Appendix B

Mathematical Background on Probabilities

Po-Ning Chen, Professor

Department of Communications Engineering

National Chiao Tung University

Hsin Chu, Taiwan 300, R.O.C.

Probability space and random variable

- A *probability space* is a triple $(\Omega, \mathcal{F}, P)$, where
  - Ω is the set of all possible outcomes (often named *sample space*), and
  - $\mathcal{F}$ is the σ-field of Ω (often named *event space*), and
  - P is a probability measure on the σ-field, satisfying:
    1. $0 \leq P(A) \leq 1$ for $A \in \mathcal{F}$;
    2. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
    3. **(countable additivity)** if A_1, A_2, ... is a disjoint sequence of sets in $\mathcal{F}$, then
$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k).$$
- A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is a *real-valued* function on Ω (i.e., $X : \Omega \to \Re$), satisfying that the set $\{\omega : X(\omega) = x\} \in \mathcal{F}$ for each real x.

Probability space and random variable

- The observation probability space $(\Omega_X, \mathcal{F}_X, P_X)$ of a random variable X defined over the probability space $(\Omega, \mathcal{F}, P)$ satisfies:

$$\begin{cases} \Omega_X = X(\Omega) \\ \mathcal{F}_X = \{X(A) \subset \Re : A \in \mathcal{F}\} \\ P_X(G) = P\big(\{\omega \in \Omega : X(\omega) \in G\}\big) \text{ for any } G \in \mathcal{F}_X, \end{cases}$$

  where $X(A) = \{x \in \Re : X(\omega) = x \text{ for some } \omega \in A\}$.

Why define random variables based on $(\Omega, \mathcal{F}, P)$?

Answer: $(\Omega, \mathcal{F}, P)$ is what truly occurs internally, but possibly **non-observable**.

- In order to infer which of the *non-observable* ω occurs, an experiment that results in observable x that is a function of ω is performed.
- So x is a function of ω, which takes real values.
- Such an experiment results in the random variable X whose probability is so defined over the probability space $(\Omega, \mathcal{F}, P)$.

Why define random variables based on $(\Omega, \mathcal{F}, P)$?

- In applications, we are more interested in the observation probability space $(\Omega_X, \mathcal{F}_X, P_X)$ than the inherited probability space $(\Omega, \mathcal{F}, P)$ on which the random variable is defined.
- It can be proved that given a real-valued non-negative function $F(\cdot)$, satisfying that $\lim_{x\downarrow-\infty} F(x) = 0$ and $\lim_{x\uparrow\infty} F(x) = 1$, there exist a random variable and an inherited probability space such that the cumulate distribution function (cdf) of the random variable defined over the probability space is equal to $F(\cdot)$.
- The above result releases us with the burden of referring to a probability space before our defining a random variable. In other words, we can indeed define a random variable X directly by its cdf, i.e., $\Pr[X \leq x]$, without bothering to refer to its inherited probability space. Nevertheless, it is better to keep in mind (and learn) that a formal mathematical notion of random variables is defined over some probability space.
- In what follows, you will notice that most of the properties of random variables (random processes) are defined based on their *observation probability space*.

Random Processes

- A random process $\boldsymbol{X} = \{X_t, t \in I\}$ is a collection of random variables that arise from *common* probability space $(\Omega, \mathcal{F}, P)$.
  - Under this definition, any finite dimensional distribution of $\{X_t, t \in I\}$ is well-defined. For example,

$$\begin{aligned} & \Pr[X_1 \le x_1, X_5 \le x_5, X_9 \le x_9] \\ = \; & P\left(\{\omega \in \Omega : X_1(\omega) \le x_1, X_5(\omega) \le x_5, X_9(\omega) \le x_9\}\right) \end{aligned}$$

  - Again, we only consider its distribution $P_{\boldsymbol{X}}$ and *observation event space* $\mathcal{F}_{\boldsymbol{X}}$ in the sequel.

Random Processes in Communications

$\boldsymbol{X} = \ldots, X_{-3}, X_{-2}, X_{-1}, X_0, X_1, X_2, X_3, \ldots$

- Memoryless = independent and identically distributed (i.i.d.)
- First-order stationary = $(\forall\ i, j \in I)\ P_{X_i} = P_{X_j}$
- Second-order stationary = $(\forall\ i, j \in I \text{ and } \forall\ k)\ P_{X_i, X_j} = P_{X_{i+k}, X_{j+k}}$
- (Strictly) stationary =k-th-order stationary for all k
- Weakly stationary

$$(\forall\ i, j \in I \text{ and } \forall\ k) \quad E[X_i] = E[X_j] \text{ and } E[X_i X_j] = E[X_{i+k} X_{j+k}]$$

