

Appendix B

Overview in Probability and Random Processes

Po-Ning Chen, Professor

Department of Electrical and Computer Engineering

National Chiao Tung University

Hsin Chu, Taiwan 30010, R.O.C.

Probability space

I: b-1

Definition B.1 (σ -Fields) Let \mathcal{F} be a collection of subsets of a non-empty set Ω . Then \mathcal{F} is called a σ -field (or σ -algebra) if the following hold:

1. $\Omega \in \mathcal{F}$.
2. *Closedness of \mathcal{F} under complementation:* If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where $A^c = \{\omega \in \Omega : \omega \notin A\}$.
3. *Closedness of \mathcal{F} under countable union:* If $A_i \in \mathcal{F}$ for $i = 1, 2, 3, \dots$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Probability space

I: b-2

- It directly follows that the empty set \emptyset is also an element of \mathcal{F} (as $\Omega^c = \emptyset$) and that \mathcal{F} is closed under countable intersection since

$$\bigcap_{i=1}^{\infty} A_i^c = \left(\bigcup_{i=1}^{\infty} A_i \right)^c .$$

- The largest σ -field of subsets of a given set Ω is the collection of all subsets of Ω (i.e., its powerset), while the smallest σ -field is given by $\{\Omega, \emptyset\}$.
- Also, if A is a proper (strict) non-empty subset of Ω , then the smallest σ -field containing A is given by $\{\Omega, \emptyset, A, A^c\}$.

Probability space

I: b-3

Definition B.2 (Probability space) A *probability space* is a triple (Ω, \mathcal{F}, P) , where Ω is a given set called *sample space* containing all possible outcomes (usually observed from an experiment), \mathcal{F} is the σ -field of subsets Ω and P is a probability measure $P : \mathcal{F} \rightarrow [0, 1]$ on the σ -field satisfying the following:

1. $0 \leq P(A) \leq 1$ for all $A \in \mathcal{F}$.
2. $P(\Omega) = 1$.
3. *Countable additivity*: If A_1, A_2, \dots is a sequence of disjoint sets (i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$) in \mathcal{F} , then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k).$$

- It directly follows from properties 1-3 of the above definition that $P(\emptyset) = 0$.
- Usually, the σ -field \mathcal{F} is called the *event space* and its elements (which are subsets of Ω satisfying the properties of Definition B.1) are called *events*.

Random variable and random process

I: b-4

- A random variable X defined over probability space (Ω, \mathcal{F}, P) is a real-valued function $X : \Omega \rightarrow \mathbb{R}$ that is *measurable* (or \mathcal{F} -*measurable*), i.e., satisfying the property that

$$X^{-1}((-\infty, t]) \triangleq \{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}$$

for each real t .

- A random process is a collection of random variables that arise from the same probability space. It can be mathematically represented by the collection

$$\{X_t, t \in I\},$$

where X_t denotes the t^{th} random variable in the process, and the index t runs over an index set I which is arbitrary.

- The index set I can be uncountably infinite — e.g., $I = \mathbb{R}$ — in which case we are effectively dealing with a continuous-time process.
- We will however exclude such a case in this chapter for the sake of simplicity. To be precise, we will only consider the following cases of index set I :
 - case a)* I consists of one index only.
 - case b)* I is finite.
 - case c)* I is countably infinite.

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

I: b-5

Answer 1: (Ω, \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 (manageable) real values.
- Such an experiment results in the random variable X whose probability is so defined over the probability space (Ω, \mathcal{F}, P) .

Answer 2: With the underlying probability space, any finite dimensional distribution of $\{X_t, t \in I\}$ is well-defined.

- Under this definition, any finite dimensional distribution of $\{X_t, t \in I\}$ is well-defined. For example,

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

Distribution functions

I: b-6

- In applications, we are perhaps more interested in the **distribution functions** of the random variables and random processes than the underlying probability space (Ω, \mathcal{F}, P) on which the random variables and random processes are defined.

