Probailbity Theory Notes

Hsi-Kang Hsu

1 Expected Values

1.1 Moments

Proposition 1. *For $j \geq k$, if X has j-th moments, then X has k-th moments.*

Proof. Since $x \mapsto x^{\frac{k}{j}}$ is concave, we have

$$(\mathbb{E}\, X^k)^{\frac{1}{k}} = \left(\mathbb{E}(X^j)^{\frac{k}{j}}\right)^{\frac{1}{k}} \leq (\mathbb{E}\, X^j)^{\frac{1}{j}}$$

, where the last inequality is by Jensen's inequality. □

Proposition 2. *For non-negative random variable X, we have the following formula to calculate moments*

$$\mathbb{E}\, X^k = \int_0^\infty kx^{k-1} P(X \geq x)\, dx$$

In particular, we have $\mathbb{E}\, X = \int_0^\infty P(X \geq x)\, dx$

Proof. By Fubini, we have

$$\begin{aligned}
\mathbb{E}\, X^k = \int X^k\, dP = \int \int_0^X kx^{k-1}\, dx\, dP &= \int \int \mathbb{1}(0 \leq x \leq X) kx^{k-1}\, dx\, dP \\
&= \int \int \mathbb{1}(0 \leq x \leq X) kx^{k-1}\, dP\, dx \quad \text{(Fubini)} \\
&= \int kx^{k-1} P(X \geq x \geq 0)\, dx \\
&= \int_0^\infty kx^{k-1} P(X \geq x)\, dx
\end{aligned}$$

□

1.2 Moment generating functions

Definition 1. For random variable X, define the moment generating function $M : \mathbb{R} \to \mathbb{R}$ of X as

$$M(t) = \mathbb{E}\, e^{tX}$$

If X is unbounded, $M(t)$ could be ∞.

Proposition 3. *We have $M(0) = 1$. The values where $M(t) < \infty$ is an interval containing 0. (may just be $\{0\}$)*

Proof. For $t > 0$, if $M(t) < \infty$, then for $s \in [0, t)$, we have $\mathbb{E}\, e^{sX} = \mathbb{E}\, e^{sX}\mathbb{1}(X \geq 0) + \mathbb{E}\, e^{sX}\mathbb{1}(X < 0) \leq 1 + \mathbb{E}\, e^{tX}\mathbb{1}(X < 0) \leq 1 + M(t) < \infty$. For $t < 0$, the proof is similar. □

Proposition 4 (Power series of MGF)**.** *Let X be a random variable s.t it's MGF $M(t) < \infty$ for $t \in (-t_0, t_0)$ for some $t_0 > 0$. Then*

$$M(t) = \sum_{n=0}^{\infty} \mathbb{E}\, X^n \frac{t^n}{n!}$$

for all $t \in (-t_0, t_0)$. Therefore, $\mathbb{E}\, X^k = M^{(k)}(0)$ for all $k \in \mathbb{N}$.

Proof. I'll first show X has all moments. For $t \in (-t_0, t_0)$, by the inequality $e^{t|x|} \leq e^{tx} + e^{-tx}$, we have $\mathbb{E}\, e^{t|X|} \leq \mathbb{E}\, e^{tX} + \mathbb{E}\, e^{-tX} < \infty$. Fix $t \in (0, t_0)$ and $k \in \mathbb{N}$. For x large enough, we have $|x|^k \leq e^{t|x|}$. Thus there exists constant C s.t

$$\mathbb{E}\, |X|^k = \mathbb{E}\, |X|^k \mathbb{1}(|X| \leq C) + \mathbb{E}\, |X|^k \mathbb{1}(|X| > C) \leq |C|^k + \mathbb{E}\, e^{t|X|} < \infty$$

If $t \in (-t_0, t_0)$, by DCT, we have

$$\mathbb{E}\, e^{tX} = \mathbb{E} \sum_{n=1}^{\infty} X^n \frac{t^n}{n!} = \sum_{n=1}^{\infty} \mathbb{E}\, X^n \frac{t^n}{n!}$$

□

Proposition 5. *Let X be a random variable s.t it's MGF $M(t) < \infty$ for $t \in (-t_0, t_0)$ for some $t_0 > 0$. Then for $k \in \mathbb{N}$*

$$M^k(t) = \mathbb{E}\, X^k e^{tX}$$

for all $t \in (-t_0, t_0)$

Proof. By the power series expansion of the previous proposition, we have

$$M^k(t) = \sum_{n=0}^{\infty} \mathbb{E}\, X^{k+n} \frac{t^n}{n!} = \mathbb{E}\, X^k \sum_{n=0}^{\infty} X^n \frac{t^n}{n!} = \mathbb{E}\, X^k e^{tX}$$

It suffices to justify the exchange of summation and expectation of the second equality. By DCT, it suffices to show $\left| X^k \sum_{n=0}^{N} X^n \frac{t^n}{n!} \right|$ is bounded by an integrable r.v for fixed $t \in (-t_0, t_0)$. Choose t' s.t $t + t' \in (-t_0, t_0)$, then $|X|^k \leq e^{t'|X|}$ for large enough X. Thus for large enough X,

$$\left| X^k \sum_{n=0}^{N} X^n \frac{t^n}{n!} \right| \leq e^{t'|X|} e^{t|X|} \leq e^{(t'+t)|X|}$$

, which is integrable. □

2 Weak law of large numbers

Theorem 1 (Weak law of large numbers - finite variance)**.** *If $X_1, X_2, \ldots$ are iid random variables with $\mathbb{E}\, X_1 = \mu$ and $\mathbb{E}\, X_1^2 < \infty$, then*

$$\frac{X_1 + \cdots + X_n}{n} \to \mu \quad \text{in probability}$$

Proof. Apply Chebyshev's equality. □

2.1 Triangular arrays

In this subsection, we consider the limiting behavior of row sums $S_n = X_{n,1} + \cdots + X_{n,n}$ of triangular array $(X_{n,k} : 1 \le k \le n, n \ge 1)$

$$\begin{array}{lll} X_{1,1} & & \\ X_{2,1} & X_{2,2} & \\ X_{3,1} & X_{3,2} & X_{3,3} \\ \cdots & & \end{array}$$

Proposition 6. *Let S_n be the row sum of a triangular array. Write $\mu_n = \mathbb{E}\, S_n$ and $\sigma_n^2 = Var(S_n)$. Let (b_n) be a sequence of positive numbers s.t $\frac{\sigma_n^2}{b_n^2} \to 0$. Then*

$$\frac{S_n - \mu_n}{b_n} \to 0$$

in probability.

Proof. Apply Chebyshev's equality. □

Theorem 2 (Weak law for triangular array)**.** *For each n, let $(X_{n,k}, 1 \le k \le n)$ be independent. Let $b_n > 0$ with $b_n \to \infty$. Define $\overline{X}_{n,k} = X_{n,k} \mathbb{1}(|X_{n,k} \le b_n|)$. Suppose that as $n \to \infty$, one has*

1. $\sum_{k=1}^n P(|X_{n,k}| > b_n) \to 0$
2. $\frac{1}{b_n^2} \sum_{k=1}^n \mathbb{E}\, \overline{X}_{n,k}^2 \to 0$

Write $S_n = X_{n,1} + \cdots + X_{n,n}$ and $a_n = \sum_{k=1}^n \mathbb{E}\, \overline{X}_{n,k}$. Then

$$\frac{S_n - a_n}{b_n} \to 0$$

in probability.

Proof. We have $P\left(\left|\frac{S_n-a_n}{b_n}\right| \geq \varepsilon\right) = P\left(\left|\frac{S_n-a_n}{b_n}\right| \geq \varepsilon,\, S_n \neq \overline{S}_n\right) + P\left(\left|\frac{S_n-a_n}{b_n}\right| \geq \varepsilon,\, S_n = \overline{S}_n\right) \leq P\left(S_n \neq \overline{S}_n\right) + P\left(\left|\frac{\overline{S}_n-a_n}{b_n}\right| \geq \varepsilon\right)$.

We have $P\left(S_n \neq \overline{S}_n\right) \leq P\left(\bigcup_{k=1}^n \{X_{n,k} \neq \overline{X}_{n,k}\}\right) \leq \sum_{k=1}^n P(|X_{n,k}| > b_n) \to 0$.