$\boldsymbol{X} = \ldots, X_{-3}, X_{-2}, X_{-1}, X_0, X_1, X_2, X_3, \ldots$

- Ergodic
  - Time-shift invariant subset/event in $\mathcal{F}_{\boldsymbol{X}}$:
    - Recall that an event is a subset of the sample space $\boldsymbol{X}(\Omega)$, and is an element of the observation event space $\mathcal{F}_{\boldsymbol{X}}$.
    - A time-shift invariance event $A \in \mathcal{F}_{\boldsymbol{X}}$ is one that the time shift counterpart of any element in A still belong to A. E.g.,

$$\begin{aligned} A \;=\; \{&(\ldots, x_{-1} = 0, x_0 = 1, x_1 = 0, x_2 = 1, \ldots), \\ &(\ldots, x_{-1} = 1, x_0 = 0, x_1 = 1, x_2 = 0, \ldots)\}. \end{aligned}$$

  - For any two time-shift invariant events, either they are disjoint, or one is contained in the other, or their intersection is also time-shift invariant.
  - Ergodic = every time-shift invariant event has either probability one or zero!

- Ergodic = every time-shift invariant event (in $\mathcal{F}_{\boldsymbol{X}}$) has either probability one or zero.
  - *Interpretations*: At most one of the following time-shift invariant events has probability one.

$$\left\{ x_{-\infty}^{\infty} \in \{0,1\}_{-\infty}^{\infty} \ : \ \lim_{n\to\infty} \frac{x_{-n} + \cdots + x_0 + \cdots + x_n}{2n+1} = 0.0 \right\}$$

$$\left\{ x_{-\infty}^{\infty} \in \{0,1\}_{-\infty}^{\infty} \ : \ \lim_{n\to\infty} \frac{x_{-n} + \cdots + x_0 + \cdots + x_n}{2n+1} = 0.1 \right\}$$

$$\vdots$$

$$\left\{ x_{-\infty}^{\infty} \in \{0,1\}_{-\infty}^{\infty} \ : \ \lim_{n\to\infty} \frac{x_{-n} + \cdots + x_0 + \cdots + x_n}{2n+1} = 0.9 \right\}$$

$$\left\{ x_{-\infty}^{\infty} \in \{0,1\}_{-\infty}^{\infty} \ : \ \lim_{n\to\infty} \frac{x_{-n} + \cdots + x_0 + \cdots + x_n}{2n+1} = 1.0 \right\}$$

  $\star$ Under boundedness, an outcome has the same limit (if it exists) as its shift counterpart.

$$\left| \frac{x_{-n} + \cdots + x_n}{2n+1} - \frac{x_{-n+1} + \cdots + x_{n+1}}{2n+1} \right| = \left| \frac{x_{-n} - x_{n+1}}{2n+1} \right| \leq \frac{1}{2n+1}.$$

- Stationarity $\Rightarrow$ Convergence of time average to a random variable
  - (One-sided) *Time average* is what we can truly observe and calculate!
  - E.g., $\{X_i\}_{i=1}^n$ i.i.d. with $\Pr\{X_i = 1\} = 1 - \Pr\{X_i = 0\} = U$, where $\Pr\{U = 0.2\} = \Pr\{U = 0.8\} = 0.5$.

$$\Rightarrow \quad \frac{X_1 + X_2 + \cdots + X_n}{n} \to U.$$

- Ergodicity and boundedness $\Rightarrow$ Time average converges to a constant
  - It does not guarantee that the *constant* is exactly the *ensemble mean*.
  - E.g. $\Pr\{(\ldots, x_{-1} = 0, x_0 = 1, x_1 = 0, x_2 = 1, \ldots)\} = 0.2$ and $\Pr\{(\ldots, x_{-1} = 1, x_0 = 0, x_1 = 1, x_2 = 0, \ldots)\} = 0.8$. Then

$$\frac{X_{-n} + \cdots + X_0 + \cdots + X_n}{2n+1} \to \frac{1}{2};$$

    but $E[X_i] =$ either 0.8 or 0.2.

- **Note**: Two alternative names for *time average* are *sample average* and *Cesáro mean*. Throughout the slides, we will use the *time average* since it gives more intuition to its formula.

- Stationary ergodicity $\Rightarrow$ Convergence of time average to ensemble average.

