

Lecture 10: Weak convergence and central limit theorem

Lecturer: Yiqiao Zhong

Last update: 12:46:36, 2024-10-09

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

In this lecture, we will learn another aspect of the asymptotic theory that characterize the distributions of estimator with a large sample size. The probability foundation is the well celebrated central limit theorem.

Quantifying uncertainty of estimation. For an estimation problem, we have learned that a basic principle for designing an estimator is consistency, namely whether $\hat{\theta}_n$ converges to the “true” θ when $n \rightarrow \infty$. But how does the distribution $\hat{\theta}_n$ look like?

Imagine that we repeat our scientific experiment or change the random seed in computational experiment, we would get many realization of our estimator $\hat{\theta}_n$. If we have a distributional information about $\hat{\theta}_n$, then we will be able to tell if we are luck or unlucky, which is a first step towards replicability.

To this end, we need to study how random variables are distributed under an appropriate limit. A powerful theoretical tool for this purpose is the central limit theorem (CLT).

1 Weak convergence

Let $(X_n)_{n=1}^\infty$ be a sequence of random variables or vectors in $\mathbb{R}^d$, and $(F_n)_{n=1}^\infty$ be their CDFs. Also, let X be a random variable or vector also in $\mathbb{R}^d$, and F be its CDF.

Definition 1. We say that $(F_n)_{n=1}^\infty$ converges weakly to F and write $F_n \xrightarrow{w} F$ if and only if, for each continuity point x of F ,

$$\lim_{n \rightarrow \infty} F_n(x) = F(x). \quad (1)$$

We say that $\{X_n\}$ converges to X in distribution and write $X_n \xrightarrow{d} X$ if and only if $F_{X_n} \xrightarrow{w} F_X$.

In this definition, we require Eqn. 1 only at continuity point x . This allows flexibility with point masses. Under this definition, the distribution of Dirac measure δ_{a_n} converges weakly to that of δ_a if $\lim_{n \rightarrow \infty} a_n = a$; however, this would be inconveniently false had we defined weak convergence by requiring Eqn. 1 for every x ([exercise](#)).

Remark 1 (On terminology and notations). There are different names for the same definition: converge in distribution, converge weakly, converge in law. You may also see various notations for weak convergence.

$$X_n \xrightarrow{d} X, \quad X_n \xrightarrow{D} X, \quad X_n \Rightarrow X$$

It appears that in the statistics literature, “converge in distribution” is the most common, while weak convergence is more broad (not limited to random variables).

Remark 2. Weak convergence is very different from previous convergence notations, as we do not require that $X, X_1, X_2, \dots$ lie in the same probability space. The only things that matter are the distributions, or measures induced by each random variable.

As an example, consider $X = G \sim \mathcal{N}(0, 1)$, then by the symmetry of normal distribution,

$$X_n \xrightarrow{d} G \iff X_n \xrightarrow{d} -G \text{ because } G \stackrel{d}{=} -G.$$

where $\stackrel{d}{=}$ means two variables having same distribution. Convergence in distribution does not track or relate to the behavior of random variables as measurable functions $X_n(\omega)$ and $G(\omega)$.

Theorem 1. Let $X, X_1, X_2, \dots$ be random vectors in $\mathbb{R}^d$.

(i) If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

(ii) If $X_n \xrightarrow{d} X$ and $X = c$ a.s., where $c \in \mathbb{R}^k$ is a constant vector, then $X_n \xrightarrow{p} c$.

(iii) If $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$.

For simplicity, we prove this theorem for $d = 1$. See [2] for the general case.

Proof. (i) Let x be a continuous point of F . Then for any $\varepsilon > 0$, we can find $\delta > 0$ such that $|F(y) - F(x)| < \varepsilon$ for all $y \in [x - \delta, x + \delta]$. By the definition of convergence in probability,

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \delta) = 0.$$

We observe that

$$F_n(x) = P(X_n \leq x) \stackrel{(i)}{\leq} P(X \leq x + \delta) + P(|X_n - X| \geq \delta) \leq F(x) + \varepsilon + P(|X_n - X| \geq \delta) \quad (2)$$

where (i) is because if $X_n \leq x$, then at least one of the two events $\{X \leq x + \delta\}$, $|X_n - X| \geq \delta$ must happen (due to triangle inequality), namely

$$\{X_n \leq x\} \subseteq \{X \leq x + \delta\} \cup \{|X_n - X| \geq \delta\}$$

Taking limit in Eqn. 2, we find $\limsup_n F_n(x) - F(x) \leq \varepsilon$. Similarly, we can prove $\liminf_n F_n(x) - F(x) \geq -\varepsilon$. Since ε is arbitrary, we must have $\lim_n F_n(x) = F(x)$.

