A Generalized Central Limit Theorem for Two–Sample U–Statistics Under Dependence

1 Introduction

This document sketches (at a high level) a proof of the asymptotic normality of two–sample U–statistics, even though the summands exhibit dependence across indices. The standard i.i.d. Central Limit Theorem (CLT) does not apply directly because each X_i is re-used against all Y_j (and vice versa), causing correlations. Nonetheless, there is a well-known *generalized* CLT (often attributed to Hoeffding [Hoe48], detailed in [Ser80], [Vaa98], and others) that shows such U–statistics still converge to a normal distribution when properly normalized.

2 Two–Sample U–Statistics

Let $\{X_i\}_{i=1}^m$ be i.i.d. random variables from a distribution F, and let $\{Y_j\}_{j=1}^n$ be i.i.d. from a distribution G, independent of $\{X_i\}$. We define a (measurable) kernel

$$\phi(x, y) : \mathcal{X} \times \mathcal{Y} \;\to\; \mathbb{R}$$

with finite second moments. The associated two–sample U–statistic is

$$U_{m,n} \;=\; \frac{1}{m\,n} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi\big(X_i, Y_j\big).$$

We want to prove that for large m, n, this statistic is asymptotically normal after an appropriate centering and scaling.

Notation and Limits

Define $N = m + n$ as the total sample size. Often one assumes $m, n \to \infty$ with

$$\frac{m}{m+n} \;\to\; \alpha \in (0,1), \quad \text{so } \frac{n}{m+n} \;\to\; 1-\alpha.$$

We will show that

$$\sqrt{m+n}\,\Big(U_{m,n} - \mathbb{E}[U_{m,n}]\Big) \;\xrightarrow{d}\; \mathcal{N}\big(0, \sigma^2\big)$$

for some $\sigma^2 > 0$ that depends on F, G, ϕ and on the asymptotic ratio α.

3 Hoeffding–Type Decomposition

A classical approach is to write a *Hoeffding decomposition* of the kernel $\phi(x, y)$ in terms of its "one–dimensional projections." That is, we write

$$\phi(x, y) \;=\; \theta \;+\; f(x) \;+\; g(y) \;+\; h(x, y),$$

where:

$$\theta \;=\; \mathbb{E}[\phi(X, Y)], \quad f(x) \;=\; \mathbb{E}[\phi(x, Y)] - \theta, \quad g(y) \;=\; \mathbb{E}[\phi(X, y)] - \theta,$$

and

$$h(x, y) \;=\; \phi(x, y) - \theta - f(x) - g(y).$$

By construction, $h(x, y)$ satisfies

$$\mathbb{E}[h(X, Y) \,\big|\, X = x] \;=\; 0, \quad \mathbb{E}[h(X, Y) \,\big|\, Y = y] \;=\; 0.$$

Hence h has mean 0 (both marginally in x and y).

U–Statistic in Decomposed Form

Then

$$U_{m,n} \;=\; \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j) \;=\; \theta \;+\; \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \Big[f(X_i) + g(Y_j) + h(X_i, Y_j) \Big].$$

We separate this into three sums:

$$U_{m,n} - \theta \;=\; \frac{1}{mn} \sum_{i,j} f(X_i) \;+\; \frac{1}{mn} \sum_{i,j} g(Y_j) \;+\; \frac{1}{mn} \sum_{i,j} h(X_i, Y_j).$$

Notice that

$$\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(X_i) \;=\; \frac{1}{m} \sum_{i=1}^{m} f(X_i) \quad \text{and} \quad \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} g(Y_j) \;=\; \frac{1}{n} \sum_{j=1}^{n} g(Y_j).$$

Hence

$$U_{m,n} - \theta \;=\; \frac{1}{m} \sum_{i=1}^{m} f(X_i) \;+\; \frac{1}{n} \sum_{j=1}^{n} g(Y_j) \;+\; \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} h(X_i, Y_j).$$