In mathematics:

$$\text{Stationarity} \;\Rightarrow\; \Pr\left\{\lim_{n\to\infty}\frac{X_{-n}+\cdots+X_0+\cdots+X_n}{2n+1}=Y\right\}=1$$

for some random variable Y

$$\text{Ergodicity and boundedness} \;\Rightarrow\; \Pr\left\{\lim_{n\to\infty}\frac{X_{-n}+\cdots+X_0+\cdots+X_n}{2n+1}=a\right\}=1$$

for some scalar a

$$\text{Stationary Ergodicity} \;\Rightarrow\; \Pr\left\{\lim_{n\to\infty}\frac{X_{-n}+\cdots+X_0+\cdots+X_n}{2n+1}=E[X_0]\right\}=1$$

Physical meaning of stationary ergodic assumption

- Stationary ergodic random source
  - Empirical distribution (relative frequency) can be used to approximate the true distribution.
  - Note that empirical distribution is what we can truly observe and calculate! E.g., observe a dice rolling experiment and obtain

$$X_1^{30} = 154326543334225632425644234443$$

Given that the experimental source is stationary ergodic, we can approximate the true distribution by (cf. pointwise ergodic theorem or Theorem B.1):

$$\Pr\{X_i = 1\} \approx \frac{1}{30} \qquad \Pr\{X_i = 2\} \approx \frac{6}{30} \qquad \Pr\{X_i = 3\} \approx \frac{7}{30}$$
$$\Pr\{X_i = 4\} \approx \frac{9}{30} \qquad \Pr\{X_i = 5\} \approx \frac{4}{30} \qquad \Pr\{X_i = 6\} \approx \frac{3}{30}$$

- Non-stationary or non-ergodic source
  - Empirical distribution (relative frequency) cannot necessarily be used to approximate the true distribution.

Physical meaning of stationary ergodic assumption

- In theories of communications, people assume that *the source is stationary* or *the source is stationary ergodic*. But you seldom find the assumption of *the source being ergodic but non-stationary*. Why?
  - Because an ergodic but non-stationary source not only does not facilitate the analytical study of communications problems, but seems no applications in practice.
  - From this, we learn that assumptions are made either to ease our analytical study of communications problem or to fit a specific need of applications. Without the two footings, an assumption becomes of minor interest.
- In other words, *ergodicity* assumption usually comes after *stationarity* assumption. A specific example is the **pointwise ergodic theorem**, where the random processes considered is presumed to be **stationary**.

Theorem B.1 (pointwise ergodic theorem) Give a discrete-time stationary random process $\{X_n\}_{-\infty<n<\infty}$. For arbitrary real-valued function $f(\cdot)$ on $\Re$ with finite mean ($|E[f(X_n)]| < \infty$), there exists a random variable $\hat{Y}$ such that

$$\Pr\left[\lim_{n\to\infty}\frac{1}{n}\sum_{n=0}^{\infty} f(X_n) = \hat{Y}\right] = 1.$$

If, in addition to stationarity, the process is also ergodic, then

$$\Pr\left[\lim_{n\to\infty}\frac{1}{n}\sum_{n=0}^{\infty} f(X_n) = E[\hat{Y}]\right] = 1.$$

Example B.2 Consider the process $\{X_i\}_{i=-\infty}^{\infty}$ consisting of a family of i.i.d. binary random variables (obviously, it is stationary and ergodic). Define the function $f(\cdot)$ by $f(0) = 0$ and $f(1) = 1$, Hence,

$$E[f(X)] = P_X(0)f(0) + P_X(1)f(1) = P_X(1)$$

is finite. By the pointwise ergodic theorem, we have

$$\lim_{n\to\infty}\frac{f(X_1)+f(X_2)+\cdots+f(X_n)}{n} = \lim_{n\to\infty}\frac{X_1+X_2+\cdots+X_n}{n} = P_X(1).$$

- First-order Markov: Denote by $X \to Y \to Z$ or $Z \to Y \to X$ or $X \leftrightarrow Y \leftrightarrow Z$.
  - In words, X and Z are conditionally independent given Y
- kth-order Markov

