Asymptotic Normality and Rates of Convergence for Random Forests via Generalized U-statistics

by

Wei Peng

B.S. in Statistics, Nanjing University, 2015

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Wei Peng

It was defended on

03/26/2021

and approved by

Lucas Mentch, Department of Statistics

Satish Iyengar, Department of Statistics

Zhao Ren, Department of Statistics

Larry Wasserman, Department of Statistics and Data Science (Carnegie Mellon University)

Asymptotic Normality and Rates of Convergence for Random Forests via Generalized U-statistics

Wei Peng, PhD

University of Pittsburgh, 2021

Random forests are among the most popular off-the-shelf supervised learning algorithms. Despite their well-documented empirical success, however, until recently, few theoretical results were available to describe their performance and behavior. In this work we push beyond recent work on consistency and asymptotic normality by establishing rates of convergence for random forests and other supervised learning ensembles. We develop the notion of generalized U-statistics and show that within this framework, random forest predictions can remain asymptotically normal for larger subsample sizes and under weaker conditions than previously established. Moreover, we provide Berry-Esseen bounds in order to quantify the rate at which this convergence occurs, making explicit the roles of the subsample size and the number of trees in determining the distribution of random forest predictions. When these generalized estimators are reduced to their classical U-statistic form, the rates we establish are faster than any available in the existing literature. We also provide a consistency estimate of the variance of random forest and illustrate that quantifying the uncertainty of random forest is typically more expensive than obtaining the random forest itself.

Table of Contents

Preface ix

1.0 Introduction 1

2.0 Background 4

3.0 Asymptotic Normality 9

3.1 Asymptotic normality of generalized U-statistics 9

3.2 Variance ratio behavior 11

4.0 Berry-Esseen Bounds 16

4.1 Bounds for generalized U-statistics 16

4.2 A tighter bound 19

5.0 Variance Estimation 21

5.1 Introduction 21

5.2 Background of the infinitesimal jackknife (IJ) 24

5.3 The infinitesimal jackknife estimate for bootstrap ($\mathrm{IJ_B}$) 26

5.3.1 The convergence of three different approaches 27

5.3.2 The bias of $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ 31

5.3.3 The consistency of $\mathrm{IJ_B}$ 36

5.4 The pseudo infinitesimal jackknife estimate for U-statistic (s-$\mathrm{IJ_U}$) 37

5.4.1 IJ for U-statistic 37

5.4.2 s-$\mathrm{IJ_U}$ for U-statistic 38

5.4.3 The consistency of s-$\mathrm{IJ_U}$ 40

5.4.4 Higher order s-$\mathrm{IJ_U}$ 43

6.0 Discussion 46

Appendix A. Proofs in Chapter 2 47

Appendix B. Proofs in Chapter 3 50

B.1 H-decomposition 50

B.2 Proofs of asymptotic normality 52

B.3 Simple base learner variance ratios 59
B.4 Variance ratios of RP trees 63
Appendix C. Proofs in Chapter 4 66
C.1 Introduction to Lemma 4 66
C.2 Berry-Esseen bounds for generalized U-statistics 70
C.3 Discussion on a tighter bound 78
Appendix D. Proofs in Chapter 5 80
D.1 $\mathrm{IJ_B}$ for bootstrap 80
D.2 $\mathrm{IJ_U}$ and s-$\mathrm{IJ_U}$ for U-statistic 83
D.3 Discussion on extensions 91
Bibliography 98

List of Figures

3.1 The function of $c(k)$ for $k = 1, ..., 50$. $c(k)$ is monotonically decreasing as k increases and bounded with $1 < c(k) \leq 2$. 13

5.1 Performance of the infinitesimal jackknife and its bias-corrected alternatives on estimating the variance of the bagged sample mean (left: $B = 100$, right: $B = 1000$). 33

5.2 Performance of the infinitesimal jackknife and its bias-corrected alternatives on estimating the variance of the bagged sample variance (left: $B = 100$, right: $B = 1000$). 35

5.3 Performance of the infinitesimal jackknife and its bias-corrected alternatives on estimating the variance of the bagged sample maximum (left: $B = 100$, right: $B = 1000$). 36

5.4 The distribution of $\mathrm{U}_{n,k,N,\omega}$ as a function of N. $N \gg n$ is required to for s-$\mathrm{IJ}_{\mathrm{U}}^{\dagger}$ to estimate $\zeta_{1,\omega}$ consistently. 42

5.5 A plot of $\{r_j(d)\}_{j=1}^{k}$, where $n = 20$ and $k = 10$. As d increases, the curve of r_j is bending further away the horizontal line. 45

Dedication

This work is dedicated to my grandparents Shanyu Peng, Tianmei Gao, Jiping Wang, Yonglan Lu, for their endless love and unconditional support

Preface

I could not have done this work without help from many people.

First and foremost, I want to thank my advisor, Lucas Mentch, for always believing in me and encouraging me to pursue my interests, to dream big and to do my best. He provided the best academic resources that he could get to help me succeed. Not only have I learned how to be a good scholar from him, but aslo how to take good care of the family. I will definitely cherish and miss the precious afternoons we spent together discussing research problems.

I also want to thank Professor Larry Wasserman, Professor Zhao Ren, Professor Satish Iyengar for being my committee members: for Professor Larry Wasserman, I benefited a lot from his generous encouragement and valuable feedback; for Professor Zhao Ren, I wouldn't be selected into the PITT Statistics Ph.D. program without his strong recommendation; for Professor Satish Iyengar, the essence of probability I learned from his class helped me look beyond surface. And I would like to thank Professor Yu cheng, Professor Kehui Chen and Professor Allan Sampson for their kind help and support.

I am so grateful for my peers, Tim Coleman, Siyu Zhou, PeIliang Zhang, Zhexuan Li and Jiasheng Lu, who have helped me think through some of the difficult concepts in the thesis. I would also like to thank Marc Richards, Jack Werner and Sam Walczak who teamed up with me to win the 2021 NFL Big Data Bowl Champion. It brought me so much fun and fulfillment. And I really appreciate my friends for making my life more enjoyable.

I especially wish to express my gratitude to my high school math teacher Dekun He. He made me fall in love with math and build my confidence to solve any challenging problems. And I would like to thank Dr. Clifford Pickover. His book *The Mathematics of Oz: Mental Gymnastics from Beyond the Edge* opened the wonderland of mathematics to a 12 year-old child. I am also deeply grateful to my high school head teacher, Piling Mao. His words -"To do the things that one thinks are correct" has always inspired me to stay true to myself.

Finally, I would like to thank my parents, Guohua Peng and Tiyu Wang for their love, support and patience, and my younger sister Zhihan Zhai for her innocence and loveliness,

which brings me so many joyful moments. Most of all, I would like to thank Yunqi (Jane) Wang, for accompanying me through some of the most difficult time of my life and supporting me in many ways behind every of my accomplishment.

1.0 Introduction

The random forest algorithm is a supervised learning tool introduced by [12] that constructs many independently randomized decision trees and aggregates their predictions by averaging in the case of regression or taking a majority vote for classification. Random forests have been shown to successfully handle high-dimensional and correlated data while exhibiting appealing properties such as fast and accurate off-the-shelf fitting without the overfitting issues that often plague related methods. They have been successfully applied in a variety of scientific fields including remote sensing [2], computational biology [59], stock price forecasting [48], and forecasting bird migration [21]. In a recent large-scale empirical study comparing 179 classifiers across the 121 datasets comprising the entire UCI machine learning repository, [34] found that random forests performed extremely well with 3 of the top 5 algorithms being some variant of the standard procedure.

Despite their wide-ranging applicability and well-documented history of empirical success, establishing formal mathematical and statistical properties for random forests has proved quite difficult, due in large part to the complex, data-dependent nature of the CART-splitting criterion [13] traditionally used to construct individual trees. [12] provided the first such result, demonstrating that the expected misclassification rate is a function of the accuracy of the individual classifiers and the correlation between them. The bound on the misclassification rate postulated in the work is loose but suggestive in the sense that interplay between these two sets the foundation for understanding the inner-workings of the procedure. [1] established a limit law for the split location in a regression tree context with independent Gaussian noise. Further analysis of the behavior of CART-style splitting was conducted by [44] who demonstrated an end-cut preference, whereby splits along non-informative variables are more likely to occur near the edges of the feature space.

A variety of other work has focused on analyzing other properties of random forest ensembles or extending the methodology to related problem types. [51] developed the idea of potential nearest neighbors and demonstrated their relationship to tree-based ensembles. More recently [52] analyzed the tradeoff between the size of the ensemble and the classi-

fication accuracy. [8], [6], and [5] studied various idealized versions of random forests and investigated consistency while [26] proved consistency for a particular type of online forest. [46] developed the idea of random survival forests and the consistency of such models is investigated in [45] and [22]. [53] extended random forest estimates to the context of quantile regression and [76] experimented with reinforcement learning trees. For a more detailed accounting of random-forest-related research, we refer readers to an excellent recent review by [9].

In recent years, many promising developments have come by considering individual trees built with subsamples rather than the more traditional bootstrap samples. [71] extended the infinitesimal jackknife estimates of variance introduced by [31] to produce confidence intervals for subsampled random forest predictions. [63] provided the first consistency result for Breiman's original forests, establishing L^2 consistency whenever the underlying regression function is additive. [54] made the connection to infinite-order U-statistics and provided the asymptotic distributions of random forest predictions. [69] showed that for large ensembles, subsampled random forests are both asymptotically unbiased and Gaussian whenever individual trees are built according to honesty and regularity conditions.

In this paper, we continue the trend of establishing mathematical properties of random forests by building on the U-statistic connection made in [54]. As in other recent theoretical analyses on the topic (e.g. [8, 6, 5, 54, 69]), we adopt a general notion of random forests, viewing this class of estimators as those producing predictions of the form

$$\mathrm{RF}(x) = \frac{1}{N}\sum_{i=1}^{N} h(x; Z_{i1}, \ldots, Z_{is}; \omega)$$

where each $Z_{i1}, \ldots, Z_{is}$ denotes a subsample taken without replacement from the available training data and ω denotes additional randomness injected into the base learner h. In particular, we do not require that base learners be trees constructed via the CART methodology as originally proposed in [12]. We establish central limit theorems for such estimators, that, to our knowledge, cover a broader range of estimators and also allow for faster subsampling rates than established in existing literature. A consistent estimate of the variance of such estimators are as well provided. More notably, we take a step forward in the theoretical analysis of random forests by providing Berry-Esseen Theorems governing the rate at

which this convergence takes place by bounding the maximal error of approximation between the Gaussian distribution and that of the random forest predictions. In establishing these, we develop the notion of *generalized* U-statistics which allow for kernels to be incomplete, randomized, and infinite-order. Importantly, when these estimators are simplified to their classical U-statistic form, the resulting bounds we provide are faster than any in the existing literature.

The remainder of this paper is organized as follows. In Chapter 2, we provide additional background on the random forest algorithm and introduce the notion of generalized U-statistics. In Chapter 3 we provide a theorem that describes the asymptotic distribution of these statistics when the rank of the kernel is allowed to grow with n. These distributional results rely on the behavior of a variance ratio and we conclude Chapter 3 by discussing its behavior for a variety of base learners. Building on these preliminary results, in Chapter 4, we provide Berry-Esseen bounds for both complete and incomplete generalized U-statistics. In Chapter 5, we analyze the properties of the infinitesimal jackknife method and propose a consistent estimate of the variance of these statistics.

2.0 Background

Suppose that we have data of the form $Z_1, \ldots, Z_n$ assumed to be independent and identically distributed (i.i.d.) from some distribution F_Z and let θ be some parameter of interest. Suppose further that there exists an unbiased estimator h of θ that is a function of $s \leq n$ arguments and without loss of generality, assume that h is permutation symmetric in those arguments. The minimum variance unbiased estimator for θ given by

$$U_{n,s} = \binom{n}{s}^{-1} \sum_{(n,s)} h(Z_{i1}, \ldots, Z_{is}) \tag{1}$$

is a U-statistic as introduced by [38] and [40], where the sum is taken over all $\binom{n}{s}$ subsamples of size s; we use the (n, s) shorthand for this quantity throughout the remainder of this paper. Standard elementary examples of U-statistics include sample mean, sample variance and covariance, and Kendall's τ-statistic. When both the kernel h and rank s are held fixed, [40] showed that $U_{n,s}$ tends toward a normal distribution with mean θ and variance $s^2\zeta_1/n$ where, for any $1 \leq c \leq s$,

$$\zeta_c = \mathrm{Cov}\left(h(Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s), h(Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_s)\right)$$

where $Z'_{c+1}, \ldots, Z'_n$ are i.i.d. from F_Z and independent of $Z_1, \ldots, Z_n$.

Throughout the remainder of this paper, we consider a regression framework where the data consist of independent pairs of random variables consisting of covariates and a response $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ $(i = 1, \ldots, n)$ sampled from a common distribution F_Z. Unless otherwise stated, we assume $\mathcal{X} = \mathbb{R}^p$ for analytical convenience.

Given some $s \leq n$, let $Z_{i1}, \ldots, Z_{is}$ denote a subsample of size s and consider a particular location $x \in \mathbb{R}^p$. The prediction at x can be written as $h_x(Z_{i1}, \ldots, Z_{is})$ where the function h_x takes the subsampled covariates and responses as inputs, forms a regression estimate, and outputs the predicted response at x. Throughout the remainder of this paper, we drop

the subscript x for notational convenience. Repeating this process on N subsamples and averaging across predictions gives

$$U_{n,s,N}(x) = \frac{1}{N}\sum_{i=1}^{N} h(Z_{i1}, \ldots, Z_{is})$$

so that our prediction now takes the form of a U-statistic with kernel h. When all subsamples are used so that $N = \binom{n}{s}$, the form is that of a *complete* U-statistic; whenever a smaller number of subsamples are utilized, it is *incomplete*. When the subsample size s grows with the sample size n, these estimators are considered *infinite-order* U-statistics as introduced by [35] and utilized by [54] to establish asymptotic normality of random forests.

In a general supervised learning framework, these kernels can be thought of as base learners in an ensemble. Decision trees are among the most popular choices of base learners and are typically built according to the CART criterion. Here, splits in each cell A are chosen to maximize

$$\begin{aligned} L(j,z) = &\frac{1}{|A|}\sum_{i=1}^{n}(Y_i - \bar{Y}_A)^2 \mathbf{1}_{X_i \in A} \\ &- \frac{1}{|A|}\sum_{i=1}^{n}(Y_i - \bar{Y}_{A_L}\mathbf{1}_{X_{j,i}<z} - \bar{Y}_{A_R}1_{X_{j,i}\geq z})^2 \mathbf{1}_{X_i \in A} \end{aligned}$$

across all covariates X_j, $1 \leq j \leq p$, where $z \in \mathbb{R}$, $A_L = \{X \in A : X_j < z\}$, $A_R = \{X \in A : X_j \geq z\}$, and for any set S, $\bar{Y}_S$ denotes the average response value for observations $X \in S$. When trees are built with bootstrap samples, the resulting ensembles produce *bagged* estimates as discussed in [11]. The random forest extension of bagging introduced by [12] inserts additional independent randomness into each tree, typically to determine the subset of $\texttt{mtry} \leq p$ features eligible for splitting at each node. The subsampled version of this procedure thus produces estimates at x of the form

$$\tilde{U}_{n,s,N,\omega}(x) = \frac{1}{N}\sum_{i=1}^{N} h(Z_{i1}, \ldots, Z_{is}; \omega). \tag{2}$$

Note that for each decision tree we consider an i.i.d. sample of randomness ω_i but for notational convenience, we refer to this as simply ω for all trees. Furthermore, in a similar fashion as above, define $\zeta_{c,\omega}$ $(c = 1, \ldots, s-1)$ and ζ_s as

$$\begin{aligned} \zeta_{c,\omega} &= \text{Cov}\left(h(\ldots, Z_c, Z_{c+1}, \ldots, Z_s; \omega), h(\ldots, Z_c, Z'_{c+1}, \ldots, Z'_s; \omega')\right) \\ \zeta_s &= \text{Cov}\left(h(\ldots, Z_c, Z_{c+1}, \ldots, Z_s; \omega), h(\ldots, Z_c, Z_{c+1}, \ldots, Z_s; \omega)\right) \end{aligned} \tag{3}$$

and note that ζ_s is simply the variance of the kernel with randomization parameter ω.

[54] provide asymptotic distributional results for $\tilde{U}_{n,s,N,\omega}$ with respect to their individual means that cover all possible growth rates of N with respect to n, though the form of the result provided has several practical limitations. In particular, the authors require that $\zeta_{1,\omega}$ does not approach 0, but for most practical base learners, the correlation between estimators with only one observation in common should vanish as the subsample size grows. Indeed, Lemma 1 in Appendix A gives that $\zeta_{1,\omega} \leq \frac{1}{s}\zeta_{s,\omega} \leq \frac{1}{s}\zeta_s$ so that when ζ_s is bounded, $\zeta_{1,\omega} \to 0$ as $s \to \infty$. In very recent work, [61] showed that the same result could be obtained under a more mild condition. In both results, however, the subsample size is limited to $s = o(n^{1/2})$ which can be quite restrictive in practice. In Appendix A, we demonstrate that this limitation is a result of a reliance on Hájek projections and in fact, whenever such an approach is taken, there is strong reason to believe that a subsampling rate of $s = o(n^{1/2})$ is the largest possible. As later discussed by [69] however, when s is small, it is possible that the squared bias decays slower than the variance, thereby producing confidence intervals which, when built according to the stated Gaussian limit distribution, may not cover the true value. [69] provide an alternative central limit theorem for averages over trees built according to honesty and regularity conditions. When base learners conform to such conditions and N is very large, the authors show that the subsampling rate can be improved to $s = o(n^{\beta})$ for $0.5 < \beta < 1$ while retaining consistent estimates.

Motivated by the form of Eq. (2) we now formalize the notion of *generalized* U-statistics.

Definition 1 (generalized U-statistic)**.** *Suppose $Z_1, \ldots, Z_n$ are i.i.d. samples from F_Z and let h denote a (possibly randomized) real-valued function utilizing s of these samples that*

is permutation symmetric in those s arguments. A generalized U-statistic with kernel h of order (rank) s refers to any estimator of the form

$$U_{n,s,N,\omega} = \frac{1}{N} \sum_{(n,s)} \rho h(Z_{i1}, \ldots, Z_{is}; \omega) \tag{4}$$

where ω denotes i.i.d. randomness, independent of the original data. The ρ denote i.i.d. Bernoulli random variables determining which subsamples are selected and $\Pr(\rho = 1) = N/\binom{n}{s}$. When $N = \binom{n}{s}$, the estimator in Eq. (4) is a generalized complete U-statistic and is denoted as $U_{n,s,\omega}$. When $N < \binom{n}{s}$, these estimators are generalized incomplete U-statistics.

Let $\hat{N}$ denote the actual number of subsamples selected. Though it is not practical to simulate $\binom{n}{s}$ Bernoulli random variables, fortunately, it is equivalent to first simulate $\hat{N} \sim \text{Binomial}(\binom{n}{s}, N/\binom{n}{s})$ and then randomly generate $\hat{N}$ subsamples without replacement. Note also that while the number of subsamples $\hat{N}$ in Eq. (4) is random, it concentrates around N.

Allowing for the possibility of a randomized kernel is of benefit here as it allows the results that follow to pertain to the kinds of randomized ensembles often considered in practice. The randomization parameter ω might, for example, perform some kind of feature subsampling as is commonly associated with random forests – much further discussion along these lines is provided in Chapter 3. We stress however that the mere inclusion of such a randomization parameter is not where the true innovation in our work lies, nor should it be viewed as the "essential ingredient" in what we refer to as generalized U-statistics. Indeed, in several of the results that follow, the theoretical details needed to establish them follow a near-identical recipe regardless of whether the kernel itself takes on additional randomness.

Rather, the real benefit of considering generalized U-statistics lies in the form of the estimator itself that allows for, in essence, a random weighting to be applied to the kernel through the use of ρ. Note, for example, that Eq. (4) has a slightly smaller variance than its fixed counterpart in Eq. (2). As a bit of a preview of what is to follow, note also that in this generalized form, an incomplete U-statistic can be viewed as merely a complete U-statistic with a different kernel. It is these kinds of realizations that provide significant benefits for theoretical analysis by allowing us to view incomplete U-statistics as merely a modified

version of its complete form, rather than as an approximation to it that inherits a remainder term that needs to be controlled. Furthermore, in the complete case, it can be shown that the variance of the U-statistic can be decomposed into a sum over s terms and that the structure of the statistic itself shrinks the higher-order terms in that sum. This careful examination of higher-order terms allows us to not only establish asymptotic normality, but to provide rates of convergence sharper than any in the existing literature, some of which are based on fundamental work dating back several decades.

In the literature on classic U-statistics, many results are derived by applying a technique called the H-decomposition, which allows the statistic to be written as a sum of uncorrelated terms. Appendix B.1 contains a detailed overview of the classic H-decomposition. The idea was first introduced by [41], but has analogues in many parts of statistics, most notably in the analysis of variance in balanced experimental designs; for a more general result, see [27]. To handle *generalized* U-statistics, we begin by extending the concept of the H-decomposition to this more general setting.

Definition 2 (H-decomposition)**.** *Suppose that $Z_1, \ldots, Z_s$ are i.i.d. samples from F_Z and $h(z_1, \ldots, z_s; \omega)$ is a (possibly randomized) real valued function that is permutation-symmetric in $(z_1, \ldots, z_s)$. Let $h_i(z_1, \ldots, z_i) = \mathbb{E}[h(z_1, \ldots, z_i, Z_{i+1}, \ldots, Z_s; \omega)] - \mathbb{E}[h]$ for $i = 1, \ldots, s$ and let*

$$h^{(i)} = h_i(z_1, \ldots, z_i) - \sum_{j=1}^{i} \sum_{(s,j)} h^{(j)}(z_{i1}, \ldots, z_{ij}), \qquad \text{for } i = 1, \ldots, s-1 \text{ and}$$

$$h^{(s)} = h(z_1, \ldots, z_s; \omega) - \sum_{j=1}^{s-1} \sum_{(s,j)} h^{(j)}(z_{i1}, \ldots, z_{ij}).$$

The H-decomposition of a generalized complete U-statistic is expressed as

$$U_{n,s,\omega} = \sum_{j=1}^{s} \binom{s}{j} \binom{n}{j}^{-1} \sum_{(n,j)} h^{(j)}(Z_{i1}, \ldots, Z_{is}). \tag{5}$$

When no extra randomness is injected into h, the above definition reduces to the classic H-decomposition. Note that the randomness ω is only involved in $h^{(s)}$; for $h^{(1)}, \ldots, h^{(s-1)}$, it is marginalized out. Note that because each subsample is associated with an i.i.d. draw of the randomness ω, each of the $h^{(s)}$ terms in Eq. (5) involves this randomness though this notation is suppressed in Eq. (5) for readability.

3.0 Asymptotic Normality

3.1 Asymptotic normality of generalized U-statistics

Before providing the asymptotic distributional results for generalized U-statistics of the form

$$U_{n,s,N,\omega} = \frac{1}{N} \sum_{(n,s)} \rho h(Z_{i1}, \ldots, Z_{is}; \omega) \tag{6}$$

we pause to emphasize the value in considering this form of estimator and to distinguish this generalization from the more classical counterparts considered in recent studies. [54] produce a central limit theorem for infinite-order U-statistics, but consider randomized kernels only insofar as establishing that when such randomness is well-behaved, the limiting distributions are equivalent. More recently, [69] analyzed random forests constructed with all possible subsamples where the kernel can thus be written in a form where the additional randomness is marginalized out. Such estimators take the form

$$\binom{n}{s}^{-1} \sum_{(n,s)} \mathbb{E}_\omega h(Z_{i1}, \ldots, Z_{is}; \omega) \tag{7}$$

so that the kernels themselves are non-random and thus the estimator is simply a complete, infinite-order U-statistic with kernel $g = \mathbb{E}_\omega h(Z_{i1}, \ldots, Z_{is}; \omega)$.

While more convenient for theoretical analysis, random forests of the form conceived in Eq. (7) are not generally utilized in practice, even in small-data settings since, by construction, such a statistic involves building every possible randomized tree on every possible subsample of the data. In practice, random forests might be loosely seen as a *double* or *nested* Monte Carlo approximation to the estimators in Eq. (7), where one source of approximation results from using $N < \binom{n}{s}$ subsamples and the other results from estimating the kernel itself $\mathbb{E}_\omega h(Z_{i1}, \ldots, Z_{is}; \omega)$ on each subsample. Recent work by [60] provides an analysis of these kinds of nested approximations.

In practice, however, random forests are nearly always constructed by selecting subsamples at random and pairing each with an independently selected randomization instance ω,

which is itself generally assumed to be selected uniformly at random. Generalized U-statistics therefore provide a direct and accurate representation of such estimators. We begin with a theorem establishing asymptotic normality for complete generalized U-statistics.

Theorem 1. *Let $Z_1, \ldots, Z_n$ be i.i.d. from F_Z and $U_{n,s,\omega}$ be a generalized complete U-statistic with kernel $h(Z_1, \ldots, Z_s; \omega)$. Let $\theta = \mathbb{E}[h]$, $\zeta_{1,\omega} = \mathrm{Var}(\mathbb{E}[h|Z_1])$ and $\zeta_s = \mathrm{Var}(h)$. If $\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$, then*

$$\frac{U_{n,s,\omega} - \theta}{\sqrt{s^2\zeta_{1,\omega}/n}} \rightsquigarrow N(0,1). \tag{8}$$

The proof of Theorem 1 is provided in the Appendix B.2. The general strategy is to find a linear statistic to approximate $U_{n,s,\omega}$, and show that the difference is negligible by applying the H-Decomposition.

Remark 1. *The condition in Theorem 1 that $\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$ can be replaced by the weaker condition that $\frac{s}{n}(\frac{\zeta_s}{s\zeta_{1,\omega}} - 1) \to 0$. In practice, this condition can be satisfied by choosing s to grow slow relative to the variance ratio $\zeta_s/s\zeta_{1,\omega}$. In particular, whenever the ratio is bounded, choosing $s = o(n)$ is sufficient. Thus, in establishing asymptotic normality, this weaker condition may be of minimal consequence. However, in quantifying the finite sample deviations from normality via the Berry-Esseen Theorems in Chapter 4, this alternative condition plays an important role in establishing the bounds provided.*

Similar results for non-generalized U-statistics have appeared in the recent works discussed earlier. Theorem 1 in [54] can be modified slightly to provide an analogous result whenever $\frac{s^2}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$. A recent result in [61] proceeds along these lines. Both results, however, could be improved by applying the H-decomposition rather than the Hájek projection. Similarly, Theorem 3.1 in [69] establishes asymptotic normality for non-generalized, complete U-statistics whenever the subsample size s grows like n^β for some $\beta < 1$. Here though the authors are concerned only with base learners that take the form of averages over honest and regular trees and in particular, with controlling the asymptotic bias of the resulting estimator. Thus, with minor modifications, Theorem 3.1 in [69] could be seen as something of a corollary to our Theorem 1 above, corresponding to the special case where the within-kernel randomness is held fixed or marginalized out.

The complete forms of these estimators are almost never utilized in practice due to

the computational burden involved with calculating $\binom{n}{s}$ base learners. Thus, armed with the results for the complete case, we now establish an analogous result for *incomplete* generalized U-statistics.

Theorem 2. *Let $Z_1, \ldots, Z_n$ be i.i.d. from F_Z and $U_{n,s,N,\omega}$ be a generalized incomplete U-statistic with kernel $h(Z_1, \ldots, Z_s; \omega)$. Let $\theta = \mathbb{E}[h]$, $\zeta_{1,\omega} = \mathrm{Var}(\mathbb{E}[h|Z_1])$ and $\zeta_s = \mathrm{Var}(h)$. Suppose that $\mathbb{E}[|h-\theta|^{2k}]/\mathbb{E}^2[|h-\theta|^k]$ is uniformly bounded for $k = 2, 3$ and for all s. If $\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$ and $N \to \infty$, then*

$$\frac{U_{n,s,N,\omega} - \theta}{\sqrt{s^2\zeta_{1,\omega}/n + \zeta_s/N}} \rightsquigarrow N(0,1). \tag{9}$$

Remark 2. *Note that the variance in the theorem above takes a different form than in Theorem 1 but the requirement that $\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$ remains the same. Indeed, whenever this condition is satisfied, the complete U-statistic analogue is also asymptotically normal and normality of the incomplete version can be established as a by-product. However, this condition and more generally, asymptotic normality of the complete version, is not necessary. In such cases, choosing a very small ensemble size (e.g. N=o(n/s)) is sufficient. More details and related results are provided in Appendix B.2 along with the proof of Theorem 2.*

Taken together, Theorem 1 and Theorem 2 provide the asymptotic distribution of generalized U-statistics for all possible growth rates on the number of subsamples N relative to n. Besides the regularity conditions on the kernel, these results require only that $\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$. This condition, similar to the notion of ν-incrementality discussed in [69], is not overly strong but may appear somewhat arbitrary. In the following subsection we investigate the behavior of this ratio for a variety of base learners.

3.2 Variance ratio behavior

For a given kernel h, let $\hat{h}$ be the projection of h onto the linear space. We have that $\hat{h} = \sum_{i=1}^{s} h_1(Z_i)$ and thus

$$\frac{\mathrm{Var}(h)}{\mathrm{Var}(\hat{h})} = \frac{\mathrm{Var}(h)}{s\mathrm{Var}(\mathbb{E}[h|Z_1])} = \frac{\zeta_s}{s\zeta_{1,\omega}}. \tag{10}$$

Since ζ_s is the overall variance and $\zeta_{1,\omega}$ can be written as the variance of the expectation of the kernel conditioning on one argument, we can view the ratio in Eq. (10) as a measure of the potential influence of one single observation on the output of the kernel. When $\zeta_s/s\zeta_{1,\omega} \to 1$, h itself is asymptotically linear. More generally though, Theorems 1 and 2 require only that $\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$ in order for the generalized U-statistic to be asymptotically normal. Thus, if the limiting behavior of the variance ratio $\zeta_s/s\zeta_{1,\omega}$ is understood, the subsampling rate can be chosen to ensure the entire term approaches 0.

For simple kernels such as the sample mean and sample variance, it is straightforward to show that the limit of this variance ratio is 1, though this can also be shown to hold for more standard regression estimates such as ordinary least squares; see Appendix B.3 for explicit calculations. Here we focus our attention more on nearest-neighbor estimators and linear smoothers as these are more directly relatable to the tree-style base learners often used in practice.

Proposition 1. *Let $Z_1, \ldots, Z_s$ denote i.i.d. pairs of random variables (X_i, Y_i) and suppose $Y_i = f(X_i) + \epsilon_i$ where f is continuous, ϵ_i has mean 0 and variance σ^2, and the X_i and ϵ_i are independent. Let φ denote the standard k-nearest neighbor (kNN) estimator. Then*

$$\limsup_{s\to\infty} \frac{\zeta_s}{s\zeta_1} \leq c(k)$$

where

$$c(k) = 2k/\left[\sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \frac{(i+j)!}{i!j!}\frac{1}{2^{i+j}}\right]$$

so that $c(k)$ is decreasing in k and $1 < c(k) \leq 2$.

A sketch of $c(k)$ for $k = 1, ..., 50$ is shown in Fig. 3.1. The proof of Proposition 1 is provided in Appendix B.3. Note that kNN is a nonadaptive linear smoother, the variance ratio of which is bounded above by a constant. The following result gives an upper bound for the more general class of all linear smoothers.

Proposition 2. *Let $Z_1, \ldots, Z_s$ denote i.i.d. pairs of random variables (X_i, Y_i) and suppose $Y_i = f(X_i) + \epsilon_i$ where f is bounded, ϵ has mean 0 and variance σ^2, and the X_i and ϵ_i are independent. Let*

$$\varphi = \sum_{i=1}^{s} w(i, x, \mathbf{X})Y_i$$

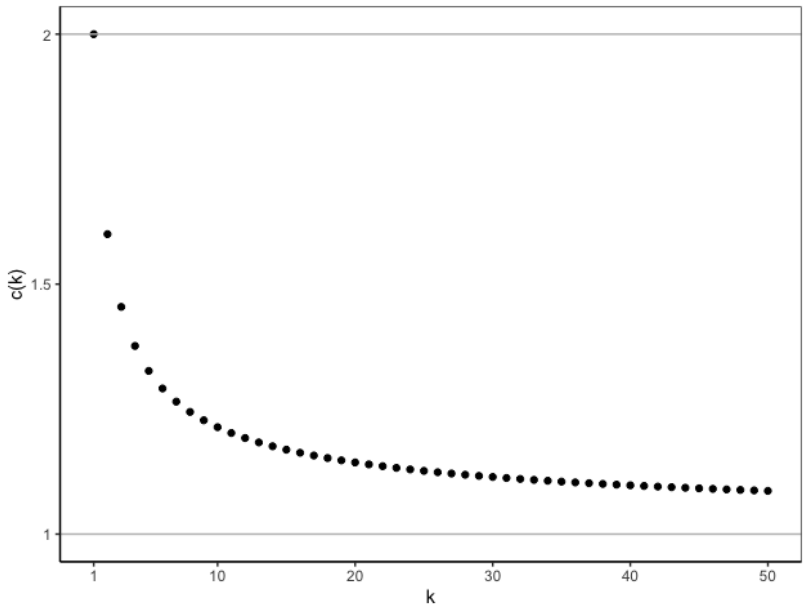

Figure 3.1: The function of $c(k)$ for $k = 1, ..., 50$. $c(k)$ is monotonically decreasing as k increases and bounded with $1 < c(k) \leq 2$.

such that $\sum_{i=1}^{s} w(i, x, \mathbf{X}) = 1$*, where* $\mathbf{X}$ *denotes* $\{X_i\}_{i=1}^{s}$*. Then* $\limsup_{s\to\infty} \frac{\zeta_s/s\zeta_1}{s} < \infty$.

The results above demonstrate that the behavior of the variance ratio is manageable for k-nearest neighbor base learners and more generally, linear smoothers. Recent work [62, 57] has sought to draw a connection between these estimators and the CART-style trees utilized in Breiman's original random forests. The purely random forest [8] that determines splits completely at random, for example, is exactly a linear smoother and thus by the above result, has a variance ratio that behaves like $O(s)$. In work dating back even further, [51] introduced the concept of *potential* nearest neighbors (PNNs) and showed that random forests can be viewed as an adaptively weighted k-PNN method.

Definition 3 ([51])**.** *A sample point* $Z_i = (X_i, Y_i)$ *is called a k-potential nearest neighbor (*k-PNN*) of a target point* x *if and only if there are fewer than* k *sample points other than* X_i *in the hyperrectangle defined by* x *and* X_i*.*

Typically, the number of potential nearest neighbors is much larger than the number

of nearest neighbors. Existing nearest-neighbor methods, both adaptive and nonadaptive, predict by selecting and averaging over k points from the set of all k-PNNs. The classical kNN procedure non-adaptively chooses the k points as those closest to x under some metric whereas commonly used tree-based methods may have a terminal size bounded by k and adaptively select points from the k-PNNs based on empirical relationships in the data.