Define

$$A_m \;=\; \sum_{i=1}^{m} \big[f(X_i) \big], \quad B_n \;=\; \sum_{j=1}^{n} \big[g(Y_j) \big], \quad C_{m,n} \;=\; \sum_{i=1}^{m} \sum_{j=1}^{n} h(X_i, Y_j).$$

So

$$U_{m,n} - \theta \;=\; \frac{A_m}{m} \;+\; \frac{B_n}{n} \;+\; \frac{C_{m,n}}{mn}.$$

Centering

We also have $\mathbb{E}[A_m] = m\,\mathbb{E}[f(X)] = 0$ by construction of f. Similarly $\mathbb{E}[B_n] = 0$, and $\mathbb{E}[C_{m,n}] = 0$. Thus

$$\mathbb{E}[U_{m,n}] \;=\; \theta.$$

4 Proof of Asymptotic Normality: Outline

We want to show

$$\sqrt{m+n}\Big(U_{m,n}-\theta\Big) \;=\; \sqrt{m+n}\Big(\frac{A_m}{m}+\frac{B_n}{n}+\frac{C_{m,n}}{mn}\Big) \;\xrightarrow{d}\; \mathcal{N}(0,\sigma^2).$$

Below is a sketch of how each term behaves and how one applies classical CLT arguments:

Term 1: $\sqrt{m+n}\,\frac{A_m}{m}$

Recall $A_m=\sum_{i=1}^m f(X_i)$ where $\{X_i\}$ are i.i.d. from F. Then

$$\mathbb{E}[f(X_i)]=0, \quad \mathrm{Var}[f(X_i)] =: \sigma_f^2<\infty.$$

A classical CLT for i.i.d. sequences shows

$$\frac{A_m}{\sqrt{m}} \;=\; \frac{1}{\sqrt{m}}\sum_{i=1}^m f(X_i) \;\xrightarrow{d}\; \mathcal{N}(0,\,\sigma_f^2).$$

But we are multiplying by $\sqrt{m+n}$ and dividing by m. Observe that

$$\sqrt{m+n}\,\frac{A_m}{m} \;=\; \frac{\sqrt{m+n}}{m}\sum_{i=1}^m f(X_i) \;=\; \sqrt{\frac{m+n}{m}}\,\frac{A_m}{\sqrt{m}}.$$

If $m/(m+n)\to\alpha\in(0,1)$, then $\frac{m+n}{m}\to\frac{1}{\alpha}$. Thus

$$\sqrt{\frac{m+n}{m}} \;\to\; \frac{1}{\sqrt{\alpha}}.$$

Hence

$$\sqrt{m+n}\,\frac{A_m}{n} \;=\; \left(\frac{1}{\sqrt{\alpha}}+o(1)\right)\frac{A_m}{\sqrt{m}}.$$

So in the limit, that term converges in distribution to $\mathcal{N}\big(0,\,\sigma_f^2/(1-\alpha)\big)$ (by Slutsky's theorem).

Term 2: $\sqrt{m+n}\,\frac{B_n}{m}$

An exactly analogous argument applies to $B_n=\sum_{j=1}^n g(Y_j)$ with $\mathrm{Var}[g(Y_j)] =: \sigma_g^2$, so

$$\sqrt{m+n}\,\frac{B_n}{n} \;=\; \sqrt{\frac{m+n}{n}}\,\frac{B_n}{\sqrt{n}} \;\xrightarrow{d}\; \mathcal{N}\Big(0,\,\frac{\sigma_g^2}{1-\alpha}\Big).$$