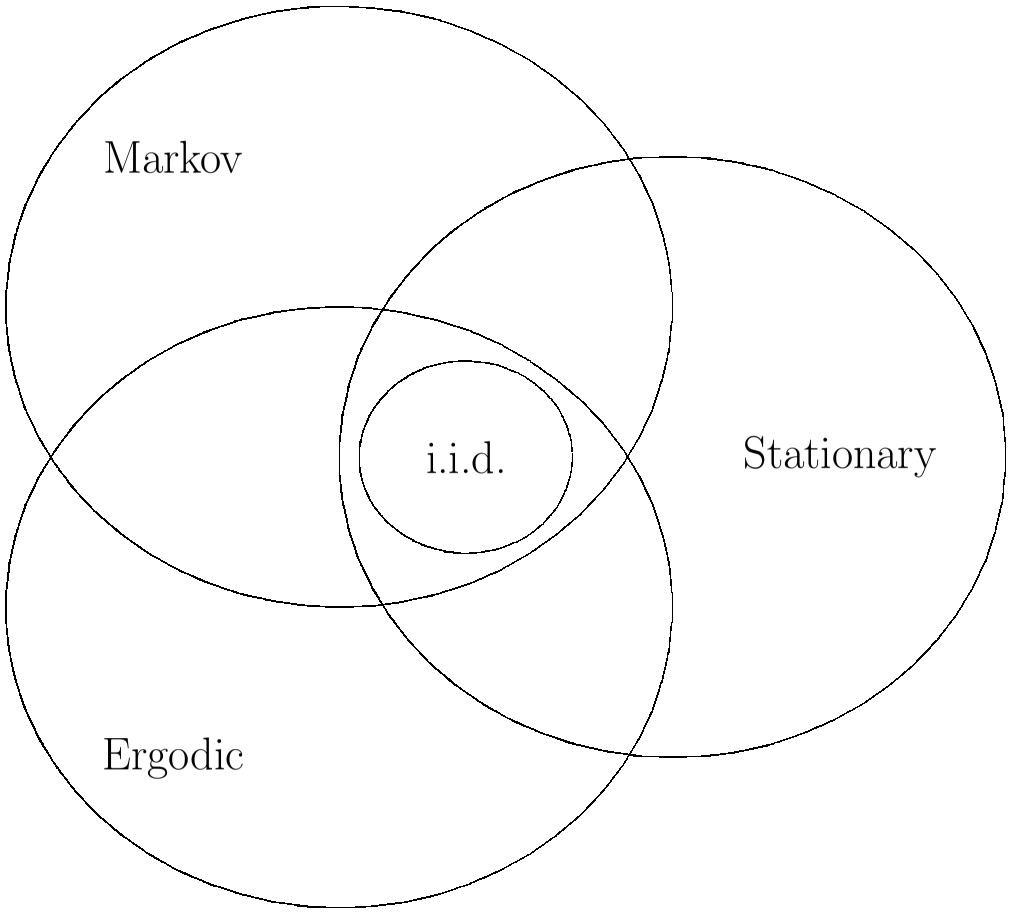
$$\begin{aligned} & \Pr\{X_n = x_n | X_{n-1} = x_{n-1}, \cdots, X_1 = x_1\} \\ = \; & \Pr\{X_n = x_n | X_{n-1} = x_{n-1}, \cdots, X_{n-k} = x_{n-k}\}. \end{aligned}$$

- (k-th order) Irreducible Markov
  - No reduction on the number of possible state outcome is rendered.

$$(\forall\ x^k, y^k \text{ and } j) \quad \Pr\left\{ X_j^{k+j-1} = x^k \middle| X_1^k = y^k \right\} > 0.$$

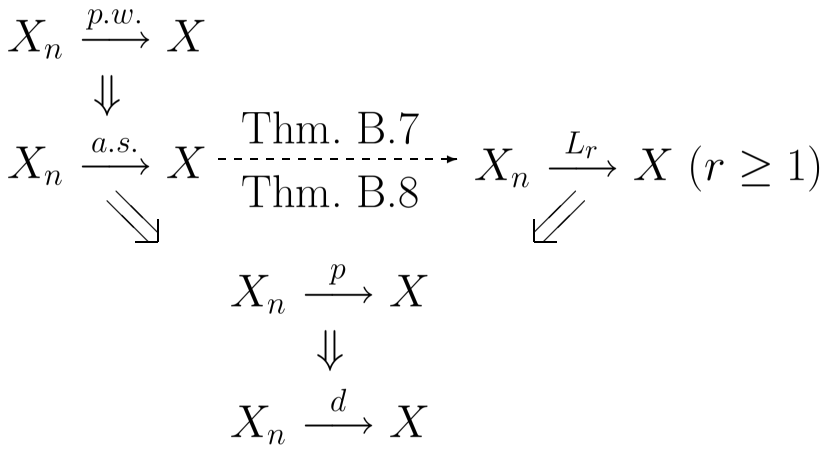
- Homogeneous or time-invariant Markov
  - The transition probability is time-invariant.
- Stationary Markov = Homogeneous with stationary initial distribution.