Moving closer to this, consider the base learner that forms a prediction at x by simply choosing k of the s observations in the subsample uniformly at random and averaging the corresponding response values. In Appendix B.4, we show that the resulting variance ratio for this naive estimator is given by

$$\frac{\zeta_s}{s\zeta_1} = \frac{s}{k} = \frac{1/k \cdot s^2}{s \cdot 1}.$$

Now reconsider the kNN base learner. We can view such an estimator as "randomly" selecting k points from the k-NNs and again predicting by taking the average. In this case, the variance ratio can be written as

$$\frac{\zeta_s}{s\zeta_1} = O(1) = O\left(\frac{1/k \cdot k^2}{s \cdot \mathbb{E}[\Pr^2(X_1 \in \text{kNN} \mid X_1)]}\right).$$

The form of this result may naturally lead one to conjecture that for any base learner that predicts by randomly selecting and averaging over points in some set A, the resulting ratio may have the form

$$\frac{\zeta_s}{s\zeta_1} = O\left(\frac{1/k \cdot |A|^2}{s \cdot \mathbb{E}[\Pr^2(X_1 \in A \mid X_1)]}\right). \tag{11}$$

Consider then a simple tree-style estimator that predicts at x by sampling k points uniformly at random from its k-PNNs and averaging the corresponding response values; we refer to these random potential nearest neighbor estimators as RP trees. The additional difficulty introduced with RP trees is that the size of this set of potential nearest neighbors is itself random, though from [51], we know that the expected number of k-PNNs is $O(k(\log s)^{p-1})$ and so extending our conjecture, we arrive at the following proposition. we have

Proposition 3. *Let $Z_1, \ldots, Z_s$ denote i.i.d. pairs of random variables (X_i, Y_i) and suppose $Y_i = f(X_i) + \epsilon_i$ where f is bounded, ϵ has mean 0 and variance σ^2, and X_i and ϵ_i are independent. Suppose further that that the density of X is bounded away from 0 and infinity in $[0,1]^p$. Then for the RP tree estimator, we have*

$$\limsup_{s \to \infty} \frac{\zeta_s / s\zeta_1}{(\log s)^{2p-2}} < \infty. \tag{12}$$

The proof of Proposition 3 is provided in Appendix B.4. Here, asymptotic normality can be ensured by insisting on the same subsample sizes put forth in [69], namely that $s = o(n^\beta)$ for some $0.5 < \beta < 1$.

Calculating the variance ratio for adaptive base learners without imposing specific constraints on the base learners and/or data is quite challenging. However, we conclude our discussion here by noting that the previous calculations offer some encouragement. Given the k-PNNs of some target point x and considering estimators that predict by averaging over some subset of these, we showed that for non-adaptive estimators like kNN, the variance ratio is bounded. On the other hand, when the samples are selected uniformly at random from all k-PNNs, the variance ratio is on the order of $(\log s)^{2p-2}$. Tree-based estimators, by definition, predict by averaging over subsets of potential nearest neighbors, though the particular fashion in which those neighbors are chosen is often data-dependent. If, however, we are in a common regression setting where the response is regressed on covariates that contain some signal, then trees may heavily weight only a subset of the potential nearest neighbors, particularly in directions that can account for some of the variability in the response. In such settings, the variance ratio may be approximately of the form in Eq. (11) for some smaller set S and therefore be smaller than that of RP trees. Since this is not the main focus of this work, we do not investigate this idea further here but leave further exploration of the variance ratio behavior as potentially interesting future work.

4.0 Berry-Esseen Bounds

Given i.i.d. random variables $Z_1, \ldots, Z_n$ with mean μ and variance σ^2, the Berry-Esseen theorem [4, 33] provides a classical result describing the rate of convergence of $S_n = \sum_i (Z_i - \mu)/\sigma\sqrt{n}$ to the normal distribution. It states that provided the third moment $v_3 = \mathbb{E}|Z - \mu|^3$ is finite,

$$\sup_{z \in \mathbb{R}} |F_n(z) - \Phi(z)| \leq \frac{C v_3}{\sigma^3 \sqrt{n}}$$

where F_n is the distribution function of S_n, Φ is the distribution function of the standard normal, and C is a constant independent of n and the Z_i. Several authors (e.g. [14, 15, 37, 17]) have since contributed various iterations of Berry-Esseen theorems for U-statistics. In the following sections, we derive bounds for *generalized* U-statistics involving n, s, N, and the moments of the base learner to lend some intuition regarding how these parameters might be chosen in practice. We utilize the H-decomposition along with novel representations of U-statistics in order to provide bounds sharper than previously established in the literature for infinite-order U-statistics as well as first-of-their-kind bounds for *generalized* U-statistics.

4.1 Bounds for generalized U-statistics

We begin with the following result on generalized, complete U-statistics.

Theorem 3. *Suppose that $Z_1, \ldots, Z_n$ are i.i.d. from F_Z and that $U_{n,s,\omega}$ is a generalized complete U-statistic with kernel $h = h(Z_1, \ldots, Z_s; \omega)$. Let $\theta = \mathbb{E}[h]$, $\zeta_s = \mathrm{Var}(h)$ and $\zeta_{1,\omega} = \mathbb{E}[g^2(Z_1)]$, where $g(z) = \mathbb{E}[h(z, Z_2, \ldots, Z_s; \omega)] - \theta$. Suppose further that $\zeta_s < \infty$ and $\zeta_{1,\omega} > 0$, then*

$$\sup_{z \in \mathbb{R}} \left| \Pr\left\{ \frac{U_{n,s,\omega} - \theta}{\sqrt{s^2 \zeta_{1,\omega}/n}} \leq z \right\} - \Phi(z) \right| \leq \frac{6.1 \mathbb{E}|g|^3}{n^{1/2} \zeta_{1,\omega}^{3/2}} + (1 + \sqrt{2}) \left\{ \frac{s}{n} \left(\frac{\zeta_s}{s \zeta_{1,\omega}} - 1 \right) \right\}^{1/2}.$$

A number of important points are worth noting here. First, when s is fixed, the bound has a rate on the order of $1/\sqrt{n}$ as should be expected since this is the standard rate associated

with classic (finite-order), complete U-statistics. Additionally, when the randomness ω is held fixed so that the estimator reduces to an infinite-order U-statistic, this bound is sharper than that provided in [17], which, to our knowledge, is the sharpest to date in the existing literature. Specifically, the bound appearing in [17] replaces the term

$$\frac{s}{n}\left(\frac{\zeta_s}{s\zeta_1} - 1\right)$$

in the bound above, which is on the order of s/n, with

$$\frac{(s-1)^2\zeta_s}{s(n-s+1)\zeta_1}$$

which is on the order of s^2/n. An immediate consequence of this tighter bound is that when kernels are employed such that the resulting terms $\zeta_s/s\zeta_1$ and $\mathbb{E}|g|^3/\zeta_1^{3/2}$ are bounded, a subsampling rate of $s = o(n)$ is sufficient to ensure the bound converges to 0, whereas a rate of $s = o(\sqrt{n})$ would be required according to the bound given in [17]. This sharper rate we obtain is ultimately a result of utilizing Lemma 4 together with the H-decomposition. Full details are provided in Appendix C.2.

Generalized incomplete U-statistics can be viewed as generalized *complete* U-statistics with an alternative kernel. Recognizing this fact, we can make use of the H-decomposition and Lemma 4 given in the appendix to obtain the following bound for incomplete, generalized U-statistics.

Theorem 4. *Suppose that $Z_1, \ldots, Z_n$ are i.i.d. from F_z and that $U_{n,s,N,\omega}$ is a generalized incomplete U-statistic with kernel $h = h(Z_1, \ldots, Z_s; \omega)$. Let $\theta = \mathbb{E}[h]$, $\zeta_s = \mathrm{Var}(h)$, and $\zeta_{1,\omega} = \mathbb{E}[g^2(Z_1)]$, where $g(z) = \mathbb{E}[h(z, Z_2, \ldots, Z_s; \omega)] - \theta$. Suppose further that $\zeta_s < \infty$ and $\zeta_{1,\omega} > 0$. Then*

$$\begin{aligned}
&\sup_{z\in\mathbb{R}}\left|\Pr\left\{\frac{U_{n,s,N,\omega} - \theta}{\sqrt{s^2\zeta_{1,\omega}/n}} \le z\right\} - \Phi(z)\right| \\
&\le \frac{6.1\mathbb{E}|g|^3}{n^{1/2}\zeta_{1,\omega}^3} + (1+\sqrt{2})\left\{\frac{s}{n}\left(\frac{\zeta_s}{s\zeta_{1,\omega}} - 1\right)\right\}^{1/2} + \left(1 + \sqrt{\frac{1}{s}}\right)\left\{\frac{n}{N}(1-p)\frac{\zeta_s}{s\zeta_{1,\omega}}\right\}^{1/2}.
\end{aligned}$$

The proof of Theorem 4 is provided in Appendix C.2. The preceding theorems indicate that for both infinite-order and generalized U-statistics, when incomplete versions of these estimators are used, these statistics remain asymptotically as efficient as the complete forms so long as $n = o(N)$. Comparing Theorems 3 and 4, note that these bounds differ only by the inclusion of an additional final term in the sum, which is close to 0 in such large-N settings. However, in small-N settings, this final term can become quite large, leading to a bound nearing or even exceeding 1, thereby making it of little use. Theorem 5 below provides improved Berry-Esseen bounds for such small-N settings where relatively few base learners are employed. To achieve this, rather than writing the estimators as linear statistics plus a small additional manageable term, we take an alternative approach that views incomplete U-statistics as complete U-statistics plus a remainder. This strategy is similar to that used in [20] who recently derived non-asymptotic Gaussian approximation error bounds for high-dimensional, incomplete U-statistics, but for kernels with fixed (finite) rank. Proofs of the following results are provided in Appendix C.2.

Theorem 5. *Suppose that $Z_1, \ldots, Z_n$ are i.i.d. from F_Z and that $U_{n,s,\omega,N}$ is a generalized incomplete U-statistic with kernel $h = h(Z_1, \ldots, Z_s; \omega)$. Let $\theta = \mathbb{E}[h]$, $\zeta_s = \mathrm{Var}(h)$, and $\zeta_{1,\omega} = \mathbb{E}[g^2(Z_1)]$, where $g(z) = \mathbb{E}[h(z, Z_2, \ldots, Z_s; \omega)] - \theta$. Suppose further that $\zeta_s < \infty$ and $\zeta_{1,\omega} > 0$. If $\mathbb{E}[|h-\theta|^{2k}]/\mathbb{E}^2[|h-\theta|^k]$ is uniformly bounded for $k = 2, 3$ and for all s. Then*

$$\sup_{z\in\mathbb{R}} \left| \Pr\left\{ \frac{U_{n,s,N,\omega} - \theta}{\sqrt{s^2\zeta_{1,\omega}/n + \zeta_s/N}} \leq z \right\} - \Phi(z) \right|$$
$$\leq C \left\{ \frac{\mathbb{E}|g|^3}{n^{1/2}(\mathbb{E}|g|^2)^{3/2}} + \frac{\mathbb{E}|h-\theta|^3}{N^{1/2}(\mathbb{E}|h-\theta|^2)^{3/2}} + \left\{ \frac{s}{n}\left(\frac{\zeta_s}{s\zeta_{1,\omega}} - 1 \right) \right\}^{1/2} + \left(\frac{s}{n} \right)^{1/3} \right\}$$

for some constant $C > 0$.

Here we see that when s is fixed, the Berry-Esseen bound is on the order of $n^{-1/3}$. When s grows with n, the bound converges to zero as long as $s/n \to 0$ and $N \to \infty$ with some mild conditions on h.

The fundamental task in producing this result is to show that the convolution of the two independent sequences approaches a normal distribution. A number of approximations

are required, though we give a nearly sharp bound on each in order to provide the Berry-Esseen bound shown. As noted earlier, [20] recently investigated a similar setup for higher-dimensional kernels assumed to be of fixed rank. In our case, the use of an infinite-order kernel injected with extra randomness introduces additional technical difficulties, though the restriction to one-dimensional settings allows us to incorporate more useful concentration inequalities. Thus, even for kernels assumed to have a fixed, finite rank, the result above is sharper than that provided in [20] for the one-dimensional setting.

As an additional benefit, we note that the bound above consisting of a four-term sum contains insightful terms not produced in [20]. In particular, the second term corresponds to the bound that would be available for an estimator that takes an average of i.i.d. random variables, while the first term plus the third term gives the bound for the complete infinite-order U-statistic setting. This leads to the very natural intuition that when N is quite small, the bound produced is approximately what would be expected by averaging over independent base learners whereas when N is large, the bound is approximately what we would expect for a complete infinite-order U-statistic. To see where the fourth and final term $(s/n)^{1/3}$ comes from, we now delve into the proof details in the following subsection.

4.2 A tighter bound

In order to obtain the previous bounds, we first condition on $Z_1, \ldots, Z_n$ and obtain a Berry-Esseen bound for the difference between the infinite-order forms of incomplete and complete U-statistics, $U_{n,s,N} - U_{n,s}$. The terms involved in this bound are themselves infinite-order U-statistics with kernels that are power functions of the original kernel h. We make use of Chebyshev's inequality to replace those infinite-order U-statistics by their population mean, the application of which requires no particular assumptions on the tail behavior of the kernel. This approach, however, leads to the non-optimal term of $(\frac{s}{n})^{1/3}$. We thus conclude our discussion on Berry-Esseen Theorems by showing in this final subsection that placing additional assumptions on the kernel h can allow the application of sharper concentration inequalities that can therefore allow the term $(\frac{s}{n})^{1/3}$ to be replaced by $(\frac{s}{n})^{1/2}$.

Theorem 6. *Suppose that $Z_1, \ldots, Z_n$ are i.i.d. from F_Z and that $U_{n,s,N,\omega}$ is a generalized incomplete U-statistic with kernel $h = h(Z_1, \ldots, Z_s; \omega)$. Let $\theta = \mathbb{E}[h]$, $\zeta_s = \mathrm{Var}(h)$ and $\zeta_{1,\omega} = \mathbb{E}[g^2(Z_1)]$, where $g(z) = \mathbb{E}[h(z, Z_2, \ldots, Z_s; \omega)] - \theta$. Suppose further that $\zeta_s < \infty$ and $\zeta_{1,\omega} > 0$. If $|h - \theta|^k$ is sub-Gaussian after standardization with variance proxy that is uniformly bounded for $k = 2, 3$ and all s, then*

$$\sup_{z \in \mathbb{R}} \left| \Pr\left\{ \frac{U_{n,s,N,\omega} - \theta}{\sqrt{s^2 \zeta_{1,\omega}/n + \zeta_s/N}} \leq z \right\} - \Phi(z) \right|$$
$$\leq C \left\{ \frac{\mathbb{E}|g|^3}{n^{1/2}(\mathbb{E}|g|^2)^{3/2}} + \frac{\mathbb{E}|h-\theta|^3}{N^{1/2}(\mathbb{E}|h-\theta|^2)^{3/2}} + \left[\frac{s}{n} \left(\frac{\zeta_s}{s\zeta_{1,\omega}} - 1 \right) \right]^{1/2} + \left(\frac{s}{n} \right)^{\eta} \right\},$$

where $C > 0$ is some constant and $0 < \eta < 1/2$.

Note that since there is a trade-off between the probability and concentration bound, larger η requires a larger n to ensure the above inequality holds. Proof details of Theorem 6 are given in Appendix C.3 for the incomplete infinite-order U-statistic setting; the extension to *generalized* incomplete U-statistics follows in an identical fashion.

5.0 Variance Estimation

5.1 Introduction

It is difficult to overstate the importance and utility of resampling methods and the bootstrap in particular for determining properties of estimators whenever exact, explicit sampling distributions cannot be readily determined. Given a sample $X_1, ..., X_n \sim F_X$, a parameter of interest θ, and an estimator $\hat{\theta} = s(X_1, ..., X_n)$, it is often of interest to estimate, for example, $\text{Var}(\hat{\theta})$. Let $\boldsymbol{x} = (x_1, ..., x_n)$ denote the observed values of the sample so that for a particular realization, $\hat{\theta} = s(\boldsymbol{x})$. To provide bootstrap estimate of the variability of the estimator, we can draw B (re)samples of size n with replacement from $\{x_1, ..., x_n\}$ to form bootstrap samples $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_B^*$ from which we calculate bootstrap estimates $\hat{\theta}_1, ..., \hat{\theta}_B$. The nonparametric bootstrap variance estimate of $\hat{\theta}$ is then taken as the empirical variance of $\hat{\theta}_1, ..., \hat{\theta}_B$ [28, 31].

Within this context, given the necessity of calculating $\hat{\theta}_1, ..., \hat{\theta}_B$, it is natural to instead consider the estimator

$$\tilde{\theta}_B = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b \tag{13}$$

as a "bootstrap smoothed" alternative to the original $\hat{\theta}$ [32]. This sort of bootstrap aggregation (bagging) was also proposed by [11] as a means by which predictive variance may be reduced when each bootstrap sample is used to construct an individual model, frequently a classification or regression tree.

The standard bootstrap approach – referred to recently as the *brute force* approach by [31] – to assessing the variability of $\tilde{\theta}_B$, though straightforward, is computationally quite burdensome, requiring several bootstrap replicates of not only the original data, but also from within the bootstrap samples themselves. This double bootstrap, first proposed in [3], is especially costly whenever the original statistic T is computationally costly.

A variety of approaches have been suggested to reduce the computational burden of these sorts of problems. [73] and [23, 24, 25] employ what is now referred to as the fast

double bootstrap whereby only a single second-level bootstrap sample is collected. [36] employ such an approach in running Monte Carlo experiments and a more careful analysis is given in [16]. [65] propose some alternative nonparametric means by which $\text{Var}(\tilde{\theta}_B)$ may be estimated, suggesting also that the number of second-level bootstrap replicates B' need only be a fraction of the original resample size B. In lieu of full bootstrap samples, subsampling, or m-out-of-n bootstrap sampling, was proposed by [58] and [10]. More recently, [64] proposed a combination of these approaches, first subsampling and then employing a single second-level resample. Similarly, [49] proposed the bag of little bootstraps which involves splitting the original dataset into a number of subsamples and then taking bootstrap samples on each subset allowing the process to more easily scale by being capable of efficiently running in parallel.

Though the above approaches can substantially reduce the computational complexity in estimating the variance of estimators based on resampling procedures, each nonetheless involves further resampling in order to obtain such an estimate. Recently, motivated by the problem of taking into account not only the sampling variability but also the variability in model selection, [31] alleviated this issue by developing an algebraically compact, closed-form estimator for the variance of a bagged estimate. Instead of additional resampling, Efron's proposal required only additional bookkeeping to recall which samples in the original data appeared how many times in each bootstrap sample. This development was particularly beneficial in estimating the variance in predictions generated via supervised learning ensembles that are relatively computationally expensive. A number of recent works, for example, have successfully applied this type of estimator in the context of random forests [72, 70, 75].

Though its final form is algebraically simple, the derivation of Efron's variance estimator is fairly involved and may appear somewhat mysterious to many readers. Its development comes from an application of the original theory for the infinitesimal jackknife involving functional derivatives. Likely as a result, studies rigorously investigating the statistical properties of this estimator as well as the contexts in which such an estimator would be appropriate are lacking in the current literature. Efron, for example, notes that the appropriateness of his nonparametric delta method (infinitesimal jackknife) approach follows from the fact that the bagged estimates represent a more smooth function of the data. Thus, while clearly an ex-

tremely significant result in and of itself, these estimates would not appear to apply to more general resampling schemes wherein such smoothness assumptions may not be reasonable.

In this chapter, we strive to take a step forward both in better understanding the intuition behind this important estimator as well as in understanding it's statistical properties. In addition to the infinitesimal jackknife derivation utilized by Efron, we consider two alternative approaches that are a bit more straightforward and easily motivated. The first of these exploits the important fact that conditional on the observed data, the bagged estimate in Eq. (13) depends only on the resampling weights. We consider a linear approximation to this function of bootstrap weights (i.e. standard linear regression) and demonstrate that this approach exactly reproduces the infinitesimal jackknife results given in [31] whenever all bootstrap samples are employed. As an additional benefit, this setup motivates a more general procedure for estimating the variance of any resampled estimate, not just one based on the bootstrap.

In addition to the linear regression and infinitesimal jackknife approaches, we also consider a classical jackknife motivation and once again demonstrate its equivalence in the full bootstrap context. Importantly, this alternative representation of the estimator allows us to explore its asymptotic properties and in particular, the bias. While the variance estimators motivated by the jackknife, infinitesimal jackknife, and linear regression approaches are shown to be identical when all bootstrap samples are used, they differ in practical settings when only a randomly selected subsample are employed, suggesting different bias corrections that might be imposed.

Finally, we derive the form of the infinitesimal jackknife estimate of variance in the U-statistic regime where the resampling is instead done by subsampling without replacement. We discover that the variance estimators commonly employed in practice for quantifying the predictive uncertainty in supervised learning ensembles like random forests are something of a "pseudo" infinitesimal jackknife in that they differ from the correct form when properly derived. However, the difference is minor when subsample size is small. We then investigate the properties of the "pseudo" infinitesimal jackknife and provide a consistent estimate of the variance of generalized U-statistics.

5.2 Background of the infinitesimal jackknife (IJ)

Let $\mathcal{D}_n$ denote a sample of observed values from real-valued random variables $X_1, \ldots, X_n$ from a distribution P. In practice, we are often interested in estimating statistical functionals – functions of the underlying distribution P, often estimated via the empirical probability $\mathbb{P}_n$. Denote this statistic as $s(X_1, \ldots, X_n) = f(\mathbb{P}_n)$ and assume that s is permutation symmetric in these n arguments. These "functions of functions" were first introduced by [68] and today are a familiar topic of advanced analysis. Any statistic that treats the samples equivalently can also be viewed as a function of $\mathbb{P}_n$, albeit without always having an explicit form of f. We can further extend the domain of f to any non-negative functions on $X_1, \ldots, X_n$ by defining

$$f(\mathbb{P}) = f(c \cdot \mathbb{P}), \quad \text{for any } c > 0. \tag{14}$$

A common task, especially in today's big data era is to find an appropriate and feasible means of estimating the variance of $f(\mathbb{P}_n)$. Historically, there have been three main methods: the infinitesimal jackknife [55], influence curves [39, 43], and the delta method [30]. Though each method was motivated differently, [29] pointed out that the three methods are identical. We thus refer the common estimator as IJ defined as

$$\text{IJ} = \frac{1}{n^2} \sum_i \text{D}_i^2, \tag{15}$$

where

$$\text{D}_i = \lim_{\epsilon \to 0} \frac{f((1-\epsilon)\mathbb{P}_n + \epsilon \delta_{X_i}) - f(\mathbb{P}_n)}{\epsilon} \tag{16}$$

and δ_x is the Dirac delta function.

We now briefly review the original derivation of the IJ, following closely the original construction by [56] and [47]. Let $\mathcal{P}$ be the set of all linear combinations of P and an arbitrary finite number of the δ_x measures. Let $\mathcal{P}^+$ be the the set of positive measures in $\mathcal{P}$, not including the zero measures and assume f is defined for the probability measures in $\mathcal{P}^+$. As above, extend f to all of $\mathcal{P}^+$ by letting $f(c \cdot \text{P}) = f(\text{P})$ for all $c > 0$. Note that $\mathcal{P}^+$ is convex and includes $\mathbb{P}_n$. We now formally define the derivative of f. We say f is differentiable at G in $\mathcal{P}^+$ if there exists a function $f'(\text{G}, x)$, defined at all x in $\mathbb{R}$, with the following property:

Definition 4 ([47]). *Let* H *be any member of* $\mathcal{P}$ *such that* $\mathrm{G}+t\mathrm{H}$ *is in* $\mathcal{P}^+$ *for all* t *in some intervals* $0 \le t \le t_{\mathrm{H}}$, $t_{\mathrm{H}} > 0$, *so that* $f(\mathrm{G}+t\mathrm{H})$ *is defined for* t *in this interval. Then for any such* H, $f'(\mathrm{G}, x)$ *satisfies*

$$\frac{df(\mathrm{G}+t\mathrm{H})}{dt}|_{t=0} = \lim_{t\to 0} \frac{f(\mathrm{G}+t\mathrm{H}) - f(\mathrm{G})}{t} = \int f'(\mathrm{G}, x)\, d\mathrm{H}(x). \tag{17}$$

If $\mathrm{H} = \mathrm{G}$, we see that $\int f'(\mathrm{G}, x)\, d\mathrm{G}(x) = 0$ since $f(c\mathrm{G}) = f(\mathrm{G})$. On the other hand, if $\mathrm{H} = \delta_x - \mathrm{G}$, we find

$$\lim_{t\to 0} \frac{f((1-t)\mathrm{G} + t\delta_x) - f(\mathrm{G})}{t} = \int f'(\mathrm{G}, x)\, d(\delta_x - \mathrm{G})(x) = f'(\mathrm{G}, x). \tag{18}$$

Indeed, [39] defined $f'(\mathrm{G}, x)$ by Eq. (18) and has called it the "influence curve", since it reflects the influence of f by adding a small mass on G at x. Additionally, the derivative of $f(\mathrm{G}+t\mathrm{H})$ at arbitrary t_0 with $0 < t_0 < t_H$ is given by

$$\frac{df(\mathrm{G}+t\mathrm{H})}{dt}|_{t=t_0} = \int f'(\mathrm{G}+t_0\mathrm{H}, x)\, d\mathrm{H}(x). \tag{19}$$

Now, we assume that f is differentiable, in the sense defined above, at all G in some convex neighbor of P in $\mathcal{P}^+$, such that $\mathbb{P}_n$ lies in the neighborhood with probability approaching one. We now describe the motivation of IJ for answering when we think IJ could be a sensible estimate of the variance of $f(\mathbb{P}_n)$. We parameterize the segment from P to $\mathbb{P}_n$ by $\mathrm{P}(t) = \mathrm{P} + t(\mathbb{P}_n - \mathrm{P})$for $0 \le t \le 1$. Then if $\mathbb{P}_n$ lies in the neighborhood of P, we hope that

$$\begin{aligned} f(\mathbb{P}_n) - f(\mathrm{P}) &= f(\mathrm{P}(1)) - f(\mathrm{P}(0)) \\ &= \int f'(\mathrm{P}, x)\, d(\mathbb{P}_n - \mathrm{P})(x) + o_p\left(\frac{1}{n}\right) \\ &= \frac{1}{n}\sum_i f'(\mathrm{P}, X_i) + o_p\left(\frac{1}{n}\right). \end{aligned} \tag{20}$$

The third equality is due to the fact that $\int f'(\mathrm{G}, x)\, d\mathrm{G}(x) = 0$. Since the first term on the right side is a sum of i.i.d. random variables, $\sqrt{n}(f(\mathbb{P}_n) - f(\mathrm{P}))$ is asymptotic normal with mean 0 and variance $\mathrm{V} = \int [f'(\mathrm{P}, x)]^2\, d\mathrm{P}(x)$. Since P is unknown and $f'(\mathrm{P}, x)$ depends on

both f and P, we generally do not know $f'(\mathrm{P}, x)$ in advance. Thus, we would estimate V by

$$\int f'^2(\mathbb{P}_n, x)\, d\mathbb{P}_n(x) = \frac{1}{n}\sum_i [f'(\mathbb{P}_n, X_i)]^2. \tag{21}$$

Then $\mathrm{Var}(f(\mathbb{P}_n))$ can be estimated by $\frac{1}{n^2}\sum[f'(\mathbb{P}_n, X_i)]^2$, which is exactly equal to IJ since $\mathrm{D}_i = f(\mathbb{P}_n, X_i)$.

In summary, to obtain the final estimate of the variance of $f(\mathbb{P}_n)$, we actually introduce two steps of approximation. Eq. (20) approximates $f(\mathbb{P}_n)$ by a linear statistic at $\mathbb{P}_n$, whereas Eq. (21) approximates P by $\mathbb{P}_n$ and $f'(\mathrm{P})$ by $f'(\mathbb{P}_n)$. Obviously, the most questionable parts are whether f is close to a linear statistic and whether $f'(\mathbb{P}_n)$ is close to $f'(\mathrm{P})$.

5.3 The infinitesimal jackknife estimate for bootstrap ($\mathrm{IJ_B}$)

In this section, we focus on a special f induced by bootstrap. Suppose that $s(X_1, ..., X_n)$ is statistic, not necessarily a function of $\mathbb{P}_n$. We take all possible bootstrap samples $(X_1^*, ..., X_n^*)$, plug in s to obtain s^*, and then take the average. We call the new statistic the bootstrap smoothed (bagged) alternative of s and denote it as $\mathbb{E}_*[s^*]$, where $\mathbb{E}_*[\cdot]$ is the expectation taken over the bootstrap sampling procedure conditioned on the data. Note that $\mathbb{E}_*[s^*]$ is now a function of $\mathbb{P}_n$. The dependence of f on $\mathbb{P}_n$ can be explicitly expressed out as

$$f(\mathbb{P}_n) = \int s\, d\mathbb{P}_n \times \cdots \times \mathbb{P}_n = \int s\, d(\mathbb{P}_n)^n. \tag{22}$$

Therefore, f depends on $(\mathbb{P}_n)^n$ and the dependence roughly exponentially grows with n. By Berry-Esseen theorem, the distance of $\mathbb{P}_n$ and P is at order of $1/\sqrt{n}$. However, the distance of $(\mathbb{P}_n)^n$ and $(\mathrm{P})^n$ is just $\mathcal{O}(1)$. Therefore, the distance of $f(\mathbb{P}_n)$ and $f(\mathrm{P})$ could be large if $f(\mathbb{P}_n)$ depends on $\mathbb{P}_n$ exponentially.

5.3.1 The convergence of three different approaches

We follow three different approaches, including the infinitesimal jackknife method, to derive an estimate of $\mathrm{Var}(\mathbb{E}_*[s^*])$ and show that they are equivalent when all bootstrap samples are taken.

The infinitesimal jackknife approach Since $\mathbb{E}_*[s^*]$ can be viewed as a function of $\mathbb{P}_n$, estimating $\mathrm{Var}(\mathbb{E}_*[s^*]) = \mathrm{Var}(f(\mathbb{P}_n))$ is a standard problem for the infinitesimal jackknife method and we call the estimate $\mathrm{IJ_B}$. By the definition of $\mathbb{E}_*[s^*]$, we have

$$\begin{aligned} f((1-\epsilon)\mathbb{P}_n + \epsilon\delta_{X_i}) &= n^{-n}\sum \frac{s(X_1^*,\ldots,X_n^*)n!}{(w_1^*!)(w_2^*!)\ldots(w_n^*!)}\left[(1-\epsilon)^{\sum_{k\neq i} w_k^*}(1+(n-1)\epsilon)^{w_i^*}\right] \\ &= n^{-n}\sum \frac{s(X_1^*,\ldots,X_n^*)n!}{(w_1^*!)(w_2^*!)\ldots(w_n^*!)}\left[1+n\epsilon(w_i^*-1)\right] + o(\epsilon^2) \\ &= f(\mathbb{P}_n) + \epsilon n\mathrm{Cov}_*(s^*, w_i^*) + o(\epsilon^2), \end{aligned} \tag{23}$$

where $w_i^* = \#\{j : X_j^* = X_i\}$. Thus, by Eq. (16), $\mathrm{D}_i = n\mathrm{Cov}_*(s^*, w_i^*)) = n\mathrm{cov}_i$ and

$$\mathrm{IJ_B} = \sum_i \mathrm{cov}_i^2 = \sum_i \mathrm{Cov}_*^2(s^*, w_i^*). \tag{24}$$

Eq. (24), following a simple application of the infinitesimal jackknife method, was first derived by [31]. The author did not discuss how good the estimation is or how bad it could possibly be. Actually, there has been no general theory answering these questions since the invention of the infinitesimal jackknife method. One possible reason could be that the process is quiet abstract. It involves with functional derivatives, which loses probability meaning. However, we found that we can derive the same estimator from other approaches, getting rid of functional derivatives. In particular, we have an explicit expression of $f'(\mathrm{P})$ and a different interpretation of $\mathrm{Cov}_*(s^*, w_i^*)$.

First, note that $\mathbb{E}_*[s^*]$ is a symmetric function of $X_1,\ldots,X_n$. Therefore, there exists a H-decomposition [42] of $\mathbb{E}_*[s^*]$. Let $t = \mathbb{E}_*[s^*]$, we have $t = \mathbb{E}[t] + \sum_j \sum_{i_1,\ldots,i_j} t_j(X_{i_1},\ldots X_{i_j})$, where

$$\begin{aligned} t_1(x_1) &= \mathbb{E}[t|X_1 = x_1] - \mathbb{E}[t] \\ t_2(x_1,x_2) &= \mathbb{E}[t|X_1 = x_1, X_2 = x_2] - t_1(x_1) - t_1(x_2) - \mathbb{E}[t] \\ &\vdots \\ t_n(x_1,\ldots,x_n) &= t - \sum_{j=1}^{n-1}\sum_{(n,j)} t_j(x_{i_1},\ldots,x_{i_j}) - \mathbb{E}[t], \end{aligned} \tag{25}$$

where $t_1, \ldots, t_n$ are uncorrelated with mean 0. Let us consider using the linear term $l_b = \mathbb{E}[t] + \sum_i t_1(X_i)$ as an approximation of t. We know that $\mathrm{Var}(l_b) = n \int t_1^2 \, d\mathrm{P}$. Since t and P are unknown, we can not get the exact value of $\mathrm{Var}(l_b)$, but we could estimate t_1 and P and hence obtain an estimation of $\mathrm{Var}(l_b)$. Firstly, $\mathbb{P}_n$ is at hand a good candidate for estimating P, and we have

$$\int t_1^2 \, d\mathrm{P} \approx \int t_1^2 \, d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} t_1^2(X_i). \tag{26}$$

Secondly, as for $t_1(X_1) = \mathbb{E}[t|X_1] - \mathbb{E}[t]$, $\mathbb{E}[\cdot]$ is again unknown, but we could substitute it with $\mathbb{E}_*[\cdot]$ instead and obtain

$$\begin{aligned} t_1(X_1) &= \mathbb{E}[t(X_1, \ldots, X_n)|X_1] - \mathbb{E}[t] \\ &\approx \mathbb{E}_*[t(X_1, X_2^*, \ldots, X_n^*)] - \mathbb{E}_*[t] \\ &= \mathbb{E}_*[s^*(X_1, X_2^*, \ldots, X_n^*)] - \mathbb{E}_*[s^*] \\ &= e_1 - s_0, \end{aligned} \tag{27}$$

where $e_1 = \mathbb{E}_*[s^*(X_1, X_2^*, \ldots, X_n^*)]$ and $s_0 = \sum e_i/n$. Putting all approximations together, we have

$$\widehat{\mathrm{Var}(l_b)} = \sum_i (e_i - s_0)^2. \tag{28}$$

Recall how we develop IJ by the infinitesimal jackknife method. We can find that $t_1 = f'(\mathrm{P})$ - the derivative of f at P. The infinitesimal jackknife method approximates $t = f(\mathbb{P})$ by $f(\mathrm{P}) + \int f'(\mathrm{P}) \, d\mathbb{P}_n$, whereas here we approximate t by $\mathbb{E}[t] + n^{-1} \sum_i t_1(X_i)$. Basically, we show how the idea behind the infinitesimal jackknife coincides with the idea of linear approximation by H-Decomposition. Indeed, we will later show that $\sum (e_i - s_0)^2 = \mathrm{IJ_B}$ unsurprisingly. Note that the approximation in Eq. (27) might not be good since $\mathbb{E}[\mathbb{E}_*[s^*]|X_1] - \mathbb{E}[\mathbb{E}_*[s^*]]$ is substituded by $\mathbb{E}_*[s^*|X_1^* = X_1] - \mathbb{E}_*[s^*]$. Also, $\mathbb{E}_*[s^*]$ might not be close to the linear statistic l_b. Therefore, we suspect that $\mathrm{IJ_B}$ is only appropriate for estimating the $\mathrm{Var}(\mathbb{E}_*[s^*])$ for in limited cases.