Distribution functions

I: b-7

- It can be proved [Billingsley, Thm. 14.1] that given a real-valued non-negative function $F(\cdot)$ that is non-decreasing and right-continuous and satisfies

$$\lim_{x \downarrow -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \uparrow \infty} F(x) = 1,$$

there exist a random variable and an underlying probability space such that the cumulative distribution function (cdf) of the random variable defined over the probability space is equal to $F(\cdot)$.

- This result releases us from the burden of referring to a probability space before defining the random variable. In other words, we can define a random variable X directly by its cdf, i.e., $\Pr[X \leq x]$, without bothering to refer to its underlying probability space.
- Nevertheless, it is better to keep in mind (and learn) that, formally, random variables and random processes are defined over underlying probability spaces.

Relation between a source and a random process

I: b-8

- In statistical communication, a *discrete-time* source

$$(X_1, X_2, X_3, \dots, X_n) \triangleq X^n$$

consists of a sequence of random quantities, where each quantity usually takes on values from a source *generic alphabet* \mathcal{X} , namely

$$(X_1, X_2, \dots, X_n) \in \mathcal{X} \times \mathcal{X} \times \dots \times \mathcal{X} \triangleq \mathcal{X}^n.$$

- The elements in \mathcal{X} are usually called *letters*.

Statistical properties of random sources

I: b-9

- **Statistical evolution in time** is an important factor for a random source.
- In particular, a “time-shift” property should be noted first.
- An event E in $\mathcal{F}_{\mathbf{X}} \triangleq \{X(A) : A \in \mathcal{F}\}$ ¹ for a random process

$$\mathbf{X} = \{X_1, X_2, X_3, \dots\}$$

is said to be \mathbb{T} -*invariant* with respect to the left-shift (or shift transformation) \mathbb{T} if

$$\mathbb{T}E \subseteq E,$$

where

$$\mathbb{T}E \triangleq \{\mathbb{T}\mathbf{x} : \mathbf{x} \in E\} \quad \text{and} \quad \mathbb{T}\mathbf{x} \triangleq \mathbb{T}(x_1, x_2, x_3, \dots) = (x_2, x_3, \dots).$$

- In other words, \mathbb{T} is equivalent to “chopping the first component.”

¹Note that by definition, only the events $E \in \mathcal{F}_{\mathbf{X}}$ is probabilistically measurable, and $P_{\mathbf{X}}(E) = P(A)$.

Statistical properties of random sources

I: b-10

Example. Applying \mathbb{T} onto an event E defined below,

$$\begin{aligned}
 E &\triangleq \{(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1, \dots), (x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, \dots), \\
 &\quad (x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1, \dots)\}, \tag{B.5.1} \\
 &= \left\{ \underbrace{1111\dots}_{\text{all one}}, \underbrace{0111\dots}_{\substack{\text{all one but} \\ \text{the first}}}, \underbrace{0011\dots}_{\substack{\text{all one but} \\ \text{the first two}}} \right\}
 \end{aligned}$$

yields

$$\begin{aligned}
 \mathbb{T}E &= \{(x_1 = 1, x_2 = 1, x_3 = 1, \dots), (x_1 = 1, x_2 = 1, x_3 = 1 \dots), \\
 &\quad (x_1 = 0, x_2 = 1, x_3 = 1, \dots)\} \\
 &= \{(x_1 = 1, x_2 = 1, x_3 = 1, \dots), (x_1 = 0, x_2 = 1, x_3 = 1, \dots)\} \\
 &= \left\{ \underbrace{1111\dots}_{\text{all one}}, \underbrace{0111\dots}_{\substack{\text{all one but} \\ \text{the first}}} \right\}
 \end{aligned}$$

We then have $\mathbb{T}E \subseteq E$, and hence E is \mathbb{T} -invariant. □

Statistical properties of random sources

I: b-11

- It can be proved (cf. the textbook) that if $\mathbb{T}E \subseteq E$, then $\mathbb{T}^2E \subseteq \mathbb{T}E$.
- By induction, we can further obtain

$$\dots \subseteq \mathbb{T}^3E \subseteq \mathbb{T}^2E \subseteq \mathbb{T}E \subseteq E. \quad (\text{B.5.2})$$