(ii) The CDF $F(x)$ has a single discontinuity point at $x = c$ and flat otherwise. So $F_n(x) \rightarrow 1$ for any $x > c$ and $F_n(x) \rightarrow 0$ for any $x < c$. To prove $X_n \xrightarrow{p} c$, observe that for any $\varepsilon > 0$,

$$P(|X_n - c| > \varepsilon) \leq P(X_n > c + \varepsilon) + P(X_n < c - \varepsilon) = 1 - F_n(c + \varepsilon) + F_n(c - \varepsilon)$$

The right-hand side converges to 0 as $n \rightarrow \infty$, so $\lim_n P(|X_n - c| > \varepsilon) = 0$.

(iii) Fix any $\varepsilon > 0$. We can choose some $C > 0$ such that $P(|X| > C) \leq \varepsilon/2$ and $F(x)$ is continuous at $x = C$ (exercise), so

$$1 - F(C) + F(-C) \leq \varepsilon/2$$

For sufficiently large n , $|F_n(x) - F(x)| \leq \varepsilon/4$ at both $x = C$ and $x = -C$. So by the triangle inequality, for sufficiently large n ,

$$P(|X_n| > C) \leq 1 - F_n(C) + F_n(-C) \leq \varepsilon.$$

□

1.1 Properties

The following result gives some useful sufficient and necessary conditions for convergence in distribution. By convention, we denote by $C_b(\mathbb{R}^d)$ the space of all continuous bounded functions on $\mathbb{R}^d$. The characteristic function of a random vector X in $\mathbb{R}^d$ is $\varphi_X(t) = \mathbb{E}[e^{it^\top X}]$, a complex-valued function defined over $t \in \mathbb{R}^d$,

Theorem 2 (Theorem 1.9 of [2]). *Let $X, X_1, X_2, \dots$ be random vectors in $\mathbb{R}^d$.*

(i) $X_n \xrightarrow{d} X$ if and only if for all $h \in C_b(\mathbb{R}^k)$,

$$\lim_{n \rightarrow \infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]. \quad (3)$$

(ii) Let $\phi_X, \phi_{X_1}, \phi_{X_2}, \dots$ be the characteristic functions of $X, X_1, X_2, \dots$, respectively. $X_n \xrightarrow{d} X$ if and only if $\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathbb{R}^d$.

(iii) (Cramér-Wold) $X_n \xrightarrow{d} X$ if and only if $c^\top X_n \xrightarrow{d} c^\top X$ for every $c \in \mathbb{R}^d$.

You may try to convince yourself about the correctness of (i) in Theorem 2 by proving simpler cases (exercise). Case a: F_n and F have bounded pdf. Case b: F_n and F have pmf.

Heuristics. Instead of showing the full proofs, which are technical for this course, we will provide some heuristics.

(i) The function $h \in C_b(\mathbb{R}^d)$ is sometimes called a test function (probing the value of a measure using a continuous measurement).

A distribution function $F_X(x)$ can be viewed as an expectation of an indicator function $F_X(x) = \mathbb{E}[\mathbf{1}\{X \leq x\}]$. The indicator function $\tilde{h}(x') = \mathbf{1}\{x' \leq x\}$ may be viewed as an improper test function (discontinuous at a point), so we need to do extra work using the assumption of continuity about x to derive (3). Conversely, we can approximate the indicator function $\tilde{h}$ except at the discontinuous point using a sequence of test functions in $C_b(\mathbb{R}^k)$.

(ii) There is a one-one mapping between characteristic functions and distribution functions. Characteristic functions are a variant of **Fourier transform**, where the “time domain” of CDF $F_X(x)$ is represented by the “frequency domain” $\varphi_X(t)$. Weak convergence in the time domain can be translated to pointwise convergence in the frequency domain under such Fourier transform (which is easily proved by (i)), and in reverse direction.