On the other hand, $P\left(\left|\frac{\overline{S}_n-a_n}{b_n}\right| \geq \varepsilon\right) \leq \frac{1}{b_n^2\varepsilon^2} Var(\overline{S}_n) = \frac{1}{b_n^2\varepsilon^2} \sum_{k=1}^n Var(\overline{X}_{n,k}) \leq \frac{1}{b_n^2} \sum_{k=1}^n \mathbb{E}\, \overline{X}_{n,k}^2 \to 0$. □

Proposition 7. *Let $X_1, X_2, \ldots$ be iid random variables with $\lim_{x\to\infty} xP(|X_1| > x) = 0$. Let $S_n = X_1 + \cdots + X_n$ and let $\mu_n = \mathbb{E}\, X_1 \mathbb{1}(|X_1| \leq n)$. Then $\frac{S_n}{n} - \mu_n \to 0$.*

Theorem 3 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be iid random variables with $\mathbb{E}\,|X_1| < \infty$. Let $S_n = X_1 + \cdots + X_n$ and let $\mu = \mathbb{E}\, X_1$. Then*

$$\frac{S_n}{n} \to \mu$$

in probability.

Proof. We have

$$xP(|X_1| > x) = \mathbb{E}\, x\mathbb{1}(|X_1| > x) \leq \mathbb{E}\,|X_1|\mathbb{1}(|X_1| > x) \to 0$$

by DCT. Thus, by the above proposition, $\frac{S_n}{n} - \mu_n \to 0$ in probability. Also, since

$$\mu_n = \mathbb{E}\, X_1 \mathbb{1}(|X_1| \leq n) \to \mathbb{E}\, X_1 = \mu$$

by DCT, we have proved the theorem. □

Remark 1. Weak law of large numbers fail for "heavy-tailed distributions", i.e, $P(|X| > x)$ is large. For example, X_i are iid Cauchy distributed random variables ($f(x) = \frac{1}{\pi(1+x^2)}$).

3 Strong law of large numbers

3.1 Maximal inequalities

Let $X_1, X_2, \ldots$ be independent random variables. Set $S_n = X_1 + \cdots + X_n$. Define $M_n = \max\{|S_1|, \ldots, |S_n|\}$. The main point is: Whenever M_n is large, it is unlikely that S_n is small.

Theorem 4 (Kolmogorov maximal inequality)**.** *Suppose $X_1, X_2, \ldots$ are independent with $\mathbb{E}\, X_i = 0$ and $\mathbb{E}\, X_i^2 < \infty$. Then for any $a > 0$,*

$$P(M_n \geq a) \leq \frac{1}{a^2} Var(S_n)$$

Corollary 1 (One series theorem). *If $X_1, X_2, \ldots$ are independent with $\mathbb{E}\, X_i = 0$ for all i, then*

$$\sum_{n}^{\infty} Var(X_n) < \infty \Rightarrow \sum_{n=1}^{\infty} X_n \text{ converges a.s}$$

Proof. By Kolmogorov's maximal inequality, we have

$$P(\max_{1\le k\le r} |S_{n+k} - S_n| > \varepsilon) \le \frac{1}{\varepsilon^2} \sum_{k=1}^{r} Var(X_{n+k})$$

Take $r \to \infty$ on both sides, we have

□

Theorem 5 (Etemadi). *If $X_1, X_2, \ldots$ are independent, then for any $a \ge 0$*

$$P(M_n \ge 3a) \le 3 \max_{1\le k\le n} P(|S_k| \ge a)$$

Corollary 2. *If $X_1, X_2, \ldots$ are independent, then S_n converges in probability if and only if S_n converges a.s*

Proof. We only prove the only if direction. Let $T_n = \sup_{k\ge n} |S_k - S_n|$. We have

$$\left\{ \sup_{k,j\ge n} |S_k - S_j| \ge \varepsilon \right\} \subset \{T_k \ge 0.5\varepsilon\} \cup \{T_j \ge 0.5\varepsilon\}$$

Define $E(n,\varepsilon) = \left\{\sup_{k,j\ge n} |S_k - S_j| \ge \varepsilon\right\}$ and let $E(\varepsilon) = \bigcap_n E(n,\varepsilon)$. Since $E(n,\varepsilon) \downarrow E(\varepsilon)$, $P(E(\varepsilon)) = \lim_n P(E(n,\varepsilon))$. If we can show $\lim_n P(E(n,\varepsilon)) = 0$, then $P(E(\varepsilon)) = 0$. Let $E = \bigcup_\varepsilon E(\varepsilon)$ where ε is taken over all rationals, then $P(E) = 0$ and thus S_n converges a.s.

Thus, it suffices to show T_n converges in probability. By Etemadi,

$$\begin{aligned} P\left(\max_{n+1\le j\le m} |X_{n+1} + \cdots + X_j| > \varepsilon\right) &\le 3 \max_{n+1\le j\le m} P\left(|X_{n+1} + \cdots + X_j| > \frac{\varepsilon}{3}\right) \\ &= 3 \max_{n+1\le j\le m} P\left(|S_j - S_n| > \frac{\varepsilon}{3}\right) \end{aligned}$$

If $m \ge n \ge N$, since S_n converges in probability, the right hand side can be made less than any δ for N large enough. Thus we have

$$P\left(\max_{n+1\le j\le m} |X_{n+1} + \cdots + X_j| > \varepsilon\right) < \delta$$

for all $n \ge N$. Take $m \to \infty$, we then have $P(T_n \ge \varepsilon) < \delta$ for all $n \ge N$. □

Theorem 6 (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be iid random variables with $\mathbb{E}\, X_1 = 0$. Then $\frac{S_n}{n} \to 0$ a.s.*

Proof. Set $Y_n = X_n \mathbb{1}(|X_n| \leq n)$. Then $\sum_n P(X_n \neq Y_n) = \sum_n P(|X_n| > n) < \infty$ since $\mathbb{E}\, X_1 < \infty$. By Borel-Cantelli, $P(Y_n \neq X_n \text{ i.o}) = 0$, with this we can show $\frac{X_1+\cdots+X_n}{n} \to 0$ a.s is equivalent to $\frac{Y_1+\cdots+Y_n}{n} \to 0$ a.s: For fixed ω, we can find K s.t $X_n = Y_n$ for all $n \geq K$. Write $\left|\frac{Y_1+\cdots+Y_n}{n}\right| \leq \left|\frac{X_1+\cdots+X_n}{n}\right| + \left|\frac{(Y_1+\cdots+Y_K)-(X_1+\cdots+X_K)}{n}\right|$, the first term can be made small with n large enough since $\frac{X_1+\cdots+X_n}{n} \to 0$. For the second term, the numerator is constant for ω fixed, thus for n large enough it can also be made small. The other direction is similar.

Define $Y_n' = Y_n - \mathbb{E}\, Y_n$. We have $\mathbb{E}\, Y_n \to \mathbb{E}\, X_1 = 0$ by DCT. I claim $\frac{Y_1+\cdots+Y_n}{n} \to 0$ a.s is equivalent to $\frac{Y_1'+\cdots+Y_n'}{n} \to 0$ a.s: We have $\left|\frac{Y_1+\cdots+Y_n}{n} - \frac{Y_1'+\cdots+Y_n'}{n}\right| = \left|\frac{\mathbb{E}\, X_1+\cdots+\mathbb{E}\, X_n}{n}\right| \leq \left|\frac{\mathbb{E}\, X_1+\cdots+\mathbb{E}\, X_K}{n}\right| + \left|\frac{\mathbb{E}\, X_{K+1}+\cdots+\mathbb{E}\, X_n}{n}\right|$ for some K chosen s.t $\mathbb{E}\, X_n < \varepsilon$ for all $n \geq K$. Then that first term's numerator is bounded, the second term is less than ε. Take $n \to \infty$, the first term becomes 0.

