem 4 is basically a multi-sample version of Theorem 1, and its result can be strengthened in the same way. The details are omitted here to conserve space. We conclude this section with a machine learning example.

Example 3. (Ranking measure, Chen et al. (2009)). The ranking measure is an important topic in machine learning research. In the commonly used pairwise approach, the loss for a given classifier score function f is given by

$$L(f) = \sum_{1 \le i < j \le K} \sum_{x \in G_i, y \in G_j} \psi(f(y) - f(x)),$$

where $G_1, \ldots, G_K$ are K groups ranked in ascending order. Here, ψ could that the form of

(i) hinge function: $\psi(z) = (1 - z)_+$, or a

(ii) logistic function: $\psi(z) = \log(1 + \exp(-z))$

for the Ranking SVM and RankNet methods, respectively. In the simulation, we set $K = 2$, that is, the two-sample case, $|G_1| = |G_2| = 10^4$, $f(G_1) \sim N(0, 4)$, and $f(G_2) \sim N(5, 4)$. Figure 1 reveals the high efficiency of $\tilde{U}_{oa}$ compared with that of U_{RND}.

3. Debiased ICUDO for Degenerate Cases

Recall the ICUDO procedure is actually biased owing to the inclusion of combinations with duplicate units. The bias square is $O(n^{-2})$ for any pair (g, F), which is negligible compared to $\mathrm{Var}(U_0) \asymp n^{-1}$ in the non-degenerate case. One can see that it is no longer negligible in the degenerate case. In this section, we propose a debiased version of the ICUDO.

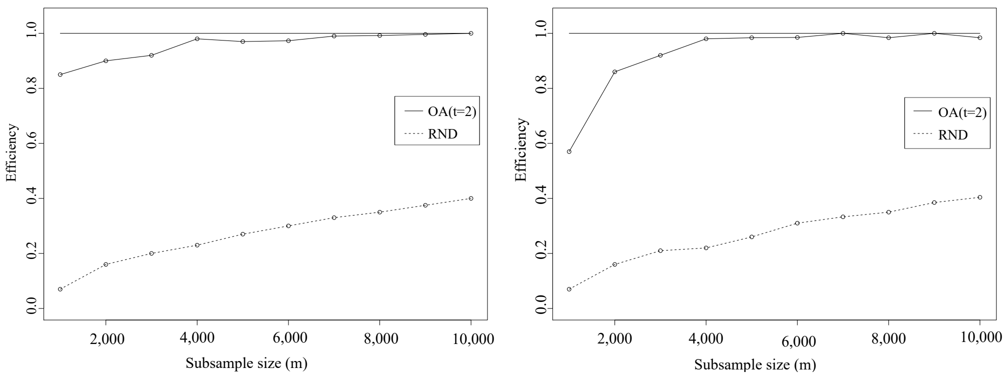

Figure 1. Comparison of efficiencies of $\tilde{U}_{oa}$ and U_{RND} with respect to subsample size m for loss function (i) (left) and (ii) (right).

We provide details for the multi-sample cases, where the one-sample cases are achieved by taking $K = 1$. To proceed, Let $S_0^* = \{(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K) : \boldsymbol{\eta}_k = (\eta_{k,1}, \ldots, \eta_{k,d_k}) \in \mathcal{Z}_{n_k}^{d_k}, \eta_{k,j_1} \neq \eta_{k,j_2} \text{ for any } j_1 \neq j_2\}$. The debiased ICUDO is constructed in the same way as the original, except that step 3 changes as follows:

Step 3$'$. For $i = 1, \ldots, m$, independently draw $\boldsymbol{\eta}^i$ from the uniform distribution on $\mathcal{G}_{\boldsymbol{a}^i} \cap S_0^*$. Adopting (2.13) with $S_{oa}^* = \{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^m\}$, we have the debiased ICUDO as

$$\tilde{U}_{oa} = \frac{1}{m} \sum_{i=1}^{m} \omega_{\boldsymbol{\eta}^i} g(X_{\boldsymbol{\eta}^i}), \tag{3.1}$$

where $\omega_{\boldsymbol{\eta}^i} = L^d |\mathcal{G}_{\boldsymbol{a}^i} \cap S_0^*| / |S_0^*|$.

Theorem 5. *$\tilde{U}_{oa}$ based on $OA(m, d, L, t)$ is an unbiased estimator, and*

$$\text{MSE}(\tilde{U}_{oa}) = \text{MSE}(U_0) + \frac{R(t)}{m} + o\left(\frac{1}{m}\right). \tag{3.2}$$

Theorem 5 is analogous to Theorems 1 and 4 for the one-sample and multi-sample cases, respectively, except that the bias square term $O(n^{-2})$ and $O(n_{\min}^{-2})$ are eliminated. Now, for an order-q degenerate U-statistic, the debiased ICUDO can be asymptotically efficient with $m \asymp n^{q+1}$, while the ICUR requires $m \succ n^{q+1}$. Moreover, we could allow m to grow more slowly for the debiased ICUDO under some mild conditions on (g, F). For example, when $d = 2$, $q = 1$, and the conditions of Theorem 2 hold, the debiased ICUDO only needs $m \succ n$ to be asymptotically efficient, while the ICUR requires $m \succ n^2$. For the general order q of degeneration, we have $m_{oa}^* = (m_{\text{RND}}^*)^{d/(d+1)}$, for all d, under the conditions

Table 3. Result of Example 4.

$m/\binom{n}{2}$	0.002	0.01	0.02	0.04	0.06	0.1	0.14	0.2
$\mathrm{Eff}(\tilde{U}_{oa_3})$	0.836%	10.9%	**15.6%**	**35.9%**	**44.9%**	**56.9%**	**75.1%**	**94.1%**
$\mathrm{Eff}(U_{oa_3})$	0.861%	9.50%	12.9%	25.2%	28.3%	29.8%	36.3%	39.0%
$\mathrm{Eff}(U_{oa_4})$	0.450%	4.96%	6.78%	10.6%	10.7%	11.9%	14.5%	15.6%
$\mathrm{Eff}(U_{\mathrm{RND}})$	0.179%	0.701%	1.50%	2.93%	4.19%	7.84%	10.9%	13.1%

in Theorem 2. Here, m^*_{oa} and m^*_{RND} represent the minimum m required for the ICUDO and ICUR, respectively, to be asymptotically efficient.