(iii) It is an easy consequence of (ii), since the characteristic function in $\mathbb{R}^d$ is defined by projecting a random vector to one dimension with $t \in \mathbb{R}^d$ first.

2 Central limit theorem

Before stating the CLT, we make a simple observation. If $X = (\xi_1, \xi_2, \dots, \xi_k)^\top \in \mathbb{R}^d$ is a random vector, then $\mathbb{E}(\xi_j^2) < \infty$ for all j implies $\mathbb{E}[|\xi_j \xi_m|] < \infty$ for all $j, m = 1, 2, \dots, k$ (by Cauchy-Schwarz). Thus, finite second moments of a random vector imply that $\Sigma = \text{Cov}(X_1)$ exists.

Theorem 3 (Central limit theorem). *Let $X_1, X_2, \dots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ and assume that the second moment exists. Denote the covariance matrix $\Sigma = \text{Cov}(X_1)$. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mathbb{E}(X_1)) \xrightarrow{d} N(0, \Sigma).$$

Note that CLT holds under quite minimal assumptions! Arguably it is the important theoretical result in probability theory and statistics. For an average of i.i.d. random vectors, the shape of the asymptotic distribution is determined purely by its covariance.

The normal or Gaussian distribution also plays a central role in various fields, such as Brownian motion and financial markets, heat equation and diffusion process in physics, etc.

3 Three proofs ideas for CLT

First, we notice that proving CLT for the general $\mathbb{R}^d$ case boils down to proving the one-dim case: by the Cramér-Wold device, it suffices to show

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n c^\top X_i - \mathbb{E}(c^\top X_i) \xrightarrow{d} N(0, c^\top \Sigma c), \quad \text{for all } c \in \mathbb{R}^k.$$

We can introduce $\tilde{X}_i = c^\top X_i$. Simple calculation shows that $\tilde{X}_i$ has finite second moment and $\text{Var}(\tilde{X}_i) = c^\top \Sigma c$ (exercise), so one-dim CLT implies the general CLT.

Second, notice that weak convergence holds after centering and rescaling the sequence of random variables (exercise). Thus, without loss of generality, we consider the 1-d case (namely $d = 1$) and assume zero mean with unit variance.

3.1 Characteristic function

Recall the definition of the characteristic function (ch.f.) $\phi_X(t)$ of a generic random variable

$$\phi_X(t) = \mathbb{E}(e^{iXt}) = \mathbb{E}[\cos(Xt) + i \sin(Xt)].$$

Weak convergence is characterized by pointwise convergence of ch.f. An important property about characteristic functions is that a probability distribution is uniquely determined by a ch.f. [1][Thm. 3.3.11]. Moreover, weak convergence of distributions can be translated to pointwise convergence of ch.f. and backwards due to **Lévy's continuity theorem**.

Theorem 4 (Optional, Thm 3.3.17 of [1]). *Let $\mu_n, 1 \leq n \leq \infty$ be probability measures with ch.f. φ_n .*

- (i) *If $\mu_n \xrightarrow{d} \mu_\infty$, then $\varphi_n \rightarrow \varphi_\infty(t)$ for all $t \in \mathbb{R}$.*
- (ii) *If $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at 0, then the sequence of distributions μ_n converges weakly to a measure μ with ch.f. φ .*

Proof idea of CLT. Suppose $X_1, X_2, \dots, X_n$ are i.i.d. random variables with zero mean and unit variance. Let $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ and denote by $\varphi_n(t)$ the ch.f. of S_n . And let $\varphi(t)$ be the ch.f. of a standard normal variable.

We know from Lecture 5 that $\varphi(t) = \exp(-t^2/2)$, which is a continuous function. By Lévy's continuity theorem (ii), we only need to show

$$\varphi_n(t) \rightarrow \varphi(t) \quad \text{for all } t \in \mathbb{R}.$$

Recall that from Lecture 5, independence of random variables is translated to products of ch.f. So $\varphi_n(t) = [\varphi_1(t/\sqrt{n})]^n$. If $n \rightarrow \infty$, we apply Taylor's expansion to $\varphi_1(t/\sqrt{n})$ around 0, so

$$\varphi_n(t) \approx \left[\varphi_1(0) + \frac{t}{\sqrt{n}} \varphi_1'(0) + \frac{t^2}{2n} \varphi_1''(0) + o\left(\frac{t^2}{n}\right) \right]^n$$

