Probability Theory I

Basics of Probability Theory; Law of Large Numbers, Central Limit Theorem and Large Deviation

Seiji HIRABA

October 27, 2023

Contents

1 Bases of Probability Theory

Probability theory is based on measure theory and Lebesgue integrals. However, in this text, we do not assume the readers have enough knowledge on measure theory and Lebesgue integrals.

1.1 Probability spaces and random variables

In the probability theory we investigate various properties of random variables $X = X(\omega)$ which are defined on an appropriate probability space $(\Omega, \mathcal{F}, P)$.

Here the probability space $(\Omega, \mathcal{F}, P)$ is that

- Ω is a (non-empty) set; $\omega \in \Omega$.
- $\mathcal{F}$ ($\subset 2^\Omega$) is a σ-additive class (σ-field) on Ω. (2^Ω is a total family of subsets of Ω. That is, it satisfies the following:
  (1) $\Omega \in \mathcal{F}$
  (2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
  (3) $A_n \in \mathcal{F}$ $(n = 1, 2, \dots) \Rightarrow \bigcup A_n \in \mathcal{F}$
- $P = P(d\omega)$ is a **probability measure** on a measurable space $(\Omega, \mathcal{F})$, i.e., $P : \mathcal{F} \to [0, 1]$ is a set function such that
  (1) $P(\Omega) = 1$
  (2) $A_n \in \mathcal{F}$ $(n = 1, 2, \dots)$ are mutually disjoint $\Rightarrow P(\bigcup A_n) = \sum P(A_n)$ (σ-**additivity**)

Question 1.1 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Show the following hold:*

(1) σ-add. class is closed under operations of a countable number of sets, i.e., if $\mathcal{F}$ is a σ-add. class, then for $A, B, A_n \in \mathcal{F}$, the following are also in $\mathcal{F}$.

$$\emptyset, \quad A \cap B, \quad A \setminus B, \quad A \triangle B := (A \setminus B) \cup (B \setminus A), \quad \bigcap_{n=1}^{\infty} A_n.$$

We also have $\overline{\lim} A_n = \limsup A_n := \bigcap_{N \geq 1} \bigcup_{n \geq N} A_n$, $\underline{\lim} A_n = \liminf A_n := \bigcup_{N \geq 1} \bigcap_{n \geq N} A_n \in \mathcal{F}$, where you may remember as $\overline{\lim} = \inf \sup$, $\underline{\lim} = \sup \inf$.

Note that we simply denote $\overline{\lim} A_n$ as A_n i.o. (= infinitely often), $\underline{\lim} A_n$ as A_n e.f. (= except finite),

(2) $P(\emptyset) = 0$, $A_k \in \mathcal{F}$ $(k = 1, 2, \dots, n)$ are mutually disjoint $\Rightarrow P(\bigcup_{k=1}^n A_k) = \sum_{k=1}^n P(A_k)$ **(finite additivity)**.

(3) $A, B \in \mathcal{F}; A \subset B \Rightarrow P(A) \leq P(B)$ **(monotonicity)**.

(4) $A_n \in \mathcal{F}$, $A_n \uparrow \Rightarrow P\left(\bigcup A_n\right) = \lim_{n \to \infty} P(A_n)$.

(5) $A_n \in \mathcal{F}$, $A_n \downarrow \Rightarrow P\left(\bigcap A_n\right) = \lim_{n \to \infty} P(A_n)$.

(4), (5) are called the **monotone continuity of the probability**

(6) $A_n \in \mathcal{F}$ $(n \geq 1) \Rightarrow P\left(\bigcup A_n\right) \leq \sum P(A_n)$ (σ-**sub additivity**).

(7) **(Borel-Cantelli's lemma)** *$A_n \in \mathcal{F}$ $(n \geq 1)$, $\sum P(A_n) < \infty \Rightarrow P\left(\limsup_{n \to \infty} A_n\right) = 0$, i.e., $P\left(\liminf_{n \to \infty} A_n^c\right) = 1$.*

Ans. (1) $\emptyset = \Omega^c$, $A \cap B = (A^c \cup B^c)^c$, $A \setminus B = A \cap B^c$, by def., and by de Morgan's law.
(2) By σ-additivity, $P(\emptyset) = \infty \cdot P(\emptyset)$. $A_k = \emptyset$ for $k > n$. (3) $B = A \cup (B \setminus A)$.
(4) $B_n = A_n \setminus A_{n-1}$ with $A_0 = \emptyset$ and $\sum_{n\geq 1} = \lim_{N\to\infty} \sum_{n=1}^{N}$. (5) complement. (6) $B_n = A_n \setminus \left(\bigcup_{k=1}^{n-1} A_k \right)$.
(7) $\bigcup_{n\geq N} A_n \downarrow \limsup A_n$ and by monotone continuity, σ-sub additivity and def. of infinite sum.

On this probability space $(\Omega, \mathcal{F}, P)$ a function $X = X(\omega) : \Omega \to \mathbf{R}$ is called a **random variable** if $\{X \leq a\} := \{\omega \in \Omega; X(\omega) \leq a\} \in \mathcal{F}$ $({}^\forall a \in \mathbf{R})$. Especially, if X takes in a countable set $S = \{a_j\}_{j\geq 1} \subset \mathbf{R}$, then the above condition is equivalent to $\{X = a_j\} \in \mathcal{F}$ $({}^\forall j \geq 1)$.

Let X_k be a real-valued random variable on $(\Omega, \mathcal{F}, P)$ for $k = 1, 2, \ldots, n$. $\{X_k\}_{k=1}^n$ is **independent** if

$$P(X_1 \leq a_1, \cdots, X_n \leq a_n) = P(X_1 \leq a_1) \cdots P(X_n \leq a_n) \quad ({}^\forall a_k \in \mathbf{R}, k = 1, \ldots, n).$$

Moreover, in case of $n = \infty$, $\{X_k\}_{k\geq 1}$ is independent if ${}^\forall N \geq 1$, $\{X_k\}_{k=1}^N$ is independent. Especially, if X_k takes in $S = \{a_j\}_{j\geq 1}$, then the above condition is equivalent to

$$P(X_1 = b_1, \cdots, X_n = b_n) = P(X_1 = b_1) \cdots P(X_n = b_n) \quad (b_k \in S, k = 1, \ldots, n).$$

Furthermore, $\mu_X(A) = P(X \in A)$ is called a **distribution** of X, and $F(x) = P(X \leq x)$ is called **distribution function** of X.

1.2 Expectations, means

The **expectation** or **means** of a random variable X on a prob. sp. $(\Omega, \mathcal{F}, P)$ is defined as a Lebesgue integral by the probability measure P;

$$EX = E[X] := \int X dP = \int_\Omega X(\omega) P(d\omega)$$

However, here, we give how to define EX for X which is $\overline{\mathbf{Z}} := \mathbf{Z} \cup \{\pm\infty\}$-valued.

(1) If $X \geq 0$, then

$$EX := \sum_{n=0}^{\infty} nP(X = n) + \infty \cdot P(X = \infty).$$

(If $P(X = \infty) = 0$, then $\infty \cdot P(X = \infty) = 0$. If $P(X = \infty) > 0$, then $EX = \infty$.)

(2) If X is in general, then let $X^+ := X \vee 0$, $X^- := (-X) \vee 0$, ($X^\pm \geq 0$, $X = X^+ - X^-$ hold $\to$ show.) and set $EX := EX^+ - EX^-$ except the case of $\infty - \infty$.

Formally, we denote $EX = \sum_{n\in\overline{\mathbf{Z}}} nP(X = n)$. Moreover, for a function $f : \overline{\mathbf{Z}} \to \mathbf{R}$, $Ef(X) = \sum_{n\in\overline{\mathbf{Z}}} f(n)P(X = n)$. (Of course, it can be defined by dividing to $\sum_{n;f(n)>0}$ and $\sum_{n;f(n)<0}$ if at least one is finite.)

For a RV X, a **variance** is defined by $V(X) := E[(X - EX)^2] = E[X^2] - (EX)^2$ (show the last equal). From this, we have $(EX)^2 \leq E[X^2]$.

Theorem 1.1 (Chebyshev's inequality) *Let $p \geq 1$. For ${}^\forall a > 0$,*

$$P(|X| \geq a) \leq \frac{E[|X|^p]}{a^p}.$$

Proof. Since $P(|X| \geq a) = P(|X|^p \geq a^p)$, we may set $p = 1$.

$$E|X| = \sum_{n\geq 1} nP(|X| = n) \geq \sum_{n\geq a} nP(|X| = n) \geq a \sum_{n\geq a} P(|X| = n) = aP(|X| \geq a).$$

More generally, (by using Lebesgue integrals)

$$E|X| = \int_\Omega |X| dP \geq \int_{\{|X|\geq a\}} |X| dP \geq aP(|X| \geq a).$$

■

Theorem 1.2 *Let $X_1, \ldots, X_n$ be $\overline{\mathbf{Z}}$-valued RVs such that $E[X_k^2] < \infty$ $(k = 1, \ldots, n)$. If $X_1, \ldots, X_n$ are independent, then $E[X_j X_k] = E[X_j]E[X_k]$ $(j \neq k)$. Moreover, if $E[X_k] = 0$, then*

$$E\left[\left(\sum_{k=1}^n X_k\right)^2\right] = \sum_{k=1}^n E[X_k^2].$$

Proof. (1) If $j \neq k$, then by the independence $P(X_j = m, X_k = n) = P(X_j = m)P(X_k = n)$. and this implies

$$E[X_j X_k] = \sum_{m,n} mnP(X_j = m, X_k = n) = \sum_{m,n} mnP(X_j = m)P(X_k = n) = E[X_j]E[X_k].$$

(2) By $\left(\sum_{k=1}^n X_k\right)^2 = \sum_{k=1}^n X_k^2 + \sum_{j\neq k} X_j X_k$ and by (1), if $j \neq k$, then $E[X_j X_k] = E[X_j]E[X_k] = 0$, and the result is clear. ■

1.3 LLN=Law of Large Numbers

In the coin tossing, the rate of the heads of the coin appear goes to $1/2$ as the times of tossing increases, This is a typical example satisfying LLN.

In order to treat in mathematics, In the tossing the coin n-th time, if the head appears, then set $X_n = 1$, if the tail appears, then set $X_n = 0$. In this case, the probabilistic mean is $EX_n = 1/2$ (and the variance is $V(X_n) = 1/4$) and the arithmetic mean is $\dfrac{1}{n}\displaystyle\sum_{k=1}^n X_k$. The LLN is that this mean "converges" to the probabilistic mean $1/2$ as $n \to \infty$.

Theorem 1.3 (Weak Law of Large Numbers) *Let $X_1, X_2, \ldots$ be independent RVs with constant means $EX_n = m$, and bounded variances $v := \sup_n V(X_n) < \infty$. It holds that for $^\forall\varepsilon > 0$,*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^n X_k - m\right| \geq \varepsilon\right) = 0, \quad \textit{i.e.,} \quad \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^n X_k - m\right| < \varepsilon\right) = 1.$$

Proof. Since $\{X_n\}$ are independent, $\{\widetilde{X}_n = X_n - m\}$ are so, too (make sure). Hence, by

$$\frac{1}{n}\sum_{k=1}^n X_k - m = \frac{1}{n}\sum_{k=1}^n (X_k - m)$$

and by considering $\widetilde{X}_n$ instead of X_n, we may set $m = 0$, i.e., $E[X_n] = 0$. Then $V(X_n) = E[X_n^2]$ and by the previous proposition, we have

$$E\left[\left(\sum_{k=1}^n X_k\right)^2\right] = \sum_{k=1}^n E[X_k^2] = \sum_{k=1}^n V(X_k) \leq n \sup_n V(X_n) = nv.$$

Therefore, for $^\forall\varepsilon > 0$,

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{k=1}^n X_k\right| \geq \varepsilon\right) &= P\left(\left|\sum_{k=1}^n X_k\right| \geq \varepsilon n\right) \leq \frac{E[(\sum_{k=1}^n X_k)^2]}{\varepsilon^2 n^2} \\ &\leq \frac{nv}{\varepsilon^2 n^2} = \frac{v}{\varepsilon^2 n} \to 0 \quad (n \to \infty). \end{aligned}$$

■

Under the same conditions as above, the strong result holds. it is the following theorem:

Theorem 1.4 (Strong Law of Large Numbers) *Let $X_1, X_2, \dots$ be independent RVs with constant means $EX_n = m$, and bounded variances $v := \sup_n V(X_n) < \infty$. It holds that*

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = m\right) = 1.$$

Remark 1.1 *In general, in case of non-constant means, it holds that*

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} (X_k - EX_k) = 0\right) = 1.$$

We give the proof in the next subsection. However, under the stronger condition, we can show it easily.