We have a lemma: If $\sum_{n=1}^{\infty} \frac{x_n}{n}$ for a sequence of real numbers (x_n), then $\frac{x_1+\cdots+x_n}{n} \to 0$. By this lemma, it suffices to show $\sum_{n=1}^{\infty} \frac{Y_n'}{n} < \infty$ a.s. By the one series theorem, it remains to show $\sum_{n=1}^{\infty} Var(\frac{Y_n'}{n}) < \infty$. We have

$$Var(\frac{Y_n'}{n}) = Var(\frac{Y_n}{n}) \leq \mathbb{E}\left(\frac{Y_n}{n}\right)^2 = \frac{1}{n^2}\mathbb{E}\, X_1^2 \mathbb{1}(|X_1 \leq n)$$

$$\begin{aligned}
\sum_n Var(\frac{Y_n'}{n}) \leq \sum_n \frac{1}{n^2}\mathbb{E}\, X_1^2 \mathbb{1}(|X_1 \leq n) &= \mathbb{E}\, X_1^2 \sum_n \frac{1}{n^2}\mathbb{1}(|X_1| \leq n) \\
&= \mathbb{E}\, X_1^2 \sum_{n \geq |X_1|} \frac{1}{n^2} \\
&\leq \mathbb{E}\, X_1^2 \frac{c}{|X_1|} \\
&= c\,\mathbb{E}\,|X_1| < \infty
\end{aligned}$$

□

4 Weak Convergence

4.1 Definitions and basic properties

Definition 2. We say that F_n converges to F weakly (F, F_n are distribution functions on $\mathbb{R}$), denoted $F_n \Rightarrow F$ if $F_n(x) \to F(x)$ for all continuity points x of F. We say that X_n converges to X weakly/in distribution, denoted $X_n \Rightarrow X$, if the associated distribution functions F_n converges weakly to the distribution function of X.

Example 4.1. (Why do we only consider continuity points of F) Consider $X_n = \frac{1}{n}$, $X = 0$, then $X_n \to X$ a.s. We would expect $X_n \Rightarrow X$. But $F_n = \mathbb{1}_{[\frac{1}{n},\infty)}, F = \mathbb{1}_{[0,\infty)}$, thus $F_n(0) = 0 \neq F(0) = 1$.

Proposition 8. *Suppose X_n, X are random variables on the same probability space. If $X_n \to X$ in probability, then $X_n \Rightarrow X$.*

Proof. For $\varepsilon > 0$, $x \in \mathbb{R}$, we have $P(X_n \leq x) \leq P(X \leq x + \varepsilon) + P(|X - X_n| > \varepsilon)$. Thus $\limsup_n P(X_n \leq x) \leq P(X \leq x + \varepsilon)$. Take $\varepsilon \to 0$, by continuity, we have $\limsup_n P(X_n \leq x) \leq P(X \leq x)$. On the other hand, $P(X \leq x - \varepsilon) \leq P(X_n \leq x) + P(|X_n - X| > \varepsilon)$, thus $P(X < x) \leq \liminf_n P(X_n \leq x)$. This gives $P(X < x) \leq \liminf_n P(X_n \leq x) \leq \limsup_n P(X_n \leq x) \leq P(X \leq x)$. If x is a continuity point of F, then inequality becomes equality. □

If $X_n \Rightarrow X$, then X_n does not converge to X point-wise or in probability in general. This is because the X_n's and X might not even be defined on the same probability space, thus convergence in probability or point-wise cannot be discussed.

There is a special case where convergence in probability can we discussed if X_n are defined on different probability spaces and is implied by weak convergence. Consider $X \equiv a$ for some constant a, then the condition

$$\lim_n P(|X_n - a| \geq \varepsilon) = 0$$

can be discussed even if X_n are not defined on the same probability space. This gives the following theorem:

Theorem 7. $\lim_n P(|X_n - a| \geq \varepsilon) = 0$ *holds for all $\varepsilon > 0$ if and only if $X_n \Rightarrow a$, that is, if and only if*

$$\lim_n P(X_n \leq x) = \begin{cases} 0, & \text{if } x < a \\ 1, & \text{if } x > a \end{cases}$$

Proof. For the only if part, remember we do not need to consider when $x = a$ since a is not a continuity point of the distribution function of a. Put $\varepsilon = |x - a|$, if

$x < a$, then $P(X_n \leq x) \leq P(|X_n - a| \geq \varepsilon) \to 0$. Else if $x > a$, then $P(X_n \leq x) \geq P(|X_n - a| \leq \varepsilon) \to 1$. This proves $X_n \Rightarrow a$.

For the if part, $P(|X_n - a| \geq \varepsilon) \leq P(X_n \leq a - \varepsilon) + P(X_n \geq a + \varepsilon) = P(X_n \leq a - \varepsilon) + 1 - P(X_n < a + \varepsilon) \leq P(X_n \leq a - \varepsilon) + 1 - P(X_n \leq a + 0.5\varepsilon) \to 0$. □

Remark 2. Note that the condition $\lim_n P(|X_n - a| \geq \varepsilon) = 0$ isn't really a special case of convergence in probability as X_n need not to be defined on the same space. Convergence in probability in this new sense will be denoted $X_n \to a$, in accordance with the theorem just proved.

Proposition 9. *Suppose $X_n \Rightarrow X$ and $\delta_n \to 0$, then $\delta_n X_n \Rightarrow 0$*

Proof. By the above theorem, it suffices to show $P(|\delta_n X_n| \geq \varepsilon) \to 0$. Given ε, η, we have $P(|\delta_n X_n| \geq \varepsilon) = P(|X_n| \geq \frac{\varepsilon}{|\delta_n|})$. Choose x s.t $P(|X| \geq x) < \eta$ and $P(X = x) = P(X = -x) = 0$ and choose N s.t $n \geq N$ implies $|\delta_n| < \varepsilon/x$ and $|P(X_n \leq y) - P(X \leq y)| < \eta$ for $y = x, -x$. Then $P(|X_n| \geq \frac{\varepsilon}{|\delta_n|}) \leq P(|X_n| > x) = P(X_n > x) + P(X_n < -x) \leq P(X > x) + \eta + P(X \leq -x) + \eta \leq P(|X| \geq x) + 2\eta < 3\eta$ for all $n \geq N$. □

Theorem 8. *If $X_n \Rightarrow X$ and $X_n - Y_n \Rightarrow 0$, then $Y_n \Rightarrow X$*

Proof. See Billingsley Theorem 25.4 □

4.2 Fundamental Theorems

Here are some basic facts about weak convergence:

- If $F_n \Rightarrow F$ and $F_n \Rightarrow G$, then $F = G$: This is obvious since continuity points of F is dense and the right continuity of distribution functions.
- If $\lim_n F_n(d) = F(d)$ for d in a set D dense in $\mathbb{R}$, then $F_n \Rightarrow F$: If F is continuous at x, then there exists $d' < x < d''$, $d', d'' \in D$ and $F(d'') - F(d') < \varepsilon$. Since $F_n(d') \leq F_n(x) \leq F_n(d'')$, we have $F(d') \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(d'')$, thus $\lim_n F_n(x) = F(x)$.

We can change all the random variables s.t they are defined on the same space and the distribution of them remains the same. Moreover, point-wise convergence holds.

Theorem 9 (Skorokhod)**.** *Suppose that $\mu_n \to \mu$ for Borel probability measures μ_n, μ on $\mathbb{R}$. There exists a sequence Y_n and a random variable Y, all defined on the same probability space (Ω, F, P) s.t*

1. $Y_n(\omega) \to Y(\omega) \quad \forall \omega \in \Omega$

2. *Y_n has distribution μ_n, Y has distribution μ.*

Many assumptions of previous theorems related to the convergence of expectation can be weakened to weak convergence instead of point-wise convergence:

- (Fatou's Lemma) If $X_n \Rightarrow X$, then $\mathbb{E}\,|X| \leq \liminf_{n\to\infty} \mathbb{E}\,|X_n|$. Proof is to first use Skorokhod's theorem to transform X_n, X to Y_n, Y on the same probability space and apply Fatuo's lemma. Since their distributions are the same, their expectations are the same.
- If $X_n \Rightarrow X$, (X_n) is uniformly integrable, then $\mathbb{E}\,|X| < \infty$ and $\mathbb{E}\,X_n \to \mathbb{E}\,X$.

