# Topic 4
# Decoupling Inequalities for (Generalized) U-Statistics

Victor H. de la Peña

Professor of Statistics, Columbia University

Artificial Intelligence Institute for Advances in Optimization
Georgia Institute of Technology 2024

# Topics Preview

1. U-statistics

2. The Generalized U-statistics with Applications

3. Decoupling Inequalities for U-statistics

# Topics Preview

1. U-statistics

2. The Generalized U-statistics with Applications

3. Decoupling Inequalities for U-statistics

# U-Statistics

Let $X_1, ..., X_n$ be a random sample (i.i.d. observations) from an unknown distribution $F$ in $\mathbb{R}$. Given a known function $h : \mathbb{R} \to \mathbb{R}$, consider the estimation of the "parameter"

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

Of course, you may interested in more complex spaces, which the random variables live in or $h$ maps to, but now let us think about the simpler case.

A natural unbiased estimator of $\theta$ you propose is $h(X_1, ..., X_m)$, and since n observations (with $n \geq m$) are available, this simple estimator can be improved. Now you decide to get the average of $h(X_{\alpha_1}, ..., X_{\alpha_m})$, where $(X_{\alpha_1}, ..., X_{\alpha_m}) \in \Pi$, the set of all permutations of $m$ integers such that

$$1 \leq \alpha_i \leq n, \quad \alpha_i \neq \alpha_j \ \ if \ \ i \neq j, \ \ (i, j = 1, ...m).$$

Congratulations! You successfully construct a U-Statistic, which in this context is defined by

$$U_n = U(X_1, ..., X_n) = \frac{1}{n(n-1)...(n-m+1)} \sum_{(X_{\alpha_1}, ..., X_{\alpha_m}) \in \Pi} h(X_{\alpha_1}, ..., X_{\alpha_m}). \tag{1}$$

If $h$ is permutation invariant ( for instance, when $r = 3$:
$h((x_1, x_2, x_3)) = h((x_2, x_1, x_3)) = h((x_3, x_1, x_2)) = h((x_1, x_3, x_1)) =$
$h((x_2, x_3, x_1)) = h((x_3, x_2, x_1))$ ), the definition (1) is equivalent to

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le \alpha_1 < \ldots < \alpha_m \le n} h(X_{\alpha_1}, \ldots, X_{\alpha_m}) \tag{2}$$

Although it may be the first time you hear U-Statistics, you have played with it for a long time. Look at equation (2), then set $h : \mathbb{R}^2 \to \mathbb{R}$ be such that $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$, you can verify that $U_n$ is exactly twice the sample variance, i.e.,

$$s_n^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1} = \frac{\sum_{1\le i<j\le n} \frac{1}{2}(X_i - X_j)^2}{\binom{n}{2}} = \frac{\sum_{1\le i<j\le n}(X_i - X_j)^2}{n(n-1)}.$$

And by doing so, it is convenient to show that $s_n^2$ is an unbiasedness estimator. This is why we call such estimators U-statistics: the letter "U" stands for unbiased.

# Examples

There are several examples of U-statistics. The sample mean is definitely a U-statistic. And when $X_1 \sim X \not\equiv 0$ is nonnegative a.s., the sample Gini mean difference (GMD), defined as

$$d = \frac{1}{n(n-1)} \sum_{i \neq j} |X_i - X_j| = \frac{2}{n(n-1)} \sum_{i<j} |X_i - X_j|,$$

is also a U-statistic. You can find several examples of U-statistics, together with many brilliant limiting theorems, in the giant paper by W. Hoeffding (1948, [4]).

# Topics Preview

# Generalized U-statistics

We now extend this notion of U-statistics. Let $\{X_i\}$ be a sequence of independent random variables in a measurable space $(S, \mathcal{S})$ and $\boldsymbol{f} = \{f_{ij},\ 1 \leq i \neq j \leq n\}$, a family of functions of two variables taking $S \times S$ into $(D, \|\cdot\|)$. Then we define the generalized U-statistic $U_n$ as

$$U_n = \sum_{1 \leq i \neq j \leq n} f_{ij}(X_i, X_j) \tag{3}$$

You can notice that the usual U-statistics can be obtained by letting $f_{ij} = f / \binom{n}{2}$. And such a generalized version may remind you more examples. For instance, the quadratic form $X^T A X = \sum_{1 \leq i \neq j \leq n} a_{ij} X_i X_j$, where the diagonal elements of the symmetric matrix $A$ are set to be zero.