This is of course only a formal expansion. A rigorous proof will need to establish differentiability of $\varphi(t)$ and work out $o(t^2)$ first (we are now dealing with complex-value functions!). It is possible ([exercise*](#)) to show

$$\varphi(0) = 1, \quad \varphi'_1(0) = i \mathbb{E} X_1 = 0, \quad \varphi''_1(0) = -\mathbb{E}[X_1^2] = -1$$

so

$$\varphi_n(t) \approx \left[1 + \frac{t^2}{2n} + o(t^2)\right]^n = \exp\left(n \log\left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)\right) = \exp\left(n\left(-\frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)\right) \rightarrow e^{-t^2/2}$$

which completes the proof (sketch). $\square$

In the above proof sketch, the dominant term is the 2-nd order term since 1-st order term vanishes due to zero-mean assumption.

3.2 The moment method*

Another way is using the moment method¹ to prove the theorem. This method is quite general and can address (weakly) dependent data, and also works for random matrices. The drawback is that it is often tedious.

The idea is to first develop a continuity theorem similar to Lévy's continuity theorem: if $\mathbb{E}[X_n^k]$ converges to $\mathbb{E}[X^k]$ for every positive integer k , then we have weak convergence $X_n \xrightarrow{d} X$. Essentially the moment exponent k plays a similar role as t in ch.f. $\varphi(t)$. Unfortunately, unlike ch.f., moments do not uniquely determine a probability distribution, and the convergence correspondence requires an additional assumption (see Sect. 3.3.5 of [1]).

Definition 2. Let $\mathcal{I}$ be a countable set. We say $(X_i)_{i \in \mathcal{I}}$ is uniformly subgaussian if there exist constants $C, c > 0$ such that

$$\mathbb{P}(|X_i| \geq t) \leq C e^{-ct^2} \quad \text{for all } t > 0, \text{ for all } i \in \mathcal{I}.$$

We say X is subgaussian if $|\mathcal{I}| = 1$.

Uniformly subgaussian condition puts a strong restriction on the tail probability behavior of a collection of random variables. It essentially requires that the tail probability decay is uniformly controlled by the Gaussian probability tail (which is often considered as *light tail*). Under this condition, we have another equivalence for convergence in distribution.

Theorem 5 (Carleman continuity theorem). Let $(X_n)_{n \geq 1}$ be uniformly subgaussian random variables, X is subgaussian. Then $X_n \xrightarrow{d} X$ if and only if

$$\lim_n \mathbb{E}(X_n^k) = \mathbb{E}(X^k) \quad \text{for all } k = 0, 1, 2, \dots$$

Proof for 'only if' part. Let $\phi(x) \in [0, 1]$ be a smooth function such that $\phi(x) = 1$ if $x \in [-1, 1]$ and $\text{supp } \phi \subset [-2, 2]$. Fix any positive integer k . The goal is to show that

$$X_n \xrightarrow{d} X \implies \text{for given } N > 0, \quad \lim_{n \rightarrow \infty} \mathbb{E}[X_n^k \phi(X_n/N)] = \mathbb{E}[X^k \phi(X/N)], \quad (4)$$

$$\text{Uniformly subgaussian} \implies \lim_{N \rightarrow \infty} \sup_n |\mathbb{E}[X_n^k (1 - \phi(X_n/N))]| = 0, \quad (5)$$

$$\text{Subgaussian} \implies \lim_{N \rightarrow \infty} |\mathbb{E}[X^k (1 - \phi(X/N))]| = 0. \quad (6)$$

¹Not to confuse with method-of-moment. In the context of probability theory, moment method refers to a proof strategy.