The jackknife approach We introduce the jackknife method here and see how it can motivate us to propose an estimator for $\mathrm{Var}(\mathbb{E}_*[s^*])$. Denote t as $\mathbb{E}_*[s^*]$. The delete-1 jackknife samples are selected by taking the original data vector and deleting one observation from the set. Thus, there are n unique jackknife samples, and the ith jackknife sample vector

is defined as $\mathcal{D}_n[i] = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. The ith jackknife replicate is defined as the value of the estimator $t(\cdot)$ evaluated at the i th jackknife sample. The jackknife variance is then defined by

$$\begin{aligned} \mathrm{JK} &= \frac{n-1}{n} \sum (t(\mathcal{D}_n[i]) - \bar{t})^2 \\ &\approx n\mathrm{Var}_* \left(\sum_i t(\mathcal{D}_n[i]) \cdot \mathbf{1}_{\{X_i \text{ is deleted}\}} \right). \end{aligned} \tag{29}$$

The idea is that we expect $t(\mathcal{D}_n[i])$ be closed to $t(\mathcal{D}_n)$ and thus use the sample variance $t(\mathcal{D}_n[i])$ to estimate the variance of $t(\mathcal{D}_n)$. And since those $t(\mathcal{D}_n[1]), \ldots, t(\mathcal{D}_n[n])$ are strongly correlated, we scale the sample variance by n. Note that $t = \int s(x_1, \ldots, x_n)\, d\mathbb{P}_n(x_1) \times \cdots \times \mathbb{P}_n(x_n)$. We can consider fixing the ith position instead of deleting ith sample. Thus we replace $t(\mathcal{D}_n[i])$ by

$$t_{(i,j)} = \int s(x_1, x_2, \ldots, x_{i-1}, X_j, x_{i+1}, \ldots, x_n)\, d\mathbb{P}_n(x_1) \cdots \mathbb{P}_n(x_{i-1}) \times \mathbb{P}_n(x_{i+1}) \cdots \mathbb{P}_n(x_n)$$

for $1 \le i, j \le n$ and propose

$$\mathrm{JK_B} = n\mathrm{Var}_* \left(\sum_j t_{(i,j)} \mathbf{1}_{\{X_i^* = X_j\}} \right) = \sum_j (e_j - s_0)^2 \tag{30}$$

as an estimate of the variance of $\mathbb{E}_*[s^*]$. The third equality in Eq. (30) is simply due to the fact that $t_{(i,j)} = e_j$.

The least squared approach Recall that in the standard approach, B equally weighted resamples of size n are independently taken from the rows of $\mathcal{D}_n$ with replacement. Thus each weight vector $\mathbf{w}^* \sim \mathrm{Multinomial}(\frac{1}{n}, \ldots, \frac{1}{n})$. Also note that the specific weight vector $\mathbf{w}^* = \mathbf{1}_n$ corresponding to the case where each original sample in $\mathcal{D}_n$ is taken exactly once and thus, continuing with the notation from the previous section, $s(\mathbf{1}_n)$ gives the complete (original) estimate. Since we are conditioned on $\mathcal{D}_n$, s^* is essentially a function of $(w_1^*, \ldots, w_n^*)$. Consider the linear space spanned by $\mathbf{w}^* = (w_1^*, \ldots, w_n^*)$ and denote the l^* as the projection of $s(\mathbf{w}^*)$ onto the linear space. We use $\mathrm{Var}_*(l^*)$ as an estimate of $\mathrm{Var}(\mathbb{E}_*[s^*])$.

We show that three different ideas converge to an identical variance estimator -$\mathrm{IJ_B}$, whenever all bootstrap samples are used.

Theorem 7. *Suppose we have data $\mathcal{D}_n$ and a statistic s. Let $(X_1^*, \ldots, X_n^*)$ be a general bootstrap sample of $\mathcal{D}_n$ and $s^* = s(X_1^*, \ldots, X_n^*)$, then*

1. $\mathbb{E}_*[s^* w_j^*] = e_j$*;*

2. $l^* = \sum_j w_j^* \beta_j$*, where* $\beta_j = (e_j - s_0)$*;*

3. $\mathrm{Var}_*(l^*) = \mathrm{JK_B} = \mathrm{IJ_B}$.

where $e_j = \mathbb{E}_*[s^* | X_1^* = X_j]$ *and* $s_0 = \mathbb{E}_*[s^*]$.

The proof of Theorem 7 can be found in Appendix D.1. In practice, limited by computational power, we typically don't have all bootstrap samples. Image that we draw B times of $(X_1^*, X_2^*, \ldots, X_n^*)$ and obtain $(X_{b1}^*, \ldots, X_{bn}^*)$ for $b = 1, \ldots, B$. For each b, we have $s_b^* = s(X_{b1}^*, \ldots, X_{bn}^*)$. Consider the bagging estimate $\overline{s^*} = \frac{1}{B}\sum_{b=1}^B s_b^*$. For the variance of $\overline{s^*}$, by law of total variance, we have

$$\begin{aligned} \mathrm{Var}(\overline{s^*}) &= \mathrm{Var}(\mathbb{E}[\overline{s^*}|\mathcal{D}_n]) + \mathbb{E}[\mathrm{Var}(\overline{s^*}|\mathcal{D}_n)] \\ &= \mathrm{Var}(\mathbb{E}_*[s^*]) + \frac{1}{B}\mathbb{E}[\mathrm{Var}_*(s^*)]. \end{aligned} \tag{31}$$

The dominant term in Eq. (31) is $\mathrm{Var}(\mathbb{E}_*[s^*])$, so the goal is to provide a good estimate of $\mathrm{Var}(\mathbb{E}_*[s^*])$. Now, since we don't have all bootstrap samples, we can not use $\mathrm{IJ_B}$ directly. However, we could estimate $\mathrm{IJ_B}$ with finite bootstrap samples. Thus, a natural estimate of cov_j is the $\widehat{\mathrm{cov}}_j$, the sample covariance of $(s_1^*, \ldots, s_B^*)$ and $(w_{1j}^*, \ldots, w_{Bj}^*)$. For $e_j - s_0$, s_0 can be estimated as $\sum_{b=1}^B s_b^*/B$. Since e_j the expected value of s^* given $X_i^* = X_j$ for $i = 1, \ldots, n$, thus a natural estimate is the weighted average of the mean of s_b^* where $X_i^* = X_j$. The weights are the proportion of $X_i^* = X_j$ in those B bootstrap samples. After simple calculation, we find that

$$\widehat{e_j} = \sum_{b=1}^B \frac{w_{bj}^*}{\sum w_{bj}^*} s_b^*, \quad \widehat{e_j - s_0} = \sum_{b=1}^B \left(\frac{w_{bj}^*}{\sum w_{bj}^*} - \frac{1}{B} \right) s_b^*. \tag{32}$$

Lastly, the estimate of $\mathrm{Var}_*(l^*)$ is suggested by $\widehat{\mathrm{Var}}(\hat{l})$, where - $\hat{l} = (\hat{l}_1, \ldots, \hat{l}_B)$ is the projection of $(s_1^*, \ldots, s_B^*)$ onto the linear space spanned by $(w_{1j}^*, \ldots, w_{Bj}^*)$ for $j = 1, \ldots, n$ and $\widehat{\mathrm{Var}}(\hat{l}) = \frac{1}{B}\sum_{b=1}^B (\hat{l}_b - \bar{\hat{l}})^2$ is the sample variance of $\hat{l}$.

In summary, we follow three different ideas to obtain $\mathrm{IJ_B}$, an identical estimation of $\mathrm{Var}(\mathbb{E}_*[s^*])$. The finite sample version of $\mathrm{IJ_B}$ will be different for these three approaches. In particular, we have

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{l}) &= \frac{1}{B}\sum_{b=1}^{B}(\hat{l}_b - \bar{\hat{l}})^2 \\
\widehat{\mathrm{JK}}_\mathrm{B} &= \sum_{j=1}^{n} \widehat{e_j - s_0}^{\,2} = \sum_{j=1}^{n}\left[\sum_{b=1}^{B}\left(\frac{w^*_{bj}}{\sum w^*_{bj}} - \frac{1}{B}\right)s^*_b\right]^2 \\
\widehat{\mathrm{IJ}}_\mathrm{B} &= \sum_{j=1}^{n}\widehat{\mathrm{cov}}_j^2 = \sum_{j=1}^{n}\left[\frac{1}{B-1}\sum_{b=1}^{B}(s^*_b - \bar{s^*})(w^*_{bj} - \bar{w^*_j})\right]^2 .
\end{aligned} \tag{33}$$

$\widehat{\mathrm{JK}}_\mathrm{B}$ was also proposed by [74] in the name of "Balanced Variance Estimation Method". We would expect that $\widehat{\mathrm{Var}}(\hat{l})$, $\widehat{\mathrm{JK}}_\mathrm{B}$ and $\widehat{\mathrm{IJ}}_\mathrm{B}$ would perform similarly in simulations.

5.3.2 The bias of $\widehat{\mathrm{IJ}}_\mathrm{B}$

Given an ensemble of size B, $\mathrm{Var}(\overline{s^*}) = \mathrm{Var}(\mathbb{E}_*[s^*]) + \mathbb{E}[\mathrm{Var}_*(s^*)]/B$. The dominant term is $\mathrm{Var}(\mathbb{E}_*[s^*])$. Therefore, we just need to understand how well $\widehat{\mathrm{IJ}}_\mathrm{B}$ estimates $\mathrm{Var}(\mathbb{E}_*[s^*])$. We consider the Monte Carlo bias and sampling bias of $\widehat{\mathrm{IJ}}_\mathrm{B}$ for estimating $\mathrm{Var}(\mathbb{E}_*[s^*])$, where $*$ refers to the bootstrap procedure . The sampling bias is considered with respect to variation of the data, whereas the Monte Carlo bias is considered with respect to bootstrap process conditioned on the data. We combine those two bias by

$$\widehat{\mathrm{IJ}}_\mathrm{B} - \mathrm{Var} \propto \underbrace{\mathbb{E}_*[\widehat{\mathrm{IJ}}_\mathrm{B}] - \mathrm{IJ}_\mathrm{B}}_{\textbf{Monte Carlo Bias}} + \underbrace{\mathbb{E}[\mathrm{IJ}_\mathrm{B}] - \mathrm{Var}}_{\textbf{Sampling Bias}} . \tag{34}$$

where $\widehat{\mathrm{IJ}}_\mathrm{B}$ is given in Eq. (33).

The Monte Carlo bias We first consider the Monte Carlo bias of $\widehat{\mathrm{IJ}}_\mathrm{B}$. We have $\mathbb{E}_*[\widehat{\mathrm{IJ}}_\mathrm{B}] - \mathrm{IJ}_\mathrm{B} = \sum_j \mathbb{E}_*[\widehat{\mathrm{cov}}_j^2] - \mathrm{cov}_j^2$ and $\mathbb{E}_*[\widehat{\mathrm{cov}}_j^2] - \mathrm{cov}_j^2 = \mathbb{E}_*[\widehat{\mathrm{cov}}_j^2] - \mathbb{E}_*^2[\widehat{\mathrm{cov}}_j] = \mathrm{Var}_*(\widehat{\mathrm{cov}}_j)$.

Next,

$$
\begin{aligned}
&\mathrm{Var}_*(\widehat{\mathrm{cov}}_j) \\
&= -\frac{B-2}{B(B-1)}\mathrm{Cov}_*^2(s^*, w_j^*) + \frac{\mathrm{Var}_*(s^*)\mathrm{Var}_*(w_j^*)}{B(B-1)} + \frac{\mathbb{E}_*[(s^* - \mathbb{E}_*[s^*])^2(w_j^* - \mathbb{E}_*[w_j^*])]^2}{B} \\
&= -\frac{(B-2)}{B(B-1)}\mathrm{Cov}_*^2(s^*, w_j^*) + \frac{(n-1)}{nB(B-1)}\mathrm{Var}_*(s^*) + \frac{1}{B}\mathbb{E}_*[(s^* - \mathbb{E}_*[s^*])^2(w_j^* - \mathbb{E}_*[w_j^*])]^2 \\
&= \frac{1}{B}\underbrace{\mathrm{Var}_*\left((s^* - \mathbb{E}_*[s^*])(w_j^* - \mathbb{E}_*[w_j^*]\right)}_{\mathrm{I}} + \frac{1}{B(B-1)}\underbrace{\left[\mathrm{Var}_*(s^*)\mathrm{Var}_*(w_j^*) + \mathrm{Cov}_*^2(s^*, w_j^*)\right]}_{\mathrm{II}}.
\end{aligned}
$$

Note that I is the dominant term and is $\mathcal{O}(1/B)$. Essentially, using $\widehat{\mathrm{cov}}_j^2$ to estimate cov_j^2 is analogous to using $\bar{X}^2$ to estimate $\mathbb{E}^2[X]$, which is biased since $\mathbb{E}[\bar{X}^2] = \mathbb{E}^2[X] + \mathrm{Var}(X)/B$. $\mathrm{Var}(X)/B$ might not be negligible, especially when B is not large and $\mathbb{E}[X]$ is small. $\mathrm{Cov}_*(s^*, w_j^*)$ is actually small, since s is permutation symmetric and thus the impact of w_j^* on the outcome of s^* is small. Therefore, a bias correction term is necessary.

Corollary 1. *A Monte Carlo bias corrected version of* $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ *is defined as*

$$
\widehat{\mathrm{IJ}}_{\mathrm{B}}^{mc} = \widehat{\mathrm{IJ}}_{\mathrm{B}} - \frac{1}{B}\sum_j \widehat{\mathrm{Var}}\left((s^* - \bar{s^*})(w_j^* - \bar{w_j^*})\right), \tag{35}
$$

where $\widehat{\mathrm{Var}}$ *denotes sample variance. The bias correction term is a sum of* n *terms. When* n *is small,* $\widehat{\mathrm{Var}}\left((s^* - \bar{s^*})(w_j^* - \bar{w_j^*})\right)$ *would not be minor. If additionally* B *is not large, then the bias correction term will be significant.*

Remark 3. *In recent work, [71] proposed the following Monte Carlo bias corrected* $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ *by*

$$
\widehat{\mathrm{IJ}}_{\mathrm{B}}^{efron} = \widehat{\mathrm{IJ}}_{\mathrm{B}} - \frac{n}{B}\widehat{\mathrm{Var}}(s^*). \tag{36}
$$

We can see that if $\mathrm{Var}_*((s^* - \mathbb{E}_*[s^*])(w_j^* - \mathbb{E}_*[w_j^*])$ *is close to* $\mathrm{Var}_*(s^* - \mathbb{E}_*[s^*])\mathrm{Var}_*(w_j^* - \mathbb{E}^*[w_j^*]) = (1 - \frac{1}{n})\mathrm{Var}_*(s^* - \mathbb{E}_*[s^*])$*, then Eq.* (35) *is close to Eq.* (36)*. We will conduct a simulation to compare Eq.* (35) *and Eq.* (36) *in the next section.*

The sampling bias Generally, the bias of IJ_{B} for bootstrap is difficult to understand, largely due to the replicates of bootstrap samples. We already found in Theorem 7 that $\mathrm{IJ}_{\mathrm{B}} = \sum_j (e_j - s_0)^2 = \mathrm{Var}_*(l^*)$. This observation enables us to move a small step further in terms of understanding the bias of IJ_{B}. First, let's take a look at how IJ_{B} behaves on some simple examples.

Example 1: sample mean Consider $s = s(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$. We have

$$s^* = \frac{1}{n}\sum_{i=1}^n X_i^*, \quad l^* = \sum_{i=1}^n \sum_{j=1}^n 1_{X_i^* = X_j}(e_j - s_0). \tag{37}$$

Then $\mathbb{E}_*[s^*] = \frac{1}{n}\sum_{i=1}^n X_i$ and $\mathrm{Var}_*(l^*) = \frac{1}{n^2}\sum_{i=1}^n (X_i - \bar{X})^2$. So

$$\mathrm{Var}(\mathbb{E}_*[s^*]) = \sigma^2/n, \quad \mathbb{E}[\mathrm{Var}_*(l^*)] = (n-1)\sigma^2/n^2. \tag{38}$$

Thus, we have $\frac{\mathbb{E}[\mathrm{Var}_*(l^*)]}{\mathrm{Var}(\mathbb{E}_*[s^*])} = \frac{n-1}{n} \to 1$ as $n \to \infty$. In Figure 5.1, $X_1, \ldots, X_n$ follow $\mathcal{N}(0, \sigma^2)$ where $n = 100$ and $\sigma^2 = 1$. Since we know that $\mathrm{Var}(\mathbb{E}_*[s^*]) = \sigma^2/n$, the oracle estimate would be $\hat{\sigma^2}/n$, where $\hat{\sigma^2}$ is the sample variance. The gray dash line is the true value of $\mathrm{Var}(\mathbb{E}_*[s^*])$. We find that $\widehat{\mathrm{IJ}}_{\mathrm{B}}^{mc}$ and $\widehat{\mathrm{IJ}}_{\mathrm{B}}^{efron}$ are quite close as expected and both perform well. The original $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ seems to be overestimating seriously when $B = 100$.

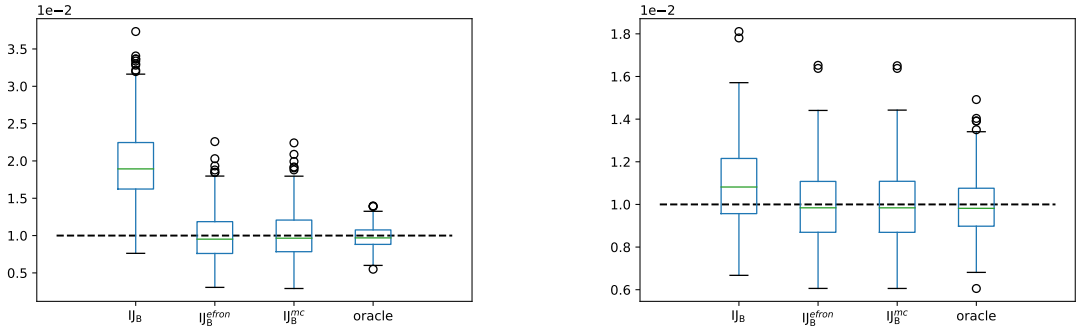

Figure 5.1: Performance of the infinitesimal jackknife and its bias-corrected alternatives on estimating the variance of the bagged sample mean (left: $B = 100$, right: $B = 1000$).

Example 2: sample variance Consider $s = \binom{n}{2}^{-1} \sum_{i<j}(x_i - x_j)^2$. We have

$$\mathbb{E}_*[s^*] = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \mathrm{Var}_*(l^*) = \frac{1}{n^2}\sum_i \left[(X_i - \bar{X})^2 - \frac{1}{n}\sum_i (X_i - \bar{X})^2\right]^2. \tag{39}$$

and then

$$\begin{aligned} \mathrm{Var}(\mathbb{E}_*[s^*]) &= \left(\frac{n-1}{n}\right)^2 \left[\frac{\mu_4}{n} - \frac{\mu_2^2}{n}\frac{n-3}{n-1}\right] \\ &= a_n\mu_4 - b_n\mu_2^2 \end{aligned} \tag{40}$$

and

$$\mathbb{E}[\mathrm{Var}_*(l^*)] = \frac{1}{n}\mathbb{E}[\mathbf{X}^T \mathrm{A}\mathbf{X}]^2, \quad \text{where } \mathrm{A} = \Sigma_1 - \frac{1}{n}\sum_i \Sigma_i. \tag{41}$$

Here $\Sigma_i = (e_i - \frac{1}{n}\mathbf{1}_n)(e_i^T - \frac{1}{n}\mathbf{1}_n^T)$, and $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$. After a careful calculation, we obtain

$$\begin{aligned} &\mathbb{E}[\mathrm{Var}_*(l^*)] \\ &= \frac{(n-1)}{n^2}\left[\mathbb{E}(X_1 - \bar{X})^4 - \mathbb{E}(X_1 - \bar{X})^2(X_2 - \bar{X})^2\right] \\ &= \left(\frac{n-1}{n}\right)^2 \left[\left(\frac{n^3 - (n-1)^2}{n^2(n-1)^2} + \frac{n}{(n-1)^5}\right)\mu_4 - \left(\frac{n^2 - 2n + 3}{(n-1)n^2} - \frac{3n^2(2n-3)}{(n-1)^5}\right)\mu_2^2\right] \\ &= a_n'\mu_4 - b_n'\mu_2^2. \end{aligned} \tag{42}$$

Thus, we have

$$\frac{a_n'}{a_n} = 1 + \tfrac{n^2+n-1}{n(n-1)^2} + \tfrac{n^2}{(n-1)^5} = 1 + \tfrac{1}{n} + o(\tfrac{1}{n}) \tag{43}$$

$$\frac{b_n'}{b_n} = 1 + \tfrac{n+3}{n(n-3)} - \tfrac{3n^3(2n-3)}{(n-1)^4(n-3)} = 1 - \tfrac{5}{n} + o(\tfrac{1}{n}). \tag{44}$$

Since $a_n'/a_n \to 1$ and $b_n'/b_n \to 1$ as $n \to \infty$, we have $\frac{\mathbb{E}[\mathrm{Var}_*(l^*)]}{\mathrm{Var}(\mathbb{E}_*[s^*])} \to 1$. $\mathrm{IJ_B}$ is asymptotically unbiased for estimating the variance of sample variance. Since sample variance is close to a linear statistic, the result is not surprising. In Figure 5.2, $X_1, \ldots, X_n$ follow $\mathcal{N}(0, \sigma^2)$ where $n = 100$ and $\sigma^2 = 1$. Since we know $\mathrm{Var}(\mathbb{E}_*[s^*]) = 2\sigma^4/n$, the oracle estimate would be $2(\hat{\sigma^2})^2/n$, where $\hat{\sigma^2}$ is the sample variance. The gray dash line is the true value of $\mathrm{Var}(\mathbb{E}_*[s^*])$. We find that $\widehat{\mathrm{IJ}}_{\mathrm{B}}^{mc}$ and $\widehat{\mathrm{IJ}}_{\mathrm{B}}^{efron}$ are quite close as expected and both perform well. The original $\widehat{\mathrm{IJ}}_{\mathrm{B}}$ seems to suffer the issue of overestimation when $B = 100$.

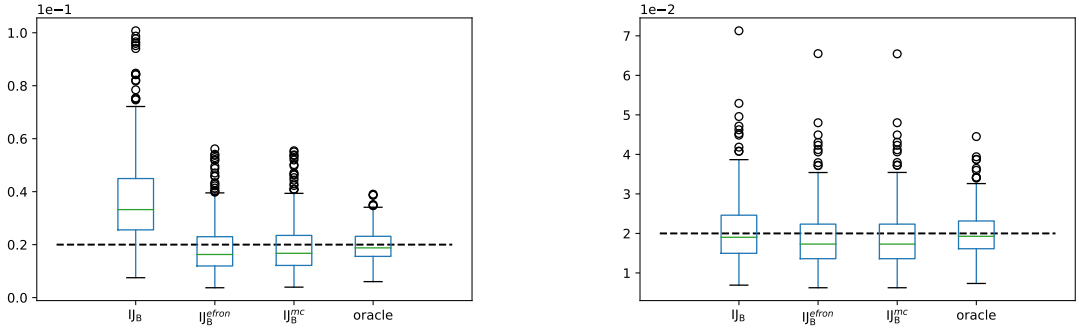

Figure 5.2: Performance of the infinitesimal jackknife and its bias-corrected alternatives on estimating the variance of the bagged sample variance (left: $B = 100$, right: $B = 1000$).

Example 3: sample maximum Consider $s = \max_i X_i$, where $X_1, \ldots, X_n$ are uniformly distributed between 0 and 1. Then,

$$\mathrm{Var}(\mathbb{E}_*[s^*]) = \frac{(n - \sum_{j=1}^n (\frac{j-1}{n})^n)(1 + \sum_{j=1}^n (\frac{j-1}{n})^n)}{(n+1)^2(n+2)} \tag{45}$$

and

$$\mathbb{E}[\mathrm{Var}_*(l^*)] = \frac{\sum_i [\sum_{j=1}^n (\frac{j-1}{n})^n - \sum_{j=i+1}^n (\frac{j-1}{n})^{n-1}]^2}{(n+1)(n+2)}. \tag{46}$$

The details of the calculation can be found in Appendix D.1. We have

$$\begin{aligned} \frac{\mathbb{E}[\mathrm{Var}_*(l^*)]}{\mathrm{Var}(\mathbb{E}_*[s^*])} &= \frac{(n+1)\sum_i [\sum_{j=1}^n (\frac{j-1}{n})^n - \sum_{j=i+1}^n (\frac{j-1}{n})^{n-1}]^2}{(n - \sum_j (\frac{j-1}{n})^n)(1 + \sum_j (\frac{j-1}{n})^n)} \\ &\to c \in [0.24, 0.25] \quad \text{as } n \to \infty. \end{aligned} \tag{47}$$

We can see that $\mathrm{IJ_B}$ is seriously underestimating of $\mathrm{Var}(\mathbb{E}_*[s^*])$. $\mathbb{E}_*[s^*]$ in this case is not close to a linear statistic, so $\mathrm{IJ_B}$ should not be expected to perform well. In Figure 5.3, $X_1, \ldots, X_n$ follow Unif(0, 1), where $n = 100$. The dash line is the true value of $\mathrm{Var}(\mathbb{E}_*[s^*])$. We don't have an oracle estimation for $\mathbb{E}_*[s^*]$ this time. We find that $\widehat{\mathrm{IJ}}_{\mathrm{B}}^{mc}$ and $\widehat{\mathrm{IJ}}_{\mathrm{B}}^{efron}$ are quite close as expected, but all three suffer the issue of underestimation, even when $B = 1000$.

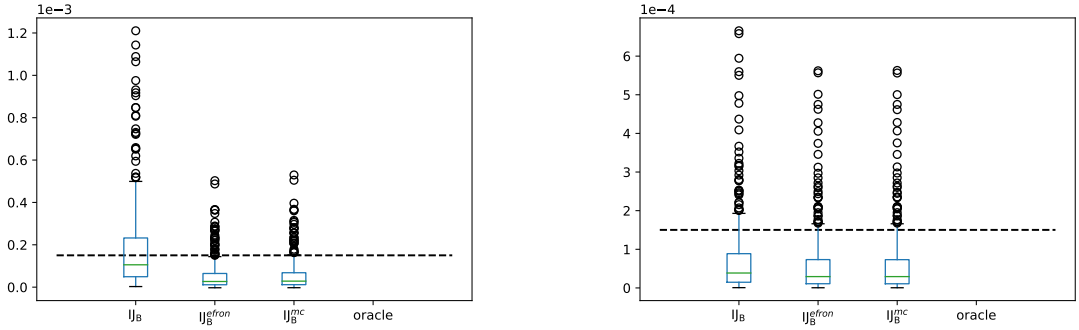

Figure 5.3: Performance of the infinitesimal jackknife and its bias-corrected alternatives on estimating the variance of the bagged sample maximum (left: $B = 100$, right: $B = 1000$).

5.3.3 The consistency of IJ_{B}

Generally, how well does $\mathrm{IJ}_{\mathrm{B}} = \mathrm{Var}_*(l^*)$ estimate $\mathrm{Var}(\mathbb{E}_*[s^*])$? Recall that the key step of IJ_{B} is essentially approximating $\mathbb{E}_*[s^*]$ by l_b and estimating $\mathrm{Var}(l_b)$ by $\mathrm{Var}_*(l^*)$. It turns out $\mathbb{E}_*[s^*] \approx l_b$ is sufficient for IJ_{B} to be consistent as shown in the following theorem.

Theorem 8. *If* $\mathbb{E}_*[s^*] = l_b + o_p\left(\frac{1}{n}\right)$, *then* $\mathrm{IJ}_{\mathrm{B}} \xrightarrow{p} \mathrm{Var}(\mathbb{E}_*[s^*])$.

Proof of Theorem 8 is provided in Appendix D.1. As shown in the above examples, for the cases of sample mean and sample variance, $\mathbb{E}_*[s^*] \approx l_b$ and $\mathrm{Var}_*(l^*)$ turns out to be a good estimate of the variance of $\mathbb{E}_*[s^*]$. For the case of sample maximum, $\mathbb{E}_*[s^*] \not\approx l_b$ and $\mathrm{Var}_*(l^*)$ turns out underestimate the variance of $\mathbb{E}_*[s^*]$. The following theorem suggests that it is necessary for $\mathbb{E}_*[s^*]$ to be asymptotic linear to make IJ_{B} consistent.

Theorem 9. *Let* $\mathbb{E}_*[s^*]$ *be the bootstrap smoothed alternative of* s, *then*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\mathrm{Var}_*(l^*)]}{\mathrm{Var}(\mathbb{E}_*[s^*])} = 1 \iff \lim_{n\to\infty} n(1-\rho) = 1, \tag{48}$$

where ρ *is the correlation between* e_1 *and* e_2 *and* $e_i = \mathbb{E}_*[s^*|X_1^* = X_i]$.

Consider the case that $s = \bar{X}$, which is linear. We obtain that $\rho = \frac{n^2-1}{n^2+n-1} = 1 - 1/n + o(1/n)$ and $\mathbb{E}[\mathrm{Var}_*(l^*)]/\mathrm{Var}(\mathbb{E}_*[s^*]) = \frac{n-1}{n} \to 1$. We suspect that to make $\rho = \mathrm{Cov}(e_1, e_2) = 1 - 1/n + o(1/n)$, $\mathbb{E}_*[s^*]$ requires to be equal to $l_b + o_p(1/n)$.

5.4 The pseudo infinitesimal jackknife estimate for U-statistic (s-IJ$_{\mathrm{U}}$)

5.4.1 IJ for U-statistic

The idea of the infinitesimal jackknife method can be extended to subsampling without replacement. In this case, $\mathbb{E}_*[s^*]$ is a U-statistic, which is more convenient for theoretical analysis and also more likely to be close to linear. Here s is a function of k i.i.d. random variables. The U-statistic can be written as

$$\mathrm{U} = \binom{n}{k}^{-1} \sum_{(n,k)} s(X_{i_1}, \ldots, X_{i_k}). \tag{49}$$

How does U depend on $\mathbb{P}_n$, such that $\mathrm{U} = f(\mathbb{P}_n)$ for some f? The dependence is abstract so that the subsampling proceeds according to the probabilities determined by $\mathbb{P}_n$. Following the definition of IJ, we have the following theorem.

Theorem 10. *The IJ estimator of the variance of a U-statistic is*

$$\mathrm{IJ}_{\mathrm{U}} = \frac{k^2}{n^2} \sum_{j=1}^{n} [\alpha e_j - \beta s_0]^2, \tag{50}$$

where $e_j = \mathbb{E}_*[s^* | X_1^* = X_j]$, $s_0 = \mathbb{E}_*[s^*]$ *and*

$$\alpha = 1 + \frac{1}{n}\left\{ \frac{k-1}{2} - \frac{1}{k} \sum_{j=0}^{k-1} \frac{j^2}{(n-j)} \right\}, \quad \beta = 1 + \frac{1}{k} \sum_{j=0}^{k-1} \frac{j}{n-j}.$$

If we write $\mathrm{Var}(\mathrm{U})$ *and* $\mathbb{E}[\mathrm{IJ}_{\mathrm{U}}]$ *in terms of* $V_1, \ldots, V_k$*, then the ratio of the coefficients of* V_j *in* $\mathbb{E}[\mathrm{IJ}_{\mathrm{U}}]$ *and that in* $\mathrm{Var}(\mathrm{U})$ *is* r_j*, where*

$$r_j = \frac{(n-k)^2}{n^2} \left[\frac{j}{1 - j/n} \alpha^2 + \frac{n}{k^2 (n-k)^2} (\alpha - \beta)^2 \right], \quad \textit{for } j = 1, \ldots, k. \tag{51}$$

Remark 4. *If* k *is fixed,* $\mathrm{Var}(\mathrm{U})$ *is dominated by the* V_1 *term. Since* $r_j \to j$ *for* $j = 1, \ldots, k$*,* IJ_{U} *is asymptotically unbiased.*

5.4.2 s-IJ$_{\mathrm{U}}$ for U-statistic

In recent work, [69] proposed another estimate of the variance of U-statistic. Since U-statistic is just subsampling without replacement, which is just slightly different from bootstrap, they copied the format of the IJ for bootstrap to U-statistic and obtained

$$\sum_j \mathrm{Cov}_*^2(s^*, w_j^*), \tag{52}$$

where $*$ refers to the subsampling procedure. However, we would call it the pseudo infinitesimal jackknife estimate (s-IJ$_{\mathrm{U}}$), since it is not derived from the orgininal definition of infinitesimal jackknife. A more rigorous motivation for s-IJ$_{\mathrm{U}}$ is provided as following. Recall that from the derivative of IJ, we assume that $f(\mathbb{P}_n) - f(\mathrm{P}) = \frac{1}{n}\sum_{i=1}^n f'(\mathrm{P}, X_i) + o_p(1/n)$, where the dominated term is a sum of i.i.d. random variables. And we estimate the variance of $f'(\mathrm{P}, X_i)$ by $\frac{1}{n}\sum_{i=1}^n f'^2(, \mathbb{P}_n, X_i)$. Now consider that we rewrite $f(\mathbb{P}_n) - f(\mathrm{P}) = \sum_{i=1}^n \sum g(X_i) + o_p(1/n)$, where $g(X_i)$ is not necessarily $\frac{1}{n}f'(\mathrm{P}, X_i)$. From the theory of U-statistic, we know that there is a natural candidate for $g(X_i)$, which is the Hájek projection - $\mathbb{E}[f(\mathbb{P}_n) - f(\mathrm{P})|X_i] = \frac{k}{n}\mathbb{E}[s - \mathbb{E}[s]|X_i]$. Since $\mathrm{Var}(\sum g(X_i)) = \frac{k^2}{n}V_1$, where $V_1 = \mathrm{Var}(\mathbb{E}[s|X_1])$, we just need to propose a reasonable estimate $\hat{V}_1$ for V_1 and use

$$\frac{k^2}{n}\hat{V}_1 \tag{53}$$

as an estimate of the variance of U-statistic. Since $V_1 = \mathbb{E}[\mathbb{E}[s|X_1] - \mathbb{E}[s]]^2$, a natural candidate for $\hat{V}_1$ would be

$$\frac{1}{n}\sum_j (\mathbb{E}_*[s^*|X_1^* = X_j] - \mathbb{E}_*[s^*])^2 = \frac{1}{n}\sum_j (e_j - s_0)^2. \tag{54}$$

It turns out that Eq. (53) is the same as Eq. (52).

Proposition 4. *Let $\mathcal{D}_n^* = (X_1^*, \ldots, X_k^*)$ be a general subsample of size k of $\mathcal{D}_n$ and $w_j^* = \mathbf{1}_{X_j \in \mathcal{D}_n^*}$, then*

$$\mathrm{Cov}_*(s^*, w_j^*) = \frac{k}{n}(e_j - s_0),$$

where $*$ refers to the procedure of subsampling without replacement, $e_j = \mathbb{E}_*[s^*|X_1^* = X_j]$ and $s_0 = \mathbb{E}_*[s^*]$. And then

$$\begin{aligned} \text{s-IJ}_{\text{U}} &= \sum_j \mathrm{Cov}_*^2(s^*, w_j^*) \\ &= \left(\frac{\binom{k}{1}}{\binom{n}{1}}\right)^2 \sum_j (e_j - s_0)^2. \end{aligned} \tag{55}$$

Analogous to IJ_{U}, we have the following theorem.