# Random Graph

We can also link the generalized U-Statistic to random colored graph theory. Let $\{X_i\}_{i=1}^n$ a independent sequence of i.i.d. random variables, i.e., $X_i \stackrel{\mathcal{D}}{=} X$ for some random variables $X$. Consider the complete graph $G = (V, E)$, where $|V| = n$ and $X_i$ is the color of the vertex $i$. Now we let $f_{ij} = f$ for some $f$ fixed, and if $f$ is symmetric, then

$$S_n(f) = \sum_{1 \leq i \neq j \leq n} f(X_i, X_j)$$

is a U-statistic (not averaged) representing some color information of vertices.

If we let $X \sim Ber(p)$, where the vertex $X_i = 1$ (resp, 0) indicates that this vertex is black (resp, white), and $f(x_1, x_2) = (1 - x_1)x_2$, which is not symmetric, then

$$S'_n(f) = \sum_{1 \leq i < j \leq n} f(X_i, X_j)$$

counts patterns beginning with a white vertex and ending with a black vertex in this random sequence. And with $f(x_1, x_2) = \mathbb{I}_{\{x_1 \neq x_2\}}$, the statistic

$$S_n\text{''}(f) = \sum_{1 \leq i < j \leq n} f(X_i, X_j)$$

counts the edges with one black and one white end-point.

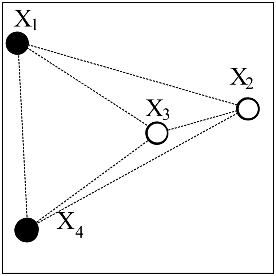



Figure: A example of a random graph, where $S'_n = 2$ and $S_n\text{''} = 4$.

# Topics Preview

You may notice that, although $X_1, ..., X_n$ are mutually independent, the random variables $f_{ij}(X_i, X_j)$'s are dependent, if $i$ or $j$ is fixed. This cause a difficulty in evaluating the expectation of $\|\sum_{1 \le i < j \le n} f(X_i, X_j)\|$ and

$\Phi\left(\|\sum_{1 \le i < j \le n} f(X_i, X_j)\|\right)$ for some $\Phi : \mathbb{R}_0^+ \to \mathbb{R}$ convex increasing.

Let us make the problem more complex, but give a formal statement: Let $X_1, ..., X_n$ be a sequence of independent random variables in a measurable space $(S, \mathcal{S})$ and let $\{f_{ij}\}$ be a family of integrable functions such that $f_{ij} : S \times S \mapsto D$ with $(D, \|\cdot\|)$ a Banach space. Let $\Phi : \mathbb{R}_0^+ \to \mathbb{R}$ be convex such that

$$\max_{1 \le i \ne j \le n} \mathbb{E}\Phi(\|f_{ij}(X_i, X_j)\|) < \infty.$$

Then how can we bound

$$\mathbb{E}\Phi(\| \sum_{1 \le i \ne j \le n} f_{ij}(X_i, X_j)\|)?$$

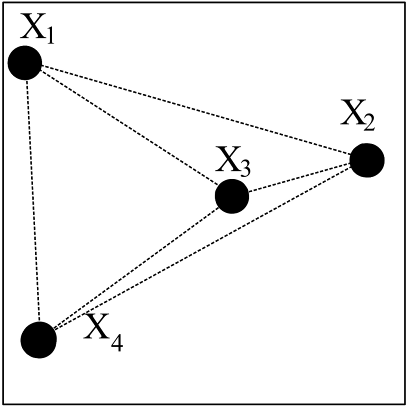
Remember that in the last lecture, I briefly introduced tangent decoupling. Think about the filtration $\mathcal{F}_i = \sigma(X_1, ..., X_i)$, and you can write