Eqn. 4 holds because $x^k \phi(x/N)$ is continuous and bounded. Eqn. 6 holds because of DCT. To prove (5), note

$$\begin{aligned} |\mathbb{E}[X_n^k(1 - \phi(X_n/N))]| &\leq \mathbb{E}[|X_n|^k \mathbb{I}_{\{|X_n| > N\}}] && (\text{since } 1 - \phi(X_n/N) = 0 \text{ if } |X_n| \leq N) \\ &\leq C_1 \int_N^\infty kt^{k-1} \mathbb{P}(|X_n| > t) dt && (\text{by Fubini's theorem; recall Lecture 4}) \\ &\leq C_2 \int_N^\infty kt^{k-1} e^{-ct^2} dt \rightarrow 0 \quad \text{as } N \rightarrow \infty. \end{aligned}$$

where $C_1, C_2 > 0$ are certain constants. The proof is completed by combining Eqns 4–6. $\square$

Proof idea for ‘if’ part. Use Lévy continuity theorem, we expand $\mathbb{E}(e^{iX_n t})$ and use Taylor expansion (admittedly, we use the standard chf stuff here). $\square$

Moments calculation and combinatorics. Back to the proof of the CLT. For simplicity, we only prove the case where X is bounded². The general case can be proved with an additional truncation argument. Define $S_n = \frac{1}{\sqrt{n}}(X_1 + \dots + X_n)$, then $(S_n)_{n \geq 1}$ is uniformly subgaussian³. Assume $\mathbb{E}(X) = 0$, $\text{Var}(X) = 1$. Now it suffices to show that $\mathbb{E} S_n^k \rightarrow \mathbb{E} G^k$ for every fixed $k = 0, 1, 2, \dots$ where $G \sim \mathcal{N}(0, 1)$ is standard Gaussian variable.

- To better understand the expression, we first do some warm-up analysis. (i) For $k = 0$, $\mathbb{E} S_n^0 = 1 = \mathbb{E} G^0$.
(ii) For $k = 1$, X_n has zero mean $\Rightarrow \mathbb{E} S_n = 0 = \mathbb{E} G$.
(iii) For $k = 2$, $\mathbb{E} S_n^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{E} X_i^2 = 1 = \mathbb{E} G^2$.
(iv) For $k = 3$, we have

$$\begin{aligned} \mathbb{E} S_n^3 &= n^{-3/2} \sum_{i=1}^n \mathbb{E} X_i^3 + n^{-3/2} \sum_{1 \leq i < j \leq n} \mathbb{E}(3X_i^2 X_j + 3X_i X_j^2) + n^{-3/2} \sum_{1 \leq i < j < k \leq n} \mathbb{E}[6X_i X_j X_k] \\ &= n^{-1/2} \mathbb{E} X_i^3 = o(1). \end{aligned}$$

- (v) For $k = 4$, we have

$$\mathbb{E} S_n^4 = n^{-2} \sum_{1 \leq i \leq n} \mathbb{E} X_i^4 + n^{-2} \sum_{1 \leq i < j \leq n} \mathbb{E}[6X_i^2 X_j^2] = 3 + o(1).$$

In general,

$$\mathbb{E} S_n^k = n^{-k/2} \sum_{m=1}^k \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} \sum_{a_1 + \dots + a_m = k} \frac{k!}{a_1! a_2! \dots a_m!} \mathbb{E}(X_{i_1}^{a_1} \dots X_{i_m}^{a_m}). \quad (7)$$

Let us take a closer look at this seemingly scary combinatorial expression.

- Observation 1: If $a_i = 1$, then $\mathbb{E}(X_i) = 0$. So we only need to consider $a_1, \dots, a_m \geq 2$.
- Observation 2: There are $\binom{n}{m} = O(n^m)$ choices of $(i_1, i_2, \dots, i_m)$ satisfies $1 \leq i_1 < i_2 < \dots < i_m \leq n$.

The leading term of the RHS of (7) is when $m = \lfloor k/2 \rfloor$, and $a_1 = a_2 = \dots = a_m = 2$ (if k is even). Then we have

$$\mathbb{E} S_n^k = n^{-k/2} \binom{n}{m} \frac{k!}{2^{k/2}} + o(1) = \begin{cases} \frac{k!}{(k/2)! 2^{k/2}} + o(1) & \text{if } k \text{ is even} \\ o(1) & \text{if } k \text{ is odd.} \end{cases}$$

The leading term coincides with the moments of the normal distribution. This implies that $\mathbb{E}(S_n^k) \rightarrow \mathbb{E}(G^k)$ as $n \rightarrow \infty$.

²Note: This is a strong assumption since a bounded random variable implies the subgaussian property.

³Note: This will be proved in the next lecture.