Theorem 10 (Mapping theorem)**.** *Let $h : \mathbb{R} \to \mathbb{R}$ be Borel measurable and set D_h be the set of discontinuities of h. If $X_n \Rightarrow X$ and $P(X \in D_h) = 0$, then $h(X_n) \Rightarrow h(X)$.*

Proof. By Skorokhod's theorem, there exists Y_n, Y with the same distribution as X_n, X and $Y_n \to Y$ point-wise. If $Y \notin D_h$, then $h(Y_n) \to h(Y)$, thus we have

$$P(h(Y_n) \to h(Y)) \geq P(h(Y_n) \to h(Y), Y \notin D_h) = P(Y \notin D_h) = P(X \notin D_h) = 1$$

Hence $h(Y_n) \to h(Y)$ a.s. This implies $h(Y_n) \Rightarrow h(Y)$ and thus $h(X_n) \Rightarrow h(X)$. □

Proposition 10. *If $X_n \Rightarrow X$ and $a_n \to a$, $b_n \to b$. Then $a_n X_n + b_n \Rightarrow aX + b$.*

Proof. $aX_n + b \Rightarrow aX + b$ follows immediately from the mapping theorem. We also have $(a_n - a)X_n \Rightarrow 0$ by previous proposition. We claim if $X_n \Rightarrow 0$ and $\delta_n \to 0$, then $X_n + \delta_n \Rightarrow 0$. Indeed, $P(|X_n + \delta_n| \geq \varepsilon) \leq P(|X_n| + |\delta_n| \geq \varepsilon) \leq P(|X_n| \geq 0.5\varepsilon) + P(|\delta_n| \geq 0.5\varepsilon) \to 0$. Hence $(a_n X_n + b_n) - (aX_n + b) \Rightarrow 0$. Combined with the fact that $aX_n + b \Rightarrow aX + b$, by the above theorem, we have $a_n X_n + b_n \Rightarrow aX + b$. □

4.3 Equivalent forms of weak convergence

Theorem 11 (Portmanteau theorem on $\mathbb{R}$)**.** *The following are equivalent:*

1. $\mu_n \Rightarrow \mu$
2. $\int f\, d\mu_n \to \int f\, d\mu$ *whenever $f : \mathbb{R} \to \mathbb{R}$ is bounded and continuous.*
3. $\mu_n(A) \to \mu(A)$ *for all Borel $A \subset \mathbb{R}$ with $\mu(\partial A) = 0$*

Proof. (1 => 2): By Skorokhod's theorem, there exists $Y_n \sim \mu_n, Y \sim \mu$ on the same probability space and $Y_n \to Y$ a.s. If f is a bounded function with $\mu(D_f) = 0$, then $P(Y \in D_f) = \mu(D_f) = 0$, thus $f(Y_n) \to f(Y)$ a.s. By the bounded convergence theorem, we have $\int f\, d\mu_n = \mathbb{E}\, f(Y_n) \to \mathbb{E}\, f(Y) = \int f\, d\mu$. Thus for a bounded function f, if $\mu(D_f) = 0$, we have $\int f\, d\mu_n \to \int f\, d\mu$.

(1 => 3) This is a special case of the proof in (1 => 2). Let $f = \mathbb{1}_A$, then $D_f = \partial A$, thus $\mu(D_f) = 0$. This give $\mu_n(A) = \int f\, d\mu_n \to \int f\, d\mu = \mu(A)$.

(3 => 1) This is trivial

(2 => 1) Fix x. Let f_ε be the following function

$$f_\varepsilon(y) = \begin{cases} 1 & \text{if } y \le x \\ 0 & \text{if } y \ge x \\ \text{linear} & \text{otherwise} \end{cases}$$

We have $\mathbb{1}_{(-\infty,x]} \le f_\varepsilon \le \mathbb{1}_{(-\infty,x+\varepsilon]}$. Since f is bounded and continuous, $\limsup_n \mu_n(-\infty, x] \le \limsup_n \int f_\varepsilon\, d\mu_n = \int f_\varepsilon\, d\mu \le \mu(-\infty, x+\varepsilon]$. Let $\varepsilon \downarrow 0$, we have $\limsup_n F_n(x) \le F(x)$. Similarly, we can shoe $\liminf_n F_n(x) \ge F(x-)$. If F is continuous at x, then $\lim_n F_n(x) = F(x)$

□

4.4 Tightness and subsequences

It is often useful to extract weakly convergent subsequences from a sequence of probability measures.

Theorem 12 (Helly's selection theorem)**.** *Let (F_n) be a sequence of distribution functions on $\mathbb{R}$. There exists a subsequence (F_{n_k}) that converges to a non-negative, non-decreasing, right continuous function F at all continuity points of F.*

Remark 3. Note that the function F might not be a probability distribution function. Consider this example: $F_n = \mathbb{1}_{[n,\infty)}$, then $F_n \to 0$ point-wise but 0 is not a probability distribution function.

The exception is the above remark can be thought as the mass of the μ_n's is "escaping" to ∞. Therefore, we can make the define a tightness condition to prevent that.

Definition 3 (Tightness)**.** A sequence of probability measures (on $\mathbb{R}$) (μ_n) is tight if for all $\varepsilon > 0$, there exists a interval $[a, b]$ s.t

$$\mu_n([a, b]) > 1 - \varepsilon \quad , \forall n$$

Remark 4. If you only have one probability measure, it is always tight. The point of this definition is that the interval $[a, b]$ works for all n. In some sense, it means all μ_n are concentrated on the same compact set.

Theorem 13. *(μ_n) is tight if and only if each subsequence (μ_{n_k}) has a further subsequence that converges weakly to a probability measure.*

Proof. For the only if direction: Given $\varepsilon > 0$, there exists $[a, b]$ s.t $\mu_n([a, b]) > 1 - \varepsilon$ for all n. Given subseqeunce (υ_n), there exists further subsequence of distribution functions that converges to F at F's continuity points by Helly's selection theorem. It suffices to show $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

Pick continuity point $a' < a$, then $F(a') = \lim_k F_{n_k}(a') = \lim_k 1 - \upsilon_{n_k}((a', \infty)) \leq \lim_k 1 - \upsilon_{n_k}([a, b]) < \varepsilon$. This shows $\lim_{x\to-\infty} F(x) = 0$. Similarly, pick continuity point $b' > b$, then $F(b') = \lim_k F_{n_k}(b') \geq \limsup_k F_{n_k}([a, b]) > 1 - \varepsilon$. This shows that $\lim_{x\to\infty} F(x) = 1$.

For the if direction: If (μ_n) is not tight, then there exists $\varepsilon > 0$ s.t for all k, there exists n_k s.t $\mu_{n_k}([-k, k]) \leq 1 - \varepsilon$. Consider this subsequence (μ_{n_k}), if there exists a sub-subsequence $(\mu_{n_{kj}})$ s.t it converges weakly to a probability measure μ, then by Pormanteau theorem, we can find $[a, b]$ s.t $\mu(\{a\}) = \mu(\{b\}) = 0$ and $\lim_j \mu_{n_{kj}}([a, b]) = \mu([a, b]) > 1 - \varepsilon$. Since $\lim_j \mu_{n_{kj}}([a, b]) \leq \liminf_j \mu_{n_{kj}}([-k_j, k_j]) \leq 1 - \varepsilon$, a contradiction. □

Corollary 3. *If (μ_n) is a tight sequence of probability measures, and if each subsequence that converges weakly at all converges weakly to the probability measure μ. Then $\mu_n \Rightarrow \mu$.*

Proof. By the above theorem, every subsequence (μ_{n_k}) contains a further subsequence $(\mu_{n_{k(j)}})$ that converges. By the assumption, it converges to μ_n. If $\mu_n \not\Rightarrow \mu$, there exists x, $\mu\{x\} = 0$ and $\mu_n(-\infty, x] \not\to \mu(-\infty, x]$, that is, $|\mu_{n_k}(-\infty, x] - \mu(-\infty, x]| \geq \varepsilon$ for some sequence $\{n_k\}$. This means any subsequence of (μ_{n_k}) does not converge to μ, a contradiction. □

4.5 Integration to the Limit

This subsection contains some limiting theorems that appeared previously but with Skorokhod's Theorem, the assumptions can be weakened.

Theorem 14. *If $X_n \Rightarrow X$ and the X_n are uniformly integrable, then X is integrable and*

$$\mathbb{E}\, X_n \to \mathbb{E}\, X$$

Proof. This follows immediately from Skorokhod's Theorem and uniform integrability. □

If there exists $\varepsilon > 0$ s.t $\sup_n \mathbb{E}\,|X_n|^{1+\varepsilon} < \infty$, then (X_n) is uniformly integrable. Indeed, we have

$$\int_{|X_n|\geq a} |X_n|\,dP \leq \frac{1}{a^\varepsilon}\,\mathbb{E}\,|X_n|^{1+\varepsilon}$$

Since $X_n \Rightarrow X$ implies $X_n^r \Rightarrow X$, we have the following theorem.

Corollary 4. *Let r be a positive integer. If $X_n \Rightarrow X$ and $\sup_n \mathbb{E}\,|X_n|^{r+\varepsilon}| < \infty$ where $\varepsilon > 0$, then $\mathbb{E}\,|X|^r < \infty$ and $\mathbb{E}\,X_n^r \to \mathbb{E}\,X^r$*

We state another condition for uniform integrability.