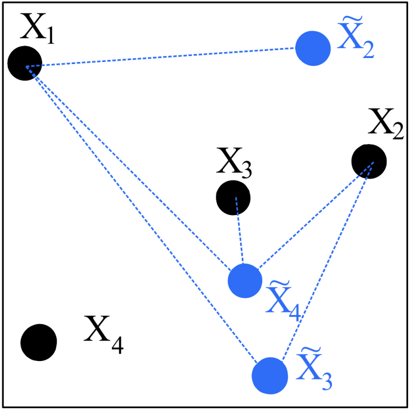
$$U_n = \sum_{1 \leq i \neq j \leq n} f_{ij}(X_i, X_j) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} f_{ij}(X_i, X_j),$$

where $\sum_{i=1}^{j-1} f_{ij}(X_i, X_j)$ is adapted to $\mathcal{F}_j$. Suppose that we have $\{\tilde{X}_i\}_{i=1}^{n}$ an independent copy of $\{X_i\}_{i=1}^{n}$. Then

$$\sum_{j=2}^{n} \sum_{i=1}^{j-1} f_{ij}(X_i, \tilde{X}_j) = \sum_{j=2}^{n} T_j(\tilde{X}_j)$$

is a sum of conditionally independent variables given $\sigma(X_1, ..., X_n)$.


$X_1$
$X_2$
$X_3$
$X_4$


$X_1$
$\tilde{X}_2$
$X_2$
$X_3$
$\tilde{X}_4$
$X_4$
$\tilde{X}_3$

## Theorem (de la Peña, 1992 [1])

*With the aforementioned setting ($\Phi : \mathbb{R}_0^+ \to \mathbb{R}$ convex increasing),*

$$\begin{aligned} M &:= \mathbb{E}\Phi(\| \sum_{1 \le i \ne j \le n} f_{ij}(X_i, X_j) \|) \\ &\le \mathbb{E}\Phi(8 \| \sum_{1 \le i \ne j \le n} f_{ij}(X_i, \tilde{X}_j) \|). \end{aligned} \tag{4}$$

*And if $f_{ij} \in \Pi_{ij}$ satisfy the symmetry conditions*

$$f_{ij} = f_{ji} \text{ and } f_{ij}(X_i, X_j) = f_{ij}(X_j, X_i),$$

*then the reverse bound holds:*

$$\mathbb{E}\Phi(\frac{1}{4} \| \sum_{1 \le i \ne j \le n} f_{ij}(X_i, \tilde{X}_j) \|) \le M. \tag{5}$$

## Remark

*The fact that the lower bound does not hold for general $f_{ij}$ follows trivially by using*

$$f_{ij}(X_i, X_j) = X_j - X_i$$

*because then $\sum_{i \neq j} f_{ij}(X_i, X_j) = 0$. But one may still obtain a lower bound by using the symmetrized kernels $\hat{f}_{ij} = [f_{ij}(X_i, X_j) + f_{ij}(X_j, X_i)]/2$ for $i \neq j$ and letting $\hat{f}_{ji} = \hat{f}_{ij}$*

## Remark

*Considering the situation of quadratic forms, $X^T AX$, where the diagonal entries of $A$ are zero and $A = A^T$, we have inequalities (4) and (5) as follows when $X_i$'s are mean-zero:*

$$\mathbb{E}\Phi(\frac{1}{4}|X^T A\tilde{X}|) \leq \mathbb{E}\Phi(|X^T AX|) \leq \mathbb{E}\Phi(4|X^T A\tilde{X}|).$$

*I will explain the smaller constant 4 soon.*

# Warm-up Lemma

We demonstrate only the first equation (4) here, with a trivial lemma. But we first, for simplicity, denote by $\mathbb{E}_\sigma Y = \mathbb{E}[Y|\sigma]$, where $Y$ is an r.v. and $\sigma$ is a $\sigma$-field.