Term 3: $\sqrt{m+n}\,\frac{C_{m,n}}{mn}$

Recall

$$C_{m,n} \;=\; \sum_{i=1}^m\sum_{j=1}^n h(X_i,Y_j), \quad \mathbb{E}[C_{m,n}]=0,$$

where $h(x,y)$ has $\mathbb{E}[h(X,Y)\mid X]=0$ and $\mathbb{E}[h(X,Y)\mid Y]=0$. Intuitively, $h(X_i,Y_j)$ is a zero-mean "interaction part" that is uncorrelated across different pairs $(i,j)\neq(i',j')$, except for

sharing X_i or Y_j. The literature on two–sample U–statistics (see [Ser80], [Vaa98, Chapter 12]) shows that

$$\mathrm{Var}\Big(\sum_{i,j} h(X_i, Y_j)\Big) \;=\; O\big(m\,n\big).$$

More precisely, the partial overlaps do not inflate the order beyond mn (they do not produce an $(mn)^2$ term!).

Thus heuristically

$$\mathrm{Var}\Big(\frac{C_{m,n}}{mn}\Big) \;=\; \frac{1}{(mn)^2}\,\mathrm{Var}\Big(\sum_{i,j} h(X_i, Y_j)\Big) \;=\; O\Big(\frac{m\,n}{(mn)^2}\Big) \;=\; O\Big(\frac{1}{mn}\Big).$$

Hence

$$\sqrt{m+n}\,\frac{C_{m,n}}{mn} \;=\; O_p\Big(\sqrt{m+n}\,\frac{1}{\sqrt{mn}}\Big) \;=\; O_p\Big(\sqrt{\frac{m+n}{mn}}\Big).$$

Since $mn/(m+n) \to \alpha(1-\alpha)\,N$, we see that

$$\sqrt{\frac{m+n}{mn}} \;=\; \sqrt{\frac{1}{m}+\frac{1}{n}} \;\to\; 0 \quad (\text{if } m, n \to \infty \text{ proportionally}).$$

Thus $\sqrt{m+n}\,\frac{C_{m,n}}{mn} \xrightarrow{p} 0$, meaning it is negligible in the $\sqrt{m+n}$ scaling.

Combining the Three Terms

Summarizing:

$$\sqrt{m+n}\,\Big(U_{m,n} - \theta\Big) \;=\; \underbrace{\sqrt{m+n}\,\frac{A_m}{m}}_{\text{CLT} \to \mathcal{N}(0,\,\sigma_f^2/\alpha)} \;+\; \underbrace{\sqrt{m+n}\,\frac{B_n}{n}}_{\text{CLT} \to \mathcal{N}(0,\,\sigma_g^2/(1-\alpha))} \;+\; \underbrace{\sqrt{m+n}\,\frac{C_{m,n}}{mn}}_{\xrightarrow{p} 0}.$$

By Slutsky's theorem (and properties of sums of asymptotically normal terms), we get an overall normal limit whose variance is $\sigma_f^2/\alpha + \sigma_g^2/(1-\alpha)$. Here,

$$\sigma_f^2 \;=\; \mathrm{Var}\big(f(X)\big), \quad \sigma_g^2 \;=\; \mathrm{Var}\big(g(Y)\big),$$

and $f(x) = \mathbb{E}[\phi(x, Y)] - \theta,\; g(y) = \mathbb{E}[\phi(X, y)] - \theta.$

That yields the main theorem:

$$\sqrt{m+n}\,\big(U_{m,n} - \theta\big) \;\xrightarrow{d}\; \mathcal{N}\Big(0,\; \sigma^2\Big) \quad \text{where} \quad \sigma^2 \;=\; \frac{\mathrm{Var}[f(X)]}{\alpha} \;+\; \frac{\mathrm{Var}[g(Y)]}{1-\alpha}.$$

(There are alternative expressions involving ξ_{10}, ξ_{01} from the "four-case" breakdown, etc. They turn out to coincide with these σ_f^2, σ_g^2 under appropriate definitions; see [Ser80; Vaa98] for details.)