[Under $\sup E[X_n^4] < \infty$, proof of Theorem 1.4] By considering $X_n - m$ instead of X_n, we may set $m = 0$, i.e., $E[X_n] = 0$. We consider the expansion of $\left(\sum_{k=1}^{n} X_k\right)^4$. By the independence and mean 0, and by noting that $E[X^2] \le (E[X^4])^{1/2}$, we have

$$E\left[\left(\sum_{k=1}^{n} X_k\right)^4\right] = \sum_{k=1}^{n} E[X_k^4] + \sum_{i\neq j, 1\le i,j\le n} E[X_i^2]E[X_j^2] \le n^2 \sup_k E[X_k^4].$$

Hence, by monotone convergence theorem or by Fubini's theorem, we get

$$E\left[\sum_{n=1}^{\infty}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^4\right] = \sum_{n=1}^{\infty}\frac{1}{n^4}E\left[\left(\sum_{k=1}^{n} X_k\right)^4\right] \le \sum_{n=1}^{\infty}\frac{1}{n^2}\sup_k E[X_k^4] < \infty.$$

This implies $P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = 0\right) = 1.$ ■

More important result is the following Central Limit Theorem (CLT).

Theorem 1.5 (CLT) *Let $\{X_n\}$ be independent identically distributed = i.i.d.) RVs. Set $EX_1 = m$ and $V(X_1) = v$. Then the distribution of $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_k - m)$ converges to the normal distribution with mean 0 and variance v, i.e., for all $a < b$,*

$$\lim_{n\to\infty} P\left(a < \frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_k - m) \le b\right) = \frac{1}{\sqrt{2\pi v}}\int_a^b e^{-\frac{x^2}{2v}}\,dx.$$

In another words, the distribution of $\frac{1}{\sqrt{nv}}\sum_{k=1}^{n}(X_k - m)$ converges to the normal distribution $N(0,1)$ with mean 0 and variance 1.

Here, we describe on the relations of independence and distributions. Let $\mathcal{B}^1 = \mathcal{B}(\mathbf{R}^1)$ be a 1-dimensional Borel-field. For real-valued RVs $X_1, \dots, X_n$, set $\boldsymbol{X} = (X_1, \dots, X_n)$ and define the joint distribution by $\mu_{\boldsymbol{X}}(A_1 \times \cdots \times A_n) = P(X_1 \in A_1, \dots, X_n \in A_n)$ $(A_i \in \mathcal{B}^1)$.

Theorem 1.6 *If real-valued RVs $X_1, \dots, X_n$ are indep., then for $\boldsymbol{X} = (X_1, \dots, X_n)$,*

$$\mu_{\boldsymbol{X}} = \bigotimes_{i=1}^{n} \mu_{X_i} \quad \textit{i.e.,} \quad \mu_{\boldsymbol{X}}(A_1 \times \cdots \times A_n) = \mu_{X_1}(A_1) \cdots \mu_{X_n}(A_n).$$

This is easily seen by that the σ-add. class generated by the family of all half-lines $(-\infty, a]$ is $\mathcal{B}^1$.

Theorem 1.7 *If real-valued RVs X, Y are indep., then for a bounded Borel function $f(x, y)$,*

$$E[f(X,Y)] = E\left[E[f(x,Y)]|_{x=X}\right] = E\left[E[f(X,y)]|_{y=Y}\right].$$

This is the same as

$$\int_{\mathbf{R}^2} f(x,y)\mu_{(X,Y)}(dx,dy) = \int_{\mathbf{R}^2} f(x,y)\mu_X(dx)\mu_Y(dy)$$

and it is clear by the above result.

Example 1.1 If X, Y are indep., then

$$P(X < Y) = \int_{\mathbf{R}} P(x < Y)\mu_X(dx).$$

1.4 Proof of LLN

We describe two convergence notions.

Let X_n, X be RVs.

$X_n \to X$ in pr., i.e., X_n converge to X **in probability** if for $^\forall\varepsilon > 0$, $P(|X_n - X| \geq \varepsilon) \to 0$ $(n \to \infty)$.

$X_n \to X$, P-a.s., i.e., X_n converges to X **almost surely** if $P(X_n \to X) = 1$.

Question 1.2 *Show $X_n \to X$, P-a.s. $\Longrightarrow X_n \to X$ in pr.*

Hint. $P(X_n \to X) = 1 \iff$

$$P\left(\bigcap_{k\geq 1}\bigcup_{N\geq 1}\bigcap_{n\geq N}\left\{|X_n - X| < \frac{1}{k}\right\}\right) = 1 \iff P\left(\bigcup_{k\geq 1}\bigcap_{N\geq 1}\bigcup_{n\geq N}\left\{|X_n - X| \geq \frac{1}{k}\right\}\right) = 0$$

$$\iff {}^\forall k \geq 1,\ \lim_{N\to\infty} P\left(\bigcup_{n\geq N}\left\{|X_n - X| \geq \frac{1}{k}\right\}\right) = P\left(\bigcap_{N\geq 1}\bigcup_{n\geq N}\left\{|X_n - X| \geq \frac{1}{k}\right\}\right) = 0$$

$$\Longrightarrow {}^\forall k \geq 1,\ \lim_{N\to\infty} P\left(|X_N - X| \geq \frac{1}{k}\right) \leq \lim_{N\to\infty} P\left(\bigcup_{n\geq N}\left\{|X_n - X| \geq \frac{1}{k}\right\}\right) = 0$$

This is equivalent to to the convergence in prob. (That is, it is possible to change $1/k$ to $\varepsilon > 0$. Why?

Remark 1.2 *In general, the inverse of the above question does not hold. That is, it is possible to make an example which converges in probability, however, which does not converge a.s.*

Question 1.3 *Show the following: If a sequence of RV's converges in pr., then there exists a suitable sub-sequence which converges a.s., i.e., "$X_n \to X$ in pr. $\Rightarrow {}^\exists\{n_k\}; X_{n_k} \to X$ a.s."*

Hint. We can see that $^\exists\{n_k\}$; $P\left(|X_{n_k} - X| \geq \dfrac{1}{2^k}\right) \leq \dfrac{1}{2^k}$. Since the sum converges, we can use Borel-Cantelli's lemma and we have $P\left(\bigcup_{N\geq 1}\bigcap_{k\geq N}\left\{|X_{n_k} - X| < \dfrac{1}{2^k}\right\}\right) = 1$. the result is easily obtained.)

Now we proceed the subject of strong law of large numbers. We describe the theorem again.

Theorem 1.8 (Strong Law of Large Numbers) *Let $X_1, X_2, \dots$ be independent random variables and have constant means; $EX_n = m$, and bounded variances; $v := \sup_n V(X_n) < \infty$. Then, the following holds:*

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = m\right) = 1.$$

[Outline of the proof] We may set $EX_n = 0$. Let $\overline{S}_n = \sum_{k=1}^{n}(X_k/k)$.

(1) By **Kolmogorov's maximal inequality**, $\sup_{k\geq n} |\overline{S_k} - \overline{S_n}| \to 0$ $(n \to \infty)$ in pr.

(2) By the result of "Convergence in pr. implies convergence a.s. of a suitable sub-sequence", we have $\{\overline{S_n}\}$ is a Cauchy sequence a.s., thus, it converges a.s.

(3) By **Kronecker's Lemma**, $\frac{1}{n}\sum_{k=1}^{n} X_k \to 0$ P-a.s.

Lemma 1.1 (Kronecker's lemma) *For a numerical sequences $\{x_n\}$, $\{a_n\}; 0 < a_n \uparrow \infty$,*

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{x_k}{a_k} \quad \textit{exists} \quad \Longrightarrow \quad \lim_{n\to\infty} \frac{1}{a_n}\sum_{k=1}^{n} x_k = 0$$

Proof. Set $s_0 = 0$ and $s_n = \sum_{k=1}^{n}(x_k/a_k) \to s$. Since

$$\frac{1}{a_n}\sum_{k=1}^{n} x_k = \sum_{k=1}^{n} \frac{a_k}{a_n}\frac{x_k}{a_k} = \sum_{k=1}^{n} \frac{a_k}{a_n}(s_k - s_{k-1}) = s_n - \sum_{k=1}^{n-1} \frac{a_{k+1} - a_k}{a_n} s_k,$$

the result is reduced to

$$s_n \to s \quad \Rightarrow \quad \frac{1}{a_n}\sum_{k=1}^{n-1}(a_{k+1} - a_k)s_k \to s$$

$s^* = \sup_m |s_m| < \infty$ and ${}^{\forall}\varepsilon > 0, {}^{\exists}N; {}^{\forall}k \geq N, |s_k - s| < \varepsilon$ imply that for $n > N$, dividing the sum at N, we have

$$\begin{aligned}
&\left|\frac{1}{a_n}\sum_{k=1}^{n-1}(a_{k+1} - a_k)s_k - s\right| \quad \left(\text{by } s = \frac{1}{a_n}\sum_{k=N}^{n-1}(a_{k+1} - a_k)s + \frac{a_N}{a_n}s \quad \text{we have}\right) \\
&\leq \quad \frac{1}{a_n}\sum_{k=N}^{n-1}(a_{k+1} - a_k)|s_k - s| + \frac{1}{a_n}\sum_{k=1}^{N-1}(a_{k+1} - a_k)(\sup_m |s_m|) + \frac{a_N}{a_n}|s| \\
&\leq \quad \varepsilon\frac{a_n - a_N}{a_n} + s^*\frac{a_N - a_1}{a_n} + \frac{a_N}{a_n}|s| \\
&\to \quad \varepsilon \quad (n \to \infty).
\end{aligned}$$

Hence, $\varepsilon > 0$ is arbitrary, the limit is 0. ∎

Lemma 1.2 (Kolmogorov's maximal inequality) *Let $\{X_n\}$ be independent RVs with means $EX_n = 0$. For $S_n = \sum_{k=1}^{n} X_n$, it holds that*

$$a > 0 \quad \Longrightarrow \quad a^2 P(\max_{1\leq n\leq N} |S_n| \geq a) \leq E[|S_N|^2; \max_{1\leq n\leq N} |S_n| \geq a] \leq E[|S_N|^2]$$

Proof. Let $A_k = \{|S_k| \geq a, |S_1| < a, \ldots, |S_{k-1}| < a\}$, and $S^{(k+1)} = X_{k+1} + \cdots + X_N$. Then, $S^{(k+1)}$ and$S_k 1_{A_k}$ are independent, and $E[S_k S^{(k+1)}; A_k] = E[S_k 1_{A_k}]E[S^{(k+1)}] = 0$. $A = \bigcup_{k=1}^{N} A_k$ (disjoint union).

$$\begin{aligned} E[|S_N|^2; \max_{1\leq n\leq N} |S_n| \geq a] &= \sum_{k=1}^{N} E[(S_k + S^{(k+1)})^2; A_k] \\ &= \sum_{k=1}^{N} E[S_k^2 + 2S_k S^{(k+1)} + (S^{(k+1)})^2; A_k] \\ &\geq \sum_{k=1}^{N} E[S_k^2; A_k] \\ &\geq \sum_{k=1}^{N} a^2 P(A_k) \quad (\text{by } |S_k| \geq a \text{ on } A_k) \\ &= a^2 P(\max_{1\leq n\leq N} |S_n| \geq a) \end{aligned}$$