Let us first see the following warm-up lemma:

**Lemma**

*For $X_1, X_2$ i.i.d., we have $\mathbb{E}(X_1|Z_1) = \frac{X_1+X_2}{2}$, where $Z_1 = (X_1, X_2)$ w.p. $1/2$ and $Z_1 = (X_2, X_1)$ w.p. $1/2$.*

We extend this result to the bi-variate case in the following lemma:

## Lemma

*Let* $\mathscr{Z} = \sigma(Z_i, i = 1, ..., n)$, *where* $\{Z_i\}$ *is a sequence of independent random vectors with* $Z_i = (X_i, \tilde{X}_i)$ *w.p.* $\frac{1}{2}$ *and* $Z_i = (\tilde{X}_i, X_i)$ *w.p.* $\frac{1}{2}$. *Then,*

$$\begin{aligned} \mathbb{E}_{\mathscr{Z}} f_{ij}(X_i, X_j) &= \mathbb{E}_{\mathscr{Z}} f_{ij}(X_i, \tilde{X}_j) \\ &= \frac{1}{4}[f_{ij}(X_i, X_j) + f_{ij}(X_i, \tilde{X}_j) + f_{ij}(\tilde{X}_i, X_j) + f_{ij}(\tilde{X}_i, \tilde{X}_j)] \end{aligned} \tag{6}$$

It is not hard to verify this lemma, by applying the same conditional law of $f_{ij}(X_i, \tilde{X}_j)$ and $f_{ij}(\tilde{X}_i, X_j)$ given $\mathscr{Z}$, and noticing that the sum of those four terms is measurable w.r.t. $\mathscr{Z}$.

Setting $\mathscr{X} = \sigma(X_1, ..., X_n)$, we use the following identity (remember we denote by $\mathbb{E}_\sigma Y = \mathbb{E}[Y|\sigma]$):

$$\begin{aligned}\sum_{1\le i\ne j\le n} f_{ij}(X_i, X_j) = &\sum_{1\le i\ne j\le n} [\mathbb{E}_{\mathscr{X}} f_{ij}(X_i, X_j) + \mathbb{E}_{\mathscr{X}} f_{ij}(X_i, \tilde{X}_j)\\ &+ \mathbb{E}_{\mathscr{X}} f_{ij}(\tilde{X}_i, X_j) + \mathbb{E}_{\mathscr{X}} f_{ij}(\tilde{X}_i, \tilde{X}_j)]\\ &- \sum_{1\le i\ne j\le n} [\mathbb{E}_{\mathscr{X}} f_{ij}(X_i, \tilde{X}_j) + \mathbb{E}_{\mathscr{X}} f_{ij}(\tilde{X}_i, X_j)\\ &+ \mathbb{E}_{\mathscr{X}} f_{ij}(\tilde{X}_i, \tilde{X}_j)].\end{aligned}$$

# A Simpler Version

Recall the Lemma 6, that

$$\mathbb{E}_{\mathscr{Z}} f(X_i, X_j) = \frac{1}{4}[f(X_i, X_j) + f(X_i, \tilde{X}_j) + f(\tilde{X}_i, X_j) + f(\tilde{X}_i, \tilde{X}_j)].$$

We **assume** that $\mathbb{E}_{\mathscr{X}} f(X_i, \tilde{X}_j) = \mathbb{E}_{\mathscr{X}} f(\tilde{X}_i, X_j) = \mathbb{E}_{\mathscr{X}} f(\tilde{X}_i, \tilde{X}_j) = 0$ (e.g., $f(x_1, x_2) = ax_1x_2$ for some constant $a$).
For the U-statistic $\sum_{1 \le i \ne j \le n} f(X_i, X_j)$ with symmetric kernel $f$, we have

$$\begin{aligned}
\mathbb{E}\Phi(|\sum f(X_i, X_j)|) &= \mathbb{E}\Phi(|\sum f(X_i, X_j) + \mathbb{E}_{\mathscr{X}}[f(X_i, \tilde{X}_j) + f(\tilde{X}_i, X_j) + f(\tilde{X}_i, \tilde{X}_j)]|) \\
&\le \mathbb{E}\Phi(|\sum \left(f(X_i, X_j) + f(X_i, \tilde{X}_j) + f(\tilde{X}_i, X_j) + f(\tilde{X}_i, \tilde{X}_j)\right)|) \\
&= \mathbb{E}\Phi(|\sum 4\mathbb{E}_{\mathscr{Z}} f(X_i, \tilde{X}_j)|) \\
&\le \mathbb{E}\Phi(4|\sum f(X_i, \tilde{X}_j)|).
\end{aligned}$$