Proposition 11. *If there exists integrable random variable Z s.t $P(|X_n| \geq t) \geq P(|Z| \geq t)$ for $t > 0$, then (X_n) is uniformly integrable.*

Proof. We recall a few facts: If X is a non-negative random variable, then $\mathbb{E}\,X = \int_0^\infty P(X > t)\,dt = \int_0^\infty P(X \geq t)\,dt$. For $a \geq 0$, replace X with $X\mathbb{1}(X > a)$ (Here X can be any random variable, not only non-negative), then we have

$$\int_{\{X>a\}} X\,dP = aP(X > a) + \int_a^\infty P(X > t)\,dt$$

Back to the proposition. With the equality above, we have

$$\begin{aligned}\int_{\{|X_n|>a\}} |X_n|\,dP &= aP(|X_n| > a) + \int_a^\infty P(|X_n| > t)\,dt \\ &\leq aP(|Z| > a) + \int_a^\infty P(|Z| > t)\,dt = \int_{\{|Z|>a\}} |Z|\,dP\end{aligned}$$

Then uniform integrability follows from DCT. □

5 Characteristic Functions

5.1 Lebesgue integral of complex functions

Definition 4. A complex-valued function on Ω has the form $f(\omega) = g(\omega) + ih(\omega)$, where g, h are ordinary finite-valued real functions on Ω. The integral of f is defined as

$$\int f\,d\mu = \int g\,d\mu + i\int h\,d\mu$$

f is integrable if g and h are integrable by definition.

Theorem 15. *Here are some properties of complex-valued integrals extended from real-valued integrals:*

- *f is integrable if and only if $\int |f|\,d\mu < \infty$*
- *$\left|\int f\,d\mu\right| \le \int |f|\,d\mu$*

Proof. Because $\max\{|g|,|h|\} \le |f| \le |g| + |h|$, the first point follows. For the second point, we start with when g, h are simple functions corresponding to the same partitions, then the result follows by triangle inequality. For general integrable f, we have simple functions $g_k \to g$, $h_k \to h$ and $|g_k|, |h_k| \le |f|$. By DCT, we have

$$\begin{aligned}\left|\int f\,d\mu\right| = \left|\int g\,d\mu + i\int h\,d\mu\right| &= \left|\lim_k \int g_k\,d\mu + i\int h_k\,d\mu\right| \\ &= \lim_k \left|\int g_k\,d\mu + i\int h_k\,d\mu\right| \\ &\le \lim_k \int |f_k|\,d\mu = \int |f|\,d\mu\end{aligned}$$

We can swap limit with $|\cdot|$ because if $a_k \to a$, $b_k \to b$, then $\lim_k |a_k + ib_k| = \lim_k \sqrt{a_k^2 + b_k^2} = \sqrt{a^2+b^2} = |a+ib|$. □

5.2 Basic Properties

Definition 5. The characteristic function of a probability measure μ on $\mathbb{R}$ is defined as

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx}\,\mu(dx) = \int_{-\infty}^{\infty} \cos tx\,\mu(dx) + i\int_{-\infty}^{\infty} \sin tx\,\mu(dx)$$

For a random variable X, it's characteristic function is $\mathbb{E}\,e^{itX}$

Theorem 16 (Basic properties of characteristic function)**.** *Here are some basic properties:*

1. *$\phi(0) = 1$ and $|\phi(t)| \le 1$ for all t.*
2. *ϕ is uniformly continuous on $\mathbb{R}$.*
3. *(Riemann-Lebesgue Lemma) If μ has a density, then $\phi(t) \to 0$ as $|t| \to \infty$.*
4. *If X, Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.*

Proof. 1. Use the fact that $|\int f| \le \int |f|$.

2. $|\phi(t+h)-\phi(t)| \le \mathbb{E}(|e^{i(t+h)X}-e^{itX}|) = \mathbb{E}(|e^{itX}||e^{ihX}-1|) = \mathbb{E}(|e^{ihX}-1|) \to 0$ as $h \to 0$ by bounded convergence theorem. Notice the latter does not depend on t, thus it is uniform.

□

We would like to expand ϕ as a Taylor series. This is possible in many cases, unlike moment generating functions. By Taylor's theorem and integration by part, we have for all $x \in \mathbb{R}, n \ge 0$

$$\left| e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right| \le \min\left\{ \frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right\}$$

Using this, if X has n moments, then

$$\left| \phi(t) - \sum_{k=0}^{n} \frac{(it)^k}{k!} \mathbb{E}(X^k) \right| \le \mathbb{E} \min\left\{ \frac{|tX|^{n+1}}{(n+1)!}, \frac{2|tX|^n}{n!} \right\}$$

If X has all moments and if RHS$\to 0$, then $\phi(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \mathbb{E}(X^k)$. If $\lim_{n\to\infty} \mathbb{E}\,|X|^n \frac{t^n}{n!} = 0$, then the RHS$\to 0$, and we have the Taylor's expansion. This is a condition on the growth of the moments of X. Moreover, if $\lim_{n\to\infty} \mathbb{E}\,|X|^n \frac{t^n}{n!} = 0$ holds for some $t \neq 0$, then we can differentiate and obtain

$$\phi^{(k)}(0) = i^k \,\mathbb{E}(X^k), \quad k \ge 0$$

The above in fact holds under a much weaker assumption

Proposition 12. *If $\mathbb{E}\,|X|^k < \infty$, then $\phi^k(0) = i^k \,\mathbb{E}\, X^k$*

5.3 The Continuity Theorem

Theorem 17. *Let μ_n, μ be probability measures with characteristic functions ϕ_n, ϕ. Then $\mu_n \Rightarrow \mu$ if and only if $\phi_n(t) \to \phi(t)$ for each t.*

Proof. For the necessity part, it follows from the fact that $\sin, \cos$ are continuous and bounded in x. Apply Portmanteau theorem and this to the real and imaginary parts of ϕ_t.

For the sufficient part, we first assume (μ_n) is tight. If (μ_n) is tight, then there exists a weakly convergent subsequence. For any subsequence (μ_{n_k}) that converges weakly at all, i.e $\mu_{n_k} \Rightarrow \upsilon$ for some υ, by the necessity part, $\phi_{n_k}(t) \to \phi_\upsilon(t)$ for each t. Therefore, since $\phi_n(t) \to \phi(t)$, $\phi_\upsilon(t) = \phi(t)$ for each t. This shows $\mu = \upsilon$. Hence, for any subsequence μ_{n_k} that converges weakly at all converges weakly to μ. By previous theorem, $\mu_n \Rightarrow \mu$

I now show (μ_n) is tight. By Fubini's Theorem, we have

$$\begin{aligned}\frac{1}{u}\int_{-u}^{u}(1-\phi_n(t))\,dt &= \frac{1}{u}\int_{-\infty}^{\infty}\int_{-u}^{u}1-e^{itx}\,dt\,\mu_n(dx)\\ &= 2\int_{-\infty}^{\infty}\left(1-\frac{\sin ux}{ux}\right)\mu_n(dx)\\ &\geq 2\int_{-\infty}^{\infty}1-\frac{1}{|ux|}\,\mu_n(dx)\\ &\geq 2\int_{\{|x|\geq 2/u\}}\frac{1}{2}\,\mu_n(dx)=\mu_n\{x:|x|\geq 2/u\}\end{aligned}$$

Since ϕ is continuous at 0 and $\phi(0)=1$, for $\varepsilon>0$, there exists $u>0$ small enough s.t $\frac{1}{u}\int_{-u}^{u}(1-\phi(t))\,dt<\varepsilon$. Also, because $\phi_n(t)\to\phi(t)$, by the bounded convergence, there exists n_0 s.t $\forall\, n\geq n_0$, $\frac{1}{u}\int_{-u}^{u}(1-\phi_n(t))\,dt<2\varepsilon$. Let $a=2/u$, then for all $n\geq n_0$, $\mu_n\{x:|x|\geq a\}<2\varepsilon$. By increasing a, we can ensure this inequality also holds for the finitely many n preceding n_0. □

Corollary 5. *Suppose that* $\lim_n\phi_n(t)=g(t)$ *for each* t*, where the limit function* g *is continuous at 0. Then there exists a* μ *s.t* $\mu_n\Rightarrow\mu$*, and* μ *has characteristic function* g*.*

Proof. Notice in the proof above, to establish tightness of (μ_n), we only need the continuity of ϕ at $t=0$. (we also need $\phi(0)=1$ but this is established already because $1=\phi_n(0)\to g(0)$). Hence we can prove (μ_n) is tight.