■

[Proof of String Law of Large Numbers (Theorem 1.4)] We may assume $EX_n = 0$. Because by considering $\widetilde{X}_n = X_n - EX_n$ instead of X_n, $\{\widetilde{X}_n\}$ are independent and $V(\widetilde{X}_n) = V(X_n) \leq v$. Moreover, we have $E[X_n X_m] = E[X_n]E[X_m] = 0$ $(m \neq n)$ and $E[X_n^2] = V(X_n) \leq v$. Let $\overline{S}_n = \sum_{k=1}^{n} \frac{X_k}{k}$. By Kolmogorov's maximal inequality, for all $a > 0$,

$$a^2 P(\max_{n<k\leq N} |\overline{S_k} - \overline{S_n}| \geq a) \leq E[|\overline{S}_N - \overline{S_n}|^2] = \sum_{k=n+1}^{N} \frac{E[X_k^2]}{k^2} \leq \sum_{k>n} \frac{v}{k^2}.$$

As $N \to \infty$ and $n \to \infty$, we have

$$\lim_{n\to\infty} P(\sup_{k>n} |\overline{S_k} - \overline{S_n}| \geq a) = 0, \quad \text{i.e.,} \quad \sup_{k>n} |\overline{S_k} - \overline{S_n}| \to 0 \ (n \to \infty) \text{ in pr.}$$

Hence, for a suitable subsequence $\{n_j\} \subset \mathbf{N}; n_j \uparrow \infty$,

$$\lim_{j\to\infty} \sup_{k\geq n_j} |\overline{S_k} - \overline{S}_{n_j}| = 0 \quad P\text{-a.s.}$$

Thus, if $n, m \geq n_j$, then $|\overline{S_n} - \overline{S_m}| \leq |\overline{S_n} - \overline{S}_{n_j}| + |\overline{S_m} - \overline{S}_{n_j}| \to 0$ $(j \to \infty)$ P-a.s., that is, $\{\overline{S_n}\}$ is a Cauchy sequence a.s. Thus, $\lim_{n\to\infty} \sum_{k=1}^{n} \frac{X_k}{k} = \lim_{n\to\infty} \overline{S_n}$ exists a.s. Therefore, by Kronecker's lemma, we have $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k = 0$ a.s. ■

From the above proof we have

Corollary 1.1 *Let $\{X_n\}$ be independent random variables with mean 0. If $\sum_{k=1}^{\infty} V(X_k) < \infty$, then the limit $\lim_{n\to\infty} \sum_{k=1}^{n} X_k$ exists with probability one.*

Corollary 1.2 *Under the same conditions as in the strong LLN, for arbitrary $\delta > 0$, the following holds*

$$\lim_{n\to\infty} \frac{1}{\sqrt{n^{1+\delta}}} \sum_{k=1}^{n} (X_k - EX_k) = 0 \quad P\text{-a.s.}$$

In the proof of LLN, it is enough to consider $\sum_{k=1}^{n}(X_k/\sqrt{k^{1+\delta}})$ instead of $\overline{S_n}$.

Now, if we let δ be 0 in the above, then what is the result?

The answer of this question is Central Limit Theorem (CLT). In the proof we use characteristic functions which are Fourier transforms of probability measures. Moreover, we use the result such that the convergence of characteristic functions implies convergence of distributions.

1.5 Characteristic functions & convergence of distributions

For a RV X, the following function $\varphi = \varphi_X : \mathbf{R}^1 \to \mathbf{C}$ is called a **characteristic function** of X;

$$\varphi(z) = \varphi_X(z) := E\left[e^{izX}\right] \quad (z \in \mathbf{R}^1).$$

For a **distribution** of X; $\mu(A) = \mu_X(A) := P(X \in A)$, it is also expressed as

$$\varphi(z) = \int_{\mathbf{R}} e^{izx}\mu(dx)$$

On the other hand, if μ is a probability measure on $\mathbf{R}^1$ (which is simply called a "distribution"), then the above $\varphi(z)$ is called a characteristic function of μ.

We first give the definition of normal distributions.

A distribution $\mu(dx) = g(x)dx$ on $\mathbf{R}$ with

$$g(x) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-m)^2}{2v}\right]$$

is called a **normal distribution** or **Gaussian distribution** with mean m, variance v, and it is denoted as $N(m, v)$ (this notion is also used as a random variable with the normal distribution).

The characteristic function of this distribution is given as

$$\varphi(z) = \int_{-\infty}^{\infty} e^{izx} \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-m)^2}{2v}\right] dx = \exp\left[imz - \frac{vz^2}{2}\right].$$

Question 1.4 Make sure the above calculation.

Tent functions: Let $T(x)$ be a function on $\mathbf{R}$ such that it has a graph which connected three points of $(-1, 0), (0, 1), (1, 0)$ in a segment of a line, and that it is 0 outside of $(-1, 1)$, i.e.,

$$T(x) = \frac{1}{2}(|x+1| + |x-1| - 2|x|).$$

This is called a **tent function on an interval** $(-1, 1)$ **with a height** 1. Moreover, for $-\infty < a < b < \infty, h > 0$, we define a **tent function on an interval** (a, b) **with a height** h as

$$T_{a,b;h}(x) = hT\left(\frac{2}{b-a}\left(x - \frac{a+b}{2}\right)\right)$$

Furthermore, for $h > 1$, we define a **trapezoid function** $D_{a,b;h}$ as

$$D_{a,b;h}(x) := (T_{a,b;h} \wedge 1)(x) = \min\{T_{a,b;h}(x), 1\} = T_{a,b;h}(x) - T_{a+(b-a)/(2h),b-(b-a)/(2h);h-1}(x).$$

This tent function appears in the following distribution:

Question 1.5 Let U, V be independent RVs with the same uniform distribution on $[0, a]$ $(a > 0)$. Show the density function of $X = U - V$ is $T_{-a,a;1/a}$. Show the characteristic function is given as

$$\varphi_X(z) = \frac{2(1 - \cos az)}{a^2 z^2}.$$

Hint. It is enough to show that For any bounded Borel functions f,

$$E[f(X)] = \int_{-a}^{a} f(x)T_{-a,a;1/a}(x)dx$$

In that use the joint dist. of (U, V) is the product of each distributions by their independence. That is,

$$P(U \in du, V \in dv) = P(U \in du)P(V \in dv) = \frac{1}{a}1_{[0,a]}(u)du\frac{1}{a}1_{[0,a]}(v)dv.$$

By this the above calculus is reduced to

$$\frac{1}{a^2}\int_{\mathbf{R}} 1_{[0,a]}(v)1_{[0,a]}(x+v)dv = T_{-a,a;1/a}(x)$$

On the later half, use that the characteristic function of X is a product of each c.f.'s of $U, -V$.

Proposition 1.1 *For a characteristic function $\varphi(z)$ of a RV X, it holds that*

$$E[T(X)] = \frac{1}{\pi}\int_{-\infty}^{\infty} \varphi(z)\frac{1-\cos z}{z^2}dz,$$

$$E[T_{a,b;h}(X)] = \frac{2h}{\pi(b-a)}\int_{-\infty}^{\infty} \varphi(z)e^{-i(a+b)z/2}\frac{1-\cos\frac{(b-a)z}{2}}{z^2}dz.$$

Proof. By the previous question,

$$\int_{-\infty}^{\infty} e^{izx}T(x)dx = \frac{2(1-\cos z)}{z^2}.$$

Moreover, it is can be seen that

$$\int_{-\infty}^{\infty} e^{izx}\frac{1-\cos z}{z^2}dz = \pi T(x). \tag{1.1}$$

If we admit this result, then by substituting X for x and taking expectation, and by Fubini's Theorem, we have

$$E[T(X)] = \frac{1}{\pi}E\left[\int_{-\infty}^{\infty} e^{izX}\frac{1-\cos z}{z^2}dz\right] = \frac{1}{\pi}\int_{-\infty}^{\infty}\varphi_X(z)\frac{1-\cos z}{z^2}dz.$$

On $E[T_{a,b;h}(X)]$, it is easy to get by a change of variables. Finally, we show (1.1). Since $(1-\cos z)/z^2$ is an even function and

$$\cos zx(1-\cos z) = \cos zx - \frac{1}{2}(\cos z(x+1) + \cos z(x-1))$$

and by a change of variables, the left-hand of (1.1) is

$$\begin{aligned}\int_{-\infty}^{\infty}\cos zx\frac{1-\cos z}{z^2}dz &= \frac{1}{2}\int_{-\infty}^{\infty}\frac{1-\cos z(x+1)}{z^2}dz + \frac{1}{2}\int_{-\infty}^{\infty}\frac{1-\cos z(x-1)}{z^2}dz \\ &\quad -\int_{-\infty}^{\infty}\frac{1-\cos zx}{z^2}dz \\ &= \int_{-\infty}^{\infty}\frac{1-\cos z}{z^2}dz\left(\frac{1}{2}(|x+1|+|x-1|)-|x|\right).\end{aligned}$$

From this and by the equation $\displaystyle\int_{-\infty}^{\infty}\frac{1-\cos z}{z^2}dz = \pi$, we have (1.1). ∎

Question 1.6 (i) Find $I(t) = \int_0^\infty e^{-tz} \sin z dz \quad (t > 0)$ by using integration by parts. $1/(1+t^2)$

(ii) Show the equation $\int_0^\infty I(t)dt = \int_0^\infty \frac{\sin z}{z} dz$, and find the integral. $\pi/2$

(iii) By using integration by parts, show $\int_{-\infty}^\infty \frac{1-\cos z}{z^2} dz = \pi$.

RVs X, Y are identically distributed means for all $a \in \mathbf{R}$, $P(X > a) = P(Y > a)$. We denote by $X \overset{(\mathrm{d})}{=} Y$ (which means $X = Y$ in the sense of distribution).

Theorem 1.9 *For characteristic functions φ_X, φ_Y of RVs X, Y, if $\varphi_X(z) = \varphi_Y(z)$ $(z \in \mathbf{R})$, then $X \overset{(\mathrm{d})}{=} Y$.*

Proof. By the assumption and the previous proposition, for any tent function $T_{a,b;h}$, it holds $E[T_{a,b;h}(X)] = E[T_{a,b;h}(Y)]$. Thus, for all trapezoid function $D_{a,b;h}$, $E[D_{a,b;h}(X)] = E[D_{a,b;h}(Y)]$. Hence, noting that $\lim_{h\to\infty} D_{a,b;h}(x) = I_{(a,b)}(x)$, by Lebesgue's convergence theorem, $P(a < X < b) = P(a < Y < b)$. Therefore, we have $X \overset{(\mathrm{d})}{=} Y$. ■

Theorem 1.10 *Let X, $\{X_n\}$ be RVs and let $\varphi(z)$, $\{\varphi_n(z)\}$ be their characteristic functions. If*

$$\lim_{n\to\infty} \varphi_n(z) = \varphi(z) \quad (z \in \mathbf{R}^1) \quad \textit{[pointwise]},$$

then for all $a \in \mathbf{R}; P(X = a) = 0$, $\lim_{n\to\infty} P(X_n > a) = P(X > a)$.

Proof. By the assumption and $|\varphi_n(z)| \le 1$, and by Lebesgue's convergence theorem,

$$\lim_{n\to\infty} \int_{-\infty}^\infty \varphi_n(z) e^{-i(a+b)z/2} \frac{1-\cos((b-a)z/2)}{z^2} dz = \int_{-\infty}^\infty \varphi(z) e^{-i(a+b)z/2} \frac{1-\cos((b-a)z/2)}{z^2} dz.$$

Hence, by Proposition 1.1, $\lim_{n\to\infty} E[T_{a,b;h}(X_n)] = E[T_{a,b;h}(X)]$. Thus, for any $D_{a,b;h}$, we have $\lim_{n\to\infty} E[D_{a,b;h}(X_n)] = E[D_{a,b;h}(X)]$. Moreover, noting that for $h > 1, a < b$,

$$I_{(a,b)}(x) \ge D_{a,b;h}(x) \ge I_{[a+(b-a)/(2h), b-(b-a)/(2h)]}(x) \quad (x \in \mathbf{R}),$$

we have

$$\begin{aligned} \liminf_{n\to\infty} P(a < X_n < b) &\ge \lim_{n\to\infty} E[D_{a,b;h}(X_n)] \\ &= E[D_{a,b;h}(X)] \ge P\left(a + \frac{b-a}{2h} \le X \le b - \frac{b-a}{2h}\right). \end{aligned}$$

By letting $h \to \infty, b \to \infty$, we have for $^\forall a \in \mathbf{R}$,

$$\liminf_{n\to\infty} P(X_n > a) \ge P(X > a).$$

On the other hand, by letting $h \to \infty, a \to -\infty$, and by changing b to a, we have that for $^\forall a \in \mathbf{R}$, $\liminf_{n\to\infty} P(X_n < a) \ge P(X < a)$. Furthermore, by this,

$$\limsup_{n\to\infty} P(X_n > a) \le 1 - \liminf_{n\to\infty} P(X_n < a) \le 1 - P(X < a) = P(X \ge a),$$

and hence, for $^\forall a \in \mathbf{R}; P(X = a) = 0$,

$$\limsup_{n\to\infty} P(X_n > a) \le P(X > a).$$

Therefore, we get $\lim_{n\to\infty} P(X_n > a) = P(X > a)$. ■

1.6 CLT=Central Limit Theorem

Theorem 1.11 (CLT) *Let RVs $\{X_n\}$ be i.i.d. Set the mean $EX_1 = m$ and the variance $V(X_1) = v$. Then the distribution of $\frac{1}{\sqrt{nv}}\sum_{k=1}^{n}(X_k - m)$ converges to a normal distribution $N(0,1)$ with mean 0 and variance 1. i.e., for all $a < b$,*

$$\lim_{n\to\infty} P\left(a < \frac{1}{\sqrt{nv}}\sum_{k=1}^{n}(X_k - m) \le b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}dx.$$

We first give several lemmas.

