Chapter 8: U-statistics and U-processes

Empirical Processes (MATH-522)

Myrto Limnios

May, 2025

So far we have studied collections of statistics that took the form of empirical averages based on i.(i.d.) r.v.s. This Chapter focus on estimators taking more complicated form, known as U-statistics, for which the theory goes back to the fundamental works of Halmos (1946) and Hoeffding (1948). This Chapter focuses on their properties: unbiased, lower variance, some asymptotic guarantees, and importantly their concentration properties.

1 General Definitions and Properties

Let θ be a functional defined on a set $\mathcal{F}$ of real-valued distribution functions, such that

$$\theta : F \in \mathcal{F} \mapsto \theta(F) \ .$$

Based on a i.i.d. sample $X_1, \ldots, X_n$ drawn from an unknown distribution F, the goal is to estimate the function $\theta(F)$.

Theorem 1.1 (Halmos)**.** *Let θ be a parameter of an unknown distribution F. Then, it admits an unbiased estimator for sufficiently large sample size n iff. for some $q \in \mathbb{N}^*$, there exists a function $h : \mathcal{X}^q \to \mathbb{R}$, such that*

$$\theta = \mathbb{E}[h(X_1, \ldots, X_q)] \ .$$

In fact, we can do slightly better than the natural estimator, and consider all possible permutations over the data, that is called U-statistic.

Definition 1.2. A symmetric one-sample U-statistic of kernel h and of order $q \in \mathbb{N}^*$, is defined by

$$U_n(h) = \binom{n}{q}^{-1} \sum_{\sigma \in \mathscr{S}_n} h(X_{\sigma(1)}, \ldots, X_{\sigma(q)}) \ , \tag{1}$$

where $\mathscr{S}_n$ is the set of all permutations of $\{1, \ldots, n\}$, i.e., the sum is over all the subsets $1 \leq i_1 < \ldots < i_q \leq n$ of the indices $\{1, \ldots, n\}$.

Remark 1.1. *Symmetric kernels are defined based on a measurable kernel $\tilde{h}$ by: $h(x_1, \ldots, x_q) = (q!)^{-1} \sum_{\sigma \in \mathscr{S}_q} \tilde{h}(x_{\sigma(1)}, \ldots, x_{\sigma(q)})$. Two fundamental references for U-statistics are Lee (1990), being the first comprehensive monograph on classic probability asymptotic theory and applications to statistical models. Korolyuk and Borovskich (1994), generalize the results by relating/decomposing the U-statistics to/in reverse martingales valued in different types of spaces (Banach and Hilbert spaces).*

Example 1.2. *Basic examples for estimating the parameters of a i.i.d. random sample $X_1, \ldots, X_n$, under some basic moment-based assumptions, and with X, $X' \overset{i.i.d.}{\sim} F$, are as follows:*

1. *mean:* $\theta(F) = \mathbb{E}_{X\sim F}[X]$, *then* $U_n = (1/n)\sum_{i=1}^n X_i$
2. *variance:* $\theta(F) = \mathrm{Var}[X]$, *then* $U_n = \binom{n}{2}^{-1}\sum_{1\le i<j\le n}(X_i - X_j)^2$
3. *covariance:* $\theta(F) = Cov[X, X']$, *then* $U_n = \binom{n}{2}^{-1}(1/2)\sum_{1\le i<j\le n}(X_i - X_j)(X'_i - X'_j)$

We can similarly define two-sample U-statistics as follows.