General Relation of Random Processes



Convergences of Sequences of Random Variables

- Relation of five modes of convergence



Convergences of Sequences of Random Variables

- Pointwise convergence and almost surely convergence

  E.g. Give a probability space
  $(\Omega = \{0, 1, 2, 3\}, 2^{\Omega}, P(0) = P(1) = P(2) = 1/3)$.

  – A random variable X_n is a mapping from a probability space to $\Re$. Let the mapping be

  $$X_n(\omega) = \frac{\omega}{n} \;\Rightarrow\; \Pr\{X_n = 0\} = \Pr\left\{X_n = \frac{1}{n}\right\} = \Pr\left\{X_n = \frac{2}{n}\right\} = \frac{1}{3}.$$

  – (*Pointwise convergence*) Observe that

  $$(\forall\ \omega \in \Omega)\ X_n(\omega) \to X(\omega),$$

  where $X(\omega) = 0$ for every $\omega \in \Omega$. So

  $$X_n \xrightarrow{p.w.} X.$$

  – (*Almost surely convergence*) Let $\tilde{X}(\omega) = 0$ for $\omega = 0, 1, 2$ and $\tilde{X}(\omega) = 1$ for $\omega = 3$. Then both of the following statements are true:

  $$X_n \xrightarrow{a.s.} X \quad \text{and} \quad X_n \xrightarrow{a.s.} \tilde{X},$$

  (since

  $$\Pr\left\{\lim_{n\to\infty} X_n = \tilde{X}\right\} = \sum_{\omega=0}^{3} P(\omega) \cdot \mathbf{1}\left\{\lim_{n\to\infty} X_n(\omega) = \tilde{X}(\omega)\right\} = 1.)$$

Convergences of Sequences of Random Variables

- Almost surely convergence (with probability 1) and convergence in probability

$$X_n \xrightarrow{a.s.} X \quad \equiv \quad \Pr\left\{\lim_{n\to\infty} X_n = X\right\} = 1$$

$$X_n \xrightarrow{p} X \quad \equiv \quad (\forall\ \gamma > 0)\ \lim_{n\to\infty} \Pr\left\{|X_n - X| < \gamma\right\} = 1$$

- Convergence in rth mean

$$X_n \xrightarrow{L_r} X \quad \equiv \quad \lim_{n\to\infty} E\left[|X_n - X|^r\right] = 0$$

- Convergence in distribution

$$X_n \xrightarrow{d} X \quad \equiv \quad \lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ for every continuous point of } F_X(x)$$

Convergences of Sequences of Random Variables

The next observation facilitates the finding of limiting random variable.

Observation B.5 (uniqueness of convergence)

1. If $X_n \xrightarrow{p.w.} X$ and $X_n \xrightarrow{p.w.} Y$,
   then $X = Y$ pointwisely. I.e.,
   $$(\forall\, \omega \in \Omega) \quad X(\omega) = Y(\omega).$$

2. If $X_n \xrightarrow{a.s.} X$ and $X_n \xrightarrow{a.s.} Y$
   (or $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$)
   (or $X_n \xrightarrow{L_r} X$ and $X_n \xrightarrow{L_r} Y$),
   then $X = Y$ with probability 1. I.e.,
   $$\Pr\{X = Y\} = 1.$$

3. $X_n \xrightarrow{d} X$ and $X_n \xrightarrow{d} Y$,
   then $F_X(x) = F_Y(x)$ for every continuity point of $F_X(x)$.

Convergences of Sequences of Random Variables

The next observation facilitates the proofs of convergence of sequences of random variables.

Observation B.6 (mutual convergence criteria)

1. $\{X_n\}_{n=1}^{\infty}$ converges pointwisely if, and only if,

$$(\forall\ \omega \in \Omega)\ \lim_{n\to\infty} |X_{n+1}(\omega) - X_n(\omega)| = 0.$$

2. $\{X_n\}_{n=1}^{\infty}$ converges with probability 1 if, and only if,

$$\begin{aligned} & P\left\{\omega \in \Omega\ :\ \lim_{n\to\infty} |X_{n+1}(\omega) - X_n(\omega)| = 0\right\} \\ = \ & \Pr\left\{\lim_{n\to\infty} |X_{n+1} - X_n| = 0\right\} = 1. \end{aligned}$$

3. $\{X_n\}_{n=1}^{\infty}$ converges in probability if, and only if, for every $\varepsilon > 0$,

$$\begin{aligned} & \lim_{n\to\infty} P\left\{\omega \in \Omega\ :\ |X_{n+1}(\omega) - X_n(\omega)| > \varepsilon\right\} \\ = \ & \lim_{n\to\infty} \Pr\left\{|X_{n+1} - X_n| > \varepsilon\right\} = 0. \end{aligned}$$

4. $\{X_n\}_{n=1}^{\infty}$ converges in r^{th} mean if, and only if,

$$\lim_{n\to\infty} E\left[|X_{n+1} - X_n|^r\right] = 0.$$

Convergences of Sequences of Random Variables

The following statement is not necessarily true.