Lemma 1.3 *For a RV X such that $EX = 0, V(X) = E(X^2) = 1$,*

$$\varphi_X\left(\frac{z}{\sqrt{n}}\right) - \left(1 - \frac{z^2}{2n}\right) = o\left(\frac{1}{n}\right) \quad (n \to \infty).$$

Proof. If $g(z)$ is defined by

$$e^{iz} - 1 - iz + \frac{z^2}{2} = z^2 g(z),$$

then it holds that $|g(z)| \le 1$, $\lim_{z\to 0} g(z) = 0$. In fact, by Taylor's theorem,

$$^{\exists}\theta \in (0,1); e^{iz} - 1 - iz = -\frac{z^2}{2}e^{i\theta z},$$

and it is easy to see that $|g(z)| \le 1$, $\to 0$ $(z \to 0)$. Thus,

$$\exp\frac{izX}{\sqrt{n}} = 1 + \frac{izX}{\sqrt{n}} - \frac{z^2X^2}{2n} + \frac{z^2X^2}{n}g\left(\frac{zX}{\sqrt{n}}\right)$$

and by taking expectations of both hand,

$$\varphi_X\left(\frac{z}{\sqrt{n}}\right) = 1 - \frac{z^2}{2n} + E\left[\frac{z^2X^2}{n}g\left(\frac{zX}{\sqrt{n}}\right)\right].$$

On the last expectation, since

$$X^2 g\left(\frac{zX}{\sqrt{n}}\right) \le X^2, \quad \lim_{n\to\infty} g\left(\frac{zX}{\sqrt{n}}\right) = 0,$$

we can use Lebesgue's convergence theorem it converges to 0 as $n \to \infty$. Therefore, we get the desired result. ■

[Proof of CLT]

Let $\widetilde{X}_n = (X_n - m)/\sqrt{v}$. Then $E\widetilde{X}_n = 0, V(\widetilde{X}_n) = 1$ and $\{\widetilde{X}_n\}$ are i.i.d. Hence, it is enough to show the case of $m = 0, v = 1$. Let $Y_n := (\sum_{k=1}^{n} X_k)/\sqrt{n}$. Since $\{X_k\}$ are i.i.d., its characteristic function is

$$\varphi_n(z) = E\left[\exp\left(\frac{iz}{\sqrt{n}}\sum_{k=1}^{n}X_k\right)\right] = \prod_{k=1}^{n}\varphi_{X_k}\left(\frac{z}{\sqrt{n}}\right) = \varphi_{X_1}\left(\frac{z}{\sqrt{n}}\right)^n.$$

By the above lemma, for each $z \in \mathbf{R}$, we have

$$\lim_{n\to\infty}\varphi_n(z) = \lim_{n\to\infty}\left(1 - \frac{z^2}{2n} + o\left(\frac{1}{n}\right)\right)^n = \exp[-z^2/2],$$

where for the last equation, if we define $R_n(z)$ by

$$\left(1 - \frac{z^2}{2n} + o\left(\frac{1}{n}\right)\right)^n = \left(1 - \frac{z^2}{2n}\right)^n + R_n(z),$$

then we can show $|R_n(z)| = o(1)$ $(n \to \infty)$ (see the next quest.) Therefore, $\varphi_n(z)$ converges pointwise to the characteristic function $\varphi(z) = \exp[-z^2/2]$ of a normal distribution $N(0,1)$. By the previous theorem (Theorem 1.10), the proof is end. ■

Question 1.7 At the end of the above proof, show $|R_n(z)| = o(1)$ $(n \to \infty)$.

1.7 Properties of characteristic functions

Proposition 1.2 *For a characteristic function $\varphi = \varphi_\mu$ of a distribution μ on $\mathbf{R}$, the following holds.*
(1) $\varphi(0) = 1$, $|\varphi(z)| \le 1$, $\overline{\varphi(z)} = \varphi(-z)$.
(2) φ is uniform continuous on $\mathbf{R}$.
(3) **[Positive Definite]** $\displaystyle\sum_{j,k=1}^{n} \alpha_j \overline{\alpha_k} \varphi(z_j - z_k) \ge 0$ *for* $^\forall n \ge 1$, $^\forall \alpha_k \in \mathbf{C}$, $^\forall z_k \in \mathbf{R}$ $(k = 1, \dots, n)$.

Proof. (1) is easy. (2) For $^\forall z, h \in \mathbf{R}$, $|e^{i(z+h)x} - e^{izx}| \le |e^{ihx} - 1| \to 0$ $(h \to 0)$ and $|e^{ihx} - 1| \le 2$. Hence by Lebesgue's convergence theorem,

$$\sup_z |\varphi(z+h) - \varphi(z)| \le \int |e^{ihx} - 1| \mu(dx) \to 0 \quad (h \to 0).$$

(3) $$\sum_{j,k=1}^{n} \alpha_j \overline{\alpha_k} \varphi(z_j - z_k) = \int \sum_{j,k=1}^{n} \alpha_j \overline{\alpha_k} e^{i(z_j - z_k)x} \mu(dx) = \int \left| \sum_{j=1}^{n} \alpha_j e^{iz_j x} \right|^2 \mu(dx) \ge 0.$$ ■

Theorem 1.12 *For a characteristic function φ, the following holds: Let $L^1(d\mu) = L^1(\mathbf{R}, \mathcal{B}, \mu)$.*
(1) If $x \in L^1(d\mu)$, then $\varphi \in C^1$ and $\varphi'(z) = i \displaystyle\int x e^{izx} \mu(dx)$.
(2) If $^\exists \varphi''(0)$, then $x^2 \in L^1(d\mu)$.

Proof. (1) is easy to show by Lebesgue's convergence theorem and noting the following: For $h \ne 0$,

$$\frac{e^{ix(z+h)} - e^{ixz}}{h} = i\frac{x}{h} \int_0^h e^{ix(z+s)} ds, \quad \left| \frac{e^{ix(z+h)} - e^{ixz}}{h} \right| \le \frac{|x|}{|h|} \int_0^{|h|} |e^{ix(z+s)}| ds = |x|.$$

(2) For $h \ne 0$, let

$$\psi_h(z) := (\varphi(z+h) + \varphi(z-h) - 2\varphi(z))/h^2$$

(a symmetric difference). It can be shown that

$$\psi_h(z) = \int_{\mathbf{R}} e^{izx} \left(\frac{i \sin(hx/2)}{h/2} \right)^2 \mu(dx). \tag{1.2}$$

By $\lim_{h \to 0} \psi_h(0) = \varphi''(0)$, and by Fatou's lemma, we have

$$|\varphi''(0)| = \lim_{h \to 0} \int_{\mathbf{R}} \left(\frac{\sin(hx/2)}{h/2} \right)^2 \mu(dx) \ge \int_{\mathbf{R}} x^2 \mu(dx).$$

■

Question 1.8 In the above proof, show the equation (1.2) and $\lim_{h \to 0} \psi_h(0) = \varphi''(0)$.

Hint for the later half Show and use $\varphi(z \pm h) = \varphi(z) + \varphi'(z)h + \varphi''(z)h^2/2 + o(h^2)$ $(h \to 0)$.

1.8 Lévy's inversion formula

Theorem 1.13 (Lévy's inversion formula) *Let μ be a distribution on $\mathbf{R}$ and φ be its characteristic function. For $a < b$ such that $\mu(\{a\}) = \mu(\{b\}) = 0$, it holds that*

$$\mu((a,b)) = \frac{1}{2\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{e^{-iza} - e^{-izb}}{iz} \varphi(z) dz.$$

More general, it holds that

$$\mu((a,b)) = \frac{1}{2\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{e^{-iza} - e^{-izb}}{iz} \varphi(z) dz - \frac{1}{2}[\mu(\{a\}) + \mu(\{b\})].$$

Proof. Noting that $|(e^{-iza}-e^{-izb})/iz| \le (b-a)$ for $z \neq 0$, by Fubini's theorem,

$$\int_{-T}^{T} \frac{e^{-iza}-e^{-izb}}{iz}\varphi(z)dz = \int_{\mathbf{R}} \mu(dx) \int_{-T}^{T} \frac{e^{iz(x-a)}-e^{iz(x-b)}}{iz}dz.$$

If we let $J(T,x,a,b)$ be the integral in z at the last term, then odd function disappears. Hence,

$$J(T,x,a,b) = 2\int_0^T \frac{\sin(x-a)z}{z}dz - 2\int_0^T \frac{\sin(x-b)z}{z}dz.$$

Now it is well-known that (see Question 1.6)

$$\int_0^\infty \frac{\sin z}{z}dz = \frac{\pi}{2} \quad \text{implies} \quad \int_0^\infty \frac{\sin zx}{z}dz = \begin{cases} \pi/2 & (x>0) \\ 0 & (x=0) \\ -\pi/2 & (x<0). \end{cases}$$

Thus,

$$\lim_{T\to\infty} J(T,x,a,b) = \begin{cases} 0 & (x<a \text{ or } b<x) \\ \pi & (x=a \text{ or } x=b) \\ 2\pi & (a<x<b). \end{cases}$$

Moreover, by the graph of sin, we have $|J(T,x,a,b)| \le 4\int_0^\pi \frac{\sin z}{z}dz$. Therefore, by Lebesgue's convergence theorem,

$$\lim_{T\to\infty} \int_{\mathbf{R}} J(T,x,a,b)\mu(dx) = \pi \int_{\mathbf{R}} 1_{\{a,b\}}(x)\mu(dx) + 2\pi \int_{\mathbf{R}} 1_{(a,b)}(x)\mu(dx).$$

Hence, the desired result is obtained. ■

Question 1.9 In the above, show $|J(T,x,a,b)| \le 4\int_0^\pi \frac{\sin z}{z}dz$.

(Hint) If $x>0$, then for $^\forall T>0$,

$$\int_0^T \frac{\sin xt}{t}dt = \int_0^{Tx} \frac{\sin z}{z}dz \le \int_0^\pi \frac{\sin z}{z}dz.$$

Theorem 1.14 (Uniqueness Theorem) *For distributions μ,ν on $\mathbf{R}$ and their characteristic functions φ_μ,φ_ν, if $\varphi_\mu=\varphi_\nu$, then $\mu=\nu$.*

Proof. Let $\mathcal{I}$ be a family of all intervals of $(a,b); \mu(\{a\}) = \mu(\{b\}) = \nu(\{a\}) = \nu(\{b\}) = 0$. By the inverse formula, we have $\mu=\nu$ on $\mathcal{I}$, Note that the number of intervals not satisfying the above conditions is countable (see the next question). For any intervals $(a,b]$ by upper approximating we have $\mu((a,b]) = \nu((a,b])$. Hence, it holds on a family of all finite unions $\bigcup_{k=1}^n (a_k,b_k]$ of disjoint intervals (we dnote as $\mathcal{A}$ which is an add. class). Therefore $\mathcal{M} = \{A \subset \mathbf{R}; \mu(A) = \nu'(A)\}$ contains $\mathcal{A}$ and this is a monotone class. Thus $\mathcal{M}$ contains $m(\mathcal{A}) = \sigma(\mathcal{A}) = \mathcal{B}^1$ (by montone class th.), that is $\mu=\nu$ on $\mathcal{B}^1$. ■

Question 1.10 Show that for a distribution μ on $\mathbf{R}$, the number of $a \in \mathbf{R}$ such that $\mu(\{a\}) > 0$ is countable at most.