Definition 1.3. Let two independent and i.i.d. random samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$, drawn from two different p.d.s F, G. A symmetric two-sample U-statistic of kernel h and of order $(q, s) \in \mathbb{N}^* \times \mathbb{N}^*$, is defined by

$$U_{n,m}(h) = \binom{n}{q}^{-1}\binom{m}{s}^{-1}\sum_{\sigma\in\mathscr{S}_n}\sum_{\sigma'\in\mathscr{S}_m} h(X_{\sigma(1)}, \dots, X_{\sigma(q)}, Y_{\sigma'(1)}, \dots, Y_{\sigma'(s)}) \ . \tag{2}$$

Example 1.3. *Mann-Whitney statistic is an important example of two-sample U-statistic of order $(1, 1)$. Suppose we want to test the equality of F, G, if F is stochastically larger that G. We want to estimate $\mathbb{P}(X \ge Y)$. The corresponding U-statistic takes the form*

$$U_{n,m} = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m 1\{X_i \ge Y_j\} \ .$$

It is of kernel $h(x, y) = 1\{X \ge Y\}$, and one can construct a test statistic based on $U_{n,m}$.

Projection methods. It is clear that U-statistic take a complicated form, that differs from what we have studied so far. This paragraph presents two important methods for decomposition any U-statistic into a main term being empirical averages, for which we can use classic theorems and concentration results, while the rest of the terms will be proved to be negligeable for large enough sample size and under some conditions.

The first method is known as Hajèk's projection, and is applicable to any square-integrable function of the data. The idea is to project onto a set, say $\mathcal{C}$, spanned by the data and composed of functions of the form $\sum_{i=1}^n g_i(X_i)$, where the functions $g_i : \mathcal{X} \to \mathbb{R}$ are square-integrable and measurable.

Theorem 1.4 (Hajèk's projection method)**.** *Consider the set $\mathcal{C}$ as defined above, based on a set of independent r.v.s $X_1, \dots, X_n$. Then, for any statistic T defined on the data sample, its projection onto $\mathcal{C}$ is defined by*

$$\widehat{T}(X_1, \dots, X_n) = \sum_{i=1}^n \mathbb{E}[T|X_i] - (n-1)\mathbb{E}[T] \ .$$

Proof. The RHS is clearly an element of $\mathcal{C}$, it remains to prove the orthogonality property ($T - \widehat{T}$ orthogonal to $\mathcal{C}$). □

Remark 1.4. *If the data is i.i.d., and T is measurable and symmetric in its arguments, then $\mathbb{E}[T|X_i = x] = \mathbb{E}[T(x, X_2, \dots, X_n)]$.*

We know formulate a complete decomposition tailored for U-statistics of any order q.

Theorem 1.5 (Hoeffding decomposition)**.** *Consider the U-statistic as defined in Eq. (2). For $j \in \{2, \dots, q\}$, let $h_j(x_1, \dots, x_j) = \mathbb{E}[h(X_1, \dots, X_q) \mid X_1 = x_1, \dots, X_j = x_j]$, such that the kernels are recursively defined:*

$$\begin{aligned} h^{(1)}(x_1) &= h_1(x_1) - \theta \ , \\ h^{(j)}(x_1, \dots, x_j) &= h_j(x_1, \dots, x_j) - \sum_{c=1}^{j-1}\sum_{\sigma\in\mathscr{S}_c} h^{(c)}(x_{\sigma(1)}, \dots, x_{\sigma(c)}) - \theta \ , \quad \forall j \in \{2, \dots, q\} \ . \end{aligned}$$

Then the U-statistic of degree q is decomposed as

$$U_n(h) = \theta + \sum_{j=1}^{q} \binom{q}{j} U_n^{(j)}(h^{(j)}), \tag{3}$$

where the $U_n^{(j)}$ are the U-statistics based on kernel $h^{(j)}$ of degree j.

An important characterization is the order of *degeneracy* (also related to the *rank* of the statistic). Indeed, a U-statistic is said to be *degenerate* or order c w.r.t. a probability measure, if the first c terms of the decomposition equal to zero a.s.. Its variance is of order $n^{-(c+1)}$. In particular, the order of degeneracy controls the limit distribution of the statistic, see e.g. Serfling (1980).