> $\{X_n\}_{n=1}^{\infty}$ converges in distribution if, and only if, $\lim_{n\to\infty} \left|F_{X_{n+1}}(x) - F_{X_n}(x)\right| = 0$ for every x.

E.g. $\Pr\{X_n = n\} = 1$ for every n. Then

$$(\forall\, x \in \Re) \lim_{n\to\infty} |F_{X_{n+1}}(x) - F_{X_n}(x)| = 0.$$

But $\{X_n\}_{n=1}^{\infty}$ does not converge in distribution to any random variable.

Theorem B.7 (monotone convergence theorem)

$$\left.\begin{array}{l}(i)\ X_n \xrightarrow{a.s.} X \\ (ii)\ (\forall\ n)\ Y \le X_n \le X_{n+1} \\ (iii)\ E[|Y|] < \infty\end{array}\right\} \Rightarrow\ X_n \xrightarrow{L_1} X \quad \Rightarrow \quad E[X_n] \to E[X].$$

Theorem B.8 (dominated convergence theorem)

$$\left.\begin{array}{l}(i)\ X_n \xrightarrow{a.s.} X \\ (ii)\ (\forall\ n)\ |X_n| \le Y \\ (iii)\ E[|Y|] < \infty\end{array}\right\} \Rightarrow\ X_n \xrightarrow{L_1} X \quad \Rightarrow \quad E[X_n] \to E[X].$$

Law of Large Numbers

Theorem B.9 (weak law of large number) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of uncorrelated random variables with common mean $E[X_i] = \mu$. If the variables also have common variance, or more generally,

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[X_i] = 0, \quad (\text{equivalently}, \frac{X_1 + \cdots + X_n}{n} \xrightarrow{L_2} \mu)$$

then

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{p} \mu.$$

proof: By Chebyshev's inequality,

$$\Pr\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \geq \varepsilon \right\} \leq \frac{1}{n^2 \varepsilon^2} \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

□

Note: $X_n \xrightarrow{L_2} X$ implies $X_n \xrightarrow{p} X$.

Law of Large Numbers

Theorem B.10 (Kolmogorov's strong law of large number) Let $\{X_n\}_{n=1}^{\infty}$ be an independent sequence of random variables with common mean $E[X_n] = \mu$. If either

1. X_n's are identically distributed; or
2. X_n's are square-integrable with

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}[X_i]}{i^2} < \infty,$$

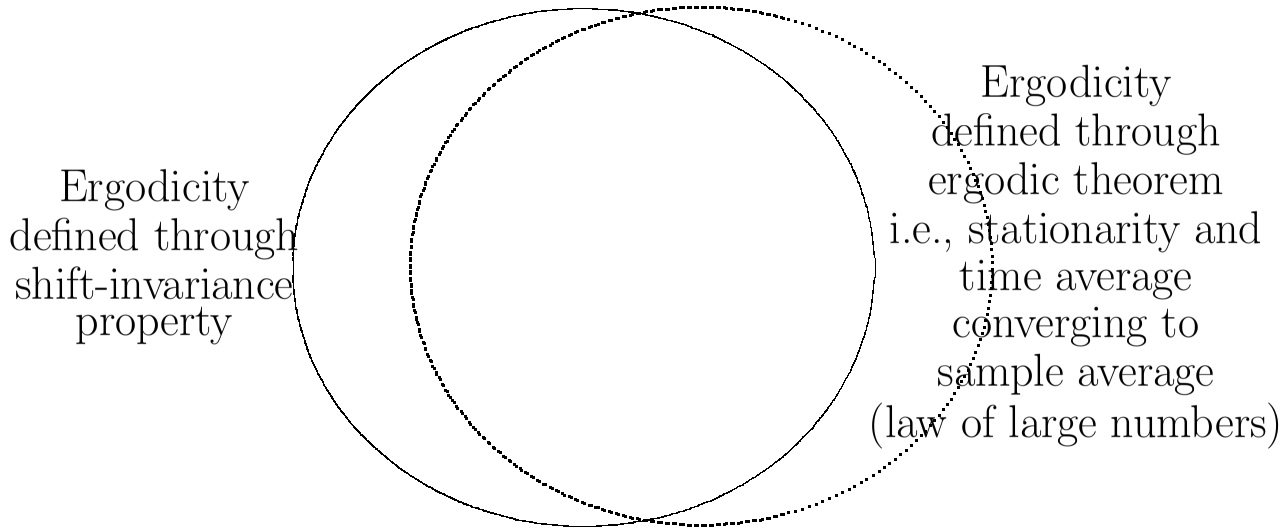
Then

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{a.s.} \mu.$$

Note: The difference of *weak* and *strong* laws of large number is that the former is convergence *in probability*, while the latter is *almost surely* convergence.