Theorem 11. *The pseudo infinitesimal jackknife estimate of the variance of a U-statistic is*

$$\text{s-IJ}_{\text{U}} = \frac{k^2}{n^2} \sum_{j=1}^{n} [e_j - s_0]^2, \tag{56}$$

where $e_j = \mathbb{E}_*[s^*|X_1^* = X_j]$ and $s_0 = \mathbb{E}_*[s^*]$. If we write the $\mathrm{Var(U)}$ *and* $\mathbb{E}[\text{s-IJ}_{\text{U}}]$ *in terms of $V_1, \ldots, V_k$, then the ratio of the coefficients of V_j in* $\mathbb{E}[\text{s-IJ}_{\text{U}}]$ *and that in* $\mathrm{Var(U)}$ *is r_j, where*

$$r_j = \left(\frac{n-k}{n}\right)^2 \frac{j}{1 - j/n}, \quad \textit{for } i = 1, \ldots, k. \tag{57}$$

Note that although our goal is using $\frac{k^2}{n}\hat{V}_1$ to estimate $\frac{k^2}{n}V_1$, $\mathbb{E}[\frac{k^2}{n}\hat{V}_1] = \mathbb{E}[\text{s-IJ}_{\text{U}}]$ involves higher order terms with $V_2, \ldots, V_k$. This is not what we want, but it is unavoidable since we are don't have new data generated from the underlying distribution. If we simply multiply s-IJ_{U} by $(\frac{n}{n-k})^2 \frac{n-1}{n}$ as proposed by [69], then only the first term is unbiased, but doubles the quadratic term, triples the cubic term and etc. This can explain why this estimation is inflated in practice. In many applications, k is not that small, and thus the higher order terms of the $\mathrm{Var(U)}$ is not negligible, so the effect of r_j cannot be ignored. A similar phenomenon was discovered by [29] for the jackknife estimation of variance.

5.4.3 The consistency of s-$\mathrm{IJ_U}$

If $k/n \to 0$, then $\alpha \to 1$ and $\alpha - \beta \to 0$, and thus $\mathrm{IJ_U} \to \text{s-}\mathrm{IJ_U}$. Other than that, s-$\mathrm{IJ_U}$ looks slighter simpler in format. Due to the nice structure of U-statistic, besides the bias of s-$\mathrm{IJ_U}$, we can actually talk about the consistency of s-$\mathrm{IJ_U}$. Moreover, not only for the U-statistics, we can even talk about generalized U-statistics - $\mathrm{U}_{n,k,N,\omega}$ as defined in Definition 1. Indeed, let $e_i^\omega = \binom{n-1}{k-1}^{-1} \sum s(X_i, \ldots, ; \omega)$ and $s_0^\omega = \binom{n}{k}^{-1} \sum s(\ldots; \omega)$. Note that each collection of subsamples is paired with an i.i.d. ω. Then the s-$\mathrm{IJ_U}$ is defined as

$$\text{s-}\mathrm{IJ}_\mathrm{U}^\omega = \frac{k^2}{n^2} \sum [e_i^\omega - s_0^\omega]^2. \tag{58}$$

Theorem 1 states that if $\frac{k}{n}(\zeta_k/k\zeta_{1,\omega} - 1) \to 0$, then the complete generalized U-statistic - $\mathrm{U}_{n,k,\omega}$ is asymptotic normal with variance of $\frac{k^2}{n}\zeta_{1,\omega}$, where $\zeta_k = \mathrm{Var}(s)$ and $\zeta_{1,\omega} = V_1 = \mathrm{Var}(\mathbb{E}[s|X_1])$. When the same conditions are met, s-$\mathrm{IJ}_\mathrm{U}^\omega$ is consistent. In other words, if the U-statistic is almost linear, then s-$\mathrm{IJ}_\mathrm{U}^\omega$ will be consistent, i.e. s-$\mathrm{IJ}_\mathrm{U}^\omega \xrightarrow{p} \mathrm{Var}(\mathrm{U}_{n,k,\omega})$.

Theorem 12. *Let $X_1, \ldots, X_n$ be i.i.d. from F_X and $U_{n,k,\omega}$ be a generalized complete U-statistic with kernel $s(X_1, \ldots, X_k; \omega)$. Let $\theta = \mathbb{E}[s]$, $\zeta_{1,\omega} = \mathrm{Var}(\mathbb{E}[s|X_1])$ and $\zeta_k = \mathrm{Var}(s)$. If $\frac{k}{n}(\frac{\zeta_k}{k\zeta_{1,\omega}} - 1) \to 0$, then*

$$\text{s-}\mathrm{IJ}_\mathrm{U}^\omega \xrightarrow{p} \mathrm{Var}(\mathrm{U}_{n,k,\omega}). \tag{59}$$

Corollary 2. *If the conditions in Theorem 12 are held, then*

$$\begin{aligned} \mathrm{U}_{n,k,\omega} \pm z_{\alpha/2}\frac{n}{n-k}\sqrt{\text{s-}\mathrm{IJ}_\mathrm{U}^\omega} &= \mathrm{U}_{n,k,\omega} \pm z_{\alpha/2}\frac{n}{n-k}\sqrt{\sum \mathrm{Cov}_*^\omega(s^*, w_i^*)^2} \\ &= \mathrm{U}_{n,k,\omega} \pm z_{\alpha/2}\frac{k}{n-k}\sqrt{\sum (e_i^\omega - s_0^\omega)^2} \end{aligned} \tag{60}$$

provides an asymptotically valid confidence interval for θ with confidence level $1-\alpha$. The $\frac{n}{n-k}$ there is for correcting the bias for non-asymptotic situation. Note that $\mathrm{Cov}_*^\omega(s^*, w_i^*)$ can be defined similarly and $\mathrm{Cov}_*^\omega(s^*, w_i^*) = \frac{k}{n}(e_i^\omega - s_0^\omega)$.

However, the complete forms of these estimators are almost never utilized in practice due to the computational burden involved with calculating $\binom{n}{k}$ base learners. We established the asymptotic normality for incomplete generalized U-statistics in Theorem 2. We need to estimate the asymptotic variance of $\mathrm{U}_{n,k,N,\omega}$ well such that a valid confidence interval can be conducted. Let

$$\text{s-IJ}_{\mathrm{U}}^{\dagger} = \frac{k^2}{n^2}\sum_i [e_i^{\dagger} - s_0^{\dagger}]^2, \tag{61}$$

where

$$s_0^{\dagger} = \frac{1}{N}\sum s(...;\omega), \quad e_i^{\dagger} = \frac{n}{Nk}\sum s(X_i, ...;\omega). \tag{62}$$

$\sum s(...;\omega)$ denotes the sum of all kernels that builds the incomplete U-statistic, whereas $\sum s(X_i, ...;\omega)$ denotes the sum of all kernels that builds the incomplete U-statistic and includes X_i in their subsamples.

Theorem 13. *Let $X_1, \ldots, X_n$ be i.i.d. from F_X and $\text{s-IJ}_{\mathrm{U}}^{\dagger}$ be as Eq. (61). Let $\theta = \mathbb{E}[s]$, $\zeta_{1,\omega} = \mathrm{Var}(\mathbb{E}[s|X_1])$ and $\zeta_k = \mathrm{Var}(s)$. Then if $\frac{k}{n}(\frac{\zeta_k}{k\zeta_{1,\omega}} - 1) \to 0$ and $\frac{n}{N}\frac{\zeta_k}{k\zeta_{1,\omega}} \to 0$, we have*

$$\text{s-IJ}_{\mathrm{U}}^{\dagger} \xrightarrow{p} \frac{k^2}{n}\zeta_{1,\omega}. \tag{63}$$

Remark 5. *Consideingr the case that $1 < c \le \zeta_k/k\zeta_{1,\omega} \le C$, Theorem 13 states that $N \gg n$ is required to make $\text{s-IJ}_{\mathrm{U}}^{\dagger} \xrightarrow{p} \frac{k^2}{n}\zeta_{1,\omega}$.*

Corollary 3. *In the literature of random forest, $\sum_i \widehat{\mathrm{Cov}}^2(s^*, w_i^*)$ has been proposed to estimate the variance of random forest. Actually, $\text{s-IJ}_{\mathrm{U}}^{\dagger}$ has strong connection to it. Indeed,*

$$\begin{aligned}
\text{s-IJ}_{\mathrm{U}}^{\dagger} &= \frac{k^2}{n^2}\sum_i [e_i^{\dagger} - s_0^{\dagger}]^2 \\
&= \frac{k^2}{n^2}\cdot\frac{n^2}{k^2}\sum [\frac{1}{N}s(...;\omega)w_i^* - \frac{1}{N}\sum s(...;\omega)\frac{k}{n}]^2 \\
&= \sum_i \left[\frac{1}{N}\sum s(...;\omega)w_i^* - \frac{1}{N}\sum s(...;\omega)\mathbb{E}_*[w_i^*]\right]^2 \\
&\approx \frac{\hat{N}^2}{N^2}\sum_i \widehat{\mathrm{Cov}}^2(s^*, w_i^*) \\
&\approx \sum_i \widehat{\mathrm{Cov}}^2(s^*, w_i^*).
\end{aligned} \tag{64}$$

We can show that under the same condition of Theorem 13, $\sum_i \widehat{\mathrm{Cov}}^2(s^*, w_i^*) \xrightarrow{p} \frac{k^2}{n}\zeta_{1,\omega}$.

To our knowledge, Corollary 3 is the first set of results to show the consistency of $\sum_i \widehat{\mathrm{Cov}}^2(s^*, w_i^*)$ in estimating the variance of random forests that built on subsamples.

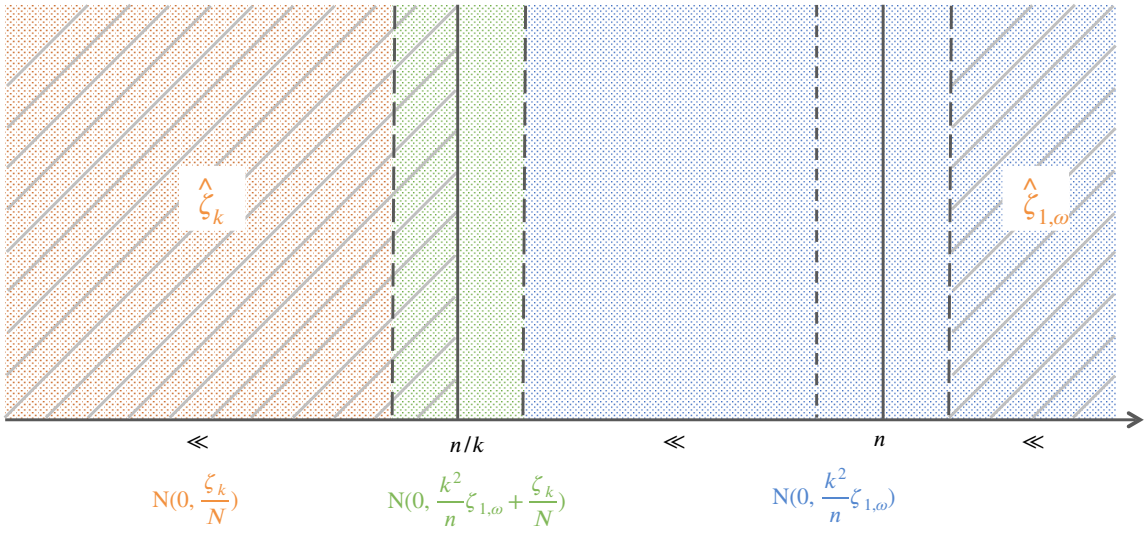

Figure 5.4: The distribution of $\mathrm{U}_{n,k,N,\omega}$ as a function of N. $N \gg n$ is required to for s-IJ$_{\mathrm{U}}^{\dagger}$ to estimate $\zeta_{1,\omega}$ consistently.

According to Theorem 2, $\mathrm{U}_{n,k,N,\omega}$ has different asymptotic distributions bases on the value of N. When $N \ll n/k$, $\mathrm{U}_{n,,N,\omega} \sim \mathcal{N}(0, \zeta_k/N)$; when $N = O(n/k)$, $\mathrm{U}_{n,,N,\omega} \sim \mathcal{N}(0, \frac{k^2}{n}\zeta_{1,\omega} + \frac{\zeta_k}{N})$ and when $N \gg n/k$ and $\mathcal{N}(0, \frac{k^2}{n}\zeta_{1,\omega})$. We can estimate ζ_k simply by calculating the sample variance of base learners built on non-overlapping subsamples. Base on the above argument, $\zeta_{1,\omega}$ can be estimated by s-IJ$_{\mathrm{U}}^{\dagger}$. Therefore, it is guaranteed that

$$\text{s-IJ}_{\mathrm{U}}^{\dagger} \xrightarrow{p} \frac{k^2}{n}\zeta_{1,\omega} \quad \text{and} \quad \hat{\zeta}_k \xrightarrow{p} \zeta_k. \tag{65}$$

Interestingly, if a random forest is built with N decision trees, where $N = O(n)$, then we can not estimate the variance of the random forest consistently if just use the trees that build the random forest. We actually need $\gg n$ many decision trees. This results shed light on the intuition that it is always more computational intensive to estimate the variance of ensembles then obtaining the ensemble itself.

5.4.4 Higher order s-IJ$_{\mathrm{U}}$

Recall that $\mathrm{Var(U)} = \sum_{j=1}^{k} \binom{k}{j}^2 \binom{n}{j}^{-1} V_k$. In the previous discussion, we assume that the U-statistic is close to a linear statistic, thus the variance of U-statistic is dominated by its first order term k^2/nV_1 and we propose an estimate of V_1 accordingly. People might wonder if the remaining terms are not negligible, can we propose some estimates of V_j for $j = 2, \ldots, k$ and obtain an estimate of Var(U)?

We first consider the results for V_2 and extend them to all j, for $3 \leq j \leq k$ in a corollary. Since $V_2 = \mathrm{Var}(\mathbb{E}[s|X_1, X_2] - \mathbb{E}[s|X_1] - \mathbb{E}[s|X_2] + \mathbb{E}[s])$, a natural estimate for the second order term in Var(U) -$\binom{k}{2}\binom{n}{2}^{-1} V_2$ would be

$$\left(\binom{k}{2} / \binom{n}{2}\right)^2 \sum_{i,j} [e_{ij} - e_i - e_j + s_0]^2, \tag{66}$$

where $e_{ij} = \mathbb{E}_*[s^*|X_1^* = X_i, X_2^* = X_j]$. Before analyzing the property of this quantity, it is interesting to point out its connection to the s-IJ$_{\mathrm{U}}$.

Proposition 5. *Let $\mathcal{D}_n^* = (X_1^*, \ldots, X_k^*)$ be a general subsample of $\mathcal{D}_n$ and $w_{ij}^* = 1_{X_i, X_j \in \mathcal{D}_n^*} - \frac{k}{n} 1_{X_i \in \mathcal{D}_n^*} - \frac{k}{n} 1_{X_j \in \mathcal{D}_n^*} + \frac{k(k-1)}{n(n-1)}$, then*

$$\mathrm{Cov}_*(s^*, w_{ij}^*) = \binom{k}{2} / \binom{n}{2} (e_{ij} - e_i - e_j + s_0) \tag{67}$$

where $*$ refers the procedure of subsampling without replacement and $e_{ij} = \mathbb{E}_*[s^*|X_1^* = X_i, X_2^* = X_j]$. We call Eq. (66) *the second order pseudo-IJ estimator of U-statistics:*

$$\text{s-IJ}_{\mathrm{U}}(2) = \sum_{i,j} \mathrm{Cov}_*^2(s^*, w_{ij}^*) = \binom{k}{2}^2 / \binom{n}{2}^2 \sum_{i,j} [e_{ij} - e_i - e_j + s_0]^2. \tag{68}$$

Note that s-IJ$_{\mathrm{U}}$ involves the covariance of s^* and w_j^*, the count of the single variable in a subsample, whereas s-IJ$_{\mathrm{U}}(2)$ involves the of s^* and w_{ij}^*, the count of pairs of variables. Therefore, s-IJ$_{\mathrm{U}}(2)$ is a natural extension of s-IJ$_{\mathrm{U}}$. For notational convenience, we also write s-IJ$_{\mathrm{U}}$ as s-IJ$_{\mathrm{U}}(1)$. Similarly,

Corollary 4. *For $d = 1, \ldots, k$, we define that*

$$\text{s-IJ}_{\text{U}}(d) = \sum_{(n,d)} \text{Cov}_*^2(s^*, w^*_{i_1,\ldots,i_d}) = \binom{k}{d}^2 / \binom{n}{d}^2 \sum_{(n,d)} \left[\sum_{j=0}^{d} (-1)^{d-j} \sum_{(d,j)} e_{i_1,\ldots,i_j} \right]^2, \qquad (69)$$

where $w^*_{i_1,\ldots,i_d} = \sum_{j=0}^{d} (-1)^{d-j} \frac{\binom{n-d+j}{k-d+j}}{\binom{n}{k}} \left[\sum_{(d,j)} \prod w^*_{i_j} \right]$. The expression for $w^*_{i_1,\ldots,i_d}$ is involved because we are considering subsampling without replacement. If it is subsampling with replacement, then $w^*_{i_1,\ldots,i_d} = \prod (w^*_{i_j} - 1)$.

Like $\mathbb{E}[\text{s-IJ}_{\text{U}}]$, $\mathbb{E}[\text{s-IJ}_{\text{U}}(d)]$ is a linear combination of the V_js. We derive $\mathbb{E}[\text{s-IJ}_{\text{U}}(2)]$ in the appendix. The expression for $d \geq 3$ can be derived in the same spirit. Let $a_i = \binom{n-i}{k-i}^{-1}$ for $i = 0, 1, \ldots, d$, and define b_i for $i = 0, 1, \ldots, d$ by

$$\begin{aligned}
&b_0 = a_0 \\
&b_1 = a_1 - a_0 = a_1 - b_0 \\
&\vdots \\
&b_d = a_d - \binom{d}{1} a_{d-1} + \binom{d}{2} a_{d-2} + \ldots a_0 = a_d - \binom{d}{1} b_{d-1} - \binom{d}{2} b_{d-2} - \cdots - b_0.
\end{aligned}$$

Let $c_i = b_i \binom{n-d}{k-i}$ and $m_i = c_{d-i}$ for $i = 0, \ldots, d$. Then for $j = 1, \ldots, k$, the coefficient of V_j in $\mathbb{E}[\text{s-IJ}_{\text{U}}(d)]$ is $\binom{k}{d}^2 / \binom{n}{d} \lambda_j(d)$, where

$$\begin{aligned}
\lambda_j(d) = &\binom{d}{0} \binom{n-d}{j-d}^{-1} \left(m_0 \binom{n-d}{j-d} \right)^2 \\
&+ \binom{d}{1} \binom{n-d}{j-d+1)}^{-1} \left[m_1 \binom{k-d+1}{j-d+1} - m_0 \binom{k-d}{j-d+1} \right]^2 \\
&+ \binom{d}{2} \binom{n-d}{j-d+2}^{-1} \left[m_2 \binom{k-d+2}{j-d+2} - \binom{2}{1} m_1 \binom{k-d+1}{j-d+2} + m_0 \binom{k-d}{j-d+2)} \right]^2 \\
&\vdots \\
&+ \binom{d}{d} \binom{n-d}{j}^{-1} \left[m_d \binom{k}{j} - \binom{d}{d-1} m_{d-1} \binom{k-1}{j} \right. \\
&\quad \left. + \binom{d}{d-2} m_{d-2} \binom{k-2}{j} - \ldots \binom{d}{1} m_0 \binom{k-d}{j} \right]^2.
\end{aligned}$$

Putting all together, we have

Proposition 6. *By writing the* $\mathrm{Var(U)}$ *and* $\mathbb{E}[\text{s-IJ}_{\mathrm{U}}]$ *in terms of* $V_1, \ldots, V_k$*, the ratio of the coefficients of* V_j *in* $\mathbb{E}[\text{s-IJ}_{\mathrm{U}}(d)]$ *and that in* $\mathrm{Var(U)}$ *is* $r_j(d)$*, where*

$$r_j(d) = \frac{\lambda_j(d)\binom{k}{d}^2\binom{n}{d}^{-1}}{\binom{k}{j}^2\binom{n}{j}^{-1}}, \quad j = 1, \ldots, k. \tag{70}$$

And $r_j(d)$ *is monotonically increasing w.r.t.* j*.*

Letting $n = 20$ and $k = 10$, we plot the curve of $r_j(d)$ for to get a glimpse of how it behaves. We hope r_j be close to 1, at least for small j, because the Var(U) should be

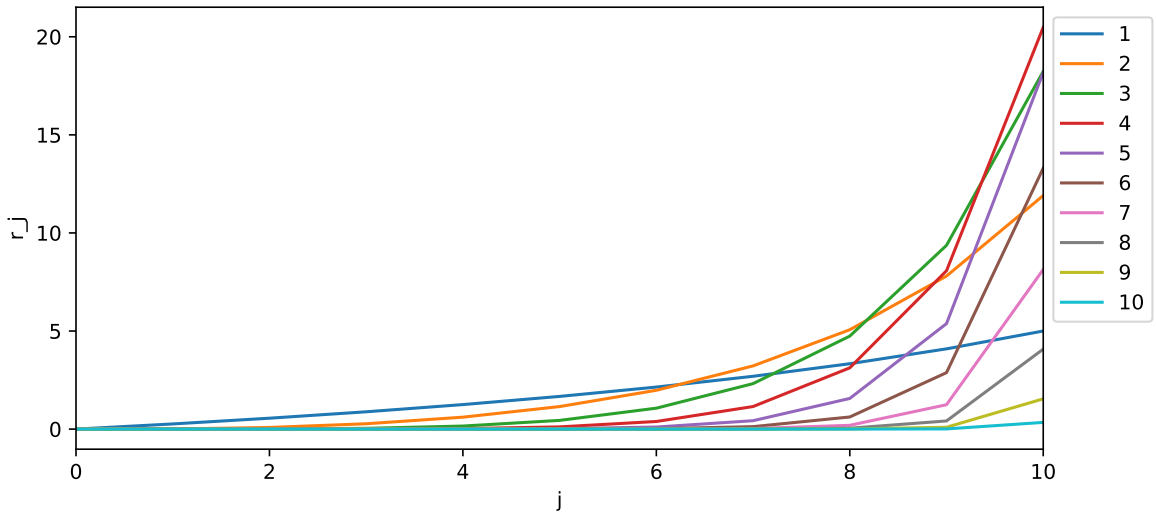

Figure 5.5: A plot of $\{r_j(d)\}_{j=1}^k$, where $n = 20$ and $k = 10$. As d increases, the curve of r_j is bending further away the horizontal line.

dominated by the first several terms. From Figure 5.5, it seems like $\text{s-IJ}_{\mathrm{U}}(1)$ performs better than others. It would be interesting to see whether combining $\text{s-IJ}_{\mathrm{U}}(d)$ for $d = 1, \ldots, k$ in some way outer performs $\text{s-IJ}_{\mathrm{U}}(1)$ or not for future research.

6.0 Discussion

We establish distributional results for random forest estimators, which take the form of *generalized* U-statistics. We showed that under mild regularity conditions, such estimators tend to a normal distribution so long as $\frac{s}{n}\left(\frac{\zeta_s}{s\zeta_{1,\omega}} - 1\right) \to 0$. When kernels are well-behaved, this thus implies that subsamples may be taken on the order of n while retaining the asymptotic normality of the estimator. In practice, we expect that this condition is often most naturally satisfied by subsampling at a slower rate with $s = o(n)$ and ensuring that the corresponding variance ratio $\zeta_s/s\zeta_{1,\omega}$ is bounded. In Chapter 3 we showed that the variance ratio is well-behaved for a number of nearest-neighbor-type base learners. In general though, such behavior is not well-understood in theory, particularly for adaptive learners. However, in Chapter 5, we propose consistent estimates of $\zeta_{1,\omega}$ and ζ_s. So the behavior of $\zeta_s/s\zeta_{1,\omega}$ can be understood in simulation. More importantly, we can make predictions with theoretically supported prediction intervals, shedding insights into the accuracy of our prediction. In Chapter 4 we provide Berry-Esseen bounds to quantify the proximity of these estimators to the normal distribution. Theorem 3 provides the sharpest bound to date on this rate for complete, infinite-order U-statistics, while the bounds that follow are each the first of their kind.

Appendix A Proofs in Chapter 2

Here we provide a fuller discussion of the previously established central limit theorems for randomized, incomplete, infinite-order U-statistics, paying particular attention to the relationship between the projection method utilized and the resulting subsampling rate necessary in order to retain asymptotic normality. As noted in Chapter 2, [54] provided one such theorem, but with somewhat strict conditions. First, the authors require that for all $\delta > 0$,

$$\frac{1}{\zeta_{1,\omega}} \int_{|h_1(z)| \geq \delta \sqrt{n\zeta_{1,\omega}}} h_1^2(z)\, dP \to 0 \quad (n \to \infty)$$

where $h_1(z) = \mathbb{E}[h(z, Z_2, \ldots, Z_s; \omega)] - \theta$. Note however that so long as $\mathbb{E}[h^2(Z_1, \ldots, Z_s; \omega)] < \infty$,

$$\frac{1}{\zeta_{1,\omega}} \int_{|h_1| \geq \delta \sqrt{n\zeta_{1,\omega}}} h_1^2(Z)\, dP = \int_{\left|\frac{h_1}{\sqrt{\zeta_{1,\omega}}}\right| \geq \delta\sqrt{n}} \left(\frac{h_1}{\sqrt{\zeta_{1,\omega}}}\right)^2 dP$$

automatically tends to 0 as $n \to \infty$ and thus this condition is redundant for kernels assumed to have finite second moment.

In Section 2, we noted that there is strong reason to suspect that a subsampling rate of $s = o(n^{1/2})$ is the largest possible when the results are established via Hájek projections. We now elaborate on that point here.

Let $\mathcal{S}$ denote the set of all variables of the form $\sum_{i=1}^n g_i(Z_i)$ for arbitrary measurable functions $g_i : \mathbb{R}^d \mapsto \mathbb{R}$ with $\mathbb{E}[g_i^2(Z_i)] < \infty$ $(i = 1, \ldots, n)$. The Hájek projection of $U_{n,s}$ onto $\mathcal{S}$ is

$$\hat{U}_{n,s} = \theta + \frac{s}{n} \sum_{i=1}^n h_1(Z_i).$$

Now, by the central limit theorem for i.i.d case, we have $\sqrt{n}\hat{U}_{n,s}/\sqrt{s^2\zeta_1} \rightsquigarrow N(0,1)$ and thus by Theorem 11.2 in [67], to obtain the asymptotic normality of U-Statistic, it is sufficient to demonstrate that $\mathrm{Var}(U_{n,s})/\mathrm{Var}(\hat{U}_{n,s}) \to 1$. This is straightforward when the rank of the kernel is fixed but requires more careful attention whenever s is allowed to grow with n. The

variance of the U-statistic is

$$\begin{aligned}\mathrm{Var}(U_{n,s}) &= \binom{n}{s}^{-1} \sum_{\beta} \sum_{\beta'} \mathrm{Cov}(h(Z_{\beta_1}, \ldots, Z_{\beta_s}), h(Z_{\beta'_1}, \ldots, Z_{\beta'_s})) \\ &= \binom{n}{s}^{-1} \sum_{j=1}^{s} \binom{s}{j} \binom{n-s}{s-j} \zeta_j \\ &= \sum_{j=1}^{s} \frac{s!^2}{j!(s-j)!^2} \frac{(n-s)\cdots(n-2s+j+1)}{n(n-1)\cdots(n-s+1)} \zeta_j\end{aligned}$$

where β indexes subsamples of size s, and the variance of $\hat{U}_{n,s}$ is

$$\mathrm{Var}(\hat{U}_{n,s}) = \frac{s^2}{n} \mathrm{Var}\left(h_1(Z_1)\right) = \frac{s^2}{n} \zeta_1.$$

The variance ratio is then $\mathrm{Var}(U_{n,s})/\mathrm{Var}(\hat{U}_{n,s}) = (a_n + b_n)/c_n$, where

$$\begin{aligned}a_n &= \frac{s^2}{n} \frac{(n-s)\cdots(n-2s+2)}{(n-1)\cdots(n-s+1)} \zeta_1, \\ b_n &= \binom{n}{s}^{-1} \sum_{j=2}^{s} \binom{s}{j} \binom{n-s}{s-j} \zeta_j, \\ c_n &= \frac{s^2}{n} \zeta_1.\end{aligned}$$

Thus, in order for the variance ratio to converge to 1, it suffices to show $a_n/c_n \to 1$ and $b_n/c_n \to 0$. To transform these two conditions with respect to s and n, we introduce the following lemmas.

Lemma 1 ([50])**.** *For $1 \le c \le d \le s$, $\zeta_s/c \le \zeta_d/d$.*

Lemma 2. *Let $H(n,s) = \left[\frac{(n-s)\cdots(n-2s+2)}{(n-1)\cdots(n-s+1)}\right]$, then $s/\sqrt{n} \to 0$ if and only if $H(n,s) \to 1$.*

Proof. When $s/\sqrt{n} \to 0$, we have

$$\begin{aligned}H(n,s) &\ge \left[\frac{n-2s+2}{n-1}\right]^{s-1} \\ &= \exp\left[(s-1)\log\left(1 - \frac{2s-3}{n-1}\right)\right] \\ &\approx \exp\left[-\frac{2s^2}{n}\right] \to 1.\end{aligned}$$

If there exists a subsequence $\{s'\}$ such that $s'/\sqrt{n'} \geq c$ for some constant $c > 0$, then

$$\begin{aligned} H(n', s') &\leq \left[\frac{n' - 3s'/2 + 1}{n - s'/2}\right]^{s'-1} \\ &= \exp\left[(s'-1)\log\left(1 - \frac{s'-1}{n - s'/2}\right)\right] \\ &\approx \exp\left[-\frac{s'^2}{n'}\right] < 1. \end{aligned}$$

□

Now, we can transform the conditions on a_n, b_n and c_n into conditions on n and s. Note that

$$a_n/c_n = H(n, s)$$

and

$$\begin{aligned} b_n/c_n &= \binom{n-1}{s-1}^{-1} \left\{\sum_{j=1}^{s-1} \frac{1}{j+1}\binom{s-1}{j}\binom{(n-1)-(s-1)}{(s-1)-j}\frac{\zeta_{j+1}}{\zeta_1}\right\} \\ &= \binom{n-1}{s-1}^{-1} \left\{\sum_{j=1}^{s-1} \binom{s-1}{j}\binom{(n-1)-(s-1)}{(s-1)-j}\frac{\zeta_{j+1}}{(j+1)\zeta_1}\right\} \\ &\geq 1 - \left[\frac{(n-s)\cdots(n-2s+2)}{(n-1)\cdots(n-s+1)}\right] \\ &= 1 - H(n, s). \end{aligned}$$

Due to Lemma 2, $s/\sqrt{n} \to 0$ is the necessary condition for $b_n/c_n \to 0$ and $a_n/c_n \to 1$. Thus, if we utilize the Hájek projection and follow the above approach in establishing that the variance ratio converges to 1, there is no apparent way to relax the condition that $s/\sqrt{n} \to 0$. On the other hand, the H-decomposition we use in Chapter 3 provides a finer approach and a better method for comparing the variance of $U_{n,s}$ and $\hat{U}_{n,s}$ thereby allowing for a faster subsampling rate to be employed.

Appendix B Proofs in Chapter 3

B.1 H-decomposition

Distributional results for U-statistics are typically established via projection methods whereby some projection $\hat{U}$ is shown to be asymptotically normal with $|U - \hat{U}| \to 0$ in probability. The most popular projections are the Hájek projection and the H-decomposition. We show in Appendix A that the approach of Hájek projection always requires $s/\sqrt{n} \to 0$ undesirably. Alternatively, the H-decomposition provides a representation of U-statistics in terms of sums of other uncorrelated U-statistics of rank $1, \ldots, s$. The form of this decomposition presented here is derived by [41]. We illustrate those techniques in the setting of the original U-statistic $U_{n,s}$ for simplicity and then extend them to the generalized complete U-statistic $U_{n,s,\omega}$. Let

$$h_c(z_1, \ldots, z_c) = \mathbb{E}[h(z_1, \ldots, z_c, Z_{c+1}, \ldots, Z_s)] - \theta,$$

and define kernels $h^{(1)}, h^{(2)}, \ldots, h^{(s)}$ of degree $1, \ldots, s$ recursively as

$$\begin{aligned} h^{(1)} &= h_1(z_1) \\ h^{(2)} &= h_2(z_1, z_2) - h_1(z_1) - h_1(z_2) \\ &\vdots \\ h^{(s)} &= h_s(z_1, \ldots, z_s) - \sum_{j=1}^{s-1} \sum_{(s,j)} h^{(j)}(z_{i1}, \ldots, z_{ij}). \end{aligned} \tag{71}$$

These kernel functions have many important and desirable properties, a sample of which are enumerated in the following proposition.

Proposition 7 ([50]). *For $h^{(j)}$, $j = 1, \ldots, s$ defined as above, we have*

1. *For $c = 1, \ldots, j-1$, $\mathbb{E}[h^{(j)}(z_1, \ldots, z_c, Z_{c+1}, \ldots, Z_j)] = 0$.*
2. *$\mathbb{E}[h^{(j)}(Z_1, \ldots, Z_j)] = 0$.*
3. *Let $j < j'$ and S_1 and S_2 be a j-subset of $\{Z_1, \ldots, Z_n\}$ and a j'-subset of $\{Z_1, \ldots, Z_n\}$ respectively, then $\mathrm{Cov}(h^{(j)}(S_1), h^{(j')}(S_2)) = 0$.*

4. *Let $S_1 \neq S_2$ be two distinct j-subsets of $\{Z_1, \ldots, Z_n\}$, then* $\mathrm{Cov}(h^{(j)}(S_1), h^{(j')}(S_2)) = 0$.