Theorem 1.6 (Theorem 1.6.2, Lee (1990))**.** *Consider the decomposition of Theorem 1.5. Then, for all $j \leq q$ and $c \leq j-1$,*

$$\mathbb{E}[h^{(j)}(X_1, \ldots, X_j) \mid X_1, \ldots, X_c] = 0 \tag{4}$$

and the kernels $\mathbb{E}[h^{(j)}(X_1, \ldots, X_j)] = 0$.

It results that the sequence $U_n^{(j)}$ is of *rank* j, for $j \leq q$.

Asymptotic properties. We now state a version of CLT for U-statistics based on one sample. A similar extension can be obtained for two-sample U-statistics.

We consider the kernel h to be fixed, and define the projection of $U_n - \theta$ onto $\mathcal{C}$ defined by

$$\widehat{U_n} = \sum_{i=1}^{n} \mathbb{E}[U_n - \theta | X_i] = \frac{q}{n} \sum_{i=1}^{n} h_1(X_i) \ ,$$

with $h_1(X_i) = \mathbb{E}[h(x, X_2, \ldots, X_q)] - \theta$.

Theorem 1.7. *If the kernel function h is square-integrable, then $\sqrt{n}(U - \theta - \widehat{U_n}) \xrightarrow{\mathbb{P}} 0$.*

Thus $\sqrt{n}(U-\theta) \rightsquigarrow Z$, with $Z \sim \mathcal{N}(0, q^2\xi)$, where $\xi = Cov(h(X_1, \ldots, X_q), h(X_1', \ldots, X_q'))$, where the $X_1, \ldots, X_n, X_1', \ldots, X_n'$ are i.i.d.

2 Concentration Results

This section states concentration properties of U-statistics around their mean. We first establish for one sample and fixed kernel, then extend to one- and two-sample U-processes.

2.1 Concentration properties for fixed kernel

We start with a simple Lemma achieving slow rate of convergence of order $O_{\mathbb{P}}(q/\sqrt{n})$.

Lemma 2.1. *Consider U_n to be of order $q \in \mathbb{N}^*$ and of kernel h based on an independent sample $X_1, \ldots, X_n$. Suppose that for all $x = (x_1, \ldots, x_q)$, $|h(x)| \leq B < \infty$. Then, for any $t > 0$,*

$$\mathbb{P}(|U_n - \theta| > t) \leq 2e^{-t^2 n/(2B^2q^2)} \ .$$

Proof. Exercise. □

We prove that this rate can be sharper and achieves $O_{\mathbb{P}}(\sqrt{q/n})$.

Lemma 2.2. *Consider U_n to be of order $q \in \mathbb{N}^*$ and of kernel h based on an independent sample $X_1, \ldots, X_n$. Suppose that for all $x = (x_1, \ldots, x_q)$, $|h(x)| \leq B < \infty$. Then, for any $t > 0$,*

$$\mathbb{P}(|U_n - \theta| > t) \leq 2e^{-t^2 n/(2B^2 q)} \ .$$

Proof. Board. □

2.2 Maximal inequalities for VC-classes of kernels

Definition 2.3 (One-sample U-statistics of degree 2)**.** Let $n \geq 2$. Consider a i.i.d. sequence $X_1, \ldots, X_n$ drawn from a probability distribution μ on a measurable space $\mathcal{X}$ and $k : \mathcal{X}^2 \to \mathbb{R}$ a square integrable function w.r.t. $P \otimes P$. The one-sample U-statistic of degree 2 and kernel function k based on the X_i's is defined as:

$$U_n(k) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} k(X_i, X_j) \ . \tag{5}$$

Example 2.1. *(Gaussian chaos of order 2) Consider $X = (X_1, \ldots, X_n)$ to be a centered Gaussian vector of covariance matrix I_n, and let $A = (a_{i,j})_{i,j \leq n}$ be a symmetric real-valued matrix, s.t., $a_{ii} = 0$ for all $i \leq n$. Define the quadratic form*

$$Z = X^T A X = \sum_{i,j \leq n} a_{i,j} X_i X_j \ .$$

Definition 2.4 (Two-sample U-statistics of degree $(1,1)$)**.** Let $n,\ m$ in $\mathbb{N}^*$. Consider two independent i.i.d. sequences $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ respectively drawn from the probability distributions P and Q on the measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a square integrable function w.r.t. $P \otimes Q$. The two-sample U-statistic of degree $(1,1)$, with kernel function ℓ and based on the X_i's and the Y_j's is defined as:

$$U_{n,m}(\ell) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(X_i, Y_j) \ . \tag{6}$$