# Proof

From the preceding and the triangle inequality,

$$
\begin{aligned}
&\mathbb{E}\Phi(\|\sum_{1\le i\ne j\le n} f_{ij}(X_i, X_j)\|) \\
&\le \mathbb{E}\Phi(\|\sum_{1\le i\ne j\le n} \mathbb{E}_{\mathscr{X}}[f_{ij}(X_i, X_j) + f_{ij}(X_i, \tilde{X}_j) + f_{ij}(\tilde{X}_i, X_j) + f_{ij}(\tilde{X}_i, \tilde{X}_j)]\| \\
&+ \|\mathbb{E}_{\mathscr{X}}[f_{ij}(X_i, \tilde{X}_j) + f_{ij}(\tilde{X}_i, X_j) + f_{ij}(\tilde{X}_i, \tilde{X}_j)]\|) \\
&\le \frac{1}{2}\mathbb{E}\Phi(2\|\sum_{1\le i\ne j\le n} \mathbb{E}_{\mathscr{X}}[f_{ij}(X_i, X_j) + f_{ij}(X_i, \tilde{X}_j) + f_{ij}(\tilde{X}_i, X_j) + f_{ij}(\tilde{X}_i, \tilde{X}_j)]\|) \\
&+ \frac{1}{2}\mathbb{E}\Phi(2\|\sum_{1\le i\ne j\le n} \mathbb{E}_{\mathscr{X}}[f_{ij}(X_i, \tilde{X}_j) + f_{ij}(\tilde{X}_i, X_j) + f_{ij}(\tilde{X}_i, \tilde{X}_j)]\|) \\
&\qquad\qquad \text{[by the convexity of } \Phi\text{]}
\end{aligned}
$$

# Proof

$$\leq \frac{1}{2}\mathbb{E}\Phi(2\|\sum_{1\leq i\neq j\leq n}[f_{ij}(X_i,X_j)+f_{ij}(X_i,\tilde{X}_j)+f_{ij}(\tilde{X}_i,X_j)+f_{ij}(\tilde{X}_i,\tilde{X}_j)]\|)$$

$$+\frac{1}{2}\mathbb{E}\Phi(2\|\sum_{1\leq i\neq j\leq n}\mathbb{E}_{\mathscr{X}}[f_{ij}(X_i,\tilde{X}_j)+f_{ij}(\tilde{X}_i,X_j)+f_{ij}(\tilde{X}_i,\tilde{X}_j)\|)$$ [conditional JENSEN inequality]

$$\leq \frac{1}{2}\mathbb{E}\Phi(8\|\sum_{1\leq i\neq j\leq n}\mathbb{E}_{\mathscr{X}}f_{ij}(X_i,\tilde{X}_j)\|)+\frac{1}{6}[\mathbb{E}\Phi(6\|\sum_{1\leq i\neq j\leq n}\mathbb{E}_{\mathscr{X}}f_{ij}(X_i,\tilde{X}_j)\|)$$

$$+\mathbb{E}\Phi(6\|\sum_{1\leq i\neq j\leq n}\mathbb{E}_{\mathscr{X}}f_{ij}(\tilde{X}_i,X_j)\|)+\mathbb{E}\Phi(6\|\sum_{1\leq i\neq j\leq n}\mathbb{E}f_{ij}(\tilde{X}_i,\tilde{X}_j)\|)]$$ [by (6) and $\Phi$ convex]

$$\leq \frac{1}{2}\mathbb{E}\Phi(8\|\sum_{1\leq i\neq j\leq n}\mathbb{E}_{\mathscr{X}}f_{ij}(X_i,\tilde{X}_j)\|)+\frac{2}{6}\mathbb{E}\Phi(6\|\sum_{1\leq i\neq j\leq n}\mathbb{E}f_{ij}(X_i,\tilde{X}_j)\|)$$