- Thus, if an element say $(1, 0, 0, 1, 0, 0, \dots)$ is in a \mathbb{T} -invariant set E , then all its left-shift counterparts (i.e., $(0, 0, 1, 0, 0, 1, \dots)$ and $(0, 1, 0, 0, 1, 0, \dots)$) should be contained in E .
- As a result, for a \mathbb{T} -invariant set E , an element and all its left-shift counterparts should be all in E .
- Hence, a “ \mathbb{T} -invariant group” such as one containing

$$(1, 0, 0, 1, 0, 0, \dots), \quad (0, 0, 1, 0, 0, 1, \dots) \text{ and } (0, 1, 0, 0, 1, 0, \dots)$$

should be treated as an indecomposable group in \mathbb{T} -invariant sets.

Statistical properties of random sources

I: b-12

- In the probabilistically literature, researchers are in particular interested in these “ \mathbb{T} -invariant indecomposable groups” (especially when defining an ergodic random process).
- It is however possible that some single “transient” element, such as $(0, 0, 1, 1, \dots)$ in (B.5.1), is included in a \mathbb{T} -invariant set, and will be excluded after applying left-shift operation \mathbb{T} .
- This can be resolved by introducing the inverse operation \mathbb{T}^{-1} . (This is the main reason why \mathbb{T}^{-1} is more commonly used rather than \mathbb{T} in defining the statistical property of random processes.)
- Note that \mathbb{T} is a many-to-one mapping, so its inverse operation does not exist in general.
- Similar to taking the closure of an open set, the definition adopted below [45, p. 3] allows us to “enlarge” the \mathbb{T} -invariant set such that all right-shift counterparts of the single “transient” element are included.

$$\mathbb{T}^{-1}E \triangleq \{\mathbf{x} \in \mathcal{X}^\infty : \mathbb{T}\mathbf{x} \in E\}.$$

Statistical properties of random sources

I: b-13

- We then notice from the above definition that if

$$\mathbb{T}^{-1}E = E, \tag{B.5.3}$$

then

$$\mathbb{T}E = \mathbb{T}(\mathbb{T}^{-1}E) = E,$$

and hence E is constituted only by the \mathbb{T} -invariant groups because

$$\dots = \mathbb{T}^{-2}E = \mathbb{T}^{-1}E = E = \mathbb{T}E = \mathbb{T}^2E = \dots .$$

- The sets that satisfy (B.5.3) are sometimes referred to as *ergodic sets* because as time goes by (the left-shift operator \mathbb{T} can be regarded as a shift to a future time), the set always stays in the state that it has been before.

Statistical properties of random sources

I: b-14

- As we (and most communication systems) only deal with one-sided random processes in the above, the discussion on \mathbb{T} -invariance only focuses on sets of one-sided sequences.
- When a two-sided random process $\dots, X_{-2}, X_{-1}, X_0, X_1, X_2, \dots$ is considered, the left-shift operation \mathbb{T} of a two-sided sequence actually has a unique inverse. Hence, $\mathbb{T}E \subseteq E$ implies $\mathbb{T}E = E$. Also, $\mathbb{T}E = E$ if, and only if, $\mathbb{T}^{-1}E = E$. Ergodicity for two-sided sequences can therefore be directly defined using $\mathbb{T}E = E$.

Statistical properties of random sources

I: b-15

We now classify several useful statistical properties of (one-sided) random process

$$\mathbf{X} = \{X_1, X_2, \dots\}.$$

- *Memoryless*: A random process is said to be *memoryless* if the sequence of random variables X_i is *independent and identically distributed* (i.i.d.).
- *First-order stationary*: A process is *first-order stationary* if the marginal distribution is unchanged for every time instant.
- *Second-order stationary*: A process is *second-order stationary* if the joint distribution of any two (not necessarily consecutive) time instances is invariant before and after left (time) shift.
- *Weakly stationary process*: A process is *weakly stationary* (or *wide-sense stationary* or *stationary in the weak or wide sense*) if the mean and autocorrelation function are unchanged by a left (time) shift.
- *Stationary process*: A process is *