Example 2.2. *A classic example of two-sample U-statistic of degree $(1,1)$ is the Mann-Whitney statistic, of kernel $\ell(x,y) = \mathbb{I}\{y < x\} + (1/2)\mathbb{I}\{y = x\}$ on $\mathbb{R}^2$. It is a natural (unbiased) estimator of the* AUC*: when computed from univariate samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ with distributions P and Q on $\mathbb{R}$, its Hoeffding decomposition yields*

$$\mathrm{AUC}_{\widehat{\mathrm{P}}_\mathrm{n}, \widehat{\mathrm{Q}}_\mathrm{m}} = \mathbb{P}\{\mathrm{X} \geq \mathrm{Y}\} + \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\widehat{\mathrm{Q}}_\mathrm{m}(\mathrm{X_i}) - \mathbb{E}[\mathrm{Q}(\mathrm{X})]) - \frac{1}{\mathrm{m}} \sum_{\mathrm{j} \leq \mathrm{m}} (\widehat{\mathrm{P}}_\mathrm{n}(\mathrm{Y_j}) - \mathbb{E}[\mathrm{P}(\mathrm{Y})]) + \mathrm{o}_{\mathbb{P}_{\mathrm{P},\mathrm{Q}}}\left(\frac{1}{\mathrm{n}} + \frac{1}{\mathrm{m}}\right) \ . \tag{7}$$

can be thus viewed as an affine transformation of the rank-sum Wilcoxon statistic.

We obtain concentration bounds for U-processes using Hoeffding's decomposition, and treat each of the terms from the obtained decomposition separately as we will see in the next section.

2.3 Concentration inequalities for U-processes

Similar to concentration bounds for empirical processes, this section encompasses concentration bounds for U-processes defined as collections of U-statistics indexed by classes of kernels.

Let $\mathcal{K}$ a class of kernel functions of order (2), U-processes based on a i.i.d. sample $\{X_1, \ldots, X_n\}$ are referred to as the mapping

$$k \in \mathcal{K} \mapsto U_n(k) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} k(X_i, X_j) \tag{8}$$

and similarly to empirical process, the goal is to control the uniform deviations of $\{U_n(k) - \theta(k)\}_{k \in \mathcal{K}}$. The selected results give insight into the control of such random object, depending on the type of class of kernels and on the measurability assumption for the uniform bound. Similarly, we refer to U-processes of degree $(1,1)$, based on the two i.i.d. and independent samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$, indexed by a class of kernels $\mathcal{L}$ the collection $\{U_{n,m}(\ell)\}_{\ell \in \mathcal{L}}$.

We start with a maximal inequality proved by Nolan and Pollard (1987), for degenerate U-processes of degree 2 for general classes of symmetric kernels, later extended to two-sample degenerate U-processes of degree $(1,1)$ by Neumeyer (2004).

Lemma 2.5 (Consequence of Theorem 6, Nolan and Pollard (1987)). *Let $n \geq 2$ and $X_1, \ldots, X_n$ be i.i.d. random variables drawn from a probability distribution P on a measurable space $\mathcal{X}$. Let $\mathcal{K}$ be a class of measurable kernels $k : \mathcal{X}^2 \to \mathbb{R}$ such that $\sup_{x,x' \in \mathcal{X}^2} |k(x,x')| \leq D < +\infty$ and $\int_{\mathcal{X}^2} k^2(x,x') P(dx)\mu(dx') \leq \sigma^2 \leq D^2$, that defines a degenerate one-sample U-process of degree 2, based on the X_i's: $\{U_n(k) \;\; k \in \mathcal{K}\}$. Suppose in addition that the class $\mathcal{K}$ is of* VC*-type with parameters $(A, \mathcal{V})$. There exists a constant $C > 0$, such that:*

$$\mathbb{E}\left[\sup_{k \in \mathcal{K}} |U_n(k)|\right] \leq \frac{2\sigma C}{n-1}\left(\frac{1}{4} + \mathcal{V}\log(A)\right) \;. \tag{9}$$