1.9 Lebesgue-Stielties measures

For a real-valued RV X on a probability space, set $F(x) := P(X \le x)$ and it is called as a **distribution function of** X. Then $F : \mathbf{R} \to [0,1]$ satisfies the following:

(1) $F \uparrow$, i.e., it is monotone increasing, i.e., $x<y \Rightarrow F(x) \le F(y)$.

(2) F is rcll, i.e., it is right-continuous and has left-hand-limits, i.e.,
$F(x) = F(x+) := \lim_{y \to x, y > x} F(y)$, $^{\exists} F(x-) := \lim_{y \to x, y < x} F(y)$.

(3) $F(+\infty) = 1, F(-\infty) = 0$

A function $F : \mathbf{R} \to [0,1]$ satisfying the above three properties is simply called a **distribution function on R**.

On the other hand, when a distribution function F is given, we can consider the question whether there exist a probability space $(\Omega, \mathcal{F}, P)$ and a random variable X such that $F(x) = P(X \leq x)$. The answer is given by the following result:

Theorem 1.15 (Lebesgue-Stielties measure) *For a distribution function $F : \mathbf{R} \to [0,1]$, $^{\exists_1}\mu : \mathcal{B}^1 \to [0,1]$ a distribution; $\mu((-\infty, x]) = F(x)$.*

This distribution μ on $(\mathbf{R}, \mathcal{B}^1)$ is called a **Lebesgue-Stielties measure (distribution)** by a distribution function $F(x)$. Moreover, we can define an integral by the measure;

$$\int_{\mathbf{R}} f(x) dF(x) := \int_{\mathbf{R}} f(x) \mu(dx); \quad \text{is called a } \textbf{Lebesgue-Stielties integral}.$$

We denote as $dF(x) = \mu(dx)$.

1.10 Weak convergence of measures

Let $\mathcal{P}(\mathbf{R})$ be a total family of distributions on $\mathbf{R}$. Let $\mu_n, \mu \in \mathcal{P}(\mathbf{R})$. μ_n **(weak) converges to** μ is defined as

$$\mu_n \to \mu \overset{\text{def}}{\iff} {}^{\forall} f \in C_b(\mathbf{R}), \langle \mu_n, f \rangle \to \langle \mu, f \rangle,$$

where $C_b(\mathbf{R})$ is a family of all bounded continuous functions on $\mathbf{R}$, and $\langle \mu, f \rangle = \int f d\mu$.

For RVs X_n, X, X_n **converges to** X **in law**, that is,

$$X_n \to X \text{ in law } \overset{\text{def}}{\iff} {}^{\forall} f \in C_b(\mathbf{R}),\ E[f(X_n)] \to E[f(X)].$$

Question 1.11 *Show that a.s. convergence implies convergence in law and that convergence in pr. implies convergence in law.*

The first half is clear. On the later half, by the property of lim sup, $^{\exists}\{n_k\}; \limsup E[f(X_{n_k})] = \limsup Ef(X_n)]$. Moreover, since convergence in pr. implies there exists a suitable sub sequence which converges a.s., $^{\exists}\{Y_j\} \subset \{X_{n_k}\}; Y_j \to X$ a.s. Hence, we have $\limsup Ef(X_n)] = \lim E[f(Y_j)] = E[f(X)]$ and for lim inf it is the same. Thus, desired result is obtained.

Some reader may think the convergence of distributions should be defined as $^{\forall} A \in \mathcal{B}, \mu_n(A) \to \mu(A)$. However, it is slightly strong, so it is not useful.

Theorem 1.16 *For distributions μ_n, μ on $\mathbf{R}$, the following are equivalent:*
(1) $\mu_n \to \mu$
(2) $^{\forall} U \subset \mathbf{R}$: an open set, $\liminf \mu_n(U) \geq \mu(U)$.
(3) $^{\forall} F \subset \mathbf{R}$: a closed set, $\limsup \mu_n(F) \leq \mu(F)$.
(4) $^{\forall} A \in \mathcal{B}^1; \mu(\partial A) = 0$, $\lim \mu_n(A) = \mu(A)$.
(5) Let F_n, F be distribution functions of μ_n, μ. $^{\forall} x$: a continuous point of F, i.e., $F(x-) = F(x)$, $F_n(x) \to F(x)$.
(6) $^{\forall} f \in C_c(\mathbf{R}), \langle \mu_n, f \rangle \to \langle \mu, f \rangle$, where $C_c(\mathbf{R})$ is a family of all continuous functions with compact supports on $\mathbf{R}$ and a support of f is $\operatorname{supp} f = \overline{\{f \neq 0\}}$.

Proof. (1) $\Rightarrow$ (2) It is possible to make continuous functions $0 \leq h_k \uparrow 1_U$ pointwise. Then, $\mu_n(h_k) \leq \mu_n(U)$ and $\lim_{n\to\infty}\langle \mu_n, h_k\rangle = \langle \mu, h_k\rangle \uparrow \langle \mu, 1_U\rangle = \mu(U)$. Hence, for the first enquality, by taking $\liminf_{n\to\infty}$ and by $k \to 0$, we get (2). In order to make h_k, let $U_k = \{x \in U; d(x, U^c) \geq 1/k\}$ be the closed subset which is shortened U by $1/k$. Since a metric function to a non-empty set is continuous, we define $h_k = d(x, U^c)/(d(x, U_k) + d(x, U^c))$, where $d(x, A) = \inf_{y\in A} d(x, y)$ is a metric function.

(2) $\iff$ (3) Take the complement. (2),(3) $\Rightarrow$ (4) is easy. (4) $\Rightarrow$ (5) It is enough to consider $A = (-\infty, x]$ with a continuous point x of F.

(5) $\Rightarrow$ (6) can be shown by using that for $f \in C_c(\mathbf{R})$, the approximating simple functions f_k for f are step functions with intervals which have the continuous points as their endpoints. In fact, we have for each $k \geq 1$, $\langle \mu_n, f_k\rangle \to \langle \mu, f_k\rangle$. Moreover, we can take f_k such that $\|f - f_k\|_\infty \to 0 (k \to \infty)$. Thus, it holds that $|\langle \mu_n, f\rangle - \langle \mu, f\rangle| \leq \langle \mu_n, |f - f_k|\rangle + |\langle \mu_n, f_k\rangle - \langle \mu, f_k\rangle| + \langle \mu, |f - f_k|\rangle \leq 2\|f - f_k\|_\infty + |\langle \mu_n, f_k\rangle - \langle \mu, f_k\rangle|$. By letting $n \to \infty$ and $k \to \infty$, we have the desired result.

(6) $\Rightarrow$ (1) It is enough to approximate $f \in C_b(\mathbf{R})$ uniformly on compact sets by the elements of $C_c(\mathbf{R})$. Concretely, let $g_k \in C_b(\mathbf{R}); 0 \leq g_k \leq 1$ be a continuous function such that $g = 1$ on $[-k, k]$, $g = 1$ on $[-k-1, k+1]^c$, and for a sufficiently large $K > 1$, we may consider fg_K. Moreover, we use $^\forall\varepsilon > 0, {}^\exists K; \sup_{n\geq 1} \mu_n([-K, K]^c) < \varepsilon$ which is obtained by (6). In fact, for this, by $1_{[-k,k]} \leq g_k \leq 1_{[-k-1,k+1]}$ and (6), we have $\mu([-k, k]) \leq \langle \mu, g_k\rangle = \lim_{n\to\infty}\langle \mu_n, g_k\rangle \leq \underline{\lim}_{n\to\infty} \mu_n([-k-1, k+1])$. Hence, $^\forall\varepsilon > 0, {}^\exists K_0; \overline{\lim}_{n\to\infty} \mu_n([-K_0 - 1, K_0 + 1]^c) \leq \mu([-K_0, K_0]^c) < \varepsilon$. Furthermore, we can get $^\exists K \geq K_0; \sup_{n\geq 1} \mu_n([-K, K]^c) < \varepsilon$. ∎

Some reader may still think that it is inadequate for the validity of the definition of the convergence of the distributions. However, we have more results related to the convergence of characteristic functions, and from these the readers may think it is necessary and sufficient.

Theorem 1.17 *Let φ_n, φ be characteristic functions of $\mu_n, \mu \in \mathcal{P}(\mathbf{R})$. If $\mu_n \to \mu$, then $\varphi_n \to \varphi$ (uniform on compact sets)*

Since $f(x) = e^{izx}$ is bounded continuous, the pointwise convergence is clear. In order to show the uniform convergence on compact sets, we need the result such that if $\{\mu_n\}$ is relatively compact, then it is tightness. (we describe at the end of this section.)

Theorem 1.18 (Lévy's Continuity Theorem) *Let φ_n be a characteristic function of $\mu_n \in \mathcal{P}(\mathbf{R})$. If $^\exists\varphi; \varphi_n \to \varphi$ (pointwise) and φ is continuous at the origin, then $^\exists\mu$: a distribution on $\mathbf{R}$; $\varphi = \varphi_\mu$ is a characteristic function of μ and $\mu_n \to \mu$. Moreover, $\varphi_n \to \varphi$ (uniform on compact sets).*

Corollary 1.3 (Glivenko's Theorem) *Let φ_n, φ be characteristic functions of $\mu_n, \mu \in \mathcal{P}(\mathbf{R})$. If $\varphi_n \to \varphi$ (pointwise), then $\mu_n \to \mu$.*

Note that in Lévy's theorem, we can not omit the continuity at the origin of φ. For instance, the characteristic function of $N(0, n)$ satisfies $\varphi_n(z) = \exp(-nz^2/2) \to 1_{\{0\}}(z)$, however, the limit is not continuous at the origin, and hence, it is not a characteristic function.

To show this result, we use the following result:

Theorem 1.19 *$\Lambda \subset \mathcal{P}(\mathbf{R})$ is tight $\overset{\text{def}}{\iff}$ $^\forall\varepsilon > 0, {}^\exists K_\varepsilon \subset \mathbf{R}$: a compact set; $^\forall\mu \in \Lambda, \mu(K_\varepsilon) > 1 - \varepsilon$. $\iff$ Λ is relatively compact, i.e., $^\forall\{\mu_n\} \subset \Lambda$, $^\exists\{n_k\}, \mu \in \mathcal{P}(\mathbf{R}); \mu_{n_k} \to \mu$*

Proof. It is easy to show that "relatively compact is tightness". If it is not tight, then $^\exists\varepsilon_0 > 0, {}^\forall K$: a compact set, $^\exists\mu_K \in \Lambda; \mu_K(K) < 1 - \varepsilon_0$. Let $K = [-n, n]$ and $\mu_n = \mu_K$. By the assumption $^\exists\mu_{n_k} \to {}^\exists\mu \in \mathcal{P}(\mathbf{R})$. If $n_k \geq n$, then $\mu_{n_k}([-n, n]) \leq \mu_{n_k}([-n_k, n_k]) < 1 - \varepsilon$. Letting $k \to \infty$, we have $\mu((-n, n)) \leq \liminf \mu_{n_k}([-n, n]) \leq 1 - \varepsilon$. On the other hand, since $n \geq 1$ is arbitrary, we have $\mu(\mathbf{R}) \leq 1 - \varepsilon$. This contradicts.

On the inverse, "tightness implies relatively compact", we give the outline.