Take any subsequence (μ_{n_k}) that converges weakly to some υ (Existence of this subsequence is implied by tightness). Then $\phi_\upsilon=\lim_n\phi_n=g$. Thus g is the c.f of ϕ_υ. This shows for any subsequence that converges weakly at all converges to a distribution with c.f g. Let μ be the measure associated with g. Then $\mu_n\Rightarrow\mu$. □

In this proof, the continuity of g is used to establish tightness. Hence if (μ_n) is assumed to be tight in the first place, the hypothesis of continuity can be suppressed:

Corollary 6. *Suppose that* $\lim_n\phi_n(t)=g(t)$ *exists for each* t *and that* (μ_n) *is tight. Then there exists a* μ *s.t* $\mu\Rightarrow\mu$*, and* μ *has characteristic function* g*.*

6 Central Limit Theorem

6.1 Classic central limit theorem

Lemma 1. *Let* u,v *be complex numbers and* $|u|,|v|\leq 1$*. Then we have*

$$|u^n-v^n|\leq n|u-v|$$

for all $n \in \mathbb{N}$.

Proof. □

Theorem 18 (Classic CLT)**.** *Let* (X_n) *be iid random variables with mean 0 and variance 1. Then we have*

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \Rightarrow N(0,1)$$

Proof. The idea is to show that cf of $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ converges point-wise to the cf of $N(0,1)$ and use the continuity theorem. The cf of $\frac{X_i}{\sqrt{n}}$ is $\mathbb{E}\, e^{i\frac{t}{\sqrt{n}}X_i} = \phi(\frac{t}{\sqrt{n}})$, thus the cf of $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ is $\phi(\frac{t}{\sqrt{n}})^n$. Since $\mathbb{E}\,|X_i|^2 < \infty$, ϕ is twice differentiable, hence by Taylor's theorem, we have

$$\begin{aligned}\phi(s) &= \phi(0) + \phi'(0)s + \phi''(0)\frac{s^2}{2} + o(s^2)\\ &= 1 - \frac{s^2}{2} + o(s^2)\end{aligned}$$

For fix t, plug $\frac{t}{\sqrt{n}}$ into s, we have $\phi(\frac{t}{\sqrt{n}})^n = \left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n$. Since $\left(1 - \frac{t^2}{2n}\right)^n \to e^{-\frac{t^2}{2}}$ as $n \to \infty$, which is the cf of $N(0,1)$, it suffices to show $\left|\left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n - \left(1 - \frac{t^2}{2n}\right)^n\right| \to 0$ as $n \to \infty$

By the above lemma, letting $u = \left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n, v = \left(1 - \frac{t^2}{2n}\right)^n$, we have $|u - v| \le n \cdot o(n^{-1}) \to 0$. □

6.2 Lindeberg CLT

Lemma 2. *Let* u_i, v_i *be complex numbers and* $|u_i|, |v_i| \le 1$*. Then* $|\prod_{i=1}^n u_i - \prod_{i=1}^n v_i| \le \sum_{i=1}^n |u_i - v_i|$.

We would like to weaken the assumption of the classic CLT. Suppose the X_i's all have mean 0 and variance 1 but are not necessary identically distributed, then do we still have CLT?

If we inspect the proof of the classic CLT, we would realize the important step is show the error term $o(n^{-1})$ is small. Suppose ϕ_i is the cf of X_i, then $\left|\phi_i(t) - \sum_{k=0}^n \mathbb{E}\,|X_i|^k \frac{(it)^k}{k!}\right| \le \mathbb{E}\min\{\frac{|tX_i|^{n+1}}{(n+1)!}, \frac{2|tX_i|^n}{n!}\}$. Plug $n = 2$, we have $\left|\phi_i(t) - (1 - \frac{t^2}{2})\right| \le \mathbb{E}\min\{\frac{|tX_i|^3}{6}, |tX_i|^2\}$. Combining this with the above lemma, we have the following:

Theorem 19 (Lindeberg CLT - Simplified version)**.** *Let X_i be independent with mean 0 and variance 1. Assume for each $\epsilon > 0$, we have*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\, X_i^2 \mathbb{1}_{\{|X_i| \geq \epsilon\sqrt{n}\}} \to 0, \quad n \to \infty$$

then we have

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \Rightarrow N(0,1)$$

Proof. Fix n and plug $\frac{t}{\sqrt{n}}$ in ϕ_i in the inequality in the discussion above, we have

$$\begin{aligned}
\left|\phi_i(\frac{t}{\sqrt{n}}) - (1 - \frac{t^2}{2n})\right| &\leq \mathbb{E} \min\{\frac{|tX_i|^3}{6n^{3/2}}, n^{-1}|tX_i|^2\} \\
&\leq \mathbb{E}\, \frac{|tX_i|^3}{6n^{3/2}} \mathbb{1}_{\{|X_i| < \epsilon\sqrt{n}\}} + \mathbb{E}\, n^{-1}|tX_i|^2 \mathbb{1}_{\{|X_i| \geq \epsilon\sqrt{n}\}} \\
&\leq \frac{t^3\epsilon}{6\sqrt{n}}\, \mathbb{E}\, \frac{X_i^2}{n} \mathbb{1}_{\{|X_i| < \epsilon\sqrt{n}\}} + \frac{t^2}{n}\, \mathbb{E}\, X_i^2 \mathbb{1}_{\{|X_i| \geq \epsilon\sqrt{n}\}} \\
&\leq \frac{t^3\epsilon^3}{6n} + \frac{t^2}{n}\, \mathbb{E}\, X_i^2 \mathbb{1}_{\{|X_i| \geq \epsilon\sqrt{n}\}}
\end{aligned}$$

Therefore, by the above lemma

$$\begin{aligned}
\left|\prod_{i=1}^{n} \phi_i(\frac{t}{\sqrt{n}}) - (1 - \frac{t^2}{2n})^n\right| &\leq \sum_{i=1}^{n} \frac{t^3\epsilon^3}{6n} + \frac{t^2}{n}\, \mathbb{E}\, X_i^2 \mathbb{1}_{\{|X_i| \geq \epsilon\sqrt{n}\}} \\
&\leq \frac{t^3\epsilon^3}{6} + t^2 \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\, X_i^2 \mathbb{1}_{\{|X_i| \geq \epsilon\sqrt{n}\}} \\
&\to c \cdot t^3\epsilon^3, \quad n \to \infty
\end{aligned}$$

□

We can further weaken the assumption of independence by considering triangular arrays $(X_{n,k}, k = 1, \ldots, k_n)_n$, where independence only holds in each row. The variance condition can also be generalized to arbitrary variance, which in turns gives us the following CLT:

Theorem 20 (Lindeberg CLT)**.** *For $n \geq 1$, $(X_{n,k}, k = 1, \ldots, k_n)$ are independent random variables. Suppose $\mathbb{E}\, X_{n,k} = 0$ for all n, k, $Var(X_{n,k}) = \sigma_{n,k}^2$, $Var(S_n) = s_n^2$ where $S_n = \sum_{k=1}^{k_n} X_{n,k}$. Assume for all $\varepsilon > 0$, $\frac{1}{s_n^2} \sum_{k=1}^{k_n} \mathbb{E}\, X_{n,k}^2 \mathbb{1}_{\{|X_{n,k}| \geq \varepsilon s_n\}} \to 0$ as $n \to \infty$, then we have*

$$\frac{S_n}{s_n} \Rightarrow N(0,1)$$

Proof. The cf for $\frac{X_{n,k}}{s_n}$ is $\phi_{n,k}(\frac{t}{s_n})$, where $\phi_{n,k}$ is the cf of $X_{n,k}$. Similar to the proof of the simplified version of Lindeberg CLT, we can show $\sum_{k=1}^{k_n}\left|\phi_{n,k}(\frac{t}{s_n})-\left(1-\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right)\right|\to 0$ as $n\to\infty$. Therefore, according to the lemma at the beginning of this subsection, it suffices to show $\max_{1\le k\le k_n}\left|1-\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right|\le 1$ for large enough n to invoke the lemma, and then show $\prod_{k=1}^{k_n}\left(1-\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right)\to e^{\frac{-t^2}{2}}$ as $n\to\infty$.