$h = h(Z_1, \ldots, Z_n)$ can be written as $h = \sum_{j=1}^{s} \sum_{(s,j)} h(Z_{i1}, \ldots, Z_{ij})$ and the expression of $U_{n,s}$ now follows easily as

$$\begin{aligned} U_{n,s} - \theta &= \binom{n}{s}^{-1} \sum_{(n,s)} h_s(Z_{i1}, \ldots, Z_{is}) \\ &= \binom{n}{s}^{-1} \sum_{(n,s)} \left\{ \sum_{j=1}^{s} \sum_{(s,j)} h^{(j)}(Z_{i1}, \ldots, Z_{ij}) \right\} \\ &= \sum_{j=1}^{s} \binom{n}{s}^{-1} \sum_{(n,s)} \sum_{(s,j)} h^{(j)}(Z_{i1}, \ldots, Z_{ij}) \\ &= \sum_{j=1}^{s} \binom{s}{j} H_n^{(j)} \end{aligned}$$

where $H_n^{(j)} = \binom{n}{j}^{-1} \sum_{(n,j)} h^{(j)}(Z_{i1}, \ldots, Z_{ij})$ is itself a U-statistic, the usefulness of which lies in the fact that $H_n^{(j)}$ $(j = 1, \ldots, n)$ are uncorrelated and the terms in $H_n^{(j)}$ are also uncorrelated. Because of the properties above, the variance of the kernel is

$$\mathrm{Var}(h) = \mathrm{Var} \left\{ \sum_{j=1}^{s} \sum_{(s,j)} h^{(j)}(Z_{i1}, \ldots, Z_{ij}) \right\} = \sum_{j=1}^{s} \binom{s}{j} V_j \tag{72}$$

where $V_j = \mathrm{Var}(h^{(j)}(Z_{i1}, \ldots, Z_{ij}))$. Similarly, the variance of the U-statistic can be written as

$$\mathrm{Var}(U_{n,s}) = \mathrm{Var} \left\{ \sum_{j=1}^{s} \binom{s}{j} H_n^{(j)} \right\} = \sum_{j=1}^{s} \binom{s}{j}^2 \binom{n}{j}^{-1} V_j. \tag{73}$$

Note that the first-order term $sH_n^{(1)}$ is exactly the same as in the Hájek projection $\hat{U}_{n,s}$, but the H-decomposition provides a convenient alternative representation of U-statistics as well as their variance. In Chapter 3, we exploit this fact to derive a tighter and more general central limit theorem for generalized U-statistics.

B.2 Proofs of asymptotic normality

Proof of Theorem 1: The generalized complete U-statistic and the base learner can be written in terms of the new kernel functions $h^{(1)}, \ldots, h^{(s)}$ defined in relation to the H-decomposition. Let $V_{i,\omega} = \mathrm{Var}(h^{(i)})$ for $i = 1, \ldots, s-1$, $V_s = \mathrm{Var}(h^{(s)})$ and define

$$V_{s,\omega} = \mathrm{Var}\left\{ h_s(Z_1, \ldots, Z_s) - \sum_{j=1}^{s-1} \sum_{(s,j)} h^{(j)}(Z_{i1}, \ldots, Z_{ij}) \right\}.$$

These new kernels $h^{(1)}, \ldots, h^{(s)}$ still retain the desirable properties in Proposition 7. Thus, similar to Eq. (72), Eq. (73), we have the following expressions for the variance of the kernel and generalized U-statistic:

$$\begin{aligned} \mathrm{Var}(h) &= \zeta_s = \sum_{j=1}^{s-1} \binom{s}{j} V_{j,\omega} + V_s, \\ \mathrm{Var}(\hat{U}_{n,s,\omega}) &= \frac{s^2}{n} V_{1,\omega} = \frac{s^2}{n} \zeta_{1,\omega}, \\ \mathrm{Var}(U_{n,s,\omega}) &= \sum_{j=1}^{s-1} \binom{s}{j}^2 \binom{n}{j}^{-1} V_{j,\omega} + \binom{n}{s}^{-1} V_s. \end{aligned} \tag{74}$$

The sequence $\hat{U}_{n,s,\omega} / \sqrt{s^2 \zeta_{1,\omega}/n}$ converges weakly to $N(0,1)$ by the central limit theorem since $\hat{U}_{n,s,\omega} = \frac{s}{n} \sum_{i=1}^{n} h_1(Z_i)$ is a sum of i.i.d. random variables for each s, which satisfies Lindeberg's condition automatically. From Eq. (74), we have

$$\begin{aligned} \frac{\mathrm{Var}(U_{n,s,\omega})}{\mathrm{Var}(\hat{U}_{n,s,\omega})} &= \left(\frac{s^2}{n} V_{1,\omega} \right)^{-1} \left\{ \sum_{j=1}^{s-1} \binom{s}{j}^2 \binom{n}{j}^{-1} V_{j,\omega} + \binom{n}{s}^{-1} V_s \right\} \\ &\leq 1 + \left(\frac{s^2}{n} V_{1,\omega} \right)^{-1} \frac{s^2}{n^2} \left\{ \sum_{j=2}^{s-1} \binom{s}{j} V_{j,\omega} + V_s \right\} \\ &\leq 1 + \frac{s}{n} \frac{\zeta_s}{s \zeta_{1,\omega}} \to 1. \end{aligned} \tag{75}$$

Thus by Theorem 11.2 in [67], we obtain $\frac{U_{n,s,\omega} - \theta}{\sqrt{s^2 \zeta_{1,\omega}/n}} \rightsquigarrow N(0,1)$.

Proof of Theorem 2: Without loss of generality, let $\theta = 0$ and observe that

$$\begin{aligned} U_{n,s,N,\omega} &= \frac{1}{N}\sum_{(n,s)} \rho h(Z_{i1},\ldots,Z_{is};\omega) \\ &= U_{n,s,\omega} + \frac{1}{N}\sum_{(n,s)} (\rho - p) h(Z_{i1},\ldots,Z_{is};\omega) \\ &= A_n + B_n \end{aligned}$$

where A_n and B_n are uncorrelated and $\mathrm{Var}(B_n) = d_{n,S,N}^2 = (1-p)\zeta_s/N$.

First, consider the case that $p = N/\binom{n}{s} \not\to 0$. Since $\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$, by Theorem 1 we have $A_n/\sqrt{s^2\zeta_{1,\omega}/n} \rightsquigarrow N(0,1)$. Moreover, we have

$$\frac{\mathrm{Var}(B_n)}{\mathrm{Var}(A_n)} \to \frac{N^{-1}(1-p)\zeta_s}{s^2\zeta_{1,\omega}/n} \leq \frac{n^2}{Ns^2}\cdot\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0.$$

Thus $U_{n,s,N,\omega}/\sqrt{s^2\zeta_{1,\omega}/n} \to N(0,1)$, implying Eq. (9).

Now suppose $p \to 0$ and define

$$\begin{aligned} \phi_{A_n+B_n}(t) &= \mathbb{E}\left[\exp\left(it\left(\frac{s^2\zeta_{1,\omega}}{n} + \frac{\zeta_s}{N}\right)^{-1/2}(A_n + B_n)\right)\right] \\ &= \mathbb{E}\left[\exp\left(it\left(\frac{s^2\zeta_{1,\omega}}{n} + \frac{\zeta_s}{N}\right)^{-1/2} A_n\right)\right. \\ &\qquad \left.\mathbb{E}\left[\exp\left(it\left(\frac{s^2\zeta_{1,\omega}}{n} + \frac{\zeta_s}{N}\right)^{-1/2} B_n\right) \mid Z_1,\ldots,Z_n;\omega\right]\right] \\ &= \mathbb{E}\left[\hat{\phi}_{A_n}(t)\hat{\phi}_{B_n}(t)\right]. \end{aligned}$$

The strategy is to show that $\hat{\phi}_{B_n}(t)$ is well behaved and then show that $\phi_{A_n+B_n} \to e^{-t^2/2}$. Note that $U_2 = \binom{n}{s}^{-1}\sum_{(n,s)} h^2(Z_{i1},\ldots,Z_{is};\omega)$ is a complete U-statistic with kernel h^2. For any $\epsilon > 0$, by Chebyshev's inequality, we have

$$\mathbb{P}\left\{|U_2 - \mathbb{E}[h^2]| \geq \epsilon\mathbb{E}[h^2]\right\} \leq \frac{s}{n}\frac{\mathbb{E}[h^4]}{\epsilon^2\mathbb{E}^2[h^2]} \leq \frac{s}{n}\frac{C}{\epsilon^2},$$

which indicates that $U_2/\zeta_s (= U_2/\mathbb{E}[|h|^2]) \xrightarrow{p} 1$. $U_3/\mathbb{E}[|h|^3] \xrightarrow{p} 1$ also holds by a similar argument, where $U_3 = \binom{n}{s}^{-1}\sum_{(n,s)} |h(Z_{i1},\ldots,Z_{is};\omega)|^3$. Let $D = \{U_2/\mathbb{E}[h^2] \in [1-\delta, 1+\delta]\} \cap \{U_3/\mathbb{E}[|h|^3] \in [1-\delta, 1+\delta]\}$. Then for any $\delta, \epsilon > 0$, D holds with probability at least $1-\epsilon$ for n sufficiently large. Let $\hat{d}_{n,s,N} = [(1-p)U_2/N]^{1/2}$ and consider $B_n/\hat{d}_{n,s,N} \mid Z_1, Z_2, \ldots, Z_n;\omega$,

which is a sum of independent random variables. Thus, to establish asymptotic normality, it suffices to check the Lyapounov's condition. We have

$$\mathcal{L} = \frac{\binom{n}{s}^{-1} \sum_{(n,s,)} |h|^3}{\left(\binom{n}{s}^{-1} \sum_{(n,s)} |h|^2\right)^{3/2}} \frac{1-2p+2p^2}{\sqrt{N(1-p)}} = \frac{1-2p+2p^2}{\sqrt{N(1-p)}} \frac{U_3}{U_2^{3/2}}.$$

Thus, as $N \to \infty$ and $p \to 0$,

$$\mathcal{L} \leq \frac{1+\delta}{(1-\delta)^{3/2}} \frac{\mathbb{E}[|h|^3]}{\mathbb{E}^{3/2}[|h|^2]} \frac{1-2p+2p^2}{\sqrt{N(1-p)}} \to 0$$

uniformly with respect to $Z_1, \ldots, Z_n$ and ω over D and hence we have

$$\mathbb{E}\left[\exp\left(iuB_n/(\hat{d}_{n,s,N})\right) \mid Z_1, \ldots, Z_n; \omega\right] \to e^{-\frac{u^2}{2}}$$

uniformly over any finite interval of u and uniformly with respect to $Z_1, \ldots, Z_n$ and ω over D. Letting the interval be $[0, t\sqrt{(1+\delta)}]$, we have

$$\begin{aligned}
&\left|\hat{\phi}_{B_n}(t) - \exp\left(\frac{t^2}{2}\left((1-p)\cdot\frac{\zeta_s/N}{s^2\zeta_{1,\omega}/n + \zeta_s/N}\right)\zeta_s^{-1}U_2\right)\right| \\
&= \left|\mathbb{E}\left[\exp\left(t\left((1-p)\cdot\frac{\zeta_s/N}{s^2\zeta_{1,\omega}/n + \zeta_s/N}\right)^{1/2}(\zeta_s^{-1}U_2)^{1/2}\cdot B_n/\hat{d}_{n,s,N}\right) \mid Z_1, \ldots, Z_n; \omega\right]\right. \\
&\left. - \exp\left(-\frac{t^2}{2}\left((1-p)\cdot\frac{\zeta_s/N}{s^2\zeta_{1,\omega}/n + \zeta_s/N}\right)\zeta_s^{-1}U_2\right)\right| \leq \epsilon.
\end{aligned}$$

over D for n, N sufficiently large. Now, let

$$\phi_B(t) = \exp\left(-\frac{t^2}{2}\left((1-p)\cdot\frac{\zeta_s/N}{s^2\zeta_{1,\omega}/n + \zeta_s/N}\right)\right).$$

Then by the uniform continuity of e^x over any finite interval, there exists $\delta' = O(\delta)$ such that

$$\left|\exp\left(-\frac{t^2}{2}\left((1-p)\cdot\frac{\zeta_s/N}{s^2\zeta_{1,\omega}/n + \zeta_s/N}\right)\zeta_s^{-1}U_2\right) - \phi_B(t)\right| \leq \delta'$$

over D. Finally, for n and N sufficiently large, we have

$$\left|\hat{\phi}_{B_n}(t) - \phi_B(t)\right| 1_D \leq (\epsilon + \delta')1_D. \tag{76}$$

Next, consider $\hat{\phi}_{A_n}$. Since $A_n/\sqrt{s^2\zeta_{1,\omega}/n} \rightsquigarrow N(0,1)$ by Theorem 1, then

$$\mathbb{E}\left[\exp\left(iuA_n/\sqrt{s^2\zeta_{1,\omega}/n}\right)\right] \to e^{-\frac{u^2}{2}}$$

uniformly over any finite interval of u. Let the interval be $[0,t]$, then for n sufficiently large, we have

$$\left|\mathbb{E}\left[\exp\left(it\left(\frac{s^2\zeta_{1,\omega}/n}{s^2\zeta_{1,\omega}/n+\zeta_s/N}\right)^{1/2}\cdot A_n/\sqrt{s^2\zeta_{1,\omega}/n}\right)\right]-e^{-\frac{t^2}{2}\frac{s^2\zeta_{1,\omega}/n}{s^2\zeta_{1,\omega}/n+\zeta_s/N}}\right|\leq\epsilon$$

Let $\phi_A(t)=e^{-\frac{t^2}{2}\frac{s^2\zeta_{1,\omega}/n}{s^2\zeta_{1,\omega}/n+\zeta_s/N}}$ and consequently, we have

$$\left|\mathbb{E}\left[\hat{\phi}_{A_n}(t)\right]-\phi_A(t)\right|\leq\epsilon. \tag{77}$$

Combining Eq. (76) and Eq. (77) gives

$$\begin{aligned}
|\phi_{A_n+B_n}(t)-\phi_A(t)\phi_B(t)| &= \left|\mathbb{E}\left[\hat{\phi}_A(t)\hat{\phi}_{B_n}(t)\right]-\phi_A(t)\phi_B(t)\right| \\
&\leq \left|\mathbb{E}\left[\hat{\phi}_{A_n}(t)\phi_B(t)1_D\right]-\phi_A(t)\phi_B(t)\right| \\
&\quad \left|\mathbb{E}\left[\hat{\phi}_{A_n}(t)\left(\hat{\phi}_{B_n}(t)-\phi_B(t)\right)1_D\right]\right|+\epsilon \\
&\leq \left|\mathbb{E}\left[\hat{\phi}_{A_n}(t)\phi_B(t)1_D\right]-\phi_A(t)\phi_B(t)\right|+(\epsilon+\delta')+\epsilon \\
&\leq \left|\mathbb{E}\left[\hat{\phi}_{A_n}(t)1_D\right]-\phi_A(t)\right|+2\epsilon+\delta' \\
&\leq \left|\mathbb{E}\left[\hat{\phi}_{A_n}(t)\right]-\phi_A(t)\right|+\left|\mathbb{E}\left[\hat{\phi}_{A_n}(t)1_{D^c}\right]\right|+2\epsilon+\delta' \\
&\leq \epsilon+\epsilon+2\epsilon+\delta' \\
&= 4\epsilon+\delta'.
\end{aligned} \tag{78}$$

Moreover, we have

$$\begin{aligned}
\phi_A(t)\phi_B(t) &= \exp\left(-\frac{t^2}{2}\left[\frac{s^2\zeta_{1,\omega}/n}{s^2\zeta_{1,\omega}/n+\zeta_s/N}+\left(\left(1-\frac{N}{\binom{n}{s}}\right)\cdot\frac{\zeta_s/N}{s^2\zeta_{1,\omega}/n+\zeta_s/N}\right)\right]\right) \\
&= e^{-\frac{t^2}{2}}\cdot\exp\left(-\frac{t^2}{2}\left[\frac{N}{\binom{n}{s}}\frac{\zeta_s/N}{s^2\zeta_{1,\omega}/n+\zeta_s/N}\right]\right) \\
&\to e^{-\frac{t^2}{2}}
\end{aligned}$$

which implies that for n, N sufficiently large, we have

$$\left|\phi_A(t)\phi_B(t)-e^{-\frac{t^2}{2}}\right|\leq\epsilon. \tag{79}$$

Combining Eq. (78) and Eq. (79) yields that $\phi_{A_n+B_n}(t)\to e^{-\frac{t^2}{2}}$, which implies Eq. (9).

Following the statement of Theorem 2 in the main text, we remarked that the form of the result provided included the condition that $\frac{s}{n}\frac{\zeta_s}{s\zeta_{1,\omega}} \to 0$, which thus implies that the complete analogue of the incomplete U-statistic is also asymptotically normal, but that such a condition is not necessary. We elaborate on this point here by providing an alternative result.

Theorem 14. *Let $Z_1, \ldots, Z_n$ be i.i.d. from F_Z and $U_{n,s,N,\omega}$ be a generalized incomplete U-statistic with kernel $h = h(Z_1, \ldots, Z_s; \omega)$. Let $\theta = \mathbb{E}[h]$ and $\zeta_s = \mathrm{Var}(h)$. Suppose that $\mathbb{E}[|h-\theta|^{2k}]/\mathbb{E}^2[|h-\theta|^k]$ is uniformly bounded for $k = 2, 3$ and for s. Then*

1. *$U_{n,s,N,\omega} - \theta = A_n + B_n$, where $B_n = N^{-1}\sum_{(n,s)}(\rho - p)(h(Z_{i1}, \ldots, Z_{is}; \omega) - \theta)$ and $A_n = U_{n,s,\omega} - \theta$, If $s/n \to 0$ and $N \to \infty$ with $p = N/\binom{n}{s} \to 0$, then*

$$\frac{B_n}{\sqrt{\zeta_s/N}} \rightsquigarrow N(0,1). \tag{80}$$

2. *In addition to the conditions in* **1**, *If $\mathrm{Var}(A_n)/\mathrm{Var}(B_n) \to 0$, then*

$$\frac{U_{n,s,N,\omega} - \theta}{\sqrt{\zeta_s/N}} \rightsquigarrow N(0,1). \tag{81}$$

Proof. For **1**, without loss of generality, let $\theta = 0$. Observe that

$$\begin{aligned} U_{n,s,N,\omega} &= \frac{1}{N}\sum_{(n,s)} \rho h(Z_{i1}, \ldots, Z_{is}; \omega) \\ &= \frac{1}{N}\sum_{(n,s)} (\rho - p) h(Z_{i1}, \ldots, Z_{is}; \omega) + U_{n,s,\omega} \\ &= B_n + A_n, \end{aligned}$$

where A_n and B_n are uncorrelated, and $\mathrm{Var}(B_n) = d_{n,S,N}^2 = (1-p)\zeta_s/N$. First, $U_2 = \binom{n}{s}^{-1}\sum_{(n,s)} h^2(Z_{i1}, \ldots, Z_{is}; \omega)$ is a complete U-statistic with kernel h^2. For any $\epsilon > 0$, by Chebyshev's inequality, we have

$$\Pr\left\{\left|U_2 - \mathbb{E}[h^2]\right| \geq \epsilon \mathbb{E}[h^2]\right\} \leq \frac{s}{n}\frac{\mathbb{E}[h^4]}{\epsilon^2 \mathbb{E}^2[h^2]} \leq \frac{s}{n}\frac{C}{\epsilon^2},$$

which indicates that $U_2/\zeta_s (= U_2/\mathbb{E}[|h|^2]) \xrightarrow{p} 1$. Let $D_2 = \{\zeta_s^{-1} U_2 \in [1-\delta, 1+\delta]\}$. Thus for any $\delta, \epsilon > 0$, for n sufficiently large, $\Pr(D_2) \geq 1 - \epsilon$. Let $\hat{d}_{n,s,N} = [(1-p)U_2/N]^{1/2}$. Then

$$\begin{aligned}
\phi_{B_n}(t) &= \mathbb{E}\left[\exp\left(itB_n/d_{n,s,N}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(itB_n/d_{n,s,N}\right) \mid Z_1, \ldots, Z_n; \omega\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(it(\zeta_s^{-1}U_2)^{1/2} \cdot B_n/\hat{d}_{n,s,N}\right) \mid Z_1, \ldots, Z_n; \omega\right]\right].
\end{aligned}$$

$B_n/\hat{d}_{n,s,N} \mid Z_1, Z_2, \ldots, Z_n; \omega$ is a sum of independent random variables and thus in order to establish asymptotic normality it suffices to check the Lyapounov's condition. We have

$$\mathcal{L} = \frac{\binom{n}{s}^{-1} \sum_{(n,s)} |h|^3}{\left(\binom{n}{s}^{-1} \sum_{(n,s)} |h|^2\right)^{3/2}} \frac{1 - 2p + 2p^2}{\sqrt{N(1-p)}} = \frac{1 - 2p + 2p^2}{\sqrt{N(1-p)}} \frac{U_3}{U_2^{3/2}}$$

where $U_3 = \binom{n}{s}^{-1} \sum_{(n,s)} |h(Z_{i1}, \ldots, Z_{is}; \omega)|^3$. Since $\mathbb{E}[h^6]/\mathbb{E}^2[|h|^3] \leq C$, by a similar argument as for U_2, $D_3 = \{U_3/\mathbb{E}[|h|^3] \in [1-\delta, 1+\delta]\}$ holds with probability at least $1 - \epsilon$. Then as $N \to \infty$ and $p \to 0$,

$$\mathcal{L} \leq \frac{1+\delta}{(1-\delta)^{3/2}} \frac{\mathbb{E}[|h|^3]}{\mathbb{E}^{3/2}[|h|^2]} \frac{1 - 2p + 2p^2}{\sqrt{N(1-p)}} \to 0$$

uniformly with respect to $Z_1, \ldots, Z_n$ and ω over $D_2 \cap D_3$. Note that distance between the characteristic function $B_n/\hat{d}_{n,s,N}$ and a standard normal distribution can be controlled by $\mathcal{L}$. Thus with probability at least $1 - 2\epsilon$,

$$\mathbb{E}\left[\exp\left(iuB_n/(\hat{d}_{n,s,N})\right) \mid Z_1, \ldots, Z_n; \omega\right] \to e^{-\frac{u^2}{2}}$$

uniformly over any finite interval of u and uniformly with respect to $Z_1, \ldots, Z_n$ and ω over $D_2 \cap D_3$. Letting the interval be $[t\sqrt{(1-\delta)}, t\sqrt{(1+\delta)}]$, we have

$$\left|\mathbb{E}\left[\exp\left(it(\zeta_s^{-1}U_2)^{1/2} \cdot B_n/(\hat{d}_{n,s,N})\right) \mid Z_1, \ldots, Z_n; \omega\right] - e^{-\frac{t^2}{2}\zeta_s^{-1}U_2}\right| \leq \epsilon$$

over $D_2 \cap D_3$ for N sufficiently large. Therefore,

$$\begin{aligned}
\left|\phi_{B_n}(t) - e^{-\frac{t^2}{2}}\right| &\leq \mathbb{E}\left[\left|\mathbb{E}\left[\exp\left(it(\zeta_s^{-1}U_2)^{1/2} \cdot B_n/\hat{d}_{n,s,N}\right) \mid Z_1, \ldots, Z_n; \omega\right] - e^{-t^2/2}\right|\right] \\
&\leq \mathbb{E}\left[\left(\left|\mathbb{E}\left[\exp\left(it(\zeta_s^{-1}U_2)^{1/2} \cdot B_n/\hat{d}_{n,s,N}\right) \mid Z_1, \ldots, Z_n; \omega\right] - e^{-\frac{t^2}{2}\zeta_s^{-1}U_2}\right| \right.\right. \\
&\quad \left.\left. + \left|e^{-\frac{t^2}{2}\zeta_s^{-1}U_2} - e^{-\frac{t^2}{2}}\right|\right) 1_{D_2 \cap D_3}\right] + 4\epsilon \\
&\leq \mathbb{E}\left[\left(\epsilon + \left|e^{-\frac{t^2}{2}\zeta_s^{-1}U_2} - e^{-\frac{t^2}{2}}\right|\right) 1_{D_2 \cap D_3}\right] + 4\epsilon \\
&\leq (\epsilon + \delta') + 4\epsilon.
\end{aligned}$$

Since ϵ and δ can be arbitrarily small, we have $\phi_{B_n}(t) \to e^{-\frac{t^2}{2}}$ and thus $B_n/d_{n,s,N} \rightsquigarrow N(0,1)$, which implies Eq. (80) since $N/\binom{n}{s} \to 0$.

For **2**, $\mathrm{Var}(A_n)/\mathrm{Var}(B_n) \to 0$ implies that $\mathrm{Var}(A_n/\sqrt{\zeta_s/N}) = o(1)$. Thus, we have

$$\frac{U_{n,s,N,\omega}}{\sqrt{\zeta_s/N}} = \frac{B_n}{\sqrt{\zeta_s/N}} + o_p(1),$$

which implies Eq. (81) by applying Slutsky's theorem. □

Part **1** of Theorem 14 gives that B_n, the difference between the incomplete and complete generalized U-statistics, is asymptotically normal under quite weak conditions so long as the number of subsamples N grows slower than $\binom{n}{s}$. In particular, no specialized conditions on the resulting variance or variance ratio are required. In Part **2**, to establish asymptotic normality of the generalized incomplete U-statistic itself, we do not require the original condition on the variance ratio given in Theorem 2, though we do require that $\mathrm{Var}(A_n)/\mathrm{Var}(B_n) \to 0$. Such a condition remains difficult to verify for general kernels, but can always be satisfied, for example, by taking $N = o(n/s)$. Indeed, note that

$$\frac{s^2}{n}\zeta_{1,\omega} \leq \mathrm{Var}(A_n) \leq \frac{s}{n}\zeta_s,$$

and $\mathrm{Var}(B_n) = (1-p)\zeta_s/N$, thus a sufficient condition for $(\zeta_s/N)^{-1/2}A_n = o_p(1)$ to hold is letting $N = o(n/s)$. Thus, asymptotic normality for incomplete U-statistics can be established without requiring normality of the complete version and in particular, without requiring the specialized condition on the variance ratio discussed at length in Chapter 3.

B.3 Simple base learner variance ratios

We now provide explicit calculations for the variance ratios corresponding to the various base learners discussed in Chapter 3. We begin with simple examples where base learners take the form of a sample mean, sample variance, and least squares estimators.

Example 1 (Sample Mean)**.** *Suppose $Z_1, \ldots, Z_s$ are i.i.d. random variables with mean μ and let $h = \bar{Z}$, then*

$$\frac{\zeta_s}{s\zeta_1} = \frac{\text{Var}\left(\sum_{i=1}^s Z_i/s\right)}{s\text{Var}\left(Z_1/s + (s-1)\mu/s\right)} = 1. \tag{82}$$

Eq. (82) holds for any estimators that can be written as a sum of i.i.d. random variables. In such cases, since $\hat{h} = h$, nothing is lost after projecting.

Example 2 (Sample Variance)**.** *Suppose $Z_1, \ldots, Z_s$ are i.i.d. random variables with variance σ^2 and fourth central moment μ_4 and consider the sample variance $h = \binom{n}{2}^{-1} \sum_{i<j}(Z_i - Z_j)^2$. Then as $s \to \infty$,*

$$\frac{\zeta_s}{s\zeta_1} = 1 + \frac{2}{(s-1)} \cdot \frac{\sigma^4}{\mu_4 - \sigma^4} \to 1.$$

The kernel h can be written as $h = \sum_{i=1}^s \frac{(Z_i - \bar{Z})^2}{s-1} \approx \sum_{i=1}^s \frac{(Z_i^2 - \mu^2)}{s-1}$. Since $\bar{Z}$ is much more stable than Z_i, h is close to a sum of i.i.d random variables.

Example 3 (OLS Estimator)**.** *Let $Z_1, \ldots, Z_s$ denote i.i.d. pairs of random variables (X_i, Y_i) and $Y_i = X_i^T\beta + \epsilon_i$. Suppose that ϵ_i has mean 0 and variance σ^2, and ϵ_i is independent of X_i. Let $h = (X^TX)^{-1}X^TY$ be the ordinary least squares (OLS) estimator of β. Then as $s \to \infty$,*

$$(s\zeta_1)^{-1}\zeta_s \to I,$$

where I is the identity matrix.

Proof. Let $\hat{\beta} = GX^TY$ be the OLS estimator, where $G = (X^TX)^{-1}$, then $\zeta_s = \text{Var}(\hat{\beta}) = \mathbb{E}[G]\sigma^2$. Since X_i and ϵ_i are independent, we have $\mathbb{E}[\hat{\beta} \mid X_1, Y_1] = \beta + \mathbb{E}_1[G]X_1 \cdot \epsilon$, where $\mathbb{E}_1$ takes the expectation conditioning on X_1. Then,

$$\zeta_1 = \text{Var}\left(\mathbb{E}[\hat{\beta} \mid X_1, Y_1]\right) = \mathbb{E}[\mathbb{E}_1[G]X_1X_1^T\mathbb{E}_1[G]]\sigma^2.$$

According to law of large numbers, $\frac{1}{s}\sum_{i=1}^{s} X_i X_i^T \xrightarrow{a.s.} \Sigma$ as $s \to \infty$ where $\mathbb{E}[X_i X_i^T] = \Sigma$, and then for $\Sigma^{-1} = \Omega$,

$$s \cdot \mathbb{E}[(X^T X)^{-1}] = \mathbb{E}\left[\left(\frac{1}{s}\sum_{i=1}^{s} X_i X_i^T\right)^{-1}\right] \to \Omega.$$

Thus, $s\zeta_s \to \Omega$. Furthermore, we have

$$sG \mid X_1 = \left(\frac{1}{s}\left[X_1 X_1^T + \sum_{i\neq 1} X_i X_i^T\right]\right)^{-1} \Bigg| X_1 \xrightarrow{a.s.} \Omega$$

and

$$\begin{aligned} s^2 \cdot \zeta_1/\sigma^2 &= \mathbb{E}\left[\mathbb{E}_1[G] X_1 X_1^T \mathbb{E}_1[G]\right] \\ &= \mathbb{E}\left[\mathbb{E}_1[sG] \cdot X_1 X_1^T \cdot \mathbb{E}_1[sG]\right] \\ &\to \Omega \cdot \Sigma \cdot \Omega \\ &= \Omega. \end{aligned}$$

Hence, ζ_1 is of order s^{-2} and $(s\zeta_1)^{-1}\zeta_s \to I$. □

Here again, note that $h = \sum_{i=1}^{s} (X^T X)^{-1} X_i Y_i$, which is still close to a sum of i.i.d. random variables. These three examples suggest that perhaps for many common base learners, ζ_s is of order s^{-1} and ζ_1 of order s^{-2}; essentially, each individual observation explains roughly s^{-1} times the variance of the base learner.

Proof of Proposition 1: Denote the kNN estimator at x as $\varphi(x)$. Let A_i denote the event that X_1 is the ith closest point to the target point x and $B = \cup_{i=1}^{k} A_i$. First, by the continuity of f at x, we have $\mathrm{Var}(\varphi(x)) \to \sigma^2/k$ as $s \to \infty$. Let $X_1^*, \ldots, X_k^*$ be the k-NNs of x. Then

$$\begin{aligned} \mathbb{E}[\varphi(x) \mid X_1, Y_1] &= \frac{1}{k}\mathbb{E}\left[\mathbf{1}_B\left[Y_1 + \sum_{i=2}^{k} Y_i^*\right] + \mathbf{1}_{B^c}\left[\sum_{i=1}^{k} Y_i^*\right] \mid X_1, Y_1\right] \\ &= \frac{\epsilon_1}{k}\mathbb{E}[\mathbf{1}_B \mid X_1] + \frac{1}{k}\left[\sum_{i=1}^{k} f(X_i^*) \mid X_1\right], \end{aligned}$$

and thus $\mathrm{Var}(\mathbb{E}[\varphi(x)|X_1, Y_1]) \geq \frac{\sigma^2}{k^2}\mathbb{E}[\Pr^2(B|X_1)]$. Next,

$$\begin{aligned}
\mathbb{E}[\Pr^2(B|X_1)] &= \mathbb{E}[(\sum_{i=1}^{k} \Pr(A_i \mid X_1))^2] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \mathbb{E}\left[\Pr(A_i \mid X_1)\Pr(A_j \mid X_1)\right] \\
&= \frac{1}{2s-1}\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\binom{s-1}{i-1}\binom{s-1}{j-1}}{\binom{2s-2}{i+j-2}} \\
&= \frac{V(k,s)}{2s-1}, \qquad (83)
\end{aligned}$$

where $V(k,s) = \sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\left[\frac{\binom{s-1}{i}\binom{s-1}{j}}{\binom{2s-2}{i+j}}\right]$. We have

$$\begin{aligned}
V(k) &= \lim_{s\to\infty} V(k,s) \\
&= \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \frac{(i+j)!}{i!j!}\frac{1}{2^{i+j}} \\
&= \sum_{c=0}^{2k-2} \sum_{i+j=c, 0\leq i,j\leq k-1} \frac{(i+j)!}{i!j!}\frac{1}{2^{i+j}}.
\end{aligned}$$

We can obtain $k \leq V(k) < 2k-1$ by simply observing that

$$\sum_{c=0}^{k-1}\sum_{i+j=c}\frac{c!}{i!j!}\frac{1}{2^c} \leq V(k) \leq \sum_{c=0}^{2k-2}\sum_{i+j=c}\frac{c!}{i!j!}\frac{1}{2^c}.$$

Thus,

$$\begin{aligned}
\limsup_{s\to\infty}\frac{\zeta_s}{s\zeta_1} &= \limsup_{s\to\infty}\frac{\mathrm{Var}(\varphi(x))}{s\mathrm{Var}(\mathbb{E}[\varphi(x)|X_1,Y_1])} \\
&\leq \limsup_{s\to\infty}\frac{2s-1}{s}\frac{k}{V(k,s)} \qquad (84) \\
&= c(k),
\end{aligned}$$

and $1 < c(k) <= 2$.

Note that Proposition 1 holds without imposing any conditions on the regression function f or the distribution of X. To see why, note from the proof that both ζ_s and ζ_1 can each be decomposed into two terms, one of which comes from the variation of the regression function while the other is due to the variation of the noise. Here, since k is fixed, the

term involving the variation in the regression function is small relative to the noise term for large s. When k grows with s, more care must be taken in assessing the contribution of the regression function. [7] (Theorem 15.3) discuss the convergence rate of the variance of kNN estimators. This result could potentially enable results to be established for more general nearest neighbor estimators where the number of neighbors k is permitted to grow with s, though we do not explore this further here.

Proof of Proposition 2: First let $\tilde{\varphi} = \sum_{i=1}^{s} w(i, x, \mathbf{X}) f(X_i)$ and note that

$$\begin{aligned} \mathrm{Var}(\varphi) &= s\mathbb{E}[w^2(1, x, \mathbf{X})]\sigma^2 + \mathrm{Var}\left(\tilde{\varphi}\right) \\ &\leq \sigma^2 + ||f||_\infty/4. \end{aligned} \tag{85}$$

Next, $\mathbb{E}[\varphi \mid X_1, \epsilon_1] = \mathbb{E}\left[\tilde{\varphi} \mid X_1\right] + \epsilon_1 \mathbb{E}\left[w(1, x, \mathbf{X}) \mid X_1\right]$, and thus

$$\begin{aligned} \mathrm{Var}(\mathbb{E}[\varphi \mid X_1, \epsilon_1]) &= \mathrm{Var}\left(\mathbb{E}\left[\tilde{\varphi} \mid X_1\right]\right) + \sigma^2 \mathbb{E}\left[\mathbb{E}^2\left[w(1, x, \mathbf{X}) \mid X_1\right]\right] \\ &\geq \sigma^2 \mathbb{E}\left[\mathbb{E}^2[w(1, x, \mathbf{X}) \mid X_1]\right] \\ &\geq \sigma^2 \mathbb{E}^2[w(1, x, \mathbf{X})] \\ &= \sigma^2/s^2. \end{aligned} \tag{86}$$

Therefore,

$$\begin{aligned} \limsup_{s\to\infty} \frac{\zeta_s}{s\zeta_1}\frac{1}{s} &= \limsup_{s\to\infty} \frac{s\sigma^2\mathbb{E}[w^2(1, x, \mathbf{X})] + \mathrm{Var}(\tilde{\varphi})}{s^2\left(\sigma^2\mathbb{E}[\mathbb{E}^2[w(1, x, \mathbf{X})) \mid X_1]] + \mathrm{Var}(\mathbb{E}[\tilde{\varphi} \mid X_1])\right)} \\ &\leq \frac{\sigma^2 + ||f||_\infty^2/4}{\sigma^2} \\ &< \infty. \end{aligned} \tag{87}$$

We emphasize that the inequalities in Eq. (85) and Eq. (86) are generally quite loose in order to cover the worst case scenario. As seen in Proposition 1, the order of ζ_1 can indeed be s^{-1} rather than s^{-2}. Nonetheless, Proposition 2 indicates that $s/\sqrt{n} \to 0$ is sufficient to ensure that $\frac{s}{n}\frac{\zeta_s}{s\zeta_1} \to 0$.