We first consider the distribution functions $F_n(r_k) = \mu_n((-\infty, r_k]) \in [0, 1]$ for rational numbers $\{r_k\} = \mathbf{Q}$. For each $k = 1, 2, \ldots$, if we take a sub-sequence $\{n_j^k\}$ such that $\{n_j^{k+1}\} \subset \{n_j^k\}$ and $F_{n_j^k}(r_k)$

converges, then the diagonal sequence $n_j = n_j^j$; $F_{n_j}(r_k)$ converges for $^\forall k$. We define the limit as $F(r_k) := \lim_{j\to\infty} F_{n_j}(r_k)$. Moreover, for $^\forall x \in \mathbf{R}$, we define $F(x) = \inf_{r>x; r\in\mathbf{Q}} F(r)$. Then it is non-decreasing, right-continuous and has left-hand limits. Furthermore, by tightness, it holds $F(-\infty) = 0, F(\infty) = 1$. Therefore, $F(x)$ is a distribution function and it is possible to show that $F_{n_j}(x) \to F(x)$ for all continuous points x of F. ■

Here we give a result which is used in the following proof.

- $\{\mu_n\} \subset \mathcal{P}(\mathbf{R})$ *is relatively cpt and* $^\exists\mu \in \mathcal{P}(\mathbf{R}); ^\forall\{n_k\}, \mu_{n_k} \to \mu \Longrightarrow \mu_n \to \mu$.

In fact, if it is not so, then $^\exists g \in C_b(\mathbf{R}^d); \langle\mu_n, g\rangle \not\to \langle\mu, g\rangle$. Hence, we have $^\exists\{n_k\}; ^\exists\langle\mu_{n_k}, g\rangle \neq \langle\mu, g\rangle$. Moreover by the relativley compactness, we have $^\exists\{n_{k_j}\}; \mu_{n_{k_j}} \to {}^\exists\widetilde{\mu} \neq \mu$ because of $\langle\widetilde{\mu}, g\rangle = \lim\langle\mu_{n_{k_j}}, g\rangle \neq \langle\mu, g\rangle$. However, this contradicts $\widetilde{\mu} = \mu$. ■

[Proof of Lévy's Continuity Theorem] By the assumption, we can show $\{\mu_n\}$ is tight, and hence, it is relatively compact; $^\exists\mu_{n_k} \to {}^\exists\mu \in \mathcal{P}(\mathbf{R})$. Thus, $\varphi = \varphi_\mu$, and again by relatively compactness, we have $\mu_n \to \mu$. In fact, if it is not so, then by the above result $^\exists\mu_{n_k} \to {}^\exists\widetilde{\mu} \neq \mu$. Hence, $\varphi_{\widetilde{\mu}} = \varphi_\mu$. By the uniqueness theorem we have $\widetilde{\mu} = \mu$. This contradicts.

On the tightness, in general, for $^\forall\nu \in \mathcal{P}(\mathbf{R}), ^\forall L > 0$, it holds

$$\nu([-2L.2L]^c) \le L\int_{-1/L}^{1/L}(1-\varphi_\nu(z))dz.$$

In fact, by Fubini (note that $|1 - e^{izx}| \le 2$), we have

$$\begin{aligned}(\text{RHS}) &= L\int_{-1/L}^{1/L} dz\int_{\mathbf{R}}(1-e^{izx})\nu(dx) = \int_{\mathbf{R}}\nu(dx)L\int_{-1/L}^{1/L}(1-e^{izx})dz \\ &= \int_{\mathbf{R}}\nu(dx)2\int_0^1(1-\cos(zx/L))dz = \int_{\mathbf{R}}2\left(1-\frac{\sin x/L}{x/L}\right)\nu(dx).\end{aligned}$$

By $1 - \dfrac{\sin x/L}{x/L} \ge 0$ and by $\dfrac{\sin x/L}{x/L} \le \dfrac{1}{2}$ if $|x| \ge 2L$, we get the above.

Now, since φ is continuous at 0, we have

$$\lim_{L\to\infty} L\int_{-1/L}^{1/L}(1-\varphi(z))dz = \lim_{h\to 0}\frac{1}{h}\int_{-h}^{h}(1-\varphi(z))dz = 2(1-\varphi(0)) = 0,$$

where the last equal is due to $\varphi(0) = \lim\varphi_n(0) = 1$. Hence, noting that $\varphi_n \to \varphi$ and $|1 - \varphi_n(z)| \le 2$, we have $^\forall\varepsilon > 0, ^\exists L > 1$; $\limsup\mu_n([-2L, 2L]^c) < \varepsilon$. Therefore, $^\exists N; a$
$n \ge N, \mu_n([-2L, 2L]^c) < \varepsilon$, and by taking a large L if necessary, it also holds for $^\forall n < N$. ■

Finally, we show "$\mu_n \to \mu \Longrightarrow \phi_n \to \phi$ (uniform on compact sets)".

$\{\mu, \mu_n\}$ is relatively compact, and hence, it is tight; $^\forall\varepsilon > 0, ^\exists K \subset \mathbf{R}$: a compact set; $\mu(K), \mu_n(K) > 1 - \varepsilon$. We have

$$^\exists\delta > 0; |^\forall h| < \delta, \sup_{x\in K}|e^{ihx} - 1| < \varepsilon.$$

Thus, for the same h, $\sup_{z\in\mathbf{R}}|\varphi_n(z+h) - \varphi_n(z)| < 3\varepsilon$ holds. For any compact sets $C \subset \mathbf{R}$, by taking $\{z_1, \ldots, z_k\}$: δ-dense points, i.e., $^\forall z \in C, ^\exists j; |z - z_j| < \delta$, and by letting $N \ge 1$ such that $^\forall n \ge N, |\varphi_n(z_j) - \varphi(z_j)| < \varepsilon$, it holds that uniformly in $^\forall z \in C$,

$$|\varphi_n(z) - \varphi(z)| \le |\varphi_n(z) - \varphi_n(z_j)| + |\varphi_n(z_j) - \varphi(z_j)| + |\varphi(z_j) - \varphi(z)| < 7\varepsilon.$$

■

2 Large Deviation Principle (=LDP)

Let RVs $\{X_n\}$ be i.i.d., $EX_1 = m \in \mathbf{R}$ and $V(X_1) = v > 0$. Set $S_n = \sum_{k=1}^n X_k$. For $a > m$, we investigate $P(S_n > an) \sim$? as $n \to \infty$.

By LLN, $S_n/n \to m$, a.s. holds and this implies $P(S_n > an) \to 0$. Moreover, By CLT,

$$P\left(S_n > \sqrt{n}a + mn\right) = P\left(\frac{S_n - mn}{\sqrt{n}} > a\right) \to \frac{1}{\sqrt{2\pi v}} \int_a^\infty e^{-x^2/(2v)} dx.$$

Hence, if we change a to $\sqrt{n}(a-m)$, then roughly speaking, we get

$$P\left(S_n > na\right) \sim \frac{1}{\sqrt{2\pi v}} \int_{\sqrt{n}(a-m)}^\infty e^{-x^2/(2v)} dx$$

We investigate the decreasing order and the coefficient as $n \to \infty$.

Clearly it is a probability of a part far from the center (mean), and hence, it is "Large Deviation".

The following result holds.

Theorem 2.1 (LDP ; Cramér's Theorem) *Let $\{X_n\}$ be i.i.d. and assume that for $^\forall t \in \mathbf{R}, E[e^{t|X_1|}] < \infty$. Set $\psi(t) = E[e^{tX_1}]$ and $S_n = \sum_{k=1}^n X_k$. Then, It holds that for $^\forall a > EX_1$,*

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \geq an) = -I(a), \ \textit{i.e.}, \ P(S_n \geq an) \sim e^{-nI(a)},$$

where $I(x) = \sup_{t\in\mathbf{R}}(xt - \log\psi(t))$ is a lower semi-continuous, convex function on $\mathbf{R}$ satisfying that

$$\lim_{x\to\pm\infty} I(x) = \infty, \quad I(x) \geq 0 = I(EX_1) \quad (x \in \mathbf{R}).$$

Cramér showed under the weak condition; $^\exists t > 0; E[e^{t|X_1|}] < \infty$.

Note that

· I: **lower semi-conti.** at $x \iff {}^\forall\varepsilon > 0, {}^\exists\delta > 0; |{}^\forall y - x| < \delta, I(y) > I(x) - \varepsilon$. This implies $I(x) \leq \liminf_{y\to x} I(y)$.

· I: **convex** on an interval $J \iff {}^\forall p, q \geq 0; p + q = 1, {}^\forall x, y \in J, I(px + qy) \leq pI(x) + qI(y)$.

We first give the following useful result:

Theorem 2.2 (sub-additive theorem) *If a sequence $\{a_n\}$ satisfies sub-additivity; $a_{m+n} \leq a_m + a_n$, then $^\exists \lim(a_n/n) = \inf(a_n/n)$. On the other hand, if $\{b_n\}$ satisfies super-additivity; $b_{m+n} \geq b_m + b_n$, then $^\exists \lim(b_n/n) = \sup(b_n/n)$.*

Proof. By $\liminf(a_n/n) \geq \inf(a_n/n)$, it is enough to show that for $^\forall m \geq 1$, $\limsup(a_n/n) \leq a_m/m$. Fix an arbitrary $m \geq 1$ and for any $n \geq 1$, by dividing n by m we have $n = km + r$ with $0 \leq r < m$ and $k = k_n \geq 0$. By sub-additivity, $a_n \leq ka_m + a_r$. and dividing by n, we have

$$\frac{a_n}{n} \leq \frac{km}{km + r} \cdot \frac{a_m}{m} + \frac{a_r}{n}.$$

Hence, by taking $\limsup_{n\to\infty}$ in both sides, and by $k = k_n \to \infty$ we get the desired result. ∎

We first discuss the existence of the limit of $n^{-1}\log P(S_n \geq na)$ for $a > EX_1$.

By $\{X_n\}$ being i.i.d.

$$P(S_m \geq ma)P(S_n \geq na) = P(S_{m+n} - S_n \geq ma, S_n \geq na) \leq P(S_{m+n} \geq (m+n)a).$$

If we set $b_n = \log P(S_n \geq na)$, then it has super-additivity. Thus, by sub-additive theorem we have $^\exists \lim(b_n/n) = \sup(b_n/n) =: -\widetilde{I}(a)$ and $0 \leq \widetilde{I}(a) \leq \infty$.

In order to show LDP, we need some results such that there exists a probability measure corresponding to a distribution function, and it can be possible to construct infinitely many independent RVs with the same distribution (and a probability space). Moreover, we need Existence Theorem of Lebesgue-Stielties Measures and Kolmogorov's Extension Theorem.

We describe them later, and we first give the proof of LDP.

[Proof of LDP] Let $I(a) := \sup_t\{at - \log\psi(t)\}$. In order to show that for $^\forall a > EX_1$, $\widetilde{I}(a) = I(a)$, i.e., $\lim_{n\to\infty}\frac{1}{n}\log P(S_n \geq an) = -I(a)$, it is enough to show that if $a = 0 > EX_1$, then

$$\lim_{n\to\infty}\frac{1}{n}\log P(S_n \geq 0) = \log\rho,$$

where $\rho := \inf_{t\in\mathbf{R}}\psi(t)$ (note that $\log 0 = -\infty$). Because for $a > EX_1$, by considering $X_1 - a$ instead of X_1, we have $E[X_1 - a] < 0$ and $\psi(t)$ is changed to $e^{-ta}\psi(t)$, and $I(a)$is to $I(0) = -\log\rho$.

For simplicity, we denote $X_1 = X$. By the assumption $^\forall t \in \mathbf{R}, E[e^{t|X|}] < \infty$, we see that $^\forall n \geq 1, ^\forall t \in \mathbf{R}$, $E[|X|^n e^{t|X|}] < \infty$. (In fact, if $x > 0$, then $x^n \leq n!e^x$.) hence, by the convergence theorem (more exactly, by the following exchange theorem of differential and integral; Theorem 2.3), $\psi(t)$ is in C^∞ and $\psi'(t) = E[Xe^{tX}]$, $\psi''(t) = E[X^2e^{tX}] > 0$. (Note that $P(X = 0) < 1$, i.e., $P(X \neq 0) > 0$ by $EX < 0$.) Hence, $\psi(t)$ is strictly convex in $t \in \mathbf{R}$ and $\psi'(0) = EX < 0$.

(Case 1) $P(X \leq 0) = 1, P(X = 0) \geq 0$.