We conclude this section with the following multi-sample example. The kernel function is degenerate, and hence favors a debiased ICUDO. However, the highest order δ^2-value vanishes, which encourages a lower strength of OA. The comparison is made between the ICUR and different versions of the ICUDO.

Example 4. Let $K = 2$, $d_1 = d_2 = 2$, $d = 4$, and

$$g(x_1^{(1)}, x_2^{(1)}, x_1^{(2)}, x_2^{(2)}) = I(x_1^{(1)} < x_1^{(2)}, x_2^{(1)} < x_1^{(2)}) + I(x_1^{(2)} < x_1^{(1)}, x_2^{(2)} < x_1^{(1)}).$$

The construction of U_{oa} and the debiased $\tilde{U}_{oa}$ is based on $OA(m, 4, m^{1/3}, 3)$ and $OA(m, 4, m^{1/4}, 4)$. For continuous distributions F_1 and F_2, it can be verified that

$$Eg(X_1^{(1)}, X_2^{(1)}, X_1^{(2)}, X_2^{(2)}) = \frac{2}{3} + \frac{1}{2}\int (F_1(x) - F_2(x))^2 d(F_1(x) + F_2(x)),$$

which indicates the similarity of F_1 and F_2. The null hypothesis of $F_1 = F_2$ is rejected when the U-statistic is significantly larger than 2/3. Note that the corresponding U-statistic is degenerate under the null hypothesis. See Table 3 for the simulation results when both samples are simulated from $N(0, 1)$ with sample sizes $n_1 = n_2 = 10^3$.

Note that in the g function of Example 4, the two separate parts are all functions of three inputs. Thus, $R(4) = 0$, and we can claim that $t = 3$ works better than $t = 4$, which is verified by the results in Table 3.

4. ICUDO for Multi-Dimensional Data

Note that step 2 of the ICUDO algorithm in Section 2 does not apply to multi-dimensional data because it relies on ordering the univariate data. To remedy this, we adopt a clustering algorithm to divide the data into homogeneous groups. In this regard, the clustered group sizes may vary. This will necessitate a re-weighting procedure similar to the debiasing step in Section 3. To save space, we focus on the debiased ICUDO and adopt the notation of the multi-sample U-statistics in the study of multi-dimensional data. For $k = 1, \dots, K$, let $X_1^{(k)}, \dots, X_{n_k}^{(k)}$ be a random sample of size n_k from the multi-dimensional distribution F_k. The algorithm is given as follows.

Step 1. Let A_0 be an $OA(m, d, L, t)$. Adopt random level permutations $\{\pi_1, \dots, \pi_d\}$ of columns of A_0 independently. Specifically, for $l \in \mathcal{Z}_L$, change all elements l in the jth column of A_0 to $\pi_j(l)$. The m rows of the resulting array A are denoted by $\{\boldsymbol{a}^i = (\boldsymbol{a}_1^i, \dots, \boldsymbol{a}_K^i) : i = 1, \dots, m; \boldsymbol{a}_k^i \in \mathcal{Z}_L^{d_k}, k = 1, \dots, K\}$.

Step 2. Let $\mathcal{P}^{(k)} = \{G_1^{(k)}, \dots, G_L^{(k)}\}$ denote an L-group partition from the clustering of $\{X_1^{(k)}, \dots, X_{n_k}^{(k)}\}$. For any $\boldsymbol{a} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_K)$, with $\boldsymbol{a}_k = (a_{k,1}, \dots, a_{k,d_k}) \in \mathcal{Z}_L^{d_k}$, define

$$\mathcal{G}_{\boldsymbol{a}} = \prod_{k=1}^{K} \prod_{j=1}^{d_k} G_{a_{k,j}}^{(k)}. \tag{4.1}$$

Step 3. For $i = 1, \dots, m$, independently draw an element $\boldsymbol{\eta}^i$ uniformly from $\mathcal{G}_{\boldsymbol{a}^i}$, where $\boldsymbol{a}^i$ is the ith row of A. Let $\omega_{\boldsymbol{\eta}^i} = L^d |\mathcal{G}_{\boldsymbol{a}^i} \cap S_0^*| / |S_0^*|$.

$$\tilde{U}_{oa} = \frac{1}{m} \sum_{i=1}^{m} \omega_{\boldsymbol{\eta}^i} g(X_{\boldsymbol{\eta}^i}). \tag{4.2}$$

An example of the construction is given in the Supplementary Material.

Theorem 6. *Suppose $\omega_{\boldsymbol{\eta}^i} \to 1$ uniformly as $n, L \to \infty$. For $\tilde{U}_{oa}$ based on $OA(m, d, L, t)$, we have*

$$\mathrm{MSE}(\tilde{U}_{oa}) = \mathrm{MSE}(U_0) + \frac{R(t)}{m} + o\left(\frac{1}{m}\right). \tag{4.3}$$

The $R(t)$ in (4.3) is given by (2.9), except that the univariate distribution F is changed to a multi-dimensional distribution. The assumption in Theorem 6 naturally holds if we force balance the group size in the clustering process. By applying the full strength $t = d$ OA to Theorem 6, we have the following corollary.

Corollary 2. *For $\tilde{U}_{oa}$ based on $OA(m, d, L, d)$, for any pair of (g, F), we have*

$$\mathrm{MSE}(\tilde{U}_{oa}) = \mathrm{MSE}(U_0) + o(m^{-1}). \tag{4.4}$$

The choice of t has been discussed and is illustrated in Examples 2 and 4.

Table 4. Result of Example 5.

m	100	400	900	1,600	2,500	3,600	4,900	6,400	8,100	10,000
MSE(U_{RND})	0.765	0.191	0.0903	0.0515	0.0260	0.0195	0.0167	0.0137	0.0098	0.0089
MSE($\tilde{U}_{oa}$)	0.075	0.0096	0.0032	0.0015	0.00063	0.00035	0.00023	0.00014	0.00011	0.00009

We do not compare different t in the following examples because $d = 2$ always holds, and so $t \leq 2$. We always take $t = 2$, $L = 10, 20, \ldots, 100$, and $m = L^t$.