For $\varepsilon>0$,

$$\begin{aligned}\frac{\sigma_{n,k}^2}{s_n^2}&=\frac{1}{s_n^2}\,\mathbb{E}\,X_{n,k}^2\mathbb{1}_{\{|X_{n,k}|<\varepsilon s_n\}}+\frac{1}{s_n^2}\,\mathbb{E}\,X_{n,k}^2\mathbb{1}_{\{|X_{n,k}|\ge\varepsilon s_n\}}\\&\le\varepsilon^2+\frac{1}{s_n^2}\sum_{k=1}^{k_n}\mathbb{E}\,X_{n,k}^2\mathbb{1}_{\{|X_{n,k}|\ge\varepsilon s_n\}}\to\varepsilon^2\end{aligned}$$

Therefore, $\max_{1\le k\le k_n}\frac{\sigma_{n,k}^2}{s_n^2}\to 0$. This shows $\max_{1\le k\le k_n}\left|1-\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right|\le 1$ for large enough n.

For the second part, notice that

$$\begin{aligned}\left|\prod_{k=1}^{k_n}\left(1-\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right)-e^{\frac{-t^2}{2}}\right|&=\left|\prod_{k=1}^{k_n}\left(1-\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right)-\prod_{k=1}^{k_n}e^{\frac{-t^2\sigma_{n,k}^2}{2s_n^2}}\right|\\&\le\sum_{k=1}^{k_n}\left|1-\frac{t^2\sigma_{n,k}^2}{2s_n^2}-e^{\frac{-t^2\sigma_{n,k}^2}{2s_n^2}}\right|\end{aligned}$$

Since for any z, we have $|e^z-1-z|=\sum_{n=2}^{\infty}\frac{|z|^n}{n!}=|z|^2\sum_{n=2}^{\infty}\frac{|z|^{n-2}}{n!}\le|z|^2e^{|z|}$, by letting $z=\frac{-t^2\sigma_{n,k}^2}{2s_n^2}$, we have

$$\begin{aligned}\sum_{k=1}^{k_n}\left|1-\frac{t^2\sigma_{n,k}^2}{2s_n^2}-e^{\frac{-t^2\sigma_{n,k}^2}{2s_n^2}}\right|&\le\sum_{k=1}^{k_n}\frac{t^4\sigma_{n,k}^4}{4s_n^4}\exp\left(\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right)\\&\le\frac{t^4}{4}\cdot\max_{1\le k\le k_n}\frac{\sigma_{n,k}^2}{s_n^2}\exp\left(\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right)\cdot\sum_{k=1}^{k_n}\frac{\sigma_{n,k}^2}{s_n^2}\end{aligned}$$

The second term converges to 0 since $\frac{\sigma_{n,k}^2}{s_n^2}\to 0$, $\exp\left(\frac{t^2\sigma_{n,k}^2}{2s_n^2}\right)\to 1$ by the first part of the proof. The third term is 1. Hence, the whole term converges to 0. □

Remark 5 (Lyapunov condition). A slightly simpler condition implies the Lindeberg condition: if for some $q>2$, we have $\frac{1}{s_n^q}\sum_{k=1}^{k_n}\mathbb{E}\,|X_{n,k}|^q\to 0$, then the Lindeberg

condition holds. The proof is as follows:

$$\frac{1}{s_n^2}\sum_{k=1}^{k_n}\mathbb{E}\,X_{n,k}^2\mathbb{1}_{\{|X_{n,k}|\geq\varepsilon s_n\}} \leq \frac{1}{s_n^2}\sum_{k=1}^{k_n}\mathbb{E}\,X_{n,k}^2\frac{|X_{n,k}|^{q-2}}{|X_{n,k}|^{q-2}}\mathbb{1}_{\{|X_{n,k}|\geq\varepsilon s_n\}}$$
$$\leq \frac{1}{\varepsilon^{q-2}}\frac{1}{s_n^q}\sum_{k=1}^{k_n}\mathbb{E}\,|X_{n,k}|^q \to 0$$

Using the Lyapunov condition, we can show the following: If $|X_{n,k}| \leq M_n$ for all n and $\frac{M_n}{s_n} \to 0$, then $\frac{S_n}{s_n} \Rightarrow N(0,1)$. The proof is by taking $q = 3$, then $\frac{1}{s_n^3}\sum_{k=1}^{k_n}\mathbb{E}\,|X_{n,k}|^3 \leq \frac{M_n}{s_n}\cdot\frac{1}{s_n^2}\sum_{k=1}^{k_n}\mathbb{E}\,X_{n,k}^2 \to 0$ since the first term converges to 0 by assumption and the second term is 1.

7 Limit Theorems in $\mathbb{R}^d$

7.1 Weak convergence in $\mathbb{R}^d$

Definition 6. A sequence (μ_n) of Borel probability measures on $\mathbb{R}^d$ converges weakly to μ if $F_n(x) \to F(x)$ for all continuity points x of F, where F_n, F are distribution functions of μ_n, μ respectively.

Theorem 21 (Portmanteau in $\mathbb{R}^d$)**.** *Let $(\mu_n), \mu$ be probability measures on $\mathbb{R}^d$ with distribution functions $(F_n), F$. The following are equivalent:*

(a) $\mu_n \Rightarrow \mu$

(b) $\forall$ *bounded continuous function* $f : \mathbb{R}^d \to \mathbb{R}$, $\int f\,d\mu_n \to \int f\,d\mu$

(c) $\mu_n(A) \to \mu(A)$ *for all Borel* $A \subset \mathbb{R}^d$ *s.t* $\mu(\partial A) = 0$.

(d) $\limsup_n \mu_n(C) \leq \mu(C)$ *for all closed* $C \subset \mathbb{R}^d$.

(e) $\limsup_n \mu_n(O) \geq \mu(O)$ *for all open* $O \subset \mathbb{R}^d$.

Proof. $(a) \Leftrightarrow (b) \Leftrightarrow (c)$: refer to the book. $(d) \Leftrightarrow (e)$ is obvious.

$(d) + (e) \Rightarrow (c)$: We have $\limsup \mu_n(A) \leq \limsup \mu_n(\overline{A}) \leq \mu(\overline{A}) = \mu(A^\circ) + \mu(\partial A) = \mu(A^\circ) \leq \mu(A)$. On the otherhand, $\liminf \mu_n(A) \geq \liminf \mu_n(A^\circ) \geq \mu(A^\circ) = \mu(A) - \mu(\partial A) = \mu(A)$. This show $\mu_n(A) \to \mu(A)$ when $\mu(\partial A) = 0$. □

Theorem 22 (Mapping theorem on $\mathbb{R}^d$)**.** *Let (μ_n) be a sequence of probability measures on $\mathbb{R}^d$ that converges weakly to μ. If $f : \mathbb{R}^d \to \mathbb{R}^k$, write D_f for its set of discontinuities. If $\mu(D_f) = 0$, then $\mu_n f^{-1} \Rightarrow \mu f^{-1}$.*

Proof. It suffices to show $\limsup_n \mu_n(f^{-1}(C) \leq \mu(f^{-1}(C))$ for closed C. We have $\limsup_n \mu_n(f^{-1}(C) \leq \limsup_n \mu_n(\overline{f^{-1}(C)} \leq \mu(\overline{f^{-1}(C)}) = \mu(f^{-1}(C)) + \mu(\partial f^{-1}(C) \setminus f^{-1}(C))$. Thus, we're left to show $\mu(\partial f^{-1}(C) \setminus f^{-1}(C)) = 0$.

I'll show $\partial f^{-1}(C) \setminus f^{-1}(C) \subset D_f$. If not, then for $x \in \partial f^{-1}(C) \setminus f^{-1}(C)$, since it belongs to the boundary, there exists $x_n \to x$, $x_n \in f^{-1}(C)$. Since $x \notin D_h$, $f(x_n) \to f(x)$ and because C is closed, $f(x) \in C$, a contradiction. □

Definition 7 (Tightness in R^d)**.** A sequence (μ_n) of probability measures on $\mathbb{R}^d$ is tight if $\forall\, \varepsilon > 0$, there exists compact set $K \subset \mathbb{R}^d$ s.t

$$\mu_n(K) \geq 1 - \varepsilon$$

for all n.