ψ is decreasing and by the convergence theorem, $\psi(t) \downarrow P(X = 0) = \rho \geq 0$ $(t \uparrow \infty)$. If $\rho > 0$, then $P(S_n \geq 0) = P(X_1 = 0, \cdots, X_n = 0) = \rho^n$ and the left hand side of the desired equation is equal to $\log\rho$. Even if $\rho = 0$, then $P(S_n \geq 0) = 0$, and by $\log 0 = -\infty$, we get the desired result as $-\infty = -\infty$. (We may think this calculus is contained to the above.)

(Case 2) $P(X < 0) > 0, P(X > 0) > 0$.

We have $\psi(\pm\infty) = \infty$. In fact, by the continuity of probability measure, $^\exists\delta > 0; P(X \geq \delta) > 0$, and hence, if $0 < t \to \infty$, then $\psi(t) \geq E[e^{tX}; X \geq \delta] \geq e^{\delta t}P(X \geq \delta) \to \infty$. The case of $0 > t \to -\infty$ is the same. On the other hand, since ψ is strictly convex and $\psi'(0) = EX < 0$, $^\exists\tau > 0; \psi'(\tau) = 0, \psi(\tau) = \inf\psi = \rho > 0$. For the distribution function $F(x) = P(X \leq x)$ of X, we use a **Cramér transform**:

$$\widehat{F}(x) := \frac{1}{\rho}\int_{-\infty}^{x} e^{\tau y}dF(y) = \frac{1}{\rho}E[e^{\tau X}; X \leq x], \text{i.e., } d\widehat{F}(x) = \tfrac{1}{\rho}e^{\tau x}dF(x).$$

By $\widehat{F}(\infty) = \psi(\tau)/\rho = 1$, $\widehat{F}$ is also a distribution function. Let $\{\widehat{X}_n\}$ be i.i.d. with $\widehat{F}$ as a distribution function. Let $\widehat{S}_n = \sum_{k=1}^n \widehat{X}_k$.

Then, it holds that

$$E\widehat{X}_1 = 0, \quad \widehat{\sigma}^2 := V(\widehat{X}_1) \in (0,\infty), \quad P(S_n \geq 0) = \rho^n E\left[e^{-\tau\widehat{S}_n}1_{\{\widehat{S}_n\geq 0\}}\right] \tag{2.1}$$

Now if we admit these, then for $^\forall M > 0$,

$$e^{-\tau M\widehat{\sigma}\sqrt{n}}P\left(0 \leq \frac{\widehat{S}_n}{\widehat{\sigma}\sqrt{n}} < M\right) \leq E\left[e^{-\tau\widehat{S}_n}1_{\{\widehat{S}_n\geq 0\}}\right] \leq 1.$$

Moreover, since CLT holds for $\widehat{S}_n$, the above probability converges to $\displaystyle\int_0^M e^{-x^2/2}dx/\sqrt{2\pi} > 0$ as $n \to \infty$. Thus, in each terms, by taking log, dividing by n and taking lim inf, we have

$$\liminf_{n\to\infty}\frac{1}{n}\log E\left[e^{-\tau\widehat{S}_n}1_{\{\widehat{S}_n\geq 0\}}\right] = 0.$$

Hence, the desired result is obtained:

$$\lim_{n\to\infty}\frac{1}{n}\log P(S_n \geq 0) = \log\rho.$$

We show (2.1). Let $\widehat{\psi}(t) := E[e^{t\widehat{X}_1}]$. Then,

$$\widehat{\psi}(t) = \int_{\mathbf{R}} e^{tx} d\widehat{F}(x) = \frac{1}{\rho}\int_{\mathbf{R}} e^{tx}e^{\tau x} dF(x) = \frac{1}{\rho}\psi(t+\tau).$$

Hence, $\widehat{\psi}$ is also in C^∞ and $E\widehat{X}_1 = \widehat{\psi}'(0) = \psi'(\tau)/\rho = 0$, $V(\widehat{X}_1) = \widehat{\psi}''(0) = \psi''(\tau)/\rho \in (0,\infty)$. Moreover, by $d\widehat{F}(x) = \frac{1}{\rho}e^{\tau x}dF(x)$, we have $dF(x) = \rho e^{-\tau x}d\widehat{F}(x)$. Thus,

$$P(S_n \geq 0) = \int_{\{\sum_{k=1}^n x_k \geq 0\}} dF(x_1)\cdots dF(x_n) = \rho^n \int_{\{\sum_{k=1}^n x_k \geq 0\}} e^{-\tau\sum_{k=1}^n x_k} d\widehat{F}(x_1)\cdots d\widehat{F}(x_n)$$

Therefore, the last equation is obtained.

Finally, we investigate the properties of $I(x) = \sup_{t\in\mathbf{R}}\{xt - \log\psi(t)\}$. In general, a supremum of a family of linear functions is convex, and supremum of a family continuous functions is lower semi-continuous ($\to$ make sure.) Hence, the supremum $I(x)$ of linear functions $x \mapsto xt - \log\psi(t)$ is lower semi-continuous and convex.

On $I(x) \to \infty$ ($x \to \pm\infty$), as $x \to \infty$, if we assume it is not, then ${}^\exists L > 0, {}^\exists x_n \geq n; I(x_n) < L$, i.e., ${}^\forall t \in \mathbf{R}, x_n t - \log\psi(t) < L$. By taking $t = 2L/x_n$, we have $2L - \log\psi(2L/x_n) < L$. By $x_n \to \infty$ and by the continuity of ψ, $0 < L < \log\psi(2L/x_n) \to \log\psi(0) = \log 1 = 0$. However, this is contradict. In case of $x \to -\infty$ is the same.

On $I(x) \geq 0 = I(EX)$, $I(x) \geq -\log\psi(0) = 0$ (taking $t = 0$) and $-\log x$ is convex, and hence, by Jensen's inequality (see the next question), we have ${}^\forall t \in \mathbf{R}$, $\log\psi(t) = \log E[e^{tX}] \geq E[\log e^{tX}] = tEX$, i.e., $tEX - \log\psi(t) \leq 0$. Moreover, by $I(x) \geq 0$, $0 \leq I(EX) = \sup_t(tEX - \log\psi(t)) \leq 0$. Thus, $I(EX) = 0$. ∎

Theorem 2.3 (Exchange Theorem of Differential and Integral) *Let $(X, \mathcal{F}, \mu)$ be a general measure space. For each $t \in (a,b)$, let $f_t = f_t(x) \in L^1(d\mu)$ and for μ-a.e $x \in X$, $f_t(x)$ be differentiable in $t \in (a,b)$. If $\sup_{t\in(a,b)}|\partial_t f_t| \in L^1$, then*

$$d_t \int_X f_t(x)\mu(dx) = \int_X \partial_t f_t(x)\mu(dx).$$

Especially, if $(X,\mathcal{F}) = (\mathbf{R}, \mathcal{B}^1)$ and $\mu = \mu_X$ is a distribution of a RV X, then

$$d_t E[f_t(X)] = E[\partial_t f_t(X)],$$

where $d_t = d/dt, \partial_t = \partial/\partial t$.

Proof. It is clear by mean-value theorem and Lebesgue's convergence theorem. In fact, for $t, t+h \in (a,b); h \neq 0$, by mean-value theorem, for a.e. x,

$${}^\exists\theta \in (0,1);\ \frac{1}{h}(f_{t+h}(x) - f_t(x)) = \partial_t f_{t+\theta h}(x).$$

Therefore, by the assumption, this can be estimated by an integrable function which is independent of h and t. Hence, by the convergence theorem, we can change the limit as $h \to 0$ and the integral in x. ∎

Question 2.1 (Jensen's ineq.) Let $-\infty \leq a < b \leq \infty$. For a convex function f on a interval $I = (a,b)$ and I-valued integrable RV X; $f(X) \in L^1$, show

$$f(EX) \leq E[f(X)], \quad \text{i.e.,} \quad f\left(\int_{\mathbf{R}} x\mu_X(dx)\right) \leq \int_{\mathbf{R}} f(x)\mu_X(dx)$$

A convex function is the supremum of a family of linear functions which are lower than or equal to it, i.e., $f(x) = \sup\{cx + d; cy + d \leq f(y)({}^\forall y \in I)\}$. Hence, it is clear by $cEX + d \leq E[f(X)]$. In fact, for $a < s < t < u < b$, by convexity

$$\frac{f(t) - f(s)}{t - s} \leq \frac{f(u) - f(t)}{u - t}.$$

(if $t = ps + qu; p, q \geq 0, p + q = 1$, then $f(t) \leq pf(s) + qf(u)$, and hence, $p(f(t) - f(s)) \leq q(f(u) - f(t))$. by $p = (u - t)/(u - s), q = (t - s)/(u - s)$ and it is obtained by multiplying by $(u - s)$ and dividing by $(u-t)(t-s)$.) Let $\alpha_t = \sup_{s<t}$(the left-hand side). Then, $\alpha_t(u-t) \leq f(u)-f(t)$, i.e., $f(u) \geq \alpha_t(u-t)+f(t)$ $(u > t)$ This holds for $u \leq t$ by the definition of α_t, and thus, it holds for all $a < u < b$. Therefore, $a < t < b$ are arbitrary. Clearly, it is equal if $t = u$, and it is expressed as a supremum of linear functions.

In another way, directly, for $f(u) \geq \alpha_t(u - t) + f(t)$, by substituting $t = EX, u = X$ and by taking expectations we get the desired result. ■

Question 2.2 Show a function as the supremum of a family of continupus functions is lower semi-continuous.
$^{\forall}t \in T$, $f_t(x)$: conti. in $x \Longrightarrow g = \sup_{t\in T} f_t$ is lower semi-conti.

$^{\forall}x$: fixed. By the property of supremum, $^{\forall}\varepsilon > 0, ^{\exists}t \in T; g(x) - \varepsilon/2 < f_t(x)$ $(\leq g(x))$. Since f_t is conti. at x, $^{\exists}\delta > 0; |^{\forall}y - x| < \delta, |f_t(y) - f_t(x)| < \varepsilon/2$. Hence, $f_t(x) - \varepsilon/2 < f_t(y)$. From these, we have $g(y) \geq f_t(y) > f_t(x) - \varepsilon/2 > g(x) - \varepsilon$. ■

3 Extension Theorem of Measures and Its Applications

In this section, we describe three important theorems which are particularly necessary in probability theory, by using Extension Theorem of Measures.

We first describe the probabilistic version of Extension Theorem of Measures.

[**Extension Theorem of Probability Measures**] Let $\mathcal{A}$ be an additive class on Ω. If a finite additive set function $P_0 : \mathcal{A} \to [0, 1]$ satisfies $P_0(\Omega) = 1$ and if it is σ-additive on $\mathcal{A}$, i.e.,

$$A_n \in \mathcal{A}\ (n = 1, 2, \dots) \text{ are disjoint and } \bigcup_{n=1}^{\infty} A_n \in \mathcal{A} \implies P_0(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P_0(A_n), \tag{3.1}$$

then there exists uniquely a probability measure P on $(\Omega, \sigma(\mathcal{A}))$ such that $P|_{\mathcal{A}} = P_0$.

This P is given as an **outer measure** of P_0:

$$P(A) = \inf\left\{\sum_{n=1}^{\infty} P_0(A_n); A_n \in \mathcal{A}, \bigcup_{n=1}^{\infty} A_n \supset A\right\},$$

where $A \subset \Omega$ is an arbitrary subset

We give the equivalent condition to (3.1) "σ-additive on $\mathcal{A}$":

$$A_n \in \mathcal{A} \downarrow \emptyset \implies \lim_{n\to\infty} P_0(A_n) = 0 \tag{3.2}$$

$$A_n \in \mathcal{A}\downarrow, \lim_{n\to\infty} P_0(A_n) > 0 \implies \bigcap_{n\geq 1} A_n \neq \emptyset \tag{3.3}$$

In practical applications we often check (3.3).