Lemma 2.6 (Consequence of Lemma 2.4, Neumeyer (2004)). *Let $(n, \; m) \in \mathbb{N}^*$. Consider two independent i.i.d. random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ respectively drawn from the probability distributions P and Q on the measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\mathcal{L}$ be a class of degenerate non-symmetrical kernels $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $\sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} |\ell(x,y)| \leq L < +\infty$ and $\int_{\mathcal{X} \times \mathcal{Y}} \ell^2(x,y) P(dx) Q(dy) \leq \sigma^2 \leq L^2$, that defines a degenerate two-sample U-process of degree $(1,1)$, based on the X_i, Y_j's: $\{U_{n,m}(\ell), \;\; \ell \in \mathcal{L}\}$. Suppose in addition that the class $\mathcal{L}$ is of* VC*-type with parameters $(A, \mathcal{V})$. There exists a constant $C > 0$, such that:*

$$\mathbb{E}\left[\sup_{\ell \in \mathcal{L}} |U_{n,m}(\ell)|\right] \leq \frac{2\sigma C}{\sqrt{nm}}\left(\frac{1}{4} + \mathcal{V}\log(A)\right) \;. \tag{10}$$

Major (2006) proved a concentration bound for one-sample degenerate U-processes of arbitrary degree indexed by L_2-dense classes of non-symmetric kernels. The lemma below formulates this result adapted to the present framework.

Lemma 2.7 (Theorem 2, Major (2006)). *Suppose the conditions of Lemma 2.5 fulfilled. Then, there exist constants $C_1 > 0$, $C_2 \geq 1$ and $C_3 \geq 0$ depending on $(A, \mathcal{V})$ such that:*

$$\mathbb{P}\left\{\sup_{k \in \mathcal{K}} |U_n(k)| \geq t\right\} \leq C_2 \exp\left\{-\frac{C_3(n-1)t}{\sigma}\right\} \;, \tag{11}$$

as soon as $C_1 \log(2D/\sigma) \leq (n-1)t/\sigma \leq n\sigma^2/D^2$.

Consider two independent i.i.d. random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ respectively drawn from the probability distributions μ and ν on the measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\mathcal{L}$ be a class of degenerate non-symmetrical kernels $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for which the following assumptions are considered.

Theorem 2.8 (Lemma 16, Clémençon, Limnios and Vayatis (2021)). *Suppose the conditions of Lemma 2.6 to be fulfilled. Then, for all $t > 0$, there exists a universal constant $K > 2$ such that:*

$$\mathbb{P}\left\{\sup_{\ell \in \mathcal{L}} |U_{n,m}(\ell)| \geq t\right\} \leq K 2^{\mathcal{V}} (A/L)^{2\mathcal{V}} e^{4/L^2} \exp\left\{-\frac{nmt^2}{ML^2}\right\} \;, \tag{12}$$

for all $nmt^2 > \max(8^4 \log(2) L^2 \mathcal{V}, (\log(2) L^2 \mathcal{V}/2)^{1+\delta})$, $\delta \in (1,2)$ *constant and* $M = 16^3/2$.

We omit the proofs as being very technical.

3 Application: The two-sample problem

We exposed the two-sample problem, and introduced an unbiased estimator of the Maximum Mean Discrepancy taking the form of U-statistic. It is considered as state-of-the-art statistical method and was introduced by Gretton et al. (2012).