B.4 Variance ratios of RP trees

Finally, we turn to the analysis for RP trees, which form predictions by taking a sample uniformly at random from the potential nearest neighbors and averaging the corresponding response values. We begin with a simpler result for base learners that take a naive random average across k response values selected uniformly at random from the entire dataset.

Example 4 (Naive Random Average)**.** *Let $Z_1, \ldots, Z_s$ denote i.i.d. pairs of random variables (X_i, Y_i). For any target point x, let $\varphi(x)$ denote the estimator that forms a prediction by simply selecting k sample points uniformly at random (without replacement) and averages the selected response values, so that we can write $\varphi(x) = \frac{1}{k}\sum_{i=1}^{s} \xi_i Y_i$, where*

$$\xi_i = \begin{cases} 1, & i^{th} \text{ sample is selected} \\ 0, & i^{th} \text{ sample is not selected.} \end{cases}$$

Then $\zeta_s = \mathrm{Var}(Y_1)/k$ and $\zeta_1 = (\mathrm{Var}(Y_1))/s^2$, so that $\zeta_s/s\zeta_1 = s/k$.

Note in the above example that when k is fixed, $s/\sqrt{n} \to 0$ is sufficient to ensure that $\frac{s}{n}\frac{\zeta_s}{s\zeta_1} \to 0$. However, when k is assumed to grow with n, the subsample size s can grow more quickly. In the adaptive case, where $w(i, x, \mathbf{X})$ may depend on $\{Y_i\}_{i=1}^{s}$, tree estimators with small terminal node sizes may look less like a linear statistic and in turn may have a larger variance ratio. However, as discussed, for non-adaptive estimators like kNN, the ratio is bounded by a constant. In this way, well-behaved tree predictors can be seen as similar to kNN and are still more easily controlled than RP trees.

We turn now to the proving Proposition 3, namely that for base learners that are RP trees,

$$\limsup_{s\to\infty} \frac{\zeta_s/s\zeta_1}{(\log s)^{2d-2}} < \infty.$$

Proof of Proposition 3: Denote the RP tree by T and the set of k-PNNs by Ξ. We have

$$\mathrm{Var}(T) = \mathrm{Var}(\tilde{T}) + \sigma^2/k \leq \sigma^2/k + ||f||_\infty^2/4.$$

where $\tilde{T}$ is the RP tree prediction in the noiseless case. Let $|\Xi|$ denote the number of k-PNNs of x, then

$$\mathbb{E}[T \mid Z_1] = \epsilon_1 \mathbb{E}\left[\frac{1}{|\Xi|}\mathbf{1}_{X_1 \in \Xi} \mid X_1\right] + \mathbb{E}\left[\tilde{T} \mid X_1\right].$$

Since Ξ is independent of ϵ, we have $\mathrm{Var}(\mathbb{E}[T \mid Z_1]) \geq \sigma^2 \mathbb{E}\left[\mathbb{E}^2[S_1 \mid X_1]\right]$, where $S_1 = \frac{1}{|\Xi|}\mathbf{1}_{X_1 \in \Xi}$. Note that $\mathbb{E}[S_1] = \mathbb{E}\left[\sum_{i=1}^{s} S_i\right]/s = 1/s$, and thus we have $\mathrm{Var}(\mathbb{E}[T \mid Z_1]) \geq \sigma^2/s^2$. Furthermore, we have

$$\begin{aligned}
\mathbb{E}[S_1|Z_1] &= \Pr_1(X_1 \in \Xi)\mathbb{E}_1\left[\frac{1}{|\Xi|} \mid X_1 \in \Xi\right] \\
&= \sum_{i=0}^{k-1} \binom{s-1}{i} u^i (1-u)^{s-1-i} \cdot \mathbb{E}_1\left[\frac{1}{|\Xi|} \mid X_1 \in \Xi\right] \\
&= \mathrm{I} \cdot \mathrm{II},
\end{aligned} \tag{88}$$

where $\Pr_1(\cdot) = \Pr(\cdot \mid X_1), \mathbb{E}_1 = \mathbb{E}(\cdot \mid X_1)$, $u = \Pr_1(X_i \in R)$ and R is the hyperrectangle defined by x and X_1. Conditioning on $X_1 \in \Xi$, define the conditional probability function of $(X_2, \ldots, X_s)$ - $\Pr_1(\cdot \mid X_1 \in \Xi)$ as $\tilde{\Pr}_1$. For I,

$$\begin{aligned}
\mathrm{I}^2 &= \left[\sum_{i=0}^{k-1} \binom{s-1}{i} u^i (1-u)^{s-1-i}\right]^2 \\
&= \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \binom{s-1}{i}\binom{s-1}{j} u^{i+j}(1-u)^{2s-2-i-j},
\end{aligned}$$

thus

$$\begin{aligned}
\mathbb{E}[\mathrm{I}^2] &= \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \binom{s-1}{i}\binom{s-1}{j} \mathbb{E}\left[u^{i+j}(1-u)^{2s-2-i-j}\right] \\
&= \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \frac{\binom{s-1}{i}\binom{s-1}{j}}{\binom{2s-2}{i+j}} \mathbb{E}\left[\frac{u^{i+j}(1-u)^{2s-2-i-j}}{\mathrm{B}(i+j+1, 2s-1-i-j)}\right] \\
&= \frac{1}{2s-1} \cdot \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \frac{\binom{s-1}{i}\binom{s-1}{j}}{\binom{2s-2}{i+j}} G(i,j),
\end{aligned} \tag{89}$$

where $u = \Pr_1(X_i \in R) \in (0,1)$. If $u \sim \mathrm{Uniform}(0,1)$, then $G(i,j) = 1$ and Eq. (89) reduces to Eq. (83). Let the probability density function of u be $p(u)$, and the probability density function of beta distribution with shape parameters α and β be $g(u, \alpha, \beta)$, then $G(i,j) = \int_0^1 g(u, \alpha, \beta)p(u)\,du$, where $\alpha = i + j, \beta = s - 1 - \alpha$. Since $i + j <= 2k - 2$, when

$s \to \infty$, $g(u, \alpha, \beta)$ is almost singular at $u = 0$. Moreover, we can find that at around $u = 0$, $p(u) \geq 1$. Thus, there exists some $c_1 > 0$ such that $G(i, j) \geq c_1$. Therefore, we have

$$\mathbb{E}[\mathrm{I}^2] \geq \frac{c_1}{2s-1} V(k,s), \quad V(k,s) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \left[\frac{\binom{s-1}{i}\binom{s-1}{j}}{\binom{2s-2}{i+j}} \right]. \tag{90}$$

For II, by Jensen's Inequality, we have $\mathrm{II} = \tilde{\mathbb{E}}_1 \left[1/|\Xi|\right] \geq 1/\tilde{\mathbb{E}}_1[|\Xi|]$, and then

$$\mathbb{E}[\mathbb{E}^2[S_1|X_1]] \geq \mathbb{E}\left[\mathrm{I}^2 \cdot \frac{1}{\tilde{\mathbb{E}}_1^2 |\Xi|} \right]. \tag{91}$$

$\tilde{\mathbb{E}}_1|\Xi|$ is just the expected number of k-PNNs conditioning on $X_1 \in \Xi$, or equivalently given that there are at most $k-1$ sample points in R. [51] showed that $\mathbb{E}[|\Xi|]$ is of order $k(\log s)^{p-1}$ when the probability density function of the features is bounded away from 0 and ∞ in $[0,1]^p$. Since k-PNN depends only on the relative distance, it can be shown that exists some $c_2 > 0$ such that

$$\tilde{\mathbb{E}}_1[|\Xi|] \leq c_2 \mathbb{E}[|\Xi|], \quad \text{for } X_1 \in [0,1]^p. \tag{92}$$

Note that $V(k,s) \geq k$. Combining Eq. (90), Eq. (91) and Eq. (92), we have

$$\limsup_{s\to\infty} \frac{\zeta_s / s\zeta_1}{(\log s)^{2p-2}} \leq \limsup_{s\to\infty} \frac{(\sigma^2/k + ||f||_\infty^2/4)(\log s)^{2-2p}}{s\left(c_1 \frac{V(k,s)}{2s-1} \cdot c_2^{-2} (k(\log s)^{p-1})^{-2} \cdot \sigma^2\right)} < \infty, \tag{93}$$

thus achieving what was claimed in Proposition 3.

Appendix C Proofs in Chapter 4

C.1 Introduction to Lemma 4

In many cases of interest, a statistic T can be written as a linear statistic plus a manageable term. [17] used the K-function approach derived from Stein's method [66] to build a random concentration inequality for linear statistics. This inequality is an extension of the usual concentration inequalities but the bounds can be random. The authors then apply this randomized concentration inequality to provide a Berry-Esseen bound for T as in Lemma 4. For completeness, we begin with a brief discussion of this inequality and its derivatives.

Let $Z_1, \ldots, Z_n$ be independent random variables and T be a statistic of the form

$$T = T(Z_1, \ldots, Z_n) = W + \Delta$$

where

$$W = \sum_{i=1}^{n} g_{n,i}(Z_i), \quad \text{and} \quad \Delta = \Delta(Z_1, \ldots, Z_n)$$

for some functions $g_{n,i}$ and Δ. Note here that W is linear and thus T takes the form of a linear statistic plus a remainder. Let $\xi_i = g_{n,i}(Z_i)$ and assume that

$$\mathbb{E}[\xi_i] = 0 \quad (i = 1, \ldots, n) \quad \text{and} \quad \sum_{i=1}^{n} \mathrm{Var}(\xi_i) = 1. \tag{94}$$

The following randomized concentration inequality can be used to establish uniform Berry-Esseen bounds on T with optimal asymptotic rates.

Lemma 3 ([17]). *Let $\delta > 0$ satisfying*

$$\sum_{i=1}^{n} \mathbb{E}\left[|\xi_i| \min(\delta, |\xi_i|)\right] \geq 1/2. \tag{95}$$

Then for any real-valued random variables Δ_1 and Δ_2,

$$\begin{aligned} \Pr(\Delta_1 \leq W \leq \Delta_2) \leq 4\delta + \mathbb{E}|W(\Delta_2 - \Delta_1)| \\ + \sum_{i=1}^{n} \left[\mathbb{E}|\xi_i(\Delta_1 - \Delta_{1,i})| + \mathbb{E}|\xi_i(\Delta_2 - \Delta_{2,i})|\right] \end{aligned} \tag{96}$$

whenever ξ_i is independent of $(W - \xi_i, \Delta_{1,i}, \Delta_{2,i})$.

For completeness, we replicate the proof of Eq. (96) originally given in [17]. The spirit of the proof is to replace bounding the probability by bounding the expectation of some functions.

Proof. Let

$$f_{a,b}(w) = \begin{cases} -\frac{1}{2}(b-a) - \delta, & w < a - \delta \\ w - \frac{1}{2}(a+b), & a - \delta \le w \le b + \delta \\ \frac{1}{2}(b-a) + \delta, & w > b + \delta \end{cases}$$

and let

$$\hat{K}_i(t) = \xi_i\{1_{-\xi_i \le t \le 0} - 1_{0 < t \le -\xi_i}\}, \quad \hat{K}(t) = \sum_{i=1}^{n} \hat{K}_i(t).$$

Since ξ_i and $f_{\Delta_{1,i},\Delta_{2,i}}(W - \xi_i)$ are independent for $1 \le i \le n$, we have

$$\begin{aligned} \mathbb{E}\left[W f_{\Delta_1,\Delta_2}(W)\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\xi_i (f_{\Delta_1,\Delta_2}(W) - f_{\Delta_1,\Delta_2}(W - \xi_i))\right] \\ &\quad + \sum_{i=1}^{n} \mathbb{E}\left[\xi_i (f_{\Delta_1,\Delta_2}(W - \xi_i) - f_{\Delta_{1,i},\Delta_{2,i}}(W - \xi_i))\right] \\ &= H_1 + H_2 \end{aligned}$$

where

$$\begin{aligned} H_1 &= \sum_{i=1}^{n} \mathbb{E}\left[\xi_i \int_{-\xi_i}^{0} f'_{\Delta_1,\Delta_2}(W + t)\, dt\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\int_{-\infty}^{\infty} f'_{\Delta_1,\Delta_2}(W + t)\hat{K}_i(t)\, dt\right] \\ &\ge \mathbb{E}\left[\int_{|t| \le \delta} f'_{\Delta_1,\Delta_2}(W + t)\hat{K}(t)\, dt\right] \\ &\ge \mathbb{E}\left[1_{\Delta_1 \le W \le \Delta_2} \int_{|t| \le \delta} \hat{K}(t)\, dt\right] \\ &= \mathbb{E}\left[1_{\Delta_1 \le W \le \Delta_2} \sum_{i=1}^{n} |\xi_i| \min(\delta, \xi_i)\right] \\ &\ge H_{1,1} - H_{1,2} \end{aligned}$$

and where

$$H_{1,1} = \Pr(\Delta_1 \le W \le \Delta_2) \sum_{i=1}^{n} \mathbb{E}\left[|\xi_i| \min(\delta, \xi_i)\right] \ge 1/2 \Pr(\Delta_1 \le W \le \Delta_2)$$

and

$$H_{1,2} = \mathbb{E}\left|\sum_{i=1}^{n}\left[|\xi_i|\min(\delta,\xi_i) - \mathbb{E}\left[|\xi_i|\min(\delta,\xi_i)\right]\right]\right| \leq \mathrm{Var}\left(\sum_{i=1}^{n}|\xi_i|\min(\delta,\xi_i)\right)^{1/2} \leq \delta.$$

For H_2, we have

$$|f_{\Delta_1,\Delta_2}(w) - f_{\Delta_{1,i},\Delta_{2,i}(w)}| \leq \frac{1}{2}|\Delta_1 - \Delta_{1,i}| + \frac{1}{2}|\Delta_2 - \Delta_{2,i}|$$

which then yields

$$|H_2| \leq \frac{1}{2}\left(\mathbb{E}|\xi_i(\Delta_1 - \Delta_{1,i})| + \mathbb{E}|\xi_i(\Delta_2 - \Delta_{2,i})|\right).$$

It follows from the definition of $f_{a,b}$ that

$$|f_{\Delta_1,\Delta_2}(w)| \leq \frac{1}{2}(\Delta_2 - \Delta_1) + \delta.$$

Hence,

$$\begin{aligned}
\Pr(\Delta_1 \leq W \leq \Delta_2) &\leq 2\mathbb{E}\left[W f_{\Delta_1,\Delta_2}(W)\right] + 2\delta + \sum_{i=1}^{n}\left[\mathbb{E}|\xi_i(\Delta_1 - \Delta_{1,i})| + \mathbb{E}|\xi_i(\Delta_2 - \Delta_{2,i})|\right] \\
&\leq \mathbb{E}|W(\Delta_2 - \Delta_1)| + 2\delta\mathbb{E}|W| \\
&\quad + 2\delta + \sum_{i=1}^{n}\left[\mathbb{E}|\xi_i(\Delta_1 - \Delta_{1,i})| + \mathbb{E}|\xi_i(\Delta_2 - \Delta_{2,i})|\right] \\
&\leq \mathbb{E}|W(\Delta_2 - \Delta_1)| + 4\delta + \sum_{i=1}^{n}\left[\mathbb{E}|\xi_i(\Delta_1 - \Delta_{1,i})| + \mathbb{E}|\xi_i(\Delta_2 - \Delta_{2,i})|\right].
\end{aligned}$$

□

Now, for any estimator of the form $T = W + \Delta$, we can write

$$-\Pr(z - |\Delta| \leq W \leq z) \leq \Pr(T \leq z) - \Pr(W \leq z) \leq \Pr(z \leq W \leq z + |\Delta|).$$

Applying Eq. (96) to these bounds, we arrive at the following lemma.

Lemma 4 ([17]). *Let $\xi_1, \ldots, \xi_n$ be independent random variables satisfying Eq.* (94)*, $W = \sum_{i=1}^n \xi_i$ and $T = W + \Delta$. Let Δ_i be a random variable such that ξ_i and $(W - \xi_i, \Delta_i)$ are independent. Then for any δ satisfying Eq.* (95)*, we have*

$$\sup_{z\in\mathbb{R}} |\Pr(T \le z) - \Pr(W \le z)| \le 4\delta + \mathbb{E}|W\Delta| + \sum_{i=1}^n \mathbb{E}|\xi_i(\Delta - \Delta_i)|.$$

In particular,

$$\sup_{z\in\mathbb{R}} |\Pr(T \le z) - \Pr(W \le z)| \le 2(\beta_2 + \beta_3) + \mathbb{E}|W\Delta| + \sum_{i=1}^n \mathbb{E}|\xi_i(\Delta - \Delta_i)| \tag{97}$$

and

$$\sup_{z\in\mathbb{R}} |\Pr(T \le z) - \Phi(z)| \le 6.1(\beta_2 + \beta_3) + \mathbb{E}|W\Delta| + \sum_{i=1}^n \mathbb{E}|\xi_i(\Delta - \Delta_i)| \tag{98}$$

where

$$\beta_2 = \sum_{i=1}^n \mathbb{E}[|\xi_i^2|\mathbf{1}_{|\xi_i|>1}] \qquad \textit{and} \qquad \beta_3 = \sum_{i=1}^n \mathbb{E}[|\xi_i^3|\mathbf{1}_{|\xi_i|\le 1}].$$

Note that since $\sum_{i=1}^n \mathbb{E}\left[\xi_i^2\right] = 1$, if $\delta > 0$ satisfies

$$\sum_{i=1}^n \mathbb{E}\left[\xi_i^2 1_{|\xi_i|\ge\delta}\right] < \frac{1}{2}$$

then Eq. (95) holds. In particular, when the ξ_i are standardized i.i.d. random variables, then δ must be on the order of $1/\sqrt{n}$. Furthermore, note that when $\beta_2 + \beta_3 \le 1$ and $4\delta \le 2(\beta_2 + \beta_3)$, then Eq. (95) is automatically satisfied and thus Eq. (97) is immediate. Eq. (98) is obtained by combining Eq. (97) and the sharp Berry-Esseen bound of the sum of independent random variables in [19].

C.2 Berry-Esseen bounds for generalized U-statistics

Proof of Theorem 3: We provide the proof for $U_{n,s}$, the extension to $U_{n,s,\omega}$ follows in the same fashion with the only difference being in the H-decomposition. Without loss of generality, let $\theta = 0$. Observe that

$$U_{n,s} = \sum_{j=1}^{s} \binom{s}{j} H_n^{(j)} = \frac{s}{n} \sum_{i=1}^{n} g(Z_i) + \sum_{j=2}^{s} \binom{s}{j} H_n^{(j)},$$

where $g(z) = \mathbb{E}[h(z, X_2, \ldots, Z_n)]$ and $H_n^{(j)} = \binom{n}{j}^{-1} \sum_{(n,j)} h^{(j)}(Z_{i1}, \ldots, Z_{ij})$. Let

$$\Delta = \sqrt{\frac{n}{s^2\zeta_1}} \sum_{j=2}^{s} \binom{s}{j} H_n^{(j)}$$

and for $l \in \{1, \ldots, n\}$, let

$$\Delta_l = \Delta - \sqrt{\frac{n}{s^2\zeta_1}} \binom{n}{j}^{-1} \sum_{S_j^{(l)}} h^{(j)}(Z_{i1}, \ldots, Z_{ij}) \tag{99}$$

where $S_j^{(l)}$ denotes the collection of all subsets of variables of size j that include the l^{th} observation. The choice of Δ_l plays key role in deciding Berry-Esseen bound. The closer Δ_l is to Δ, the tighter the bound in Eq. (98). We have

$$\sqrt{\frac{n}{s^2\zeta_1}} U_{n,s} = W + \Delta \tag{100}$$

where $W = \sum_{i=1}^{n} \xi_i$ with $\xi_i = g(Z_i)/\sqrt{n\zeta_1}$. For each $i \in \{1, \ldots, n\}$, the random variable $W - \xi_i$ and Δ_i are functions of Z_j, $j \neq i$. Therefore ξ_i is independent of $(W - \xi_i, \Delta_i)$. By the Cauchy-Schwarz inequality, we have

$$\mathbb{E}[|W\Delta|] \leq \sqrt{\mathbb{E}|W|^2} \cdot \sqrt{\mathbb{E}|\Delta|^2} = \sqrt{\mathbb{E}|\Delta|^2}$$

and

$$\sum_{i=1}^{n} \mathbb{E}[|\xi_i(\Delta - \Delta_i)|] \leq \sqrt{\sum_{i=1}^{n} \mathbb{E}[\xi_i^2]} \cdot \sqrt{\sum_{i=1}^{n} \mathbb{E}|\Delta - \Delta_i|^2} \leq \sqrt{n} \max(\sqrt{\mathbb{E}|\Delta - \Delta_i|^2}).$$

Observing the terms on the right, we have

$$\begin{aligned}\frac{s^2\zeta_1}{n}\mathbb{E}|\Delta|^2 &= \mathrm{Var}\left\{\sum_{j=2}^{s}\binom{s}{j}H_n^{(j)}\right\}\\ &= \sum_{j=2}^{s}\binom{s}{j}^2\binom{n}{j}^{-1}V_j\\ &\leq \frac{s^2}{n^2}(\zeta_s - s\zeta_1).\end{aligned}$$

Similarly, we have

$$\begin{aligned}\frac{s^2\zeta_1}{n}\mathbb{E}|\Delta-\Delta_i|^2 &= \mathrm{Var}\left\{\sum_{j=2}^{s}\binom{s}{j}\binom{n}{j}^{-1}\sum_{S_j^i}h^{(j)}(Z_{i1},\ldots,Z_{ij})\right\}\\ &= \sum_{j=2}^{s}\binom{s}{j}^2\binom{n}{j}^{-2}\binom{n-1}{j-1}V_j\\ &= \sum_{j=2}^{s}\binom{s}{j}^2\binom{n}{j}^{-1}\frac{j}{n}V_j\\ &\leq \frac{2s^2}{n^3}\sum_{j=2}^{s}\binom{s}{j}V_j\\ &\leq \frac{2s^2}{n^3}(\zeta_s - s\zeta_1).\end{aligned}$$

Note that

$$\begin{aligned}\beta_2+\beta_3 &= \sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{g(Z_i)}{\sqrt{n\zeta_1}}\right|^2\mathbf{1}_{|g(Z_i)|\geq\sqrt{n\zeta_1}}\right] + \sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{g(Z_i)}{\sqrt{n\zeta_1}}\right|^3\mathbf{1}_{|g(Z_i)|\leq\sqrt{n\zeta_1}}\right]\\ &\leq \frac{1}{n^{1/2}}\frac{\mathbb{E}|g|^3}{\zeta_1^{3/2}}.\end{aligned}$$

Finally, by applying Lemma 4, we obtain

$$\sup_{z\in\mathbb{R}}\left|\Pr\left\{\frac{U_{n,s}}{\sqrt{s^2\zeta_1/n}}\leq z\right\}-\Phi(z)\right| \leq \frac{6.1\mathbb{E}|g|^3}{n^{1/2}\zeta_1^{3/2}} + (1+\sqrt{2})\left\{\frac{s}{n}\left(\frac{\zeta_s}{s\zeta_1}-1\right)\right\}^{1/2}.$$

Proof of Theorem 4: We provide a bound for incomplete, infinite-order U-statistics. An analogous result for *generalized* incomplete U-statistics $U_{n,s,N,\omega}$ can be established by

applying the extended form of the H-decomposition. As eluded to earlier, an incomplete U-statistic can be written as

$$U_{n,s,N} = \frac{1}{N} \sum_{(n,s)} \rho h(Z_{i1}, \ldots, Z_{is}) \tag{101}$$

where $\rho \sim \text{Bernoulli}(p)$ and $p = N/\binom{n}{s}$. Note however that Eq. (101) can also be written as

$$U_{n,s,N} = \frac{1}{p} \left\{ \binom{n}{s}^{-1} \sum_{(n,s)} \rho h(Z_{i1}, \ldots, Z_{is}) \right\} = \frac{1}{p} U^*_{n,s}$$

so that the incomplete U-statistic now takes the form of a scaled, *generalized* complete U-statistic. We thus now consider $U^*_{n,s}$ and can then easily extend the results to $U_{n,s,N}$. First, note that the variance terms ζ^*_c for $c = 1, \ldots, s$ of $U^*_{n,s}$ are different from those of $U_{n,s}$ in Eq. (1). For $c = 1, \ldots, s-1$, we have

$$\zeta^*_c = \text{Cov}(\rho h(Z_1, \ldots, Z_c, Z_{c+1}, \ldots, Z_s), \rho' h(Z_1, \ldots, Z_c, Z'_{c+1}, \ldots, Z'_n)) = p^2 \zeta_c$$

and

$$\zeta^*_s = \text{Cov}(\rho h(Z_1, \ldots, Z_s), \rho h(Z_1, \ldots, Z_s)) = p\zeta_s.$$

The H-decomposition will also be different. Here, we have

$$\begin{aligned}
h^{(1)*} &= \mathbb{E}[\rho h \mid Z_1] = p h^{(1)} \\
h^{(2)*} &= \mathbb{E}[\rho h \mid Z_1, Z_2] - \mathbb{E}[\rho h | Z_1] - \mathbb{E}[\rho h \mid Z_2] = p h^{(2)} \\
&\vdots \\
h^{(s)*} &= \rho h - p \sum_{j=1}^{s-1} \sum_{(s,j)} h^{(j)}(Z_{i1}, \ldots, Z_{ij})
\end{aligned}$$

where the h appearing in the earlier form is replaced here by ρh. These kernels still retain the desirable properties laid out in Proposition 7. Furthermore, we have

$$V^*_j = p^2 V_j \quad (j = 1, \ldots, s-1)$$

and

$$V^*_s = p \sum_{j=1}^{s} \binom{s}{j} V_j - p^2 \sum_{j=1}^{s-1} \binom{s}{j} V_j = p^2 V_s + p(1-p)\zeta_s.$$

Thus,

$$\begin{aligned} U^*_{n,s} &= \sum_{j=1}^{s-1} \binom{s}{j} H_n^{(j)*} + H_n^{(s))*} \\ &= \sum_{j=1}^{s-1} \binom{s}{j} p H_n^{(j)} + \binom{n}{s}^{-1} \sum_{(n,s)} h^{(s)*}(Z_{i1}, \ldots, Z_{is}) \\ &= psH_n^{(1)} + p\sum_{j=2}^{s-1} \binom{s}{j} H_n^{(j)} + \binom{n}{s}^{-1} \sum_{(n,s)} h^{(s)*}(Z_{i1}, \ldots, Z_{is}) \end{aligned}$$

and the decomposition of $U_{n,s,N}$ is

$$U_{n,s,N} = sH_n^{(1)} + \sum_{j=2}^{s-1} \binom{s}{j} H_n^{(j)} + \frac{1}{N} \sum_{(n,s)} h^{(s)*}(Z_{i1}, \ldots, Z_{is}) = sH_n^{(1)} + \Delta.$$

Now, because we have rewritten the incomplete U-statistic as a linear term plus a remainder, we can follow the same general strategy as in the complete case above in applying Lemma 4. In particular, let

$$\Delta - \Delta_i = \sum_{j=2}^{s-1} \binom{s}{j} \binom{n}{j}^{-1} \sum_{S_j^{(i)}} h^{(j)}(Z_{i1}, \ldots, Z_{ij}) + \frac{1}{N} \sum_{S_s^{(i)}} h^{(s)*}(Z_{i1}, \ldots, Z_{is}),$$

where $S_j^{(i)}$ denotes the collection of all subsets of size j that include the i^{th} observation. Then

$$\begin{aligned} \mathbb{E}|\Delta|^2 &= \sum_{j=2}^{s-1} \binom{s}{j}^2 \binom{n}{j}^{-1} V_j + \frac{1}{N^2} \binom{n}{s} V_s^* \\ &= \sum_{j=2}^{s} \binom{s}{j}^2 \binom{n}{j}^{-1} V_j + \frac{1}{N}(1-p)\zeta_s, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}|\Delta - \Delta_i|^2 &= \sum_{j=2}^{s} \binom{s}{j}^2 \binom{n}{j}^{-1} \frac{j}{n} V_j + \frac{1}{N^2} \binom{n-1}{s-1} V_s^* \\ &= \sum_{j=2}^{s} \binom{s}{j}^2 \binom{n}{j}^{-1} \frac{j}{n} V_j + \frac{s}{n} \frac{1}{N}(1-p)\zeta_s \end{aligned}$$

and thus

$$\frac{n}{s^2\zeta_1} \mathbb{E}|\Delta^2| \leq \frac{s}{n} \left[\frac{\zeta_s}{s\zeta_1} - 1 \right] + \frac{n}{Ns}(1-p)\frac{\zeta_s}{s\zeta_1}$$

so that

$$\sum_{i=1}^{n} \frac{n}{s^2\zeta_1} \mathbb{E}|\Delta_i^2| \leq \frac{2s}{n} \left[\frac{\zeta_s}{s\zeta_1} - 1 \right] + \frac{n}{N}(1-p)\frac{\zeta_s}{s\zeta_1}.$$

The result follows by applying Lemma 4.

Proof of Theorem 5: We begin with a bound for incomplete, infinite-order U-statistics. The extension of this result to the generalized setting and can be derived in the same fashion. First rewrite Eq. (101) as

$$U_{n,s,N} = \frac{1}{N} \sum_i \rho_i h(\mathbf{Z}_i) \tag{102}$$

where $\rho_i \sim \text{Bernoulli}(N/\binom{n}{s})$ are i.i.d. and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{is})$ denotes a subsample with index i and the sum is taken over all subsamples. We can rewrite $U_{n,s,N}$ in Eq. (102) as a complete U-statistic $U_{n,s}$ plus some manageable term. We have

$$\begin{aligned} U_{n,s,N} &= \binom{n}{s}^{-1} \sum_i h(\mathbf{Z}_i) + \frac{1}{N} \sum_i (\rho_i - p) h(\mathbf{Z}_i) \\ &= U_{n,s} + \left(\sqrt{1-p}\right) \frac{1}{N} \sum_i \frac{\rho_i - p}{\sqrt{1-p}} h(\mathbf{Z}_i) \\ &= A_n + \left(\sqrt{1-p}\right) B_n \\ &= W_n. \end{aligned} \tag{103}$$

Since we already know the limiting behavior of A_n, it remains only to control B_n. Note that

$$\Pr\left(\sqrt{n} W_n \leq z\right) = \Pr\left\{ \sqrt{N} B_n \leq \frac{z}{\sqrt{\alpha_n (1-p)}} - \sqrt{\frac{N}{1-p}} A_n \right\}$$

where $\alpha_n = n/N$. Conditioning on $Z_1, \ldots, Z_n$, A_n can be treated as a constant and we have

$$\sqrt{N} B_n \mid Z_1, \ldots, Z_n = \binom{n}{s}^{-\frac{1}{2}} \sum_i \left[\frac{(\rho_l - p)}{\sqrt{p(1-p)}} \right] h(\mathbf{Z}_i) \mid Z_1, \ldots, Z_n. \tag{104}$$

Now, $\sqrt{N} B_n \mid Z_1, \ldots, Z_n$ is a sum of independent random variables with variance U_2, where

$$U_2 = \binom{n}{s}^{-1} \sum_i h^2(\mathbf{Z}_i). \tag{105}$$

Let

$$\xi_i = \frac{(p(1-p))^{-1/2} (\rho_i - p) h(\mathbf{Z}_i)}{\sqrt{\sum_i h^2(\mathbf{Z}_i)}}, \quad a_i = \frac{h(\mathbf{Z}_i)}{\sqrt{\sum_i h^2(\mathbf{Z}_i)}}$$

then

$$\sum_i a_i^2 = 1 \quad \text{and} \quad \xi_i = a_i \left[\frac{(\rho_i - p)}{\sqrt{p(1-p)}} \right].$$

Applying the Berry-Esseen bound in [18] for independent random variables, we have

$$\sup_{z \in \mathbb{R}} \left| \Pr\left(\sqrt{N} B_n \leq z \mid Z_1, \ldots, Z_n \right) - \Phi\left(z / \sqrt{U_2} \right) \right| \leq 4.1(\beta_2 + \beta_3), \tag{106}$$

where $\beta_2 = \sum_i \mathbb{E}\left[|\xi_i|^2 1_{|\xi_i| \geq 1} \right]$ and $\beta_3 = \sum_i \mathbb{E}\left[|\xi_i|^3 1_{|\xi_i| \leq 1} \right]$. Next, we show that $(\beta_2 + \beta_3)$ can be uniformly bounded by a small number with high probability and in the rare case when $(\beta_2 + \beta_3)$ is large, trivially, we have $\left| \Pr\left(\sqrt{N} B_n \leq z \mid Z_1, \ldots, Z_n \right) - \Phi\left(z / \sqrt{U_2} \right) \right| \leq 2$. Indeed,

$$\begin{aligned} \beta_2 + \beta_3 &\leq \sum_i \mathbb{E}|\xi_i|^3 \\ &= \left\{ \frac{\binom{n}{s}^{-1} \sum_i |h(\mathbf{Z}_i)|^3}{\left(\binom{n}{s}^{-1} \sum_i |h(\mathbf{Z}_i)|^2 \right)^{3/2}} \right\} \cdot \binom{n}{s}^{-\frac{1}{2}} \left[\frac{2p^2 - 2p + 1}{\left(p(1-p) \right)^{1/2}} \right] \\ &= \frac{U_3}{U_2^{3/2}} \binom{n}{s}^{-1} \left[\frac{1 - 2p}{(p(1-p))^{1/2}} \right], \end{aligned} \tag{107}$$

where $U_3 = \binom{n}{s}^{-1} \sum_i |h(\mathbf{Z}_i)|^3$. The terms of U_2 and U_3 are both complete U-statistics and as such, should be concentrated around their expectations. Let

$$\kappa_1 = \frac{\mathbb{E}|h|^4}{(\mathbb{E}|h|^2)^2}, \quad \kappa_2 = \frac{\mathbb{E}|h|^6}{(\mathbb{E}|h|^3)^2}$$

and recall that κ_1, κ_2 are uniformly bounded by our assumption. Let $\nu_2 = \mathbb{E}|h|^2 (= \zeta_s)$, $\delta_2 = (\frac{s}{n})^\eta \nu_2$, where $\eta > 0$. Then by Chebyshev's inequality, we have

$$\Pr\left(|U_2 - \nu_2| \geq \delta_2 \right) \leq \frac{s/n \cdot \operatorname{Var}\left(|h|^2 \right)}{\delta_2^2} = \left(\frac{s}{n} \right)^{1-2\eta} (\kappa_1 - 1).$$

A similar inequality holds for $|U_3 - \nu_3|$ and therefore with probability of at least $1 - \pi$, where $\pi = c_0 (\frac{s}{n})^{1-2\eta}$ for some constant $c_0 > 0$, we have

$$\left| \frac{U_3}{U_2} \right| = \left| \frac{\frac{U_3}{\nu_3}}{\frac{U_2^{3/2}}{\nu_1^{3/2}}} \cdot \frac{\nu_3}{\nu_2^{3/2}} \right| \leq \left\{ \frac{\frac{\nu_3 + \delta_3}{\nu_3}}{\frac{(\nu_2 - \delta_2)^{3/2}}{\nu_2^{3/2}}} \right\} \frac{\nu_3}{\nu_2^{3/2}} \leq c_1 \frac{\nu_3}{\nu_2^{3/2}},$$

where $c_1 = \left\{ \frac{1+(\frac{s}{n})^{\eta}}{(1-(\frac{s}{n})^{\eta})^{3/2}} \right\}$. Hence, combining this with Eq. (134), with probability of at least $1-\pi$, we have

$$\begin{aligned}
\beta_2 + \beta_3 &\leq c_1 \frac{\nu_3}{\nu_2^{3/2}} \binom{n}{s}^{-\frac{1}{2}} \left\{ \frac{1-2p+2p^2}{(p(1-p))^{1/2}} \right\} \\
&\leq c_1 \frac{\nu_3}{\nu_2^{3/2}} N^{-\frac{1}{2}} \left\{ \frac{1-2p+2p^2}{(1-p)^{1/2}} \right\} \\
&\leq c_1 c_2 \frac{\nu_3}{\nu_2^{3/2}} N^{-\frac{1}{2}},
\end{aligned}$$

where $c_2 = \frac{1-2p+2p^2}{(1-p)^{1/2}}$. The next step is to substitute U_2 by ζ_s by applying Lemma 5 stated below.