Example 5. (Kendall's tau, Chen and Kengo (2019)). The Kernel function $h((x_1, y_1), (x_2, y_2)) = 2I(x_1 < x_2, y_1 < y_2) + 2I(x_2 < x_1, y_2 < y_1) - 1$. For simplicity, we assume that (X, Y) follows a normal distribution, with $\mu = (0, 0)$ and $\Sigma = \mathrm{diag}(3, 1)$. Set $n = 10^4$. The MSE when estimating the Kendall correlation using U_{RND} and $\tilde{U}_{oa}$ is shown in Table 4. As a reference, we have $\mathrm{MSE}(U_0) = 8.97 \times 10^{-5}$.

Example 6. (Testing stochastic monotonicity, Lee, Linton and Whang (2009)). Let (X, Y) be a real-valued random vector, and denote by $F_{Y|X}(y|x)$ the conditional distribution function of Y, given X. Consider the problem of testing the stochastic monotonicity hypothesis

$$H_0 : F_{Y|X}(y|x) \leq F_{Y|X}(y|x'), \forall y \in R \text{ and whenever } x \geq x'.$$

This essentially tests where an increase in X would induce an increase in Y (e.g., income vs. expenditure in a household). Lee, Linton and Whang (2009) proposed the following testing statistic:

$$\begin{aligned} U_n(x, x') = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} &(I\{Y_i \leq x'\} - I\{Y_j \leq x'\}) \\ &\mathrm{sign}(X_i - X_j) K(x - X_i) K(x - X_j), \end{aligned} \tag{4.5}$$

where $K(x) = 0.75(1 - x^2)$. We simulate (X, Y) from a normal distribution with $\mu = (0, 0)$ and $\Sigma = \mathrm{diag}(3, 1)$, and calculate (4.5) at $(x, x') = (0, 0)$. For $n = 10^4$, the comparison between $\tilde{U}_{oa}$ and U_{RND} is given in Table 5. As a reference, we have $\mathrm{MSE}(U_0) = 2.572$.

Example 7. (Clustering performance evaluation, Papa, Clémençon and Bellet (2015)). For a given distance $D : \mathcal{X} \times \mathcal{X} \to R$ defined on $\mathcal{X}$, the performance of a partition P can be evaluated from the data $X_1, \ldots, X_n \in \mathcal{X}$ using

$$W(P) = \sum_{1 \leq i < j \leq n} D(X_i, X_j) \cdot \sum_{\mathcal{C} \in P} I\{(X_i, X_j) \in \mathcal{C}^2\}. \tag{4.6}$$

Table 5. Result of Example 6.

m	100	400	900	1,600	2,500	3,600	4,900	6,400	8,100	10,000
MSE(U_{RND})	302.7	69.01	38.01	17.45	12.86	8.613	7.438	6.273	4.886	4.327
MSE($\tilde{U}_{oa}$)	33.18	15.73	8.848	4.252	3.524	3.168	2.732	2.662	2.630	2.602

Table 6. Result of Example 7.

m	100	400	900	1,600	2,500	3,600	4,900	6,400	8,100	10,000
MSE(U_{RND})	0.216	0.0625	0.0346	0.0171	0.0064	0.0047	0.0038	0.0021	0.0017	0.0010
MSE($\tilde{U}_{oa}$)	0.011	0.0064	0.0038	0.0019	0.00056	0.00051	0.00038	0.00027	0.00013	0.00012

Our purpose is to compare the different incomplete U-statistics of (4.6); here, we focus on the k-means method for the comparison. The data are generated from a normal distribution with $\mu = (0, 0)$ and $\Sigma = \mathrm{diag}(1, 2)$, and we divide the data into two groups. The MSE of U_{RND} and $\tilde{U}_{oa}$ when estimating $W(P)$ for different m is shown in Table 6. As a reference, we have $\mathrm{MSE}(U_0) = 1.043 \times 10^{-4}$.

5. Conclusion

To tackle the computational issue of U-statistics, we have introduced a new type of incomplete U-statistic called the ICUDO, which has much higher efficiency than existing methods. The required computational burden, as indexed by the number of combinations m for the ICUDO to be statistically equivalent to the complete U-statistic, is of smaller magnitude than existing methods. This was validated theoretically and empirically for degenerate and non-degenerate one- and multi-sample U-statistics on univariate and multi-dimensional data. In fact, m is allowed to grow more slowly than the data size n in the non-degenerate case.

The OA plays a critical role in the construction of the ICUDO, in light of its projective uniformity. Other space-filling design schemes exist with similar properties, such as the OA-based Latin hypercube by Tang (1993), and the strong orthogonal array by He and Tang (2012), which is used frequently in the design of computer experiments. By exhaustive simulations, we find the improvement of the efficiency by these design schemes over that of the ICUDO to be within 1%. However, this improvement is not sufficient to advocate using these structures, owing to the extra complexity of the computation. Other improvements over the OA are based on optimal criteria, such as the generalized minimum aberration OA. However, no theoretical results are available for these fixed structures.

Lastly, the following offer potential future research directions. (i) For high-dimensional data, dimension-reduction techniques need to be integrated into our current algorithm. (*ii*) For multi-sample cases, we may divide different samples into different numbers of groups in some optimal way. This will induce more complicated OA structures. (*iii*) For the purpose of statistical inference, it would be of interest to study the asymptotic distributions of the ICUDO under different conditions. (*iv*) The dimension of the kernel functions is fixed at d as n increases, and all data are generated independently. In one important type of U-statistic based on stochastic processes, d increases with n and the data can be dependent. These topics will involve quite different methodologies, and hence are left to future work.

Supplementary Material

The online Supplementary Material generalizes the result of Theorem 2 under additional conditions. It also provides details on how to choose the combination of L and t and illustrates the generation of the ICUDO for multi-sample and multi-dimensional cases.

Acknowledgments

Dr. Kong's research was partially supported by NSFC grant 11801033 and the Beijing Institute of Technology Research Fund Program for Young Scholars. Dr. Zheng's research was partially supported by the National Science Foundation, DMS-1830864.