Ergodicity and Law of Large Number

- Ergodicity is defined through *shift-invariant property*; ergodic theorem (with implicitly stationary assumption) is a *consequence* of this definition.
- However, some engineers define **ergodicity** as the sequences that satisfies **ergodic theorem**, which is not exactly correct in mathematics. Nevertheless, since engineers usually concern only those sequences that satisfy the *ergodic theorem* (for which case we can use time average to approximate ensemble average), such misinformed definition does not cause serious problem in its usage.



Ergodicity and Law of Large Number

- The concept of ergodicity does not require stationarity.
  In other words, a non-stationary process can be ergodic.
- Many perfectly good models of physical processes are not ergodic, yet they have a form of law of large numbers (i.e., time average converging to ensemble average). In other words, non-ergodic processes can be perfectly good and useful models.
- There is no finite-dimensional equivalent definition of ergodicity as there is for stationarity. This fact makes it more difficult to describe and interpret ergodicity.
  - Hard to define a *shift-invariant* event for finite dimensional system except by adopting the concept of "cyclic," which may not be apt to engineering need.
- I.i.d. processes are ergodic, i.e., ergodicity can be thought of as a (kind of) generalization of i.i.d.

Ergodicity and Law of Large Number

- As mentioned earlier, stationarity and ergodicity implies the time average converges with probability 1 to the ensemble mean. Now if a process is stationary but not ergodic, then the time average still converges, but possibly not to the ensemble mean (indeed, to a random variable).

  E.g. A stationary non-ergodic process.

  - $\{A_n\}_{n=-\infty}^{\infty}$ and $\{B_n\}_{n=-\infty}^{\infty}$ i.i.d.
  - $\Pr\{A_n = 0\} = 1 - \Pr\{A_n = 1\} = \Pr\{B_n = 1\} = 1 - \Pr\{B_n = 0\} = 1/4$.
  - $V \in \{0, 1\}$.
  - $V \perp\!\!\!\perp \{A_n\}_{n=-\infty}^{\infty} \perp\!\!\!\perp \{B_n\}_{n=-\infty}^{\infty}$
  - Define $\{X_n\}_{n=-\infty}^{\infty}$ to be:

$$X_n = \begin{cases} A_n, & \text{if } V = 1; \\ B_n, & \text{if } V = 0. \end{cases}$$

  - Then

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{a.s.} U,$$

    where $\Pr\{U = 3/4\} = 1 - \Pr\{U = 1/4\} = \Pr\{V = 1\}$.

Ergodicity and Law of Large Number

- For a stationary but non-ergodic process, pointwise ergodic theorem gives the desired **limiting behavior** of the **time average**. E.g., in the previous example,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{a.s.} U.$$

- Note that the sequences in the above example does not satisfy the *law of large number*.

- The previous example also shed the light to a famous and surprising result: *ergodic decomposition theorem* for which one of the main contributors is Prof. Davison at EE department of UMCP,[1] who is now retired.

[1] UMCP=University of Maryland, College Park

Ergodicity and Law of Large Number

- Ergodic decomposition theorem
  - Under fairly general conditions, a stationary process is a mixture of stationary ergodic processes, e.g., in the previous example,

$$X_n = V \cdot A_n + (1 - V) \cdot B_n.$$

  - *Implication*: One always observes a stationary ergodic outcome (or stationary-ergodic-like behavior) for any stationary(-only) process, i.e.,

$$\begin{aligned} &\text{if } V = 1, \quad \text{then one observes } \ldots, A_1, A_2, A_3, \ldots \\ &\text{if } V = 0, \quad \text{then one observes } \ldots, B_1, B_2, B_3, \ldots \end{aligned}$$

where $\{A_n\}_{n=-\infty}^{\infty}$ and $\{B_n\}_{n=-\infty}^{\infty}$ are both stationary ergodic.