Remark 1 (Multiple or vector-valued kernels)**.** *The above argument generalizes to the case $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, so that $(U_{m,n} - \theta)$ is in $\mathbb{R}^d$. Then f and g also take values in $\mathbb{R}^d$. The same decomposition and variance considerations apply (just use the multivariate CLT).*

5 Variance estimators

5.1 Projection-Based (Hoeffding) Variance Estimator

The variance

In the case of an uni-dimensional kernel, it is straightforward to estimate the variance σ^2 of the limiting normal distribution by replacing θ, $\mathbb{E}[\phi(x, Y)]$, and $\mathbb{E}[\phi(X, y)]$ by their empirical version.

- **Estimate** θ

  A consistent estimate is the U–statistic itself:

$$\hat{\theta} \;=\; U_{m,n} \;=\; \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi(X_i, Y_j).$$

- **Estimate** $f(x)$

  For each observed X_i, the true $f(X_i)$ is $\mathbb{E}[\phi(X_i, Y)] - \theta$. We approximate $\mathbb{E}[\phi(X_i, Y)]$ by averaging over the Y_j's:

$$\hat{f}_i \;=\; \frac{1}{n}\sum_{j=1}^{n}\phi(X_i, Y_j) \;-\; \hat{\theta}.$$

- **Estimate** $g(y)$

  By symmetry, for each Y_j,

$$\hat{g}_j \;=\; \frac{1}{m}\sum_{i=1}^{m}\phi(X_i, Y_j) \;-\; \hat{\theta}.$$

Sample-based estimates of $\mathrm{Var}[f(X)]$ and $\mathrm{Var}[g(Y)]$ are given by

$$\widehat{\mathrm{Var}}\big[f(X)\big] \;=\; \frac{1}{m}\sum_{i=1}^{m}\big(\hat{f}_i\big)^2 \qquad \text{and} \qquad \widehat{\mathrm{Var}}\big[g(Y)\big] \;=\; \frac{1}{n}\sum_{j=1}^{n}\big(\hat{g}_j\big)^2.$$

(One might prefer an $m-1$ or $n-1$ in the denominators for unbiasedness, but for large m, n it does not matter.)

Putting it all together, if

$$\alpha_{m,n} \;=\; \frac{m}{m+n} \to \alpha \in (0,1),$$

a consistent estimator for the asymptotic variance σ^2 is

$$\widehat{\sigma}^2 \;=\; \frac{\widehat{\mathrm{Var}}[f(X)]}{\alpha_{m,n}} \;+\; \frac{\widehat{\mathrm{Var}}[g(Y)]}{1-\alpha_{m,n}}.$$

5.2 Direct Covariance Decomposition Estimator

An alternative way to arrive at a variance estimator is to "unpack" the variance of the U-statistic by writing

$$U_{m,n} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi(X_i, Y_j)$$

and then writing

$$\mathrm{Var}(U_{m,n}) = \frac{1}{m^2 n^2} \sum_{i,j} \sum_{i',j'} \mathrm{Cov}\Big(\phi(X_i, Y_j), \phi(X_{i'}, Y_{j'})\Big).$$

Because the two samples are independent, we have four distinct cases for the indices:

- **Case 1:** $i = i'$ **and** $j = j'$

  There are mn terms. Each contributes

  $$\mathrm{Var}\big(\phi(X_i, Y_j)\big).$$

- **Case 2:** $i = i'$ **and** $j \neq j'$

  There are $m\,n(n-1)$ terms. For a fixed i, using the law of total expectation and by conditioning on X_i,

  $$\begin{aligned}\mathrm{Cov}\Big(\phi(X_i, Y_j), \phi(X_i, Y_{j'})\Big) &= \mathbb{E}\Big[\phi(X_i, Y_j)\phi(X_i, Y_{j'})\Big] - \theta^2 \\ &= \mathbb{E}\Big[\big(\mathbb{E}[\phi(X_i, Y_j) \mid X_i]\big)^2\Big] - \theta^2 \\ &= \sigma_1^2,\end{aligned}$$

  where $\sigma_1^2 = \mathrm{Var}\big(\mathbb{E}[\phi(X, Y) \mid X]\big)$.