Appendix

A. Proof of Theorems

Lemmas 1–3 contribute to the proof of Theorem 1. Theorems 4 and 6 can be proved similarly as Theorem 1, but only with more tedious analysis, and hence they are omitted due to the limit of space. For any $\boldsymbol{a} \in \mathcal{Z}_L^d$, we call the set $\mathcal{G}_{\boldsymbol{a}} = \prod_{j=1}^d G_{a_j}$ a *grid*. Let F_n be the empirical distribution of $\{X_1, \ldots, X_n\}$ and define $V = \int g(x_1, \ldots, x_d) dF_n(x_1) \ldots dF_n(x_d)$. For given F_n and $\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}$, define

$$\bar{g}(\mathcal{X}_{\boldsymbol{\eta}}) = |\mathcal{G}_{\boldsymbol{a}}|^{-1} \sum_{\boldsymbol{\eta}' \in \mathcal{G}_{\boldsymbol{a}}} g(\mathcal{X}_{\boldsymbol{\eta}'}).$$

For the same $S_{oa} = \{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^m\}$ in generating U_{oa}, define

$$\bar{V} = \frac{1}{m} \sum_{i=1}^m \bar{g}(\mathcal{X}_{\boldsymbol{\eta}^i}).$$

Lemma 1. *Some properties of V and $\bar{V}$ are listed as follows.*

(i) $\bar{V}$ *is an unbiased estimator of* V.

(ii) The bias of V *is of order* $O(n^{-1})$ *and* $\mathrm{MSE}(V) = \mathrm{MSE}(U_0) + O(n^{-2})$.

(iii) U_{oa} *is an unbiased estimator of* V *and so also has bias* $O(n^{-1})$.

Proof. (i) follows the unbiasedness of orthogonal arrays. (ii) can be found in Proposition 3.5 in (Shao, 2003, p.211). (iii) follows from Owen (1992).

Lemma 2.

$$E(\bar{V} - V)^2 \leq \frac{1}{m} \sum_{\boldsymbol{u}:|\boldsymbol{u}|>t} \left(\delta_{\boldsymbol{u}}^2 + O(n^{-1})\right).$$

Proof. Let $\delta_{\boldsymbol{u}}^2 = \delta_{|\boldsymbol{u}|}^2$ and $\sigma_{\boldsymbol{u}}^2 = \sigma_{|\boldsymbol{u}|}^2$.Change the F in section 2.2 to F_n, we can define $dF_{n,\boldsymbol{u}}$, $g_{n,\boldsymbol{u}}$, $h_{n,\boldsymbol{u}}$, $\sigma_{n,\boldsymbol{u}}^2$ and $\delta_{n,\boldsymbol{u}}^2$ analogously and sequentially. Again, by substituting $\bar{g}$ for g, with F_n, we define $\bar{g}_{n,\boldsymbol{u}}$, $\bar{h}_{n,\boldsymbol{u}}$, $\bar{\sigma}_{n,\boldsymbol{u}}^2$ and $\bar{\delta}_{n,\boldsymbol{u}}^2$. Adopt (3.5) in Owen (1992) to $\bar{g}$, we have

$$E[(\bar{V} - V)^2 | F_n] \leq \frac{1}{m} \sum_{\boldsymbol{u}:|\boldsymbol{u}|>t} \bar{\delta}_{n,\boldsymbol{u}}^2 \leq \frac{1}{m} \sum_{\boldsymbol{u}:|\boldsymbol{u}|>t} \delta_{n,\boldsymbol{u}}^2,$$

which leads to $E(\bar{V}-V)^2 = E(E[(\bar{V}-V)^2|F_n]) \leq (1/m) \sum_{\boldsymbol{u}:|\boldsymbol{u}|>t} E\delta_{n,\boldsymbol{u}}^2$. Consider $\sigma_{n,\boldsymbol{u}}^2 = \int g_{n,\boldsymbol{u}}^2(x_1, \ldots, x_d) dF_n(x_1) \ldots dF_n(x_n)$, which can be further written as $\int \left(\int g_{n,\boldsymbol{u}}^2 dF_{n,\boldsymbol{u}^c}\right)^2 dF_{n,\boldsymbol{u}}$. This integer can be viewed as a V-statistic with the new kernel $g(x_1, \ldots, x_{|\boldsymbol{u}|}, x_{|\boldsymbol{u}|+1}, \ldots, x_d) \cdot g(x_1, \ldots, x_{|\boldsymbol{u}|}, x_{d+1}, \ldots, x_{2d-|\boldsymbol{u}|})$, which estimates $\sigma_{\boldsymbol{u}}^2$ with bias $O(n^{-1})$.

Lemma 3. (Lusin's theorem)
For any measurable function g on R^d and arbitrary $\epsilon > 0$, there exists a continuous g_ϵ defined on R^d with compact support such that $E|g - g_\epsilon| < \epsilon$.

Proof of Theorem 1. Define $g_F(Z_1, \ldots, Z_d) = g(F^{-1}(Z_1), \ldots, F^{-1}(Z_d))$ such that $Z \sim U(0,1)$ and $F^{-1}(Z) \sim F$. With this new kernel g_F, the distribution of random variables X is assumed to be the uniform distribution on $[0,1]$.