Ergodicity and Law of Large Number

- The previous remarks hint that **ergodicity** is not required for the **strong law of large number** to be useful.
- The next question is whether or not **stationarity** is required. Again the answer is negative! In fact, the main concern of strong law of large numbers is the convergence of sample averages to its ensemble expectation. It should be reasonable to expect that random processes could exhibit transient behavior that violates the stationarity definition, yet the sample average still converges. One can then introduce the notion of *asymptotically stationary* to the strong law of large numbers. (I.e., a sequence of random variables is said to be *asymptotically stationary* if it satisfies strong law of large numbers.)
- So please do not take stationarity and ergodicity too serious. Strong- law-of-large-numbers-type behavior is really our main concern (in information theory).
- Since stationary ergodicity assumption is occasionally seen in theoretical papers, it is nonetheless better to have a basic understanding of it.

Central Limit Theorem

Theorem B.12 (central limit theorem) If $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with common marginal mean μ and variance σ^2, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\mathcal{N}(0, \sigma^2)$ represents the Gaussian distribution with mean 0 and variance σ^2.

Convexity and Concavity

Definition B.13 (convexity) A function $f(x)$ is said to be *convex* over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

Furthermore, a function f is said to be *strictly convex* if equality holds only when $\lambda = 0$ or $\lambda = 1$.

Definition B.14 (concavity) A function f is *concave* if $-f$ is convex.

Jensen's inequality

Theorem B.15 (Jensen's inequality) If f is convex and X is a random variable, then

$$E[f(X)] \geq f(E[X]).$$

Moreover, if f is strictly convex, then equality in the above inequality immediately implies $X = E[X]$ with probability 1.

Proof: Let $y = ax + b$ be a support line through the point $(E[X], f(E[X]))$, where a support line[2] (for a convex function) at x_0 is by definition a line passing through the point $(x_0, f(x_0))$ and is lying entirely below the graph of $f(\cdot)$. Thus,

$$(\forall x \in \mathcal{X}) \quad ax + b \leq f(x).$$

By taking the expectation value of both sides, we obtain

$$a \cdot E[X] + b \leq E[f(X)],$$

but we know that $a \cdot E[X] + b = f(E[X])$. Consequently,

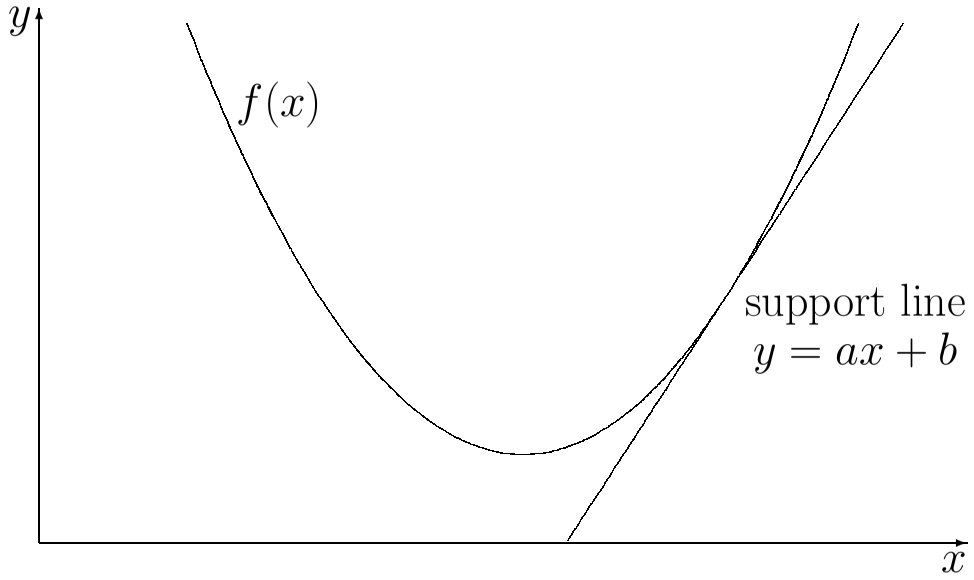
$$f(E[X]) \leq E[f(X)].$$

□

[2] A line $y = ax + b$ is said to be a support line of function $f(x)$ if among all lines of the same slope a, it is the largest one satisfying $ax + b \leq f(x)$ for every x. Hence, a support line may not necessarily pass the point $(x_0, f(x_0))$ for every x_0. Here, since we only consider convex functions, the validity of the support line at x_0 passing $(x_0, f(x_0))$ is therefore guaranteed.

Jensen's inequality

The support line $y = ax + b$ of the convex function $f(x)$.



Key Notes

- Definitions of (weakly, strictly) stationary, ergodic and Markov (irreducible, homogeneous)
- Mode of convergences (almost surely or with probability 1, in probability, in distribution, in L_r mean)
- Two Laws of large numbers
- Central limit theorem
- Jensen's inequality (convexity and concavity)