- **Case 3:** $i \neq i'$ **and** $j = j'$

  Similarly, there are $n\,m(m-1)$ terms, each with covariance

  $$\mathrm{Cov}\Big(\phi(X_i, Y_j), \phi(X_{i'}, Y_j)\Big) = \sigma_2^2,$$

  where $\sigma_2^2 = \mathrm{Var}\big(\mathbb{E}[\phi(X, Y) \mid Y]\big)$.

- **Case 4:** $i \neq i'$ **and** $j \neq j'$

  In this case the pairs (X_i, Y_j) and $(X_{i'}, Y_{j'})$ are independent so the covariance is zero.

Thus, in population terms we have

$$\mathrm{Var}(U_{m,n}) = \frac{1}{m^2 n^2}\Big\{\, m\,n\,\mathrm{Var}\big(\phi(X, Y)\big) + m\,n(n-1)\,\sigma_1^2 + n\,m(m-1)\,\sigma_2^2\Big\}.$$

A natural plug-in estimator replaces the population quantities by their sample analogues. Define the overall U–statistic mean

$$U_{m,n} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j),$$

and then form the following estimates:

- For the variance term (Case 1):

  $$\widehat{V} = \frac{1}{mn - 1} \sum_{i=1}^{m} \sum_{j=1}^{n} \Big[\phi(X_i, Y_j) - U_{m,n}\Big]^2.$$

- For the covariance among terms sharing the same X_i (Case 2):

  $$\widehat{C}_1 = \frac{1}{m\,n(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{\substack{j'=1 \\ j' \neq j}}^{n} \Big[\phi(X_i, Y_j) - U_{m,n}\Big]\Big[\phi(X_i, Y_{j'}) - U_{m,n}\Big].$$

- For the covariance among terms sharing the same Y_j (Case 3):

$$\widehat{C}_2 = \frac{1}{n\,m(m-1)} \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{\substack{i'=1 \\ i' \neq i}}^{m} \Big[\phi(X_i, Y_j) - U_{m,n}\Big] \Big[\phi(X_{i'}, Y_j) - U_{m,n}\Big].$$

Then a direct plug-in estimator for the variance of $U_{m,n}$ is given by

$$\widehat{\mathrm{Var}}(U_{m,n}) \;=\; \frac{1}{m^2 n^2} \Big\{ mn\,\widehat{V} \;+\; m\,n(n-1)\,\widehat{C}_1 \;+\; n\,m(m-1)\,\widehat{C}_2 \Big\}.$$

Equivalently, after canceling a factor of mn in the numerator,

$$\widehat{\mathrm{Var}}(U_{m,n}) \;=\; \frac{\widehat{V}}{mn} \;+\; \frac{n-1}{mn}\,\widehat{C}_1 \;+\; \frac{m-1}{mn}\,\widehat{C}_2.$$

This estimator directly reflects the contribution to the variance from the four covariance configurations in the double sum, and under standard conditions it provides a consistent estimate of $\mathrm{Var}(U_{m,n})$.

Matrix Representation

For compactness and ease of computation, one can arrange the centered kernels into an $m \times n$ matrix

$$C = \Big[\phi(X_i, Y_j) - U_{m,n}\Big]_{i=1,\ldots,m;\; j=1,\ldots,n}.$$

Then, by introducing the $m \times m$ and $n \times n$ matrices of ones, J_m and J_n, and using some matrix algebra, one can show that the variance estimator is equivalent to

$$\widehat{\mathrm{Var}}(U_{m,n}) = \frac{1}{m^2 n^2}\,\mathrm{trace}\Big(C\big[J_m C + C J_n - C\big]^\top\Big),$$

where I_m (and I_n) is the identity matrix of the appropriate size.