Write $U_{oa} - \Theta$ as $(U_{oa} - \bar{V}) + (\bar{V} - V) + (V - \Theta)$. Simple analysis reveals the following relationships among of V_{oa}, $\bar{V}$ and V. Conditional on F_n, V is constant and so $E(U_{oa} - \bar{V})(V - \Theta) = 0$, $E(\bar{V} - V)(V - \Theta) = 0$ since $E(U_{oa} - \bar{V}) = E(\bar{V} - V) = 0$. Conditional on both V and $\bar{V}$, $E(U_{oa} - \bar{V}) = 0$ which indicates $E(U_{oa} - \bar{V})(\bar{V} - V) = 0$. Thus,

$$\mathrm{MSE}(U_{oa}) = E(U_{oa} - \bar{V})^2 + E(\bar{V} - V)^2 + \mathrm{MSE}(V) \tag{A.1}$$

whose last two terms have been addressed by Lemma 2 and Lemma 1. So we need to prove $E(U_{oa}-\bar{V})^2 = o(m^{-1})$. Since U_{oa} and $\bar{V}$ always use the same $S_{oa}=\{\boldsymbol{\eta}^1,\ldots,\boldsymbol{\eta}^m\}$,

$$E(U_{oa}-\bar{V})^2 = E\left(\frac{1}{m}\sum_{i=1}^{m} g(\mathcal{X}_{\boldsymbol{\eta}^i}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}^i})\right)^2.$$

For $i_1 \neq i_2$ $(i_1, i_2 \in \mathcal{Z}_m)$, $E(g(\mathcal{X}_{\boldsymbol{\eta}^{i_1}}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}^{i_1}}))(g(\mathcal{X}_{\boldsymbol{\eta}^{i_2}}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}^{i_2}})) = 0$. Denote $\boldsymbol{\eta} \sim \boldsymbol{\eta}'$ if $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$ belong to the same grid.

$$E(U_{oa}-\bar{V})^2 \leq 2m^{-1}E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta}\sim\boldsymbol{\eta}']. \tag{A.2}$$

For any $M>0$, define $g(\boldsymbol{x},M) = \max\{\min\{g(\boldsymbol{x}),M\},-M\}$. Obviously, we have $\lim_{M\to\infty} g(\boldsymbol{x},M) = g(\boldsymbol{x})$, and dominated convergence theorem indicates

$$E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta}\sim\boldsymbol{\eta}'] = \lim_{M\to\infty} E[(g(\mathcal{X}_{\boldsymbol{\eta}},M) - g(\mathcal{X}_{\boldsymbol{\eta}'},M))^2|\boldsymbol{\eta}\sim\boldsymbol{\eta}']. \tag{A.3}$$

Thus, for arbitrary $\epsilon>0$, we can find M_ϵ such that

$$E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta}\sim\boldsymbol{\eta}'] \leq E[(g(\mathcal{X}_{\boldsymbol{\eta}},M_\epsilon) - g(\mathcal{X}_{\boldsymbol{\eta}'},M_\epsilon))^2|\boldsymbol{\eta}\sim\boldsymbol{\eta}'] + \epsilon. \tag{A.4}$$

Note that $\{X_1,\ldots,X_n\}$ are random, so is $\mathcal{X}_{\boldsymbol{\eta}}$. Note that $Eg^2(X_1,\ldots,X_d)<\infty$. We have $Eg^2(\mathcal{X}_{\boldsymbol{\eta}})<\infty$ and so $Eg(\mathcal{X}_{\boldsymbol{\eta}})<\infty$, which indicates $Eg^2(\mathcal{X}_{\boldsymbol{\eta}},M_\epsilon)<\infty$ and $Eg(\mathcal{X}_{\boldsymbol{\eta}},M_\epsilon)<\infty$. From Lusin's theorem, there exists a continuous $g^*_{\epsilon,M_\epsilon}$ with compact support such that $E|g(\mathcal{X}_{\boldsymbol{\eta}},M) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}})| < \epsilon M_\epsilon^{-1}$. Since $|g(\mathcal{X}_{\boldsymbol{\eta}},M_\epsilon)| \leq M_\epsilon$,

$$\begin{aligned}
&E[(g(\mathcal{X}_{\boldsymbol{\eta}},M_\epsilon) - g(\mathcal{X}_{\boldsymbol{\eta}'},M_\epsilon))^2|\boldsymbol{\eta}\sim\boldsymbol{\eta}'] \\
&\leq 2M_\epsilon E[|g(\mathcal{X}_{\boldsymbol{\eta}},M_\epsilon) - g(\mathcal{X}_{\boldsymbol{\eta}'},M_\epsilon)||\boldsymbol{\eta}\sim\boldsymbol{\eta}'] \\
&\leq 2M_\epsilon E|g(\mathcal{X}_{\boldsymbol{\eta}},M_\epsilon) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}})| + 2M_\epsilon E|g(\mathcal{X}_{\boldsymbol{\eta}'},M_\epsilon) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}'})| + \\
&\quad 2M_\epsilon E[|g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}}) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}'})||\boldsymbol{\eta}\sim\boldsymbol{\eta}'] \\
&\leq 4\epsilon + 2M_\epsilon E[|g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}}) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}'})||\boldsymbol{\eta}\sim\boldsymbol{\eta}']
\end{aligned} \tag{A.5}$$

Note that $g^*_{\epsilon,M_\epsilon}$ has compact support and so is uniformly continuous. There exists $\Delta(M_\epsilon^{-1}\epsilon)$ such that $|g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}}) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}'})| \leq \epsilon M_\epsilon^{-1}$ as long as $||\mathcal{X}_{\boldsymbol{\eta}} - \mathcal{X}_{\boldsymbol{\eta}'}||_2 \leq \Delta(M_\epsilon^{-1}\epsilon)$. Define

$$\mathcal{A} = \{|\mathcal{X}_{\eta_j} - \mathcal{X}_{\eta'_j}| \geq d^{-1}\Delta(M_\epsilon^{-1}\epsilon) \text{ for some } j\in\mathcal{Z}_d\},$$

with $P(\mathcal{A}) \leq \sum_{j=1}^d P\{|\mathcal{X}_{\eta_j} - \mathcal{X}_{\eta'_j}| \geq d^{-1}\Delta(M_\epsilon^{-1}\epsilon)\}$, and $||\mathcal{X}_{\boldsymbol{\eta}} - \mathcal{X}_{\boldsymbol{\eta}'}||_2 \leq \Delta(M_\epsilon^{-1}\epsilon)$

on $\mathcal{A}^c$.