Lemma 5.

$$\lim_{a \to 1^+} \sup_{z \in \mathbb{R}} \left| \frac{\Phi(az) - \Phi(z)}{a-1} \right| < \infty. \tag{108}$$

We obtain

$$\begin{aligned}
\sup_{z \in \mathbb{R}} \left| \Phi\left(z/\sqrt{U_2}\right) - \Phi\left(z/\sqrt{\zeta_s}\right) \right| &\leq c_3 |\sqrt{\zeta_s} \wedge \sqrt{U_2}|^{-1} |\sqrt{U_2} - \sqrt{\zeta_s}| \\
&\leq c_3 |\zeta_s \wedge U_2|^{-1} |U_2 - \zeta_s|.
\end{aligned}$$

Since we already derived that with probability of at least $1-\pi$, $|U_2 - \zeta_s| \leq \delta_2$, and thus

$$\sup_{z \in \mathbb{R}} \left| \Phi\left(z/\sqrt{U_2}\right) - \Phi\left(z/\sqrt{\zeta_s}\right) \right| \leq c_3 \frac{\delta_2}{\zeta_s - \delta_2} = c_3 \frac{\left(\frac{s}{n}\right)^{\eta}}{1 - \left(\frac{s}{n}\right)^{\eta}}.$$

Next, since A_n is a complete U-statistic, by Theorem 3, we have

$$\sup_{z \in \mathbb{R}} \left| \Pr\left(\sqrt{n} A_n \leq z\right) - \Pr\left(Y_A \leq z\right) \right| \leq \epsilon_2 \tag{109}$$

where $\epsilon_2 = \frac{6.1 \mathbb{E}|g|^3}{n^{1/2} \zeta_1^{3/2}} + (1+\sqrt{2}) \left\{ \frac{s}{n} \left(\frac{\zeta_s}{s \zeta_1} - 1 \right) \right\}^{1/2}$ and $Y_A \sim N(0, s^2 \zeta_1)$. Lastly,

$$\begin{aligned}
\Pr\left(\sqrt{n} W_n \leq z\right) &= \mathbb{E}\left[\Pr\left\{ \sqrt{N} B_n \leq \frac{z}{\sqrt{\alpha_n (1-p)}} - \sqrt{\frac{N}{1-p}} A_n \mid Z_1, \ldots, Z_n \right\} \right] \\
&\leq \Pr\left\{ Y_B \leq \frac{z}{\sqrt{\alpha_n (1-p)}} - \sqrt{\frac{N}{1-p}} A_n \right\} + \epsilon_1 \\
&= \Pr\left\{ \sqrt{n} A_n \leq z - \sqrt{\alpha_n (1-p)} Y_B \right\} + \epsilon_1
\end{aligned}$$

where $Y_B \sim N(0, \zeta_s)$ is independent of $Z_1, \ldots, Z_n$ and $\epsilon_1 = 4.1 \left\{ c_1 c_2 \frac{\nu_3}{\nu_2^{3/2}} N^{-1/2} + c_3 \frac{\left(\frac{s}{n}\right)^\eta}{1-\left(\frac{s}{n}\right)^\eta} \right\} + 2\pi$. Now, conditioning on Y_B, we have

$$\Pr\left\{\sqrt{n}A_n \le z - \sqrt{\alpha_n(1-p)}Y_B \mid Y_B\right\} \le \Pr\left\{Y_A \le z - \sqrt{\alpha_n(1-p)}Y_B \mid Y_B\right\} + \epsilon_2.$$

Combining Eq. (106) and Eq. (109), we conclude that

$$\begin{aligned}
\Pr\left\{\sqrt{n}W_n \le z\right\} &\le \Pr\left\{Y_A \le z - \sqrt{\alpha_n(1-p)}Y_B\right\} + \epsilon_1 + \epsilon_2 \\
&= \Pr\left\{Y_A + \sqrt{\alpha_n(1-p)}Y_B \le z\right\} + \epsilon_1 + \epsilon_2 \\
&\le \Pr\left\{Y_A + \alpha_n^{1/2} Y_B \le z\right\} + \epsilon_1 + \epsilon_2 + \epsilon_3.
\end{aligned}$$

By Lemma 5, we have

$$\begin{aligned}
\epsilon_3 &\le c_3 \left(s^2\zeta_1 + \alpha_n(1-p)\zeta_s\right)^{-1} \alpha_n p \zeta_s \\
&= c_3 \left(s^2\zeta_1 + \alpha_n(1-p)\zeta_s\right)^{-1} \binom{n}{s}^{-1} n\zeta_s \\
&\le c_3 \min\left\{ p(1-p)^{-1}, \frac{n/s}{\binom{n}{s}} \frac{\zeta_s}{s\zeta_1} \right\}.
\end{aligned}$$

Thus, in summary,

$$\sup_{z\in\mathbb{R}} \left| \Pr\left\{\sqrt{N}(U_{n,s,N}) \le z\right\} - \Pr\left\{Y_W \le z\right\} \right| \le \epsilon_1 + \epsilon_2 + \epsilon_3$$

where

$$\begin{aligned}
\epsilon_1 &= 2c_0 \left(\frac{s}{n}\right)^{1-2\eta} + 4.1 \left\{ c_1 c_2 \frac{\nu_3}{\nu_2^{3/2}} N^{-1/2} + c_3 \frac{\left(\frac{s}{n}\right)^\eta}{1-\left(\frac{s}{n}\right)^\eta} \right\} \\
\epsilon_2 &= \frac{6.1\mathbb{E}|g|^3}{n^{1/2}\zeta_1^{3/2}} + (1+\sqrt{2}) \left\{ \frac{s}{n} \left(\frac{\zeta_s}{s\zeta_1} - 1 \right) \right\}^{1/2} \\
\epsilon_3 &= c_3 \cdot \min\left\{ p(1-p)^{-1}, \frac{n/s}{\binom{n}{s}} \frac{\zeta_s}{s\zeta_1} \right\}
\end{aligned}$$

and $Y_W \sim N(0, s^2\zeta_1/n + \zeta_s/N)$. Note that ϵ_1 and ϵ_2 dominate because of the $\binom{n}{s}$ in the denominator of ϵ_3 and thus the above bound can be simplified as

$$\begin{aligned}
\epsilon_1 + \epsilon_2 + \epsilon_3 \le C \Bigg\{ &\frac{\mathbb{E}|g|^3}{n^{1/2}(\mathbb{E}|g|^2)^{3/2}} + \frac{\mathbb{E}|h|^3}{N^{1/2}(\mathbb{E}|h|^2)^{3/2}} \\
&+ \left\{ \frac{s}{n} \left(\frac{\zeta_s}{s\zeta_1} - 1 \right) \right\}^{1/2} + \left(\frac{s}{n}\right)^{1/3} \Bigg\}.
\end{aligned}$$

Proof of Lemma 5: Let $f(z) = \Phi(az) - \Phi(z)$, then $f'(z) = \frac{1}{\sqrt{2\pi}}\left(ae^{-a^2z^2/2} - e^{-z^2/2}\right)$. Solving $f'(z) = 0$, we get $z^2(a) = \frac{\log(a)}{(a^2-1)/2}$. Then

$$\begin{aligned}\lim_{a\to 1^+}\sup_{z\in\mathbb{R}}\left|\frac{\Phi(az)-\Phi(z)}{a-1}\right| &= \lim_{a\to 1^+}\left|\frac{\Phi(az(a))-\Phi(z(a))}{a-1}\right| \\ &= \lim_{a\to 1^+}\left|\Phi'(az(a))(az(a))' - \Phi'(z(a))z'(a)\right|.\end{aligned} \tag{110}$$

According to the Taylor expansion of $\log(a)$ at $a = 1$, we have $z^2(a) = \frac{2}{a+1}\left(1 - \frac{a-1}{2} + \dots\right) = 1 + o(a-1)$. Therefore, $\lim_{a\to 1^+} z^2(a) = 1$ and $\lim_{a\to 1^+} z'(a) = -\frac{1}{2}$, thus we have

$$\lim_{a\to 1^+}\sup_{z\in\mathbb{R}}\left|\frac{\Phi(az)-\Phi(z)}{a-1}\right| = \frac{e^{-1/2}}{\sqrt{2\pi}} < \infty \tag{111}$$

as desired.

C.3 Discussion on a tighter bound

Here we provide a sketch of the proof of Theorem 6. Let $m = \lfloor n/s \rfloor$ and define

$$V(Z_1, Z_2, \dots, Z_n) = \frac{1}{m}\sum_{j=0}^{m-1} h(Z_{j\cdot s+1}, Z_{j\cdot s+2}\dots, Z_{j\cdot s+s}).$$

The general form of a complete U-statistic in Eq. (1) can be rewritten as

$$U_{n,s} = \frac{1}{n!}\sum_{\beta\in S_n} V(Z_{\beta_1,\beta_2,\dots,\beta_n})$$

where S_n consists of all permutations of $(1, 2, \dots, n)$. Now, suppose that $(h-\theta)/\sigma$ is sub-Gaussian with variance proxy v^2, where $\sigma^2 = \mathrm{Var}(h)$, then by definition, we have

$$\mathbb{E}\left[\exp(\lambda(h-\theta))\right] \le \exp\left\{\frac{\lambda^2\sigma^2v^2}{2}\right\}, \quad \lambda\in\mathbb{R} \tag{112}$$

and hence we have

$$\begin{aligned}
\Pr(U_{n,s} - \theta > t) &\leq \exp(-\lambda t)\mathbb{E}[\exp\left(\lambda(U_{n,s} - \theta)\right)] \\
&\leq \exp(-\lambda t) \sum_{\beta \in S_n} \frac{1}{n!} \mathbb{E}[\exp\left(\lambda(V(Z_{\beta_1}, Z_{\beta_2}, \ldots, Z_{\beta_n}) - \theta)\right)] \\
&= \exp(-\lambda t)\mathbb{E}[\exp\left(\lambda(V - \theta)\right)] \qquad (113) \\
&\leq \exp\left\{-\frac{mt^2}{2\sigma^2 v^2}\right\}, \quad t > 0.
\end{aligned}$$

The second inequality in Eq. (113) is due to Jensen's inequality and the last inequality is due to Hoeffding inequality. Observe that $\Pr\left(U_{n,s} - \theta < t\right)$ follows in the same manner (recall that Eq. (112) holds for all $\lambda \in \mathbb{R}$), and we get

$$\Pr\left(|U_{n,s} - \theta| \geq t\right) \leq 2\exp\left\{-\frac{mt^2}{2\sigma^2 v^2}\right\}. \qquad (114)$$

Let $t = m^{-\eta}\sigma$ where $0 < \eta < 1/2$. Then with probability at least $1 - 2\exp\left(-\frac{1}{2v^2}(\lfloor n/s \rfloor)^{1-2\eta}\right)$, $|U_{n,s} - \theta| \leq (\lfloor n/s \rfloor)^{-\eta}\sigma$. Therefore if $|h - \theta|^2$ and $|h - \theta|^3$ are sub-Gaussian after being standardized, we can then apply Eq. (114) in the proof of Theorem 5 to obtain the improved result.

Appendix D Proofs in Chapter 5

D.1 IJ_{B} for bootstrap

Proof of Theorem 7:

1. By definition,

$$\begin{aligned}
\mathbb{E}_*[s^* w_j^*] &= \sum_{w_1^*+\cdots+w_n^*=n} p(w_1^*,\ldots,w_n^*) s(X_1^*,\ldots,X_n^*) w_j^* \\
&= \sum_{\substack{w_j^*\geq 1 \\ w_1^*+\cdots+w_n^*=n}} \frac{(n-1)!}{w_1^*\ldots((w_j^*-1)!)\cdots(w_n^*)!}\frac{1}{n^{n-1}} s(X_1^*,\ldots,X_n^*) \\
&= \mathbb{E}_*[s(X_1^*,\ldots,X_n^*)|X_1^*=X_j] \\
&= e_j.
\end{aligned}$$

2. Conditioned on the data, $X_1^*,\ldots,X_n^*$ are i.i.d. Consider the Hájek projection of s^*, and we have

$$\begin{aligned}
\sum_i (\mathbb{E}_*[s^*|X_i^*] - \mathbb{E}_*[s^*]) &= \sum_i \sum_j (\mathbb{E}_*[s^*|X_i^*=X_j] - \mathbb{E}_*[s^*])\mathbf{1}_{\{X_i^*=X_j\}} \\
&= \sum_i \sum_j (e_j - s_0)\mathbf{1}_{\{X_i^*=X_j\}} \\
&= \sum_j \sum_i (e_j - s_0)\mathbf{1}_{\{X_i^*=X_j\}} \\
&= \sum_j w_j^*(e_j - s_0),
\end{aligned}$$

   which is a linear function of w_j^* for $j=1,\ldots,n$ and thus $l^* = \sum_j w_j^*(e_j - s_0)$.

3. By 1, $\mathrm{IJ}_{\mathrm{B}} = \sum_j \mathrm{Cov}_*^2(s^*, w_j^*) = \sum_j (\mathbb{E}_*[s^* w_j^*] - \mathbb{E}_*[s^*]\mathbb{E}_*[w_j^*])^2 = \sum_j (e_j - s_0 \cdot 1)^2$. By 2, we have $\mathrm{Var}_*(l^*) = \mathrm{Var}_*(\sum_j w_j^*(e_j - s_0)) = \sum_j (e_j - s_0)^2$. Thus, $\mathrm{Var}_*(l^*) = \mathrm{JK}_{\mathrm{B}} = \mathrm{IJ}_{\mathrm{B}}$.

Calculations for sample maximum: Consider $s = \max_i X_i$, where $X_1, \ldots, X_n$ are uniformly distributed. The joint distribution density function of the two order statistics $X_{(i)} < X_{(j)}$ is Thus we have

$$\mathrm{Cov}(X_{(i)}, X_{(j)}) = \frac{i(n-j+1)}{(n+1)^2(n+2)}, \quad \mathbb{E}[X_{(i)}X_{(j)}] = \frac{i(j+1)}{(n+1)(n+2)}. \tag{115}$$

Note that

$$\mathbb{E}_*[s^*] = s_0 = \sum_{I=1}^{n} X_{(i)} p_i^n,$$

where $p_i^n = q_i^n - q_{i-1}^n$ and $q_i^n = \left(\frac{i}{n}\right)^n$ for $i = 1, \ldots n$. Thus,

$$\mathrm{Var}(\mathbb{E}_*[s^*]) = v^T \mathrm{Cov}(\mathbf{u}) v, \tag{116}$$

where $\mathbf{u} = (X_{(1)}, \cdots, X_{(n)})$ and $v = (p_1^n, \cdots, p_n^n)$. Let

$$\tilde{e}_i = \sum_{j=I+1}^{n} X_{(j)} p_j^{n-1} + X_{(i)} q_i^{n-1}, \quad \text{where } q_i^{n-1} = \left(\frac{i}{n}\right)^{n-1}.$$

We have $\mathrm{Var}_*(l^*) = \sum_{I=1}^n (e_i - s_0)^2 = \sum_{I=1}^n (\tilde{e}_i - s_0)^2$. Thus,

$$\mathbb{E}[\mathrm{Var}_*(l^*)] = \sum (v_i - v)^T \mathbb{E}[\mathbf{u}\mathbf{u}^T](v_i - v), \tag{117}$$

where $v_i = (\cdots, 0, \cdots, q_i^{n-1}, \cdots, p_j^{n-1} \cdots)$ and $\mathrm{V} = (v_i - v)$. Let

$$\mathrm{A} = \mathrm{Cov}(\mathbf{u}) = \left[\frac{i(n+1-j)}{(n+1)^2(n+2)}\right]_{ij}, \quad \mathrm{B} = \mathbb{E}[\mathbf{u}\mathbf{u}^T] = \left[\frac{i(j+1)}{(n+1)(n+2)}\right]_{ij}, \tag{118}$$

then $\frac{\mathbb{E}[\mathrm{Var}_*(l^*)]}{\mathrm{Var}(\mathbb{E}_*[s^*])} = \frac{\mathrm{V}^T \mathrm{B} \mathrm{V}}{v^T \mathrm{A} v}$. Next, we have

$$\begin{aligned}
v^T \mathrm{A} v = &= \frac{1}{(n+1)^2(n+2)} \sum_i \sum_j p_i^n p_j^n i(n+1-j) \\
&= \frac{1}{(n+1)^2(n+2)} \left(\sum_i i \cdot p_i^n\right) \left(\sum_j (n-j+1) p_j^n\right) \\
&= \frac{1}{(n+1)^2(n+2)} \left(\sum_i i \cdot p_i^n\right) \left(n+1-\sum_i i \cdot p_i^n\right) \\
&= \frac{1}{(n+1)^2(n+2)} (n - \sum_j (\frac{j-1}{n})^n))(1 + \sum_j (\frac{j-1}{n})^n))
\end{aligned} \tag{119}$$

by the fact that

$$\sum_{i=1}^{n} i \cdot p_i^n = \sum_{i=1}^{n} i q_i^n - \sum_{i=0}^{n-1}(i+1)q_i^n = n - \sum_{i=0}^{n-1} q_i^n. \tag{120}$$

Next, let $\mathbf{e}_n = [1, 2 \cdots, n]^T$, then

$$\begin{aligned}
\mathrm{V}^T \mathrm{B} \mathrm{V} &= \frac{\mathrm{V}^T \mathbf{e}_n \cdot (\mathbf{e}_n^T + 1_n^T)\mathrm{V}}{(n+1)(n+2)} \\
&= \frac{\mathrm{V}^T \mathbf{e}_n \cdot \mathbf{e}_n^T \mathrm{V}}{(n+1)(n+2)} \\
&= \frac{1}{(n+1)(n+2)} \sum_i \left[(n - \sum_{j=I}^{n-1} (\frac{j}{n})^{n-1}) - (n - \sum_{j=0}^{n-1} (\frac{j}{n})^n) \right]^2 \\
&= \frac{1}{(n+1)(n+2)} \sum_i \left[\sum_{j=1}^{n} (\frac{j-1}{n})^n - \sum_{j=i+1}^{n} (\frac{j-1}{n})^{n-1} \right]^2 .
\end{aligned} \tag{121}$$

In summary, we have

$$\begin{aligned}
\frac{\mathbb{E}[\mathrm{Var}_*(l^*)]}{\mathrm{Var}(\mathbb{E}_*[s^*])} &= \frac{(n+1)\sum_i [\sum_{j=1}^{n} (\frac{j-1}{n})^n - \sum_{j=i+1}^{n} (\frac{j-1}{n})^{n-1}]^2}{(n - \sum_j (\frac{j-1}{n})^n)(1 + \sum_j (\frac{j-1}{n})^n)} \\
&\to c \in [0.24, 0.25] \quad \text{as } n \to \infty.
\end{aligned} \tag{122}$$

Proof of Theorem 8: By the assumption, $\mathbb{E}_*[s^*] = l_b + \frac{1}{n} r_b$, where $l_b = \frac{1}{n} \sum \mathbb{E}[\mathbb{E}_*[s^*]|X_i]$ and $r_b = o_p(1)$. Denote $\varphi(X_i) = \mathbb{E}[\mathbb{E}_*[s^*]|X_i]$, then

$$\begin{aligned}
\mathrm{Var}_*(l^*) &= \sum (\mathbb{E}_*[s^*|X_1^* = X_i] - \mathbb{E}_*[s^*])^2 \\
&= \sum (\mathbb{E}_*[\mathbb{E}_*[s^*]|X_1 = X_i, X_2^*, \ldots, X_n^*] - \mathbb{E}_*[s^*])^2 \\
&= \sum (\mathbb{E}_*[\frac{1}{n} \sum \varphi(X_i) + \frac{1}{n} r_b | X_1 = X_i, X_2^*, \ldots, X_n^*] - \frac{1}{n} \sum \varphi(X_i) + \frac{1}{n} r_b)^2 \\
&= \frac{1}{n^2} \sum (\varphi(X_i) - \bar{\varphi} + \mathbb{E}_*[r_b | X_1 = X_i, X_2^*, \ldots, X_n^*] - r_b)^2
\end{aligned} \tag{123}$$

Since r_b is permutation symmetric, $\frac{1}{n} \sum \mathbb{E}_*[r_b | X_1 = X_i, X_2^*, \ldots, X_n^*] = \mathbb{E}[r_b | X_1^*, \ldots, X_n^*] = r_b$. Thus, $\mathbb{E}_*[r_b | X_1 = X_i, X_2^*, \ldots, X_n^*]$ for $i = 1, \ldots, n$ are i.i.d. and are also $o_p(1)$. Therefore,

$$\begin{aligned}
\mathrm{Var}_*(l^*) &= \frac{1}{n^2} \sum (\varphi(X_i) - \bar{\varphi})^2 + o_p(1) \\
&\xrightarrow{p} \mathrm{Var}(\mathbb{E}_*[l^*]) \\
&\to \mathrm{Var}(\mathbb{E}_*[s^*]),
\end{aligned} \tag{124}$$

which implies that $\mathrm{IJ_B}$ is consistent.

Proof of Theorem 9: Note that $\mathbb{E}[\mathrm{Var}_*(l^*)] = (n-1)\mathbb{E}[e_1^2 - e_1e_2]$ and $\mathrm{Var}(\mathbb{E}_*[s^*]) = \frac{1}{n}\mathrm{Var}(e_1) + \frac{n-1}{n}\mathrm{Cov}(e_1, e_2)$. Let $\rho = \mathrm{Cov}(e_1, e_2)/\mathrm{Var}(e_1)$, we have

$$\begin{aligned}\mathbb{E}[\mathrm{Var}_*(l^*)]/\mathrm{Var}(\mathbb{E}_*[s^*]) &= \frac{(n-1)(1-\rho)}{1/n + (n-1)/n \cdot \rho} \\ &= n\frac{1-\rho}{1/(n-1)+\rho}.\end{aligned} \tag{125}$$

Let $f(\rho) = \frac{n(1-\rho)}{1/(n-1)+\rho}$, we have

$$\begin{aligned}f(\rho) &= n\left(-1 + \frac{1}{1-(n-1)/n\cdot(1-\rho)}\right) \\ &= n(-1 + \frac{1}{1-r}) \\ &= n\left(-1+1+r+r^2+r^3+\dots\right) \quad (|r|<1) \\ &= n(r+r^2+r^3+\dots) \quad (|r<1|),\end{aligned} \tag{126}$$

where $r = (n-1)/n\cdot(1-\rho)$. Therefore, only if $r = \frac{1}{n} + o(\frac{1}{n})$, then $f(\rho(r)) \to 1$ as $n \to \infty$. In particular,

$$r = \frac{1}{n} + o(\frac{1}{n}) \iff 1-\rho = \frac{1}{n} + o(\frac{1}{n}). \tag{127}$$

Hence $\lim_{n\to\infty} f(\rho) = 1$ if and only if $\lim_{n\to\infty} n(1-\rho) = 1$. Thus, $\mathrm{IJ_B}$ is an asymptotic unbiased estimation of $\mathrm{Var}(\mathbb{E}_*[s^*])$ if and only if $1-\rho = 1/n + o(1/n)$.

D.2 $\mathrm{IJ_U}$ and s-$\mathrm{IJ_U}$ for U-statistic

Proof of Theorem 10: In sampling without replacement, the probability of $(x_1, \dots, x_k)$ being selected is

$$\begin{cases}\sum_{i_1,\dots,i_k} \frac{\mathbb{P}_n(x_{i_1})}{1} \times \frac{\mathbb{P}_n(x_{i_2})}{1-\mathbb{P}_n(x_1)} \times \cdots \times \frac{\mathbb{P}_n(x_{i_k})}{1-\sum_{j=1}^{k-1}\mathbb{P}_n(x_{ij})}, & x_1, \dots x_k \in \mathcal{D}_n \text{ and are distinct} \\ 0, & \text{otherwise.}\end{cases} \tag{128}$$

Note that any subsampling with a general re-weighting scheme can be derived similarly. Consider $f((1-\epsilon)\mathbb{P}_n + \epsilon\delta_{X_i})$ and let $\delta = 1-\epsilon$. We first provide the probability of obtaining $(x_1, \ldots, x_k)$. If $X_i \notin (x_1, x_2, \ldots, x_k)$, then

$$p(x_1, x_2, \ldots, x_n) = p_0 = \left[\frac{\delta}{n} \cdot \frac{\delta}{(n-\delta)} \cdots \frac{\delta}{(n-(k-1)\delta)}\right] \times k!. \tag{129}$$

If $X_i \in (x_1, \ldots, x_k)$, then $p(x_1, x_2, \ldots, x_k) = p_1 = \sum_{i=0}^{k-1} q_i$, where

$$\begin{aligned}
q_0 &= \left[\tfrac{(n-(n-1)\delta)}{n} \cdot \tfrac{1}{n-1} \cdots \tfrac{1}{n-k+1}\right] \times (k-1)! \\
q_1 &= \left[\tfrac{\delta}{n} \cdot \tfrac{n-(n-1)\delta}{n-\delta} \cdot \tfrac{1}{n-2} \cdot \tfrac{1}{n-k+1}\right] \times (k-1)! \\
&\vdots \\
q_{k-1} &= \left[\tfrac{\delta}{n}\tfrac{\delta}{n-\delta} \cdots \tfrac{\delta}{n-(k-2)\delta} \cdot \tfrac{n-(n-1)\delta}{n-(k-1)\delta}\right] \times (k-1)!.
\end{aligned}$$

Thus,

$$f((1-\epsilon)\mathbb{P}_n + \epsilon\delta_{X_i}) = \sum_{i_1, \ldots i_k} s(X_{i_1}, \ldots, X_{i_k})(p_0 \mathbf{1}_{i \notin \{i_1, \ldots, i_k\}} + p_1 \mathbf{1}_{i \in \{i_1, \ldots, i_k\}}).$$

By definition, the IJ of U-statistic is

$$\begin{aligned}
\mathrm{IJ}_{\mathrm{U}} &= \lim_{\epsilon \to 0} \frac{f((1-\epsilon)\mathbb{P}_n + \epsilon\delta_{X_i}) - f(\mathbb{P}_n)}{\epsilon} \\
&= \lim_{\delta \to 1} \frac{f(\delta\mathbb{P}_n + (1-\delta)\delta_{X_i}) - f(\mathbb{P}_n)}{1-\delta}.
\end{aligned} \tag{130}$$

We have

$$\begin{aligned}
\frac{1}{p} p_0'(\delta)|_{\delta=1} &= -\left[\frac{0}{n} + \frac{1}{n-1} + \cdots + \frac{k-1}{n-(k-1)}\right] - k, \\
\frac{1}{p} q_j'|_{\delta=1} &= -\left[\frac{0}{n} + \frac{1}{n-1} + \frac{2}{n-2} + \cdots \frac{j}{n-j}\right] + (n-j-1), \quad j = 0, \ldots, k-1,
\end{aligned}$$

and thus

$$\frac{1}{p} p_0'(\delta)|_{\delta=1} = -\left[\frac{0}{n} + \frac{1}{n-1} + \cdots + \frac{k-1}{n-(k-1)}\right] - k,$$

and

$$\begin{aligned}\frac{1}{p}p_1' &= \frac{1}{p}\sum_{j=0}^{k-1} q_j'|_{\delta=1} \\ &= \frac{1}{k}\sum_{j=0}^{k-1}\left[(n-j-1)-\left[\frac{0}{n}+\frac{1}{n-1}+\frac{2}{n-2}+\cdots\frac{j}{n-j}\right]\right] \\ &= -\frac{1}{k}\left[\frac{0\cdot k}{n}+\frac{1\cdot(k-1)}{n-1}+\cdots+\frac{(k-1)\cdot 1}{n-(k-1)}\right]-\frac{k+1}{2}+n.\end{aligned}$$

Putting all together, we have

$$\begin{aligned}\lim_{\delta\to 1}\frac{f(\delta\mathbb{P}_n+(1-\delta)\delta_{X_i})-f(\mathbb{P}_n)}{1-\delta}. &= \sum_{(n,k)}(p_0'\mathbf{1}_{w_i^*=0}+p_1'\mathbf{1}_{w_i^*=1})s(X_{i_1},\ldots,X_{i_k}) \\ &= \sum_{(n,k)} p\left[\frac{p_0'}{p}+(\frac{p_1'}{p}-\frac{{p_0}'}{p})w_i^*\right]s(X_{i_1},\ldots,X_{i_k}) \qquad (131) \\ &= \frac{k}{n}(\frac{p_1'}{p}-\frac{p_0'}{p})e_i+\frac{p_0'}{p}s_0,\end{aligned}$$

where $e_i = \mathbb{E}_*[s^*|X_1^* = X_i]$ and $s_0 = \mathbb{E}_*[s^*]$. Note that $*$ refers to the procedure of subsampling without replacement. The infinitesimal jackknife estimate is

$$\begin{aligned}\mathrm{IJ_U} &= \frac{1}{n^2}\sum_{j=1}^{n}\left[\frac{k}{n}(\frac{p_1'}{p}-\frac{p_0'}{p})e_j+\frac{p_0'}{p}s_0\right]^2. \\ &= \frac{k^2}{n^2}\sum_{j=1}^{n}\left[\frac{p_1'-p_0'}{np}e_j+\frac{p_0'}{kp}s_0\right]^2 \qquad (132) \\ &= \frac{k^2}{n^2}\sum_{j=1}^{n}[\alpha e_j+\beta s_0]^2\end{aligned}$$

where

$$\alpha = (p_1'-p_0')/(np) \quad = 1+\frac{1}{n}\left\{\frac{k-1}{2}-\frac{1}{k}\sum_{j=0}^{k-1}\frac{j^2}{(n-j)}\right\}, \qquad (133)$$

and

$$\beta = -p_0'/(kp) \quad = 1+\frac{1}{k}\sum_{j=0}^{k-1}\frac{j}{n-j}. \qquad (134)$$

We now derive $\mathbb{E}[\mathrm{IJ_U}]$. We can use the Hoeffding decomposition to rewrite U-statistic as a sum of many uncorrelated terms, so that the variance of the U-statistic can be written as

a linear combination of the variance of those terms correspondingly. Interestingly, $\alpha e_j + \beta s_0$ can be decomposed similarly. Indeed,

$$
\begin{aligned}
&(\alpha e_1 - \beta s_0) \\
&= -\beta \binom{n}{k}^{-1} \sum s(X_{i_1}, \ldots, X_{i_k}; \nexists 1) \\
&\quad + \left(\frac{\alpha \cdot (k-1)!}{(n-1)\ldots(n-k+1)} - \frac{\beta \cdot (k-1)!k}{n \cdots (n-k+1)} \right) \sum s(X_{i_1}, \ldots, X_{i_k}; \exists 1) \\
&= -(1 - \frac{k}{n})\beta \binom{n-1}{k}^{-1} \sum s(X_{i_1}, \ldots, X_{i_k}; \nexists 1) \\
&\quad + (\alpha - \frac{k}{n}\beta) \binom{n-1}{k-1}^{-1} \sum s(X_{i_1}, \ldots, X_{i_k}; \exists 1) \\
&= -(1 - \frac{k}{n})\beta \cdot \sum_{j=1}^{k} \binom{k}{j} \binom{n-1}{j}^{-1} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \nexists 1) \\
&\quad + (\alpha - \frac{k}{n}\beta) \sum_{j=1}^{k-1} \binom{k-1}{j} \binom{n-1}{j}^{-1} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \nexists 1) \\
&\quad + (\alpha - \frac{k}{n}\beta) \sum_{j=1}^{k} \binom{k-1}{j-1} \binom{n-1}{j-1}^{-1} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \exists 1) \\
&:= A_n + B_n,
\end{aligned}
\tag{135}
$$

where

$$
A_n = \sum_{j=1}^{k} \left[(\alpha - \frac{k}{n}\beta) \binom{k-1}{j} - (1 - \frac{k}{n})\beta \binom{k}{j} \right] \binom{n-1}{j}^{-1} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \nexists 1)
$$

and

$$
B_n = (\alpha - \frac{k}{n}\beta) \sum_{j=1}^{k} \binom{k-1}{j-1} \binom{n-1}{j-1}^{-1} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}).
$$

Thus,

$$
\mathrm{Var}(A_n) = \sum_{j=1}^{k} \left[(\frac{k}{j} - 1)\alpha + (\frac{k}{n} - \frac{k}{j})\beta \right]^2 \binom{k-1}{j-1}^2 \binom{n-1}{j}^{-1} V_j \tag{136}
$$

and

$$
\mathrm{Var}(B_n) = (\alpha - \frac{k}{n}\beta)^2 \sum_{j=1}^{k} \binom{k-1}{j-1}^2 \binom{n-1}{j-1}^{-1} V_j \tag{137}
$$

where $V_j = \mathrm{Var}(s^{(j)})$. Since A_n and B_n are uncorrelated, we have

$$\begin{aligned}
&\mathbb{E}\left[(\alpha e_1 - \beta s_0)^2\right] \\
&= \mathrm{Var}(A_n) + \mathrm{Var}(B_n) \\
&= \sum_{j=1}^{k} \binom{k-1}{j-1}^2 \left[\left[(\frac{k}{j} - 1)\alpha + (\frac{k}{n} - \frac{k}{j})\beta\right]^2 \binom{n-1}{j}^{-1} + (\alpha - \frac{k}{n}\beta)^2 \binom{n-1}{j-1}^{-1}\right] V_j \\
&= \sum_{j=1}^{k} \binom{k-1}{j-1}^2 \Lambda(j) V_j,
\end{aligned}$$

where $\Lambda(j) = \left[\left[(\frac{k}{j} - 1)\alpha + (\frac{k}{n} - \frac{k}{j})\beta\right]^2 \binom{n-1}{j}^{-1} + (\alpha - \frac{k}{n}\beta)^2 \binom{n-1}{j-1}^{-1}\right]$, for $j = 1, \cdots, k$.
Therefore,

$$\begin{aligned}
\mathbb{E}[\mathrm{IJ_U}] &= \frac{k^2}{n^2} \sum \mathbb{E}[(\alpha e_j - \beta s_0)^2] \\
&= \frac{k^2}{n} \sum_{j=1}^{k} \binom{k-1}{j-1}^2 \Lambda(j) V_j.
\end{aligned} \tag{138}$$

Recall that

$$\mathrm{Var(U)} = \sum_{j=1}^{k} \binom{k}{j}^2 \binom{n}{j}^{-1} V_j. \tag{139}$$

We consider the ratio of the coefficient of V_j in $\mathbb{E}[\mathrm{IJ_U}]$ and that in $\mathrm{Var(U)}$ and obtain

$$\begin{aligned}
r_j &= \frac{k^2}{n} \Lambda(j) \binom{k-1}{j-1}^2 \binom{k}{j}^{-2} \binom{n}{j} \\
&= \frac{k^2}{n} \frac{j^2}{k^2} \Lambda(j) \binom{n}{j} \\
&= \frac{(n-k)^2}{n^2} \left[\frac{j}{1 - j/n} \alpha^2 + \frac{n}{k^2 (n-k)^2} (\alpha - \beta)^2\right]
\end{aligned} \tag{140}$$

for $j = 1, \ldots, k$.