As in one dimension, we have

Theorem 23. *If (μ_n) is tight, then there is a probability measure on $\mathbb{R}^d$, μ, and a subsequence (μ_{n_k}) s.t $\mu_{n_k} \Rightarrow \mu$.*

7.2 Multidimensional characteristic functions

Definition 8. Let μ be a probability measure on $\mathbb{R}^d$. The characteristic function of μ is $\phi_\mu : \mathbb{R}^d \to \mathbb{C}$, given by

$$\phi_\mu(t) = \int e^{it\cdot x} d\mu(x)$$

Here are some similar properties to the one-dimensional characteristic function

- The characteristic function uniquely identifies the measure by an inversion formula: if $A \subset \mathbb{R}^d$ is a rectangle of the form $A = \prod_{j=1}^d (a_j, b_j]$ with $\mu(\partial A) = 0$, then
$$\mu(A) = \lim_{T\to\infty} \frac{1}{(2\pi)^d} \int_{\{|t_j|\leq T, \forall j\}} \prod_{j=1}^d \frac{e^{-it_j a_j} - e^{-it_j b_j}}{it_j} \phi_\mu(t) dt$$

- (Cramer-Wold device) Let $X = (X_1, \ldots, X_d)$ be a random vector. For fixed $t \in \mathbb{R}^d$, we can view $t \cdot X$ as a random variable. Its characteristic function is $\phi_{t\cdot X}(s) = \mathbb{E}\, e^{is(t\cdot X)}$, for $s \in \mathbb{R}$. We have the following claim:

  Claim 1. *The law of X is uniquely determined by the values it takes on half-space, i.e, sets of the form $\{x : t \cdot x \leq \alpha\}$ for all $t \in \mathbb{R}^d, \alpha \in \mathbb{R}$.*

Proof. If we know the value of the law of X on half-spaces, we know the distribution of $t \cdot X$ for all $t \in \mathbb{R}^d$. thus know their characteristic functions. Take $s = 1$ in the above formula, we obtain the value of $\mathbb{E}\, e^{it \cdot X}$ for all $t \in \mathbb{R}^d$, so we obtain the characteristic function of X. Hence, by the inversion formula, we obtain the distribution of X. □

- (Continuity theorem) $\mu_n \Rightarrow \mu$ if and only if $\phi_n(t) \to \phi(t)$ point-wise.

Proposition 13. *Let $X_n = (X_{1,n}, \ldots, X_{d,n})$ and $Y = (Y_1, \ldots, Y_d)$ be random vectors. Then $X_n \Rightarrow Y$ if and only if $X_n \cdot t \Rightarrow Y \cdot t \quad \forall t \in \mathbb{R}^d$.*

Proof. For the only if part, because inner product is a continuous function, we have $X_n \cdot t \Rightarrow Y \cdot t \quad \forall t \in \mathbb{R}^d$.

For the if part, by the one-dimensional continuity theorem, $\mathbb{E}\, e^{is(X_n \cdot t)} \to \mathbb{E}\, e^{is(Y \cdot t)} \quad \forall s \in \mathbb{R}, \forall t \in \mathbb{R}^d$. Take $s = 1$, by the d-dimensional continuity theorem, we have $X_n \Rightarrow Y$. □

7.3 Gaussian vectors

Definition 9. A random vector $X = (X_1, \ldots, X_d)$ has a (centered, or mean zero) Gaussian distribution if its characteristic function has the form

$$\phi_X(t) = \exp(-\frac{(tA) \cdot (tA)}{2}), \quad t \in \mathbb{R}^d$$

for some dxd matrix A. (Here t is viewed as a row vector, $t \cdot t = tt^T$)

Proposition 14. *Let A be a dxd matrix and X be a standard d-dimensional Gaussian vector.*

1. *The vector $Y = AX$ is a d-dimensional Gaussian associated to the matrix A.*
2. *Define $\Sigma_{ij} = \mathbb{E}\, Y_i Y_j$. Then $\Sigma = (\Sigma_{ij})_{i,j}$, the covariance matrix of Y satisfies $\Sigma = AA^T$.*
3. *If Y and Y' are d-dimensional Gaussians with the same covariance matrix, then they have the same distribution.*

Proof. 1.

$$\begin{aligned}\mathbb{E}\, e^{itX} &= \mathbb{E} \exp\left(i \sum_{j=1}^{d} t_j \left[\sum_{k=1}^{d} A_{j,k} X_k\right]\right) \\ &= \exp\left(i \sum_{k=1}^{d} X_k \left[\sum_{j=1}^{d} t_j A_{j,k}\right]\right) \\ &= \exp\left(i \sum_{k=1}^{d} X_k (tA)_k\right) = \prod_{k=1}^{d} \exp\left(\frac{(tA)_k^2}{2}\right) \\ &= \exp\left(-\frac{\sum_{k=1}^{d} (tA)_k^2}{2}\right) = \exp\left(-\frac{(tA)\cdot(tA)}{2}\right)\end{aligned}$$

2.

$$\mathbb{E}\, Y_i Y_j = \mathbb{E} \sum_{k=1}^{d} A_{ik} X_k \sum_{l=1}^{d} A_{jl} X_l = \mathbb{E} \sum_{k=1}^{d} A_{ik} A_{jk} X_k^2 = \sum_{k=1}^{d} A_{ik} A_{jk} = (AA^T)_{ij}$$

3. Assume $AX = Y, A'X = Y'$ and Σ, Σ' are covariance matrices of Y, Y'. Then $\Sigma = AA^T, \Sigma' = A'(A')^T$ by (2). $\phi_Y(t) = \exp(-\frac{(tA)\cdot(tA)}{2}) = \exp(-\frac{t\Sigma t^T}{2}) = \exp(-\frac{t\Sigma' t^T}{2}) = \exp(-\frac{(tA')\cdot(tA')}{2}) = \phi_{Y'}(t)$. Since Y, Y' have the same cf, they have the same distribution.

□

Remark 6.

- Σ is positive semi-definite since $x^T \Sigma x = (A^T x) \cdot (A^T x) \geq 0$ for all $x \in \mathbb{R}^d$
- Conversely, if Σ is an arbitrary positive semi-definite matrix, we can factorize it as AA^T for some matrix A. So we can redefine a Gaussian vector as one with characteristic function

$$\phi(t) = \exp\left(-\frac{t\Sigma t^T}{2}\right)$$

for some positive semi-definite matrix Σ. The two definitions are equivalent: if X has cf $\exp\left(-\frac{tAA^T t^T}{2}\right)$, by letting $\Sigma = AA^T$, we have the second definition; if X has cf $\exp\left(-\frac{t\Sigma t^T}{2}\right)$, by factorizing $\Sigma = AA^T$, the cf is equals to $\exp\left(-\frac{tAA^T t^T}{2}\right)$, which is the first definition.

7.4 Central Limit Theorem

We will assume the random vectors have mean 0 and well-defined covariance matrix. If $X = (X_1, \ldots, X_d)$ is a random vector, then

$$|\mathbb{E}\, X_i X_j| \leq (\mathbb{E}\, X_i^2)^{\frac{1}{2}} (\mathbb{E}\, X_j^2)^{\frac{1}{2}}$$

Therefore, if $\mathbb{E}(X_1^2 + \cdots + X_d^2) < \infty$, then X has a well-defined covariance matrix.

Theorem 24 (d-dimensional CLT)**.** *Let $X^{(1)}, X^{(2)}, \ldots$ be an iid sequence of d-dimensional random vector with mean 0 and $\mathbb{E}(X_1^2 + \cdots + X_d^2) < \infty$. Write Σ for the covariance matrix of $X^{(1)}$. Then*

$$\frac{X^{(1)} + \cdots + X^{(d)}}{\sqrt{n}} \Rightarrow N(0, \Sigma)$$

, where $N(0, \Sigma)$ is a d-dimensional Gaussian distribution with covariance matrix Σ.

Proof. By the above proposition, it suffices to show $t \cdot \left[\frac{X^{(1)}+\cdots+X^{(d)}}{\sqrt{n}}\right] \Rightarrow t \cdot Y$ for $t \in \mathbb{R}^d$ if $Y \sim N(0, \Sigma)$.

The mean of $t \cdot X^{(i)}$ is 0. For the variance, we have

$$Var(t \cdot X^{(i)}) = \sum_{k=1}^{d} \sum_{l=1}^{d} t_k t_l Cov(X^{(k)}, X^{(l)}) = t\Sigma t^T$$

Thus, by the one dimensional CLT, $t \cdot \left[\frac{X^{(1)}+\cdots+X^{(d)}}{\sqrt{n}}\right] \Rightarrow N(0, t\Sigma t^T)$. To show $t \cdot Y \sim N(0, t\Sigma t^T)$, we compare the characteristic functions.

$$\phi_{t \cdot Y}(s) = \mathbb{E}\, e^{i(st \cdot Y)} = \exp(-\frac{st\Sigma(st)^T}{2})$$
$$\phi_{N(0, t\Sigma t^T)})(s) = \exp(-\frac{s^2(t\Sigma t^T)}{2}) = \exp(-\frac{st\Sigma(st)^T}{2})$$

This completes the proof. □