$$
\begin{aligned}
&2M_\epsilon E[|g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}}) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}'})||\boldsymbol{\eta} \sim \boldsymbol{\eta}']\\
&= 2M_\epsilon P(\mathcal{A}^c)E[|g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}}) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}'})||\boldsymbol{\eta} \sim \boldsymbol{\eta}', \mathcal{A}^c]\\
&\quad +2M_\epsilon P(\mathcal{A})E[|g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}}) - g^*_{\epsilon,M_\epsilon}(\mathcal{X}_{\boldsymbol{\eta}'})||\boldsymbol{\eta} \sim \boldsymbol{\eta}', \mathcal{A}]\\
&\leq 2\epsilon + 4M_\epsilon^2 \sum_{k=1}^{d} P\{|\mathcal{X}_{\eta_j} - \mathcal{X}_{\eta'_j}| \geq d^{-1}\Delta(M_\epsilon^{-1}\epsilon)\}
\end{aligned}
\tag{A.6}
$$

Now we give the relationship among several events. For $j \in \mathcal{Z}_d$ and $\boldsymbol{\eta} \sim \boldsymbol{\eta}'$,

$$
\begin{aligned}
&\{|\mathcal{X}_{\eta_j} - \mathcal{X}_{\eta'_j}| \geq d^{-1}\Delta(M_\epsilon^{-1}\epsilon)\}\\
&= \{|\mathcal{X}_{\eta_j} - F_n(\mathcal{X}_{\eta_j}) + F_n(\mathcal{X}_{\eta_j}) - F_n(\mathcal{X}_{\eta'_j}) + F_n(\mathcal{X}_{\eta'_j}) - \mathcal{X}_{\eta'_j}| \geq d^{-1}\Delta(M_\epsilon^{-1}\epsilon)\}\\
&\subseteq \left\{\sup_{x\in(0,1)} |x - F_n(x)| \geq \frac{1}{3d}\Delta(M_\epsilon^{-1}\epsilon)\right\} \cup \left\{F_n(\mathcal{X}_{\eta_j}) - F_n(\mathcal{X}_{\eta'_j}) \geq \frac{1}{3d}\Delta(M_\epsilon^{-1}\epsilon)\right\}
\end{aligned}
$$

Note that $\boldsymbol{\eta} \sim \boldsymbol{\eta}'$, as $L \to \infty$, $P(\{F_n(\mathcal{X}_{\eta_j}) - F_n(\mathcal{X}_{\eta'_j}) \geq (1/3d)\Delta(M_\epsilon^{-1}\epsilon)\}) \to 0$. Dvoretzky-Kiefer-Wolfowitz inequality reveals $P\left(\sup_{x\in(0,1)} |F_n(x) - x| \geq \epsilon\right) \leq \exp(-2n\epsilon^2)$. So we immediately have $P(\{|\mathcal{X}_{\eta_j} - \mathcal{X}_{\eta'_j}| \geq d^{-1}\Delta(M_\epsilon^{-1}\epsilon)\}) \to 0$ as $n, L \to \infty$, and we can find n_ϵ and L_ϵ such that

$$
P(\{|\mathcal{X}_{\eta_j} - \mathcal{X}_{\eta'_j}| \geq d^{-1}\Delta(M_\epsilon^{-1}\epsilon)\}) \leq (4dM_\epsilon^2)^{-1}\epsilon \tag{A.7}
$$

as long as $n \geq n_\epsilon$ and $L \geq L_\epsilon$.

Finally, by combining (A.3)-(A.7), we know that for arbitrary $\epsilon > 0$, we can find n_ϵ and L_ϵ such that $E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta} \sim \boldsymbol{\eta}'] \leq 8\epsilon$, as long as $n \geq n_\epsilon$ and $L \geq L_\epsilon$. That means

$$
E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta} \sim \boldsymbol{\eta}'] \to 0 \tag{A.8}
$$

as $n, L \to \infty$. Theorem 1 is concluded by submitting (A.8) into (A.2) and combining (A.2) with (A.1), Lemma 1(ii) and Lemma 2.

Proof of Theorem 2. There exists $c > 0$ such that density function $f(\cdot) > c$ on $[a, b]$, and $|F(x_1) - F(x_2)| \geq c|x_1 - x_2|$ for $x_1, x_2 \in [a, b]$. In (A.1), we only analyze $E(U_{oa} - \bar{V})^2$ since the rest two terms are given by Lemma 1(ii) and Lemma 2. Dvoretzky-Kiefer-Wolfowitz inequality reveals $P\left(\sup_{x\in R} |F_n(x) - F(x)| \geq \epsilon\right) \leq \exp(-2n\epsilon^2)$. By taking $\epsilon = [\log(n)n^{-1}]^{1/2}$, we have

$$
P(\mathcal{A}) \leq \exp(-2\log n) = O(n^{-2}),
$$

where $\mathcal{A} = \{\sup_{x\in R} |F_n(x) - F(x)| \geq n^{-1/2}\log^{1/2}(n)\}$. Since g is continuous and F is bounded, we can find $M > 0$ such that $|g| \leq M$ and so $|U_{oa}|, |\bar{V}| \leq M$.

$$\begin{aligned}
&E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta} \sim \boldsymbol{\eta}']\\
&= P(\mathcal{A})E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta} \sim \boldsymbol{\eta}', \mathcal{A}] + P(\mathcal{A}^c)E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta} \sim \boldsymbol{\eta}', \mathcal{A}^c]\\
&\leq M^2 n^{-2} + E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta} \sim \boldsymbol{\eta}', \mathcal{A}^c]
\end{aligned}$$

The analysis of $E[(U_{oa} - \bar{V})^2|\mathcal{A}^c]$ is as follows. On $\mathcal{A}^c$, we have, for $1 \leq k_1, k_2 \leq nL^{-1}$,

$$\begin{aligned}
&c|X_{((l-1)nL^{-1}+k_1)} - X_{((l-1)nL^{-1}+k_2)}|\\
&\leq |F(X_{((l-1)nL^{-1}+k_1)}) - F(X_{((l-1)nL^{-1}+k_2)})|\\
&\leq |F_n(X_{((l-1)nL^{-1}+k_1)}) - F_n(X_{((l-1)nL^{-1}+k_2)})| + 2n^{-1/2}\log^{1/2} n\\
&\leq L^{-1} + 2n^{-1/2}\log^{1/2} n.
\end{aligned}$$

Since g is Lipschitz continuous, we know $(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2 = O(L^{-2} + n^{-1}\log n)$ for any $\boldsymbol{\eta} \sim \boldsymbol{\eta}'$. Then we have $E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - g(\mathcal{X}_{\boldsymbol{\eta}'}))^2|\boldsymbol{\eta} \sim \boldsymbol{\eta}', \mathcal{A}^c] = O(L^{-2} + n^{-1}\log n)$. With this equation, Theorem 8 is the direct result of (A.1) (A.2), Lemma 1(ii), Lemma 2.