Remark 3. There are n^k terms in total when we expand the sum S_n^k . However, when we evaluate the expectation of S_n^k , most terms will go away since $\mathbb{E} X_i = 0$, which leaves $O(n^{k/2})$ active terms—this is the reason for factor $\sqrt{n}$ in CLT.

3.3 Lindeberg swapping trick

The third proof idea is closely related to the *universality*. The universality idea is the following.

1. Show that some property about random variables $X_1, X_2, \dots, X_n$ (measured as $F(X_1, X_2, \dots, X_n)$ for instance) is unchanged if X_i is replaced by independent $Y_i \sim \mathcal{N}(0, 1)$, which usually has the analytic style (heavily involved with arguments about continuity and taking limits).
2. Gaussian case calculation, which usually has the flavor of algebra and combinatorics.

Applying this general strategy to prove the CLT, if we could show

$$\frac{X_1 + X_2 + \dots + X_n}{\sqrt{n}} \approx \frac{Y_1 + Y_2 + \dots + Y_n}{\sqrt{n}} \sim \mathcal{N}(0, 1),$$

then the proof is complete. Thus, we only need to justify the “ $\approx$ ” part. Below let us present a stronger form of CLT under an additional finite third moment assumption. By convention, we denote by $C^3(\mathbb{R})$ the space of functions which are third-order continuously differentiable.

Theorem 6 (a variant of Berry-Esséen theorem). *Let X be a random variable with $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) = \sigma^2$ and $\mathbb{E}(|X|^3) < \infty$, and let $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} X$. Let $\mathcal{F}_B$ be the space of function ϕ such that $\phi \in C^3(\mathbb{R})$, and $\left| \frac{d^k}{dx^k} \phi(x) \right| \leq B$ for $k = 0, 1, 2, 3$. Then*

$$\sup_{\phi \in \mathcal{F}_B} |\mathbb{E} \phi(S_n) - \mathbb{E} \phi(G)| \leq \frac{CB}{\sqrt{n}} \mathbb{E}(|X|^3).$$

Proof. Let $Y_1, Y_2, \dots, Y_n \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$, define $W_n = \frac{Y_1 + \dots + Y_n}{\sqrt{n}} \stackrel{d}{=} \mathcal{N}(0, 1)$, $S_{n,i} = \frac{X_1 + \dots + X_i + Y_{i+1} + \dots + Y_n}{\sqrt{n}}$, note $S_{n,n} = S_n$, $S_{n,0} = W_n$, $\mathbb{E}(\phi(S_n)) - \mathbb{E}(\phi(W_n)) = - \sum_{i=0}^{n-1} \mathbb{E}[\phi(S_{n,i}) - \phi(S_{n,i+1})]$. It suffices to show

$$\sup_{\phi \in \mathcal{F}_B} \sup_{0 \leq i \leq n-1} |\mathbb{E}[\phi(S_{n,i}) - \phi(S_{n,i+1})]| \leq \frac{CB}{n^{3/2}} \mathbb{E}(|X|^3).$$

Define $T_{n,i} = \frac{X_1 + \dots + X_i + Y_{i+2} + \dots + Y_n}{\sqrt{n}}$, $i = 0, 1, \dots, n-1$, then $S_{n,i} = T_{n,i} + \frac{1}{\sqrt{n}} Y_{i+1}$, and $S_{n,i+1} = T_{n,i} + \frac{1}{\sqrt{n}} X_{i+1}$. Then by Taylor expansion, we have

$$\begin{aligned} \mathbb{E}(\phi(S_{n,i})) &= \mathbb{E}(\phi(T_{n,i})) + \frac{1}{\sqrt{n}} \mathbb{E}[\phi'(T_{n,i}) Y_{i+1}] + \frac{1}{2n} \mathbb{E}[\phi''(T_{n,i}) Y_{i+1}^2] + O\left(\frac{B \cdot \mathbb{E}(|Y_{i+1}|^3)}{n^{3/2}}\right), \\ \mathbb{E}(\phi(S_{n,i+1})) &= \mathbb{E}(\phi(T_{n,i})) + \frac{1}{\sqrt{n}} \mathbb{E}[\phi'(T_{n,i}) X_{i+1}] + \frac{1}{2n} \mathbb{E}[\phi''(T_{n,i}) X_{i+1}^2] + O\left(\frac{B \cdot \mathbb{E}(|X_{i+1}|^3)}{n^{3/2}}\right). \end{aligned} \quad (8)$$