References

[Hoe48] Wassily Hoeffding. "A Class of Statistics with Asymptotically Normal Distribution". In: *The Annals of Mathematical Statistics* 19.3 (Sept. 1948), pp. 293–325. DOI: 10.1214/aoms/1177730196.

[Ser80] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics. Hoboken: John Wiley & Sons, 1980. DOI: 10.1002/9780470316481.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. DOI: 10.1017/CBO9780511802256.

A Matrix Representation of the Variance

We start by writing

$$c_{ij} \;=\; \phi(X_i, Y_j) - U_{m,n},$$

and let

$$C = \big[c_{ij}\big]_{i=1,\ldots,m;\, j=1,\ldots,n}.$$

Also, define the row- and column–sums as

$$S_{i\cdot} \;=\; \sum_{j=1}^{n} c_{ij} \quad \text{and} \quad S_{\cdot j} \;=\; \sum_{i=1}^{m} c_{ij}.$$

Expressing the Trace

Notice that

$$\operatorname{trace}\Big(C\big[J_m C + C J_n - C\big]^\top\Big) = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\Big[(J_m C)_{ij} + (C J_n)_{ij} - c_{ij}\Big].$$

Because J_m is the $m \times m$ matrix of ones and J_n is the $n \times n$ matrix of ones, we have

$$(J_m C)_{ij} = \sum_{k=1}^{m} c_{kj} \;=\; S_{\cdot j}$$

and

$$(C J_n)_{ij} = \sum_{\ell=1}^{n} c_{i\ell} \;=\; S_{i\cdot}.$$

Thus, the trace becomes

$$\begin{aligned}
\operatorname{trace}\Big(C\big[J_m C + C J_n - C\big]^\top\Big) &= \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\Big[S_{\cdot j} + S_{i\cdot} - c_{ij}\Big] \\
&= \sum_{j=1}^{n}\Big(\sum_{i=1}^{m} c_{ij}\Big)^2 + \sum_{i=1}^{m}\Big(\sum_{j=1}^{n} c_{ij}\Big)^2 - \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^2 \\
&= \sum_{j=1}^{n} S_{\cdot j}^2 + \sum_{i=1}^{m} S_{i\cdot}^2 - \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^2.
\end{aligned}$$

Rewriting the Variance Estimator

The variance estimator is given by

$$\widehat{\operatorname{Var}}(U_{m,n}) \;=\; \frac{\widehat{V}}{mn} \;+\; \frac{n-1}{mn}\,\widehat{C}_1 \;+\; \frac{m-1}{mn}\,\widehat{C}_2,$$

with

$$\widehat{V} \;=\; \frac{1}{mn-1}\sum_{i,j} c_{ij}^2.$$

To see the connection with the trace formula, let's reexpress the "covariance-terms." In fact, a short calculation shows that

$$\widehat{C}_1 \;=\; \frac{1}{m\,n(n-1)}\left[\sum_{i=1}^{m} S_{i\cdot}^2 - \sum_{i,j} c_{ij}^2\right]$$

and

$$\widehat{C}_2 \;=\; \frac{1}{n\,m(m-1)}\left[\sum_{j=1}^{n} S_{\cdot j}^2 - \sum_{i,j} c_{ij}^2\right].$$

Substitute these into the variance estimator. After some algebra (and noting that the factors combine so that all three terms have an overall denominator proportional to m^2n^2), one finds that

$$\widehat{\mathrm{Var}}(U_{m,n}) \;=\; \frac{1}{m^2n^2}\Big[\sum_{i=1}^{m} S_{i\cdot}^2 + \sum_{j=1}^{n} S_{\cdot j}^2 - \sum_{i,j} c_{ij}^2\Big].$$

But the right-hand side is exactly the trace expression we obtained above. Hence,

$$\widehat{\mathrm{Var}}(U_{m,n}) \;=\; \frac{1}{m^2n^2}\,\mathrm{trace}\Big(C\big[J_mC + CJ_n - C\big]^\top\Big).$$