Proof of Theorem 3. For convenience, we simply write g_F as g in this proof. In (A.1), we only analyze $E(U_{oa} - \bar{V})^2$ since the rest two terms are given by Lemma 1 (ii) and Lemma 2. Each row of the matrix A generated in step 1 follows the uniform distribution on $\mathcal{Z}_L^d$ since the permutation in each column of A_0 is independent. Thus,

$$\begin{aligned}
E(U_{oa} - \bar{V})^2 &= E\left(\frac{1}{m}\sum_{i=1}^{m} g(\mathcal{X}_{\boldsymbol{\eta}^i}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}^i})\right)^2\\
&= \frac{1}{mL^d}\sum_{\boldsymbol{a}\in\mathcal{Z}_L^d} E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}}))^2|\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}].
\end{aligned}$$

Analysis is now focused on $E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}}))^2|\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}]$ for every $\boldsymbol{a} \in \mathcal{Z}_L^d$. Let $X_{(0)} = 0$ and $X_{(n+1)} = 1$. For $l \in \mathcal{Z}_L$, given $X_{((l-1)nL^{-1})}$ and $X_{(lnL^{-1}+1)}$, $X_{((l-1)nL^{-1}+1)}, \ldots, X_{(lnL^{-1})}$ has the same distribution as the order statistic of L samples following the uniform distribution on $[X_{((l-1)nL^{-1})}, X_{(lnL^{-1}+1)}]$. For $\mathcal{A} = \{\sup_{x\in R} |F_n(x) - F(x)| \geq n^{-((1-c)/2)}\}$, Dvoretzky-Kiefer-Wolfowitz inequality reveals $P(\mathcal{A}) = \exp(-2n^c)$. On $\mathcal{A}^c$, we have $(X_{(lnL^{-1}+1)} - X_{((l-1)nL^{-1})})/L \to 1$ as $n \to \infty$. The analysis is now focused on $E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}}))^2|\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}, \mathcal{A}^c]$. For this given $\boldsymbol{a}$, define $\mathcal{X}_0 = (X_{0,1}, \ldots, X_{0,d})$ where $X_{0,j} = (L/n)\sum_{\eta \in G_{a_j}} X_\eta$ and so $\sum_{\eta \in G_{a_j}}(X_\eta - X_{0,j}) = 0$. Adopt the Taylor expansion on $\mathcal{X}_0$, we have

$$g(\mathcal{X}_{\boldsymbol{\eta}}) = g(\mathcal{X}_0) + \sum_{j=1}^{d} \frac{\partial g}{\partial x_j}\bigg|_{X_{0,j}} (X_{\eta_j} - X_{0,j}) + O(L^{-2}) \text{ and } \bar{g}(\mathcal{X}_{\boldsymbol{\eta}}) = g(\mathcal{X}_0) + O(L^{-2}).$$

$$\begin{aligned}
E&[(g(\mathcal{X}_{\boldsymbol{\eta}}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}}))^2 | \boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}, \mathcal{A}^c] \\
&= E\left[\left(\sum_{j=1}^{d} \frac{\partial g}{\partial x_j}\bigg|_{X_{0,j}} \cdot (X_{\eta_j} - X_{0,j}) + O(L^{-2})\right)^2 | \boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}, \mathcal{A}^c\right] \\
&= o(L^{-2}) + \sum_{j=1}^{d} E\left[\left(\frac{\partial g}{\partial x_j}\bigg|_{X_{0,j}} \cdot (X_{\eta_j} - X_{0,j})\right)^2 | \boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}, \mathcal{A}^c\right] \\
&= o(L^{-2}) + \sum_{j=1}^{d} \left(\frac{\partial g}{\partial x_j}\bigg|_{X_{0,j}}\right)^2 \frac{1}{12L^2}.
\end{aligned}$$

And then we have

$$\begin{aligned}
E(U_{oa} - \bar{V})^2 &= \frac{1}{mL^d} \sum_{\boldsymbol{a} \in \mathcal{Z}_L^d} E[(g(\mathcal{X}_{\boldsymbol{\eta}}) - \bar{g}(\mathcal{X}_{\boldsymbol{\eta}}))^2 | \boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}] \\
&= \frac{1}{12mL^2} \sum_{j=1}^{d} \left(\frac{1}{L^d} \sum_{\boldsymbol{a} \in \mathcal{Z}_L^d} \left(\frac{\partial g}{\partial x_j}\bigg|_{X_{0,j}}\right)^2\right) + o\left(\frac{1}{mL^2}\right) \\
&= \frac{1}{12mL^2} \sum_{j=1}^{d} E\left(\frac{\partial g}{\partial x_j}\right)^2 + o\left(\frac{1}{mL^2}\right),
\end{aligned}$$

Then Theorem 3 is the direct result of (A.1), Lemma 1(ii) and Lemma 2.