Note that $|\phi'''|$ is uniformly bounded by B (a crucial condition!), so it is correct to use the big-O notation in the above equations. By our definition, $T_{n,i}, X_{i+1}, Y_{i+1}$ are jointly independent! So

$$\begin{aligned} \mathbb{E}[\phi'(T_{n,i}) Y_{i+1}] &= \mathbb{E}[\phi'(T_{n,i})] \mathbb{E}[Y_{i+1}], & \mathbb{E}[\phi''(T_{n,i}) Y_{i+1}^2] &= \mathbb{E}[\phi''(T_{n,i})] \mathbb{E}[Y_{i+1}^2], \\ \mathbb{E}[\phi'(T_{n,i}) X_{i+1}] &= \mathbb{E}[\phi'(T_{n,i})] \mathbb{E}[X_{i+1}], & \mathbb{E}[\phi''(T_{n,i}) X_{i+1}^2] &= \mathbb{E}[\phi''(T_{n,i})] \mathbb{E}[X_{i+1}^2]. \end{aligned}$$

Our definition also ensures that the first two moments are matched: $\mathbb{E}[Y_{i+1}] = \mathbb{E}[X_{i+1}] = 0$, $\mathbb{E}[Y_{i+1}^2] = \mathbb{E}[X_{i+1}^2] = 1$. Taking the difference of (8), we obtain

$$|\mathbb{E}[\phi(S_{n,i}) - \phi(S_{n,i+1})]| \leq \frac{CB \cdot (\mathbb{E}(|X|^3) + \mathbb{E}(|Y_{i+1}|^3))}{n^{3/2}}.$$

□

Rate of convergence for CLT. The common Berry-Esséen theorem is stated below, which is an improvement of Theorem 6.

Theorem 7 (Berry-Esséen theorem). *Let X be a random variable with $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) = \sigma^2$ and $\mathbb{E}(|X|^3) < \infty$, let $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X$. Then*

$$\sup_{a \in \mathbb{R}} |\mathbb{P}(S_n \leq a) - \mathbb{P}(G \leq a)| \leq \frac{C}{\sqrt{n}} \mathbb{E}(|X|^3).$$

This theorem implies CLT since $F_{S_n}(a) \rightarrow F_G(a)$ for all a . The conclusion is stronger than CLT, since it provides an error bound of the uniform convergence of CDF.

4 Further reading

For more results about weak convergence and characteristic functions, a good reference is [1][Sect. 3.2—3.3].

Proofs for CLT. The three proof ideas for CLT are taken from the book [3], which provides multiple proofs for CLT.

Useful facts. As we dig into the proofs, you may find some basic facts useful.

1. If F is a CDF, then it admits at most countably many discontinuous points. It is equivalent to $\{x \in \mathbb{R} : \mu(\{x\}) > 0\}$ being countable where μ is the measure associated with F .
2. If $B \subset \mathbb{R}$ is a Borel set with a positive Lebesgue measure $m(B)$. Then for any $\varepsilon > 0$, there exist an open set O and a closed set K with $K \subseteq B \subseteq O$ such that $m(K) > m(B) - \varepsilon$ and $m(O) < m(B) + \varepsilon$. In words, a Borel set can be squeezed by an open set and a close set. More properties can be found in Wiki.
3. If $O \subset \mathbb{R}$ is an open set, then it has the form $O = \cup_{i \in \mathcal{I}} (a_i, b_i)$ for countably many suitable real numbers a_i, b_i .

The following result can be used to check whether $X_n \xrightarrow{d} X$ when X has a pdf f and X_n has a pdf f_n .

Theorem 8. (Scheffé's theorem). *Let $\{f_n\}$ be a sequence of pdf's on $\mathbb{R}^k$ w.r.t a measure ν . Suppose that $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ a.e. ν and $f(x)$ is a pdf w.r.t. ν . Then*

$$\lim_{n \rightarrow \infty} \int |f_n(x) - f(x)| d\nu = 0.$$

References

- [1] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 5th edition, 2019.
- [2] Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2nd edition, 2013.
- [3] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.