Proof of Proposition 4:

$$\begin{aligned}
\mathrm{Cov}_*(s^*, w_j^*) &= \sum_{w_1^*+\cdots+w_n^*=k} p(w_1^*, \ldots, w_n^*)[s^* - s_0]w_j^* \\
&= \sum_{w_1^*+\cdots+w_n^*=k} p(w_1^*, \ldots, w_n^*)s^* w_j^* - \frac{k}{n}s_0 \\
&= \frac{k}{n} \sum_{w_j^*=1, w_1^*+\cdots+w_n^*=k} \frac{(k-1)!}{(n-1)\cdots(n-k+1)} s^* - \frac{k}{n}s_0 \\
&= \frac{k}{n}\left[\mathbb{E}_*[s(X_1^*, \ldots, X_k^*)|X_1^* = X_j] - s_0\right] \\
&= \frac{k}{n}(e_j - s_0).
\end{aligned} \tag{141}$$

It follows that s-IJ $= \sum_{j=1}^n \mathrm{Cov}_*^2(s^*, w_j^*) = \frac{k^2}{n^2}\sum(e_j - s_0)^2$.

Proof of Theorem 11: Similar to the calculation for $\mathrm{IJ_U}$, we have

$$\mathbb{E}\left[(e_j - s_0)^2\right] = \left(\frac{n-k}{n}\right)^2 \sum_{j=1}^k \binom{k-1}{j-1}^2 \left[\binom{n-1}{j}^{-1} + \binom{n-1}{j-1}^{-1}\right] V_j.$$

Next,

$$\begin{aligned}
\mathbb{E}[\text{s-IJ}_\mathrm{U}] &= \frac{k^2}{n^2}\sum \mathbb{E}[(e_j - s_0)^2] \\
&\quad \frac{k^2}{n}\left(\frac{n-k}{n}\right)^2 \sum_{j=1}^k \binom{k-1}{j-1}^2 \left[\binom{n-1}{j}^{-1} + \binom{n-1}{j-1}^{-1}\right] V_j.
\end{aligned} \tag{142}$$

Since $\mathrm{Var(U)} = \sum_{j=1}^k \binom{k}{j}^2 \binom{n}{j}^{-1} V_j$, we have

$$\begin{aligned}
r_j &= \frac{k^2}{n}\left(\frac{n-k}{n}\right)^2 \binom{k-1}{j-1}^2 \left[\binom{n-1}{j}^{-1} + \binom{n-1}{j-1}^{-1}\right] \binom{k}{j}^{-2} \binom{n}{j} \\
&= \left(\frac{n-k}{n}\right)^2 \frac{j^2}{n}\left[2 + \frac{j}{n-j} + \frac{n-j}{j}\right] \\
&= \left(\frac{n-k}{n}\right)^2 \frac{j^2}{n} \cdot \frac{n^2}{(n-j)j} \\
&= \left(\frac{n-k}{n}\right)^2 \frac{j}{1-j/n}, \quad j = 1, \ldots, k.
\end{aligned} \tag{143}$$

Proof of Theorem 12: For simplicity, we first ignore the extra randomness ω. According to the H-decomposition,

$$\begin{aligned}
\text{s-IJ}_{\text{U}} &= \frac{k^2}{n^2}\frac{(n-k)^2}{n^2}\sum_{i=1}^{n}\left[\sum_{j=1}^{k}-\binom{k-1}{j-1}\binom{n-1}{j}^{-1}\sum s^{(j)}(X_{i_1},\ldots,X_{i_j};\nexists i)\right. \\
&\quad \left.+\binom{k-1}{j-1}\binom{n-1}{j-1}^{-1}\sum s^{(j)}(X_{i_1},\ldots,X_{i_j};\exists i)\right]^2 \\
&= \frac{k^2}{n^2}\frac{(n-k)^2}{n^2}\sum_{i=1}^{n}\left[-\frac{1}{n-1}\sum_{j\neq i}^{n}s^{(1)}(X_i)+s^{(1)}(X_i)+\frac{k^2}{n^2}\sum_{j=2}^{k}\binom{k-1}{j-1}\binom{n-1}{j}^{-1}\right. \\
&\quad \left.\sum s^{(j)}(X_{i_1},\ldots,X_{i_j};\nexists 1)+\binom{k-1}{j-1}\binom{n-1}{j-1}^{-1}\sum s^{(j)}(X_{i_1},\ldots,X_{i_j};\exists 1)\right]^2 \\
&= \frac{k^2}{n^2}\frac{(n-k)^2}{n^2}\sum_{i=1}^{n}\left[s^{(1)}(X_i)+\text{T}_i\right]^2.
\end{aligned} \tag{144}$$

We already know that $s^{(1)}(X_i)$ and T_i are uncorrelated. After some calculation, we find that

$$\begin{aligned}
&\mathbb{E}[(s^{(1)}(X_i))^2] = V_1 \\
&\qquad \mathbb{E}[\text{T}_i^2] = \frac{1}{n-1}V_1+\sum_{j=2}^{k}\binom{k-1}{j-1}^2\left[\binom{n-1}{j}^{-1}+\binom{n-1}{j-1}^{-1}\right]V_j,
\end{aligned}$$

then

$$\begin{aligned}
\mathbb{E}[\text{T}_i^2] &\approx \frac{1}{n-1}V_1+\sum_{j=2}^{k}\binom{k-1}{j-1}^2\binom{n-1}{j-1}^{-1}V_j \\
&= \frac{1}{n-1}V_1+\sum_{j=2}^{k}\frac{j}{k}\binom{k-1}{j-1}\binom{n-1}{j-1}^{-1}\left[\binom{k}{j}V_j\right]. \\
&\leq \frac{1}{n-1}V_1+\frac{2}{n}\sum_{j=2}^{k}\binom{k}{j}V_j. \\
&= \frac{1}{n-1}\zeta_1+\frac{2k}{n}(\zeta_k-k\zeta_1).
\end{aligned} \tag{145}$$

Let $\text{L} = \mathbb{E}[(s^{(1)}(X_i))^2]$ and $\text{R} = \text{T}_i^2]$. Since $\frac{k}{n}(\frac{\zeta_k}{k\zeta_1}-1)\to 0$,

$$\text{R/L} \leq \frac{2/n(\zeta_k-k\zeta_1)}{\zeta_1}+\frac{1}{n-1}\to 0. \tag{146}$$

Therefore, $s^{(1)}(X_i)$ dominates T_i and thus

$$\begin{aligned}\text{s-IJ}_{\mathrm{U}} &\xrightarrow{p} \frac{k^2}{n^2}\frac{(n-k)^2}{n^2}\sum_{i=1}^{n}[s^{(1)}(X_i)]^2 \\ &\xrightarrow{p} \frac{k^2}{n}\frac{(n-k)^2}{n^2}\mathbb{E}[s^{(1)}(X_i)]^2 \\ &\to \frac{k^2}{n}\zeta_1.\end{aligned} \tag{147}$$

Note that the extra ω does not bring more technical difficulty since structure of the U-statistic is unchanged. Similarly, let

$$e_i^{\omega} = \binom{n-1}{k-1}^{-1}\sum s(X_i,\ldots,;\omega), \quad \text{and} \quad s_0^{\omega} = \binom{n}{k}^{-1}\sum s(\ldots;\omega). \tag{148}$$

Note that each subsample is paired with an i.i.d. ω. For

$$\text{s-IJ}_{\mathrm{U}}^{\omega} = \frac{k^2}{n^2}\sum[e_i^{\omega} - s_0^{\omega}]^2, \tag{149}$$

it can be decomposed the same way as Eq. (144). Thus, we have $\text{s-IJ}_{\mathrm{U}}^{\omega} \xrightarrow{p} \frac{k^2}{n}\zeta_{1,\omega}$.

Proof of Theorem 13: Let first first ignore the extra randomness-ω for simplicity. By the definition of generalized U-statistic, the incomplete U-statistic can be viewed as a complete U-statistic with a different kernel $s^{\dagger}(X_{i_1},\ldots,X_{i_k}) = \frac{\rho}{p}s(X_{i_1},\ldots,X_{i_k})$, so that

$$\begin{aligned}\mathrm{U}_{n,k,N} &= \frac{1}{N}\sum \rho s(X_{i_1},\ldots,X_{i_k}) \\ &= \binom{n}{k}^{-1}\sum \frac{\rho}{p}s(X_{i_1},\ldots,X_{i_k}) \\ &:= \mathrm{U}_{n,k}^{\dagger}\end{aligned} \tag{150}$$

$\frac{\rho}{p}$ can be viewed as ω. Consider the H-decomposition of $\mathrm{U}_{n,k}^{\dagger}$, we have $V_j^{\dagger} = V_j$ for $j = 1,\ldots,k-1$ and $V_k^{\dagger} = V_k + \frac{1-p}{p}\zeta_k$. Similar to Eq. (144), we have

$$\begin{aligned}\text{s-IJ}_{\mathrm{U}}^{\dagger} &= \frac{k^2}{n^2}\sum_{i=1}^{n}[e_i^{\dagger} - s_0^{\dagger}]^2 \\ &= \frac{k^2}{n^2}\frac{(n-k)^2}{n^2}\sum_{i=1}^{n}\left[s^{(1)}(X_i) + \mathrm{T}_i^{\dagger}\right]^2\end{aligned} \tag{151}$$

where $s^{(1)}(x) = \mathbb{E}[s(x, X_2, \ldots, X_k)]$. Note that

$$\begin{aligned}
\mathbb{E}[(s^{(1)}(X_1))^2] &= V_1^\dagger = V_1 \\
\mathbb{E}[(\mathrm{T}_i^\dagger)^2] &= \frac{1}{n-1} V_1^\dagger + \frac{n}{k^2} \sum_{j=2}^{k} \frac{j}{1-j/n} \frac{\binom{k}{j}^2}{\binom{n}{j}} V_j^\dagger \\
&= \frac{1}{n-1} V_1 + \frac{n}{k^2} \sum_{j=2}^{k} \frac{j}{1-j/n} \frac{\binom{k}{j}^2}{\binom{n}{j}} V_j + \frac{n}{k^2} \frac{k}{1-k/n} \frac{1}{N} (1-p)\zeta_k \\
&:= \mathrm{L} + \mathrm{R} + \mathrm{M},
\end{aligned} \tag{152}$$

where $\mathrm{M} = \frac{1}{1-k/n} \frac{n}{Nk} (1-p)\zeta_k$. Since $\frac{k}{n}(\frac{\zeta_k}{k\zeta_1} - 1) \to 0$, $\mathrm{R/L} \to 0$ by Eq. (146). Next, we have

$$\begin{aligned}
\mathrm{M/L} &= \frac{\frac{1}{1-k/n} \frac{n}{Nk} (1-p)\zeta_k}{\zeta_1} \\
&\leq \frac{1}{1-k/n} \cdot \frac{n}{N} \frac{\zeta_k}{k\zeta_1} \\
&\approx \frac{n}{N} \frac{\zeta_k}{k\zeta_1} \to 0.
\end{aligned} \tag{153}$$

Therefore, $\text{s-IJ}_{\mathrm{U}}^\dagger \xrightarrow{p} \frac{k^2}{n} \frac{(n-k)^2}{n^2} \sum_i [s^{(1)}(X_i)]^2 \xrightarrow{p} \frac{k^2}{n} V_1$. Again, the extra randomness only results in an extended version of H-decomposition. Everything above can be directly applied to $\mathrm{U}_{n,k,N,\omega}$.

D.3 Discussion on extensions

Estimating the variance of U-statistic when $k = O(n)$: In application, we might choose k be a fraction of n such that we obtain a more accurate model in spite of loosing the ability to do statistical inference. Since when $k = c \cdot n$, typically we no longer have the asymptotic normality of of $\mathrm{U}_{n,k,N,\omega}$. Nonetheless, we still want estimate its variance well so that we can apply other looser concentration inequalities like Chybeshev's inequality,

Hoeffding inequality and Berstein's inequality, to provide useful confidence intervals. Recall that

$$\begin{aligned}
\mathrm{Var}(\mathrm{U}_{n,k,N}) &= \sum_{j=1}^{k} \frac{\binom{k}{j}}{\binom{n}{j}} \binom{k}{j} V_k + \frac{1}{N}(1-p)\zeta_k \\
&= \sum_{j=1}^{k} \frac{\binom{k}{j}}{\binom{n}{j}} \binom{k}{j} V_k + \frac{1}{N}(1-p) \sum_{j=1}^{k} \binom{k}{j} V_k \\
&= \sum_{j=1}^{k} \left(\frac{\binom{k}{j}}{\binom{n}{j}} + \frac{1}{N}(1-p) \right) \binom{k}{j} V_k \\
&\approx \sum_{j=1}^{k} \left(\frac{\binom{k}{j}}{\binom{n}{j}} + \frac{1}{N} \right) \binom{k}{j} V_k.
\end{aligned} \tag{154}$$

The higher order terms are not negligible. It seems like that there is no way to estimate V_1, even asymptotically. Therefore it's not possible to estimate the variance of $\mathrm{U}_{n,k,N}$ unless $s(X_1, \ldots, X_k)$ itself is almost linear, i.e. $\zeta_k / k V_1 \to 1$.

Proof of Proposition 5:

$$\begin{aligned}
&(e_{1,2} - e_1 - e_2 + s_0) \\
&= \sum_{w_1^*=1, w_2^*=1, w_1^*+\cdots+w_n^*=k} \frac{(k-2)!}{(n-2)\cdots(n-k)} s^* - \sum_{w_1^*=1, w_1^*+\cdots+w_n^*=k} \frac{(k-1)!}{(n-1)\cdots(n-k)} s^* \\
&\quad - \sum_{w_2^*=1, w_1^*+\cdots+w_n^*=k} \frac{(k-1)!}{(n-2)\cdots(n-k)} s^* + \sum_{w_1^*+\cdots+w_n^*=k} \frac{k!}{n\cdots(n-k)} s^* \\
&= \sum_{w_1^*+\cdots+w_n^*=k} \frac{n(n-1)}{k(k-1)} \binom{n}{k}^{-1} (w_1^*)(w_2^*) s^* - \sum_{w_1^*+\cdots+w_n^*=k} \frac{n}{k} \binom{n}{k}^{-1} (w_1^*) s^* \\
&\quad - \sum_{w_1^*+\cdots+w_n^*=k} \frac{n}{k} \binom{n}{k}^{-1} (w_2^*) s^* + \sum_{w_1^*+\cdots+w_n^*=k} \binom{n}{k}^{-1} s^* \\
&= \sum \binom{n}{k}^{-1} \left(\frac{n(n-1)}{k(k-1)} w_1^* w_2^* - \frac{n}{k} w_1^* - \frac{n}{k} w_2^* + 1 \right) s_0 \\
&= \frac{n(n-1)}{k(k-1)} \sum \binom{n}{k}^{-1} \left(w_1^* w_2^* - \frac{k-1}{n-1} w_1^* - \frac{k-1}{n-1} w_2^* + \frac{k(k-1)}{n(n-1)} \right) s_0.
\end{aligned} \tag{155}$$

Thus,

$$\sum_{i,j} \mathrm{Cov}_*^2(s^*, w_{ij}^*) = \left(\frac{\binom{k}{2}}{\binom{n}{2}} \right)^2 \sum_{i,j} (e_{1,2} - e_1 - e_2 + s_0)^2. \tag{156}$$

Proof of Proposition 6: Recall that

$$\begin{aligned}
\mathrm{U}_{n,k} - \theta &= \binom{n}{k}^{-1} \sum_{(n,k)} s(X_{i_1}, \ldots, X_{i_k}) \\
&= \binom{n}{k}^{-1} \sum_{(n,k)} \left\{ \sum_{j=1}^{k} \sum_{(k,j)} s^{(j)}(X_{i_1}, \ldots, x_{i_j}) \right\} \\
&= \sum_{j=1}^{k} \binom{n}{k}^{-1} \sum_{(n,k)} \sum_{(k,j)} s^{(j)}(X_{i_1}, \ldots, X_{i_j}) \\
&= \sum_{j=1}^{k} \binom{k}{j} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j})
\end{aligned}$$

The second order term is $\binom{k}{2}\binom{n}{2}^{-1} \sum_{i<j} s^{(2)}(X_i, X_j)$. Consider $e_{ij} = \mathbb{E}_*[s^* | X_1^* = X_1, X_2^* = X_2]$, then

$$\begin{aligned}
(e_{1,2} - e_1 - e_2 + s_0) &= \binom{n}{k}^{-1} \sum s(X_{i_1}, \ldots, X_{i_k}; \not\exists 1, \not\exists 2) \\
&+ \left(\binom{n-2}{k-2}^{-1} - 2\binom{n-1}{k-1}^{-1} + \binom{n}{k}^{-1} \right) \sum_{\exists 1, \exists 2} s(X_{i_1}, \ldots, X_{ik}) \\
&- \left(\binom{n-1}{k-1} - \binom{n}{k}^{-1} \right) \sum s(X_{i_1}, \ldots, X_{i_k}; \exists 1, \not\exists 2) \\
&- \left(\binom{n-1}{k-1} - \binom{n}{k}^{-1} \right) \sum s(X_{i_1}, \ldots, X_{i_k}; \not\exists 1 \exists 2) \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{I} &= \binom{n}{k}^{-1} \binom{n-2}{k} \binom{n-2}{k}^{-1} \sum s(X_{i_1}, \ldots, X_{i_k}; \not\exists 1, \not\exists 2) \\
&= \binom{n}{k}^{-1} \binom{n-2}{k} \sum_{j=1}^{k} \binom{k}{j} \binom{n-2}{j}^{-1} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \not\exists 1, \not\exists 2).
\end{aligned} \tag{157}$$

and

$$\begin{aligned}
\mathrm{II} &= -\left(\binom{n-1}{k-1}^{-1} - \binom{n}{k}^{-1}\right) \sum s(X_{i_1}, \ldots, X_{i_k}; \exists 1, \nexists 2) \\
&= -\left(\binom{n-1}{k-1}^{-1} - \binom{n}{k}^{-1}\right) \binom{n-2}{k-1}\binom{n-2}{k-1}^{-1} \sum s(X_{i_1}, \ldots, X_{i_k}; \exists 1, \nexists 2) \\
&= -\left(\binom{n-1}{k-1}^{-1} - \binom{n}{k}^{-1}\right) \binom{n-2}{k-1} \left[\sum_{j=1}^{k-1} \binom{k-1}{j}\binom{n-2}{j}^{-1}\right. \\
&\quad \left. \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \nexists 1, \nexists 2) + \sum_{j=1}^{k} \binom{k-1}{j-1}\binom{n-2}{j-1}^{-1} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \exists 1, \nexists 2)\right].
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathrm{III} &= -\left(\binom{n-1}{k-1}^{-1} - \binom{n}{k}^{-1}\right) \sum s(X_{i_1}, \ldots, X_{i_k}; \nexists 1, \exists 2) \\
&= -\left(\binom{n-1}{k-1}^{-1} - \binom{n}{k}^{-1}\right) \binom{n-2}{k-1}\binom{n-2}{k-1}^{-1} \sum s(X_{i_1}, \ldots, X_{i_k}; \nexists 1, \exists 2) \\
&= -\left(\binom{n-1}{k-1}^{-1} - \binom{n}{k}^{-1}\right) \binom{n-2}{k-1} \left[\sum_{j=1}^{k-1} \binom{k-1}{j}\binom{n-2}{j}^{-1}\right. \\
&\quad \left. \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \nexists 1, \nexists 2) + \sum_{j=1}^{k} \binom{k-1}{j-1}\binom{n-2}{j-1}^{-1} \sum s^{(j)}(X_{i_1}, \ldots, X_{i_j}; \nexists 1, \exists 2)\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{IV} &= \left(\binom{n-2}{k-2}^{-1} - 2\binom{n-1}{k-1}^{-1} + \binom{n}{k}^{-1}\right) \sum s(X_{i_1}, \ldots, X_{i_k}; \exists 1, \exists 2) \\
&= \left(\binom{n-2}{k-2}^{-1} - 2\binom{n-1}{k-1}^{-1} + \binom{n}{k}^{-1}\right) \binom{n-2}{k-2} \left[\sum_{j=1}^{k-2} \binom{k-2}{j}\binom{n-2}{j}^{-1}\right. \\
&\quad \sum s(X_{i_1}, \ldots, X_{i_j}; \nexists 1, \nexists 2) \\
&\quad + \sum_{j=1}^{k-1} \binom{k-2}{j-1}\binom{n-2}{j-1}^{-1} \sum s(X_{ij}, \ldots, X_{ij}; \exists 1, \nexists 2) \\
&\quad + \sum_{j=1}^{k-1} \binom{k-2}{j-1}\binom{n-2}{j-1}^{-1} \sum s(X_{ij}, \ldots, X_{ij}; \nexists 1, \exists 2) \\
&\quad \left. + \sum_{j=2}^{k} \binom{k-2}{j-2}\binom{n-2}{j-2}^{-1} \sum s(X_{i_1}, \ldots, X_{i_j}; \exists 1, \exists 2)\right]
\end{aligned} \tag{158}$$

In conclusion, we have $\mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV} = A+B+C+D$, where A, B, C, D are uncorrelated.

$$\begin{aligned} A = &\left(\binom{n-2}{k-2}^{-1} - 2\binom{n-1}{k-1}^{-1} + \binom{n}{k}^{-1}\right)\binom{n-2}{k-2} \\ &\sum_{j=2}^{k}\binom{k-2}{j-2}\binom{n-2}{j-2}^{-1}\sum s(X_{i_1},\ldots,X_{i_j};\exists 1,\exists 2), \end{aligned} \tag{159}$$

$$\begin{aligned} B = &-\left(\binom{n-1}{k-1} - \binom{n}{k}^{-1}\right)\binom{n-2}{k-1} \\ &\left[\sum_{j=1}^{k}\binom{k-1}{j-1}\binom{n-2}{j-1}^{-1}\sum s(X_{i_1},\ldots,X_{i_j};\exists 1,\not\exists 2)\right] \\ &+\left(\binom{n-2}{k-2}^{-1} - 2\binom{n-1}{k-1}^{-1} + \binom{n}{k}^{-1}\right)\binom{n-2}{k-2} \\ &\left[\sum_{j=1}^{k-1}\binom{k-2}{j-1}\binom{n-2}{j-1}^{-1}\sum s(X_{ij},\ldots,X_{ij};\exists 1,\not\exists 2)\right], \end{aligned} \tag{160}$$

$$\begin{aligned} C = &-\left(\binom{n-1}{k-1} - \binom{n}{k}^{-1}\right)\binom{n-2}{k-1} \\ &\left[\sum_{j=1}^{k}\binom{k-1}{j-1}\binom{n-2}{j-1}^{-1}\sum s(X_{i_1},\ldots,X_{i_j};\not\exists 1,\exists 2)\right] \\ &+\left(\binom{n-2}{k-2}^{-1} - 2\binom{n-1}{k-1}^{-1} + \binom{n}{k}^{-1}\right)\binom{n-2}{k-2} \\ &\sum_{j=1}^{k-1}\binom{k-2}{j-1}\binom{n-2}{j-1}^{-1}\sum s(X_{ij},\ldots,X_{ij};\not\exists 1,\exists 2)\ , \end{aligned} \tag{161}$$

and

$$
\begin{aligned}
D = &\binom{n}{k}^{-1}\binom{n-2}{k}\sum_{j=1}^{k}\binom{k}{j}\binom{n-2}{j}^{-1}\sum s(X_{i_1},\ldots,X_{i_j};\not{2}1,\not{2}2)\\
&-\left(\binom{n-1}{k-1}-\binom{n}{k}^{-1}\right)\binom{n-2}{k-1}\\
&\quad\left[\sum_{j=1}^{k-1}\binom{k-1}{j}\binom{n-2}{j}^{-1}\sum s(X_{i_1},\ldots,X_{i_j};\not{2}1,\not{2}2)\right]\\
&-\left(\binom{n-1}{k-1}-\binom{n}{k}^{-1}\right)\binom{n-2}{k-1}\\
&\quad\left[\sum_{j=1}^{k-1}\binom{k-1}{j}\binom{n-2}{j}^{-1}\sum s(X_{i_1},\ldots,X_{i_j};\not{2}1,\not{2}2)\right]\\
&+\left(\binom{n-2}{k-2}^{-1}-2\binom{n-1}{k-1}^{-1}+\binom{n}{k}^{-1}\right)\binom{n-2}{k-2}\cdot\\
&\quad\left[\sum_{j=1}^{k-2}\binom{k-2}{j}\binom{n-2}{j}^{-1}\sum s(X_{i_1},\ldots,X_{i_j};\not{2}1,\not{2}2)\right].
\end{aligned}
\tag{162}
$$

Let $C_2=\left(\binom{n-2}{k-2}^{-1}-2\binom{n-1}{k-1}^{-1}+\binom{n}{k}^{-1}\right)\binom{n-2}{k-2}=\frac{(n-k)^2+(n-k)}{n(n-1)}$, $C_1=\left(\binom{n-1}{k-1}^{-1}-\binom{n}{k}^{-1}\right)\binom{n-2}{k-1}=\frac{(n-k)^2}{n(n-1)}$ and $C_0=\binom{n}{k}^{-1}\binom{n-2}{k}=\frac{(n-k)(n-k-1)}{n(n-1)}=\frac{(n-k)^2-(n-k)}{n(n-1)}$, then

$$
\begin{aligned}
\mathrm{Var}(A)&=\sum_{j=2}^{k}\binom{n-2}{j-2}^{-1}\left(C_2\binom{k-2}{j-2}\right)^2V_j\\
&=\sum_{j=2}^{k}\binom{n-2}{j-2}^{-1}\left(C_2\binom{k-2}{j-2}\right)^2V_j,
\end{aligned}
\tag{163}
$$

$$
\begin{aligned}
\mathrm{Var}(B)=&\sum_{j=1}^{k-1}\binom{n-2}{j-1}\cdot\left(-C_1\binom{k-1}{j-1}\binom{n-2}{j-1}^{-1}+C_2\binom{k-2}{j-1}\binom{n-2}{j-1}^{-1}\right)^2V_j\\
&+\binom{n-2}{k-1}^{-1}\left(-C_1\binom{k-1}{k-1}\right)^2V_k\\
=&\sum_{j=1}^{k}\binom{n-2}{j-1}^{-1}\left(-C_1\binom{k-1}{j-1}+C_2\binom{k-2}{j-1}\right)^2V_j,
\end{aligned}
\tag{164}
$$

$$
\mathrm{Var}(B)=\mathrm{Var}(C),
\tag{165}
$$

and

$$\begin{aligned}\mathrm{Var}(D) = &\sum_{j=1}^{k-2} \binom{n-2}{j}^{-1} \left(C_0 \binom{k}{j} - 2C_1 \binom{k-1}{j} + C_2 \binom{k-2}{j} \right) + \\ &+ \binom{n-2}{k-1}^{-1} \left(C_0 \binom{k}{k-1} - 2C_1 \binom{k-1}{k-1} \right) \\ &+ \binom{n-2}{k} C_0 \binom{k}{k} V_k \\ = &\sum_{j=1}^{k} \binom{n-2}{j}^{-1} \left(C_0 \binom{k}{j} - 2C_1 \binom{k-1}{j} + C_2 \binom{k-2}{j} \right) V_j.\end{aligned} \tag{166}$$

Therefore, the $\mathbb{E}[\text{s-IJ}_{\text{U}}(2)] = \binom{k}{j}^2 / \binom{n}{2} \sum_{j=1}^{k} \lambda_j(2) V_j$, where

$$\begin{aligned}\lambda_j(2) = &\binom{n-2}{j-2}^{-1} \left(C_2 \binom{k-2}{j-2} \right)^2 \\ &+ 2\binom{n-2}{j-1}^{-1} \left(-C_1 \binom{k-1}{j-1} + C_2 \binom{k-2}{j-1} \right)^2 \\ &+ \binom{n-2}{j}^{-1} \left(C_0 \binom{k}{j} - 2C_1 \binom{k-1}{j} + C_2 \binom{k-2}{j} \right)^2.\end{aligned} \tag{167}$$

Bibliography

[1] Moulinath Banerjee and Ian W. McKeague. Confidence sets for split points in decision trees. *Ann. Statist.*, 35(2):543–574, 04 2007.

[2] Mariana Belgiu and Lucian Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016.

[3] Rudolf Beran. Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697, 1988.

[4] Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

[5] Gérard Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13:1063–1095, April 2012.

[6] Gérard Biau and Luc Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

[7] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

[8] Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, June 2008.

[9] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

[10] Peter J Bickel, Friedrich Götze, and Willem R van Zwet. Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica*, 7:1–31, 1997.

[11] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[12] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[13] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1st edition, 1984.

[14] Herman Callaert, Paul Janssen, et al. The berry-esseen theorem for u-statistics. *The Annals of Statistics*, 6(2):417–421, 1978.

[15] Y-K Chan and John Wierman. On the berry-esseen theorem for u-statistics. *The Annals of Probability*, pages 136–139, 1977.

[16] Jinyuan Chang and Peter Hall. Double-bootstrap methods that use a single double-bootstrap simulation. *Biometrika*, 102(1):203–214, 2015.

[17] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer Science & Business Media, 2010.

[18] Louis HY Chen and Qi-Man Shao. A non-uniform berry–esseen bound via stein's method. *Probability theory and related fields*, 120(2):236–254, 2001.

[19] Louis HY Chen, Qi-Man Shao, et al. Normal approximation under local dependence. *The Annals of Probability*, 32(3):1985–2028, 2004.

[20] Xiaohui Chen and Kengo Kato. Randomized incomplete u-statistics in high dimensions. *arXiv preprint arXiv:1712.00771*, 2017.

[21] Tim Coleman, Lucas Mentch, Daniel Fink, Frank La Sorte, Giles Hooker, Wesley Hochachka, and David Winkler. Statistical inference on tree swallow migrations with random forests. *arXiv preprint arXiv:1710.09793*, 2017.

[22] Yifan Cui, Ruoqing Zhu, Mai Zhou, and Michael Kosorok. Consistency of survival tree and forest models: splitting bias and correction. *arXiv preprint arXiv:1707.09631*, 2019.

[23] Russel Davidson and James MacKinnon. Improving the reliability of bootstrap tests. *Queens University Working paper no. 995*, 2000.

[24] Russell Davidson and James G MacKinnon. Fast double bootstrap tests of nonnested linear regression models. *Econometric Reviews*, 21(4):419–429, 2002.

[25] Russell Davidson and James G MacKinnon. Improving the reliability of bootstrap tests with the fast double bootstrap. *Computational Statistics & Data Analysis*, 51(7):3259–3281, 2007.

[26] Misha Denil, David Matheson, and Nando Freitas. Consistency of online random forests. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1256–1264, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[27] B. Efron and C. Stein. The jackknife estimate of variance. *Ann. Statist.*, 9(3):586–596, 05 1981.

[28] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

[29] Bradley Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981.

[30] Bradley Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. Siam, 1982.

[31] Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

[32] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[33] Carl-Gustaf Esseen. On the liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28:1–19, 1942.

[34] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

[35] Edward W Frees. Infinite order u-statistics. *Scandinavian Journal of Statistics*, pages 29–45, 1989.

[36] Raffaella Giacomini, Dimitris N Politis, and Halbert White. A warp-speed method for conducting monte carlo experiments involving bootstrap estimators. *Econometric theory*, 29(3):567, 2013.

[37] William F Grams, RJ Serfling, et al. Convergence rates for u-statistics and related statistics. *The Annals of Statistics*, 1(1):153–160, 1973.

[38] Paul R Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, pages 34–43, 1946.

[39] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[40] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325, 09 1948.

[41] Wassily Hoeffding. On sequences of sums of independent random vectors. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory*, pages 213–226, Berkeley, Calif., 1961. University of California Press.

[42] Wassily Hoeffding, Herbert Robbins, et al. The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15(3):773–780, 1948.

[43] Peter J Huber et al. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, 43(4):1041–1067, 1972.

[44] Hemant Ishwaran. The effect of splitting on random forests. *Mach. Learn.*, 99(1):75–118, April 2015.

[45] Hemant Ishwaran and Udaya B Kogalur. Consistency of random survival forests. *Statistics & probability letters*, 80(13-14):1056–1064, 2010.

[46] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

[47] Louis A Jaeckel. *The infinitesimal jackknife*. Bell Telephone Laboratories, 1972.

[48] Luckyson Khaidem, Snehanshu Saha, and Sudeepa Roy Dey. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*, 2016.

[49] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

[50] Justin Lee. U-statistics: Theory and practice. 1990.

[51] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

[52] Miles E Lopes. Estimating a sharp convergence bound for randomized ensembles. *Journal of Statistical Planning and Inference*, 204:35–44, 2020.

[53] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

[54] Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.*, 17(1):841–881, January 2016.

[55] Rupert G. Miller. The jackknife–a review. *Biometrika*, 61(1):1–15, 1974.

[56] R v Mises. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348, 1947.

[57] Matthew A Olson and Abraham J Wyner. Making sense of random forest probabilities: a kernel perspective. *arXiv preprint arXiv:1812.05792*, 2018.

[58] Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050, 1994.

[59] Yanjun Qi. Random forest for bioinformatics. In *Ensemble machine learning*, pages 307–323. Springer, 2012.

[60] Tom Rainforth, Robert Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. *arXiv preprint arXiv:1709.06181*, 2017.

[61] Joseph P Romano and Cyrus DiCiccio. Multiple data splitting for testing. Technical report, Technical report, 2019.

[62] Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016.

[63] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741, August 2015.

[64] Srijan Sengupta, Stanislav Volgushev, and Xiaofeng Shao. A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, 111(515):1222–1232, 2016.

[65] Joseph Sexton and Petter Laake. Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*, 53(3):801–811, 2009.

[66] Charles Stein et al. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

[67] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[68] Vito Volterra. *Sopra le funzioni che dipendono da altre funzioni*. Tip. della R. Accademia dei Lincei, 1887.

[69] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 0(0):1–15, 2018.

[70] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[71] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

[72] Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651, 2014.

[73] Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

[74] Zhengze Zhou, Lucas Mentch, and Giles Hooker. Asymptotic normality and variance estimation for supervised ensembles. *arXiv preprint arXiv:1912.01089*, 2019.

[75] Zhengze Zhou, Lucas Mentch, and Giles Hooker. V-statistics and variance estimation. *arXiv preprint arXiv:1912.01089*, 2019.

[76] Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.