Proof of Theorem 5. Consider the m rows of A, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$, generated in the step 1 of the construction in section 2.1. For any $\boldsymbol{a} \in \mathcal{Z}_L^d$, the random permutation in generating $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$ reveals that $P(\boldsymbol{a}_1 = \boldsymbol{a}) = L^{-d}$. Given F_n,

$$\begin{aligned}
E(\tilde{U}_{oa}|F_n) &= E\frac{1}{m}\sum_{i=1}^{m} \omega_{\boldsymbol{\eta}^i} g(\mathcal{X}_{\boldsymbol{\eta}^i}) = E\omega_{\boldsymbol{\eta}^1} g(\mathcal{X}_{\boldsymbol{\eta}^1}) \\
&= \sum_{\boldsymbol{a} \in \mathcal{Z}_L^d} L^{-d} E_{\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}} \omega_{\boldsymbol{\eta}} g(\mathcal{X}_{\boldsymbol{\eta}}) = \sum_{\boldsymbol{a} \in \mathcal{Z}_L^d} \frac{|\mathcal{G}_{\boldsymbol{a}^i} \cap S_0^*|}{|S_0^*|} E_{\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}} g(\mathcal{X}_{\boldsymbol{\eta}}) \\
&= \sum_{\boldsymbol{a} \in \mathcal{Z}_L^d} \frac{|\mathcal{G}_{\boldsymbol{a}^i} \cap S_0^*|}{|S_0^*|} \left(\frac{1}{|\mathcal{G}_{\boldsymbol{a}^i} \cap S_0^*|} \sum_{\boldsymbol{\eta} \in \mathcal{G}_{\boldsymbol{a}}} g(\mathcal{X}_{\boldsymbol{\eta}})\right) = \frac{1}{|S_0^*|} \sum_{\boldsymbol{\eta} \in S_0^*} g(\mathcal{X}_{\boldsymbol{\eta}}) = U_0.
\end{aligned}$$

Since U_0 is unbiased, so is $\tilde{U}_{oa}$. This proves the unbiasedness of $\tilde{U}_{oa}$. The MSE of $\tilde{U}_{oa}$ can be similar analyzed as Theorem 1, and so is omitted here.

References

Bickel, P. J. and Freedman, D.(1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **9**, 1196–1217.

Blom, G. (1976). Some properties of incomplete U-statistic. *Biometrika* **63**, 573–580.

Bretagnolle, J. (1983). Lois limits du Bootstrap de certaines fonctionnelles. *Annales de l'Institut Henri Poincare (B) Probability and Statistics* **19**, 281–296.

Brown, B. M. and Kildea, D. G. (1978). Reduced U-statistic and the Hodges-Lehmann estimator. *The Annals of Statistics* **6**, 828–835.

Chen, X. H. and Kengo Kato (2019). Randomized incomplete U-statistic in high dimensions. *The Annals of Statistics* **47**, 3127–3156..

Chen, W., Liu, T. Y., Lan, Y. Y., Ma, Z. M. and Li, H. (2009). Ranking Measures and Loss Functions in Learning to Rank. *Advances in Neural Information Processing Systems 22 (NIPS 2009).*

Dehling, H. and Mikosch, T. (1994). Random quadratic forms and the bootstrap for U-statistic. *Journal of Multivariate Analysis* **51**, 392–413.

Enqvist, E. (1985). A note on incomplete U-statistic for stationary absolutely regular processes. *Contributions to Probability and Statistics in Honour of Gunnar Blom*, 97–103. Lund University, Lund.

Eagleson, G. K. (1979). Orthogonal expansions and U-statistic. *Australian & New Zealand Journal of Statistics* **21**, 221–237.

Herrndorf, N. (1986). An invariance principle for reduced U-statistic. *Metrika* **33**, 179–188.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325.

He, Y. Z. and Tang, B. X. (2012). Strong orthogonal arrays and associated Latin hypercubes for computer experiments. *Biometrika* **100**, 254–260.

Hilbert, M. and López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science* **332**, 60–65.

Janson, S. (1984). The asymptotic distributions of incomplete U-statistic. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **21**, 1159–1169.

Lee, A. J. (1979). On the asymptotic distribution of U-statistic. *Institute of Statistics Mimeo Series* **1255**. University of North Carolina at Chapel Hill.

Lee, A. J. (1982). On incomplete U-statistic having minimum variance. *Australian & New Zealand Journal of Statistics* **24**, 275–282.

Lee, A. J. (1990). *U-Statistic: Theory and Practice*. CRC Press, Boca Raton.

Lee, S., Linton, O. and Whang, Y. (2009). Testing for stochastic monotonicity. *Econometrica* **27**, 585–602.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Statistics* **22**, 165–179.

Lin, N. and Xi, R. (2010). Fast surrogates of U-statistic. *Computational Statistics and Data Analysis* **54**, 16–24.

Huskova, M. and Janssen, P. (1993a). Consistency of the generalized bootstrap for degenerate

U-statistic. *The Annals of Statistics* **21**, 1811–1823.

Huskova, M. and Janssen, P. (1993b). Generalized bootstrap for studentized U-statistic: A rank statistic approach. *Statistics and Probability Letters* **16**, 225–233.

Owen, A. B. (1992). Orthogonal arrays for computer experiments. *Statistica Sinica* **2**, 439–452.

Papa, G., Clémençon, S. and Bellet, A. (2015). SGD algorithms based on incomplete U-statistics: Large-scale minimization of empirical risk. *Advances in Neural Information Processing Systems 28 (NIPS 2015).*

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments.* Wiley, New York.

Rempala, G. and Wesolowski, J. (2003). Incomplete U-statistic of permanent design. *Journal of Nonparametric Statistics* **15**, 221–236.

Rempala, G. and Srivastav, S. (2004). Minimum variance rectangular designs for U-statistic. *Journal of Statistical Planning and Inference* **102**, 103–118.

Sen, P. K. (1974). Weak convergence of generalized U-statistic. *The Annals of Probability* **2**, 90–102.

Sen, P. K. (1977). Almost sure convergence of generalized U-statistic. *The Annals of Probability* **5**, 287–290.

Shao, J. (2003). *Mathematical Statistics.* 2nd Edition. Springer Science-Business Media, New York.

Tang, B. X. (1993). Orthogonal array based Latin hypercubes. *Journal of the American Statistical Association* **88**, 1392–1397.

von Mises, R. (1948). On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics* **18**, 309–348.

Wang Q. (2012). Investigation of topics in U-statistics and their applications in risk estimation and cross-validation. Ph.D. dissertation, 1–187. Penn State University library electronic resource.

Wang Q and Lindsay B. (2014). Variance estimation of a general U-statistic with application to cross-validation. *Statistica Sinica* **24**, 1117–1141.

Xiangshun Kong

School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100811, China.

E-mail: kongsunday@163.com

Wei Zheng

Department of Business Analytics and Statistics, University of Tennessee, Knoxville, TN 37996, USA.

E-mail: wzheng9@utk.edu

(Received March 2019; accepted November 2019)