$$+\frac{1}{6}\mathbb{E}\Phi(6\|\sum_{1\leq i\neq j\leq n}\mathbb{E}f_{ij}(\tilde{X}_i,\tilde{X}_j)\|)$$ [by conditional JENSEN and that $\mathbb{E}f_{ij}(\tilde{X}_i,\tilde{X}_j)=\mathbb{E}f_{ij}(X_i,\tilde{X}_j)$]

$$\leq \frac{1}{2}\mathbb{E}\Phi(8\|\sum_{1\leq i\neq j\leq n}\mathbb{E}_{\mathscr{X}}f_{ij}(X_i,\tilde{X}_j)\|)+\frac{1}{2}\mathbb{E}\Phi(6\|\sum_{1\leq i\neq j\leq n}\mathbb{E}f_{ij}(X_i,\tilde{X}_j)\|)$$ [by JENSEN inequality]

$$\leq \mathbb{E}\Phi(8\|\sum_{1\leq i\neq j\leq n}f_{ij}(X_i,\tilde{X}_j)\|)$$ [by $\Phi$ increasing]

# Decoupling Inequalities for General Kernels $h$

While this lecture primarily emphasizes the (generalized) U-statistics cases with kernels of the form $h : S^2 \to D$, it's worth noting that we have also established decoupling inequalities for the more general $h : S^k \to D$ Consequently, we arrive at a frequently employed concentration inequality (See de la Peña and Montgomery-Smith [2] in Bulletin of the American Mathematical Society, or [3]).

Let $X_1, ..., X_n$ be a sequence of independent random variables on a measurable space $(S, \mathcal{S})$ and let $\{X_i^{(j)}\}$, $j = 1, ..., k$ be k independent copies of $\{X_i\}$. Let $f_{i_1,...,i_k}$ be family of functions of $k$ variables taking $S^k$ into a Banach space $(D, \|\cdot\|)$. Assume that each $f_{i_1,...,i_k}$ is permutation invariant. Then for all $n \geq k \geq 2$, $t > 0$, there exist numerical constants $C_k, \tilde{C}_k$ depending on $k$ only such that (with $P_{n,k}$ the set of all permutations $(i_1, ..., i_k) \in \{1, ..., n\}^k$)

$$\mathbb{P}\left(\|\sum_{(i_1,...,i_k)\in P_{n,k}} f_{i_1,...,i_k}(X_{i_1}, ..., X_{i_k})\| \geq t\right)$$
$$\leq C_k \mathbb{P}\left(C_k \|\sum_{(i_1,i_2,...,i_k)\in P_{n,k}} f_{i_1,...,i_k}(X_{i_1}^{(1)}, ..., X_{i_k}^{(k)})\| \geq t\right).$$

In addition,

$$\tilde{C}_k \mathbb{P}\left(\tilde{C}_k \| \sum_{(i_1,i_2,\dots,i_k)\in P_{n,k}} f_{i_1,\dots,i_k}(X_{i_1},\dots,X_{i_k})\| \geq t\right)$$
$$\geq \mathbb{P}\left(\| \sum_{(i_1,i_2,\dots,i_k)\in P_{n,k}} f_{i_1,\dots,i_k}(X_{i_1}^{(1)},\dots,X_{i_k}^{(k)})\| \geq t\right).$$

# References I

[1] V. H. de la Peña. "Decoupling and Khintchine's inequalities for U-statistics". In: *Ann. Proba.* (1992), pp. 1877–1892.

[2] V. H. de la Peña and S.J. Montgomery-Smith. "Bounds on the tail probability of U-statistics and quadratic forms". In: *AMERICAN MATHEMATICAL SOCIETY* 31.2 (1994), pp. 223–227.

[3] Victor H de la Peña and Stephen J Montgomery-Smith. "Decoupling inequalities for the tail probabilities of multivariate U-statistics". In: *The Annals of Probability* (1995), pp. 806–816.

[4] Wassily Hoeffding. "A class of statistics with asymptotically normal distribution". In: *Breakthroughs in Statistics: Foundations and Basic Theory* (1992), pp. 308–334.