3.1 Infinite-dimensional product probability spaces

Let $(\Omega_n, \mathcal{F}_n, P_n)$ be a probability space. We set an n-product probability space as follows:

$$(\Omega^{(n)}, \mathcal{F}^{(n)}, P^{(n)}) := (\prod_{k=1}^{n} \Omega_k, \bigotimes_{k=1}^{n} \mathcal{F}_k, \bigotimes_{k=1}^{n} P_k)$$

Moreover, set $\Omega := \prod_{n=1}^{\infty} \Omega_n$ and

$$\mathcal{A} = \left\{A = A_n \times \Omega_{n+1} \times \Omega_{n+2} \times \cdots ; A_n \in \mathcal{F}^{(n)}, n = 1, 2, \dots\right\}$$

Then, $\mathcal{A}$ is an additive class on Ω. For $A = A_n \times \Omega_{n+1} \times \Omega_{n+2} \times \cdots$, set $P_0(A) = P^{(n)}(A_n)$, then $P_0(\emptyset) = 0$ and P_0 can be a finite additive function on $\mathcal{A}$. It is possible to show that this satisfies (3.3). Therefore, $^{\exists_1}P$ a probability measure on $\sigma(\mathcal{A})$; $P = P_0$ on $\mathcal{A}$. That is, for each n,

$$P(A_1 \times \cdots \times A_n \times \Omega_{n+1} \times \Omega_{n+2} \times \cdots) = P_1(A_1) \cdots P_n(A_n) \quad (A_k \in \mathcal{F}_k)$$

and P is unique. We denote $\bigotimes_{n=1}^{\infty} \mathcal{F}_n := \sigma(\mathcal{A})$, $\bigotimes_{n=1}^{\infty} P_n := P$ and

$(\prod_{n=1}^{\infty} \Omega_n, \bigotimes_{n=1}^{\infty} \mathcal{F}_n, \bigotimes_{n=1}^{\infty} P_n)$ is called $(\Omega_n, \mathcal{F}_n, P_n)$ $(n = 1, 2, \dots)$ an **infinite(-dimensional) product probability space**.

[Construction of independent RVs]

By the above result we can construct infinite countable number of independent RVs such that they have given distributions. For instance, for each $n \geq 1$, let μ_n be a distribution on $(\mathbf{R}_+ = [0, \infty), \mathcal{B}(\mathbf{R}_+))$. We denote their infinite product probability space as $(\Omega, \mathcal{F}, P)$, and for $\omega = (\omega_n) \in \Omega = \mathbf{R}_+^{\infty}$, let $X_n(\omega) = \omega_n$, then X_n are RVs with distributions μ_n and independent. In fact,

$$P(X_1 \in A_1, X_2 \in A_2) = P(A_1 \times A_2 \times \mathbf{R}_+^{\infty}) = \mu_1(A_1)\mu_2(A_2)$$

In case of $\Omega^{(n)} = \mathbf{R}^n$, the above result can be extended to the following.

3.2 Kolmogorov's extension theorem

For each $n \geq 1$, let $(\mathbf{R}^n, \mathcal{B}^n, P_n)$ be an n-dimensional probability space. These satisfy the following consistency condition:

$$P_n(A) = P_{n+1}(A \times \mathbf{R}) \quad (A \in \mathcal{B}^n).$$

Let

$$\mathcal{B}^{\infty} = \mathcal{B}(\mathbf{R}^{\infty}) := \sigma(\bigcup_{n \geq 1} (\mathcal{B}^n \times \mathbf{R}^{\infty}))$$

This is called an **infinite-dimensional Borel filed**. Then, there exists uniquely a probability measure P on $(\mathbf{R}^{\infty}, \mathcal{B}^{\infty})$ such that $P(A_n \times \mathbf{R}^{\infty}) = P_n(A_n)$ $(A_n \in \mathcal{B}^n)$.

In fact, set $\mathcal{A} = \{A = A_n \times \mathbf{R}^{\infty}; A_n \in \mathcal{B}^n, n = 1, 2, \dots\}$ and for $A = A_n \times \mathbf{R}^{\infty}$, set $P_0(A) = P_n(A_n)$, Then $P_0(\emptyset) = 0$ and P_0 can be a finite additive set function on $\mathcal{A}$. It is possible to show P_0 satisfies (3.3). Therefore, there exists uniquely a probability measure P on $\mathcal{B}^{\infty} = \sigma(\mathcal{A})$ such that $P(A_n \times \mathbf{R}^{\infty}) = P_n(A_n)$ $(A_n \in \mathcal{B}^n)$.

The following two results hold in general measure spaces, however, we describe the probabilistic versions.

Theorem 3.1 (Approximating Theorem) *Let $(\Omega, \mathcal{F}, P)$ be a probability space. If $\mathcal{A} \subset \mathcal{F}$ is an additive class, then*

$$^{\forall}A \in \sigma(\mathcal{A}), ^{\exists}A_n \in \mathcal{A}; \lim_{n \to \infty} P(A \triangle A_n) = 0,$$

(where $A \triangle B = (A \setminus B) \cup (B \setminus A)$ is a symmetric difference).

Proof. Let $\mathcal{G}$ be a family of all $A \subset \Omega$ satisfying the condition of the theorem. It can be seen that this contains $\mathcal{A}$ and this is a σ-additive class. Therefore, $\sigma(\mathcal{A}) \subset \mathcal{G}$ and the desired result is obtained. ■

Question 3.1 Show the above $\mathcal{G}$ is a σ-additive class.

For a countable union, let $^\forall\varepsilon > 0$. For each $A_n \in \mathcal{G}$, fix $B_n \in \mathcal{A}; P(A_n \triangle B_n) < \varepsilon/2^{n+1}$. we approximate $A := \bigcup A_n$ by a finite sum $\bigcup_{n\leq N} A_n$ (by the upper continuity, we estimate the difference of probabilities by $\varepsilon/2$). Moreover, since it can be approximated by $C_N := \bigcup_{n\leq N} B_n \in \mathcal{A}$, we get $A \in \mathcal{G}$. In fact, by

$$A\triangle\left(\bigcup_{n=1}^N B_n\right) \subset \left(A \setminus \left(\bigcup_{n=1}^N A_n\right)\right) \cup \bigcup_{n=1}^N (A_n \triangle B_n)$$

it holds that

$$P\left(A\triangle\left(\bigcup_{n=1}^N B_n\right)\right) \leq P\left(A \setminus \left(\bigcup_{n=1}^N A_n\right)\right) + \sum_{n=1}^N P(A_n\triangle B_n) < \varepsilon.$$

■

Corollary 3.1 *Let $(\mathbf{R}^d, \mathcal{B}^d, P)$ be a probability space. For each $A \in \mathcal{B}^d$, $^\exists C_n$ a bounded closed set and $^\exists G_n$ a open set such that $C_k \subset A \subset G_k$, $\lim P(A \setminus C_n) = \lim P(G_n \setminus A) = 0$.*

Proof. Let $\mathcal{A}$ be a family of all finite unions of basic rectangles $\prod_{k=1}^d (a_k, b_k]$. It is an additive class and $\sigma(\mathcal{A}) = \mathcal{B}^d$. Since $P(A)$ is also an outer measure;

$$P(A) = \inf\left\{\sum_{n\geq 1} P(A_n); A_n \in \mathcal{A}, \bigcup_{n\geq 1} A_n \supset A\right\}$$

for $^\forall\varepsilon > 0$, A can be approximated by $\bigcup A_n$ from outer as the difference of probabilities is lower than $\varepsilon/2$ (we simply call A is $\varepsilon/2$-approximated). Each A_n is clearly $\varepsilon/2^{n+1}$-approximated by a finite unions of product sets of open intervals. The countable union of them is also an open set and A is $\varepsilon/2$-approximated. Thus, A is ε-approximated by an open set from outer. Moreover, by the complement, A is approximated by a closed set from inner, and it is approximated by a bounded closed set (it is enough to consider the intersection with $\{|x| \leq n\}$). ■

3.3 Topics on independence of infinitely many numbers

In this section, for each $n \geq 1$, let X_n be a real-valued RV and $\mathcal{F}_n$ be a sub σ-additive class of a σ-additive class $\mathcal{F}$.

In the Borel-Cantelli's lemma, we showed that if the sum of probabilities of infinitely many events is finite, then the probability of the upper limit is 0;

$$A_n \in \mathcal{F}, \sum P(A_n) < \infty \Longrightarrow P(\limsup A_n) = 0.$$

However, when the sum is infinite, does it hold that the probability is positive or 1?

One of the answer for this question is the following result:

Theorem 3.2 (Borel-Cantelli's Theorem) *When $A_n \in \mathcal{F}$ are independent, if $\sum P(A_n) = \infty$, then $P(\limsup A_n) = 1$.*

Proof. $P(\liminf A_n^c) = \lim_{N\to\infty} P(\bigcap_{n\geq N} A_n^c)$, and noting that $\{A_n^c\}$ are also independent, we have $P(\bigcap_{n\geq N} A_n^c) = \prod_{n\geq N}(1 - P(A_n))$. By $1 - x \leq e^{-x}$ and $\sum_{n\geq N} P(A_n) = \infty$, we have $P(\liminf A_n^c) = 0$. ■

A element of σ-additive class $\bigcap_{N\geq 1} \sigma\left(\bigcup_{n\geq N} \mathcal{F}_n\right)$ is called a **tail event**. Moreover, a RV which is measurable w.r.t.(=with respect to) this, is called a **tail function**. Furthermore, if $\mathcal{F}_n = \sigma(X_n) = X_n^{-1}\mathcal{B}^1)$, then they are called a **tail event, tail function w.r,t.** $\{X_n\}$.

Question 3.2 Show the following: For $A_n \in \mathcal{F}_n$, $\limsup A_n, \liminf A_n$ are tail events. $\{X_n \to 0\}$ is also a tail event w.r.t. $\{X_n\}$.

Roughly speaking, a tail event or a tail function is an event or a RV if it is independent of finite number of $\mathcal{F}_1, \ldots, \mathcal{F}_n$ or $X_1, \ldots, X_n$, respectively.

Theorem 3.3 (Kolmogorov's 0-1 law) *If sub σ-additive classes $\mathcal{F}_n \subset \mathcal{F}$ are independent, then the probabilities of all tail events are* 0 *or* 1.

Proof. Note that $P(A) = 0$ or $1 \iff P(A)^2 = P(A) \iff A$ is independent of its self. In order to show this, we use Approximating Theorem and we can take an approximating event which is independent of A. Set

$$\mathcal{G}_n = \sigma\left(\bigcup_{k\leq n} \mathcal{F}_k\right), \quad \mathcal{A} = \bigcup_{n\geq 1} \mathcal{G}_n,$$

then $\mathcal{A}$ is an additive class and $\sigma(\mathcal{A}) = \sigma\left(\bigcup_{k\geq 1} \mathcal{F}_k\right)$ holds. By $A \in \sigma(\mathcal{A})$ and Approximating Theorem, ${}^{\forall}\varepsilon > 0, A_\varepsilon \in \mathcal{A}; P(A \triangle A_\varepsilon) < \varepsilon$. By the definition of $\mathcal{A}$, ${}^{\exists}n; A_\varepsilon \in \mathcal{G}_n$ On the other hand, $A \in \sigma\left(\bigcup_{k\geq n+1} \mathcal{F}_k\right)$, is independent of A_ε. Therefore, A is independent of A, i.e., $P(A) = P(A \cap A) = P(A)^2$. In fact, by $A \subset (A \cap B) \cup (A \triangle B) \subset B \cup (A \triangle B)$, we have

$$P(A) \leq P(A \cap A_\varepsilon) + P(A \triangle A_\varepsilon) < P(A)P(A_\varepsilon) + \varepsilon \leq P(A)(P(A) + P(A \triangle A_\varepsilon)) + \varepsilon < P(A)^2 + 2\varepsilon.$$

By a similar way,

$$P(A) \geq P(A \cap A_\varepsilon) - P(A \triangle A_\varepsilon) > P(A)P(A_\varepsilon) - \varepsilon \geq P(A)(P(A) - P(A \triangle A_\varepsilon)) - \varepsilon > P(A)^2 - 2\varepsilon.$$

■

References

[1] KUMAGAI, Takashi. "Probability Theory" (in Japanese), Kyoritsu, (2003).

[2] SHIGA, Tokuzo, "From Lebesgue integrals to Probability Theory" (in Japanese), Kyoritsu (2000).

[3] NISHIO, Makiko, "Probability Theory" (in Japanese), Jikkyo (1978, 1st ed., 1985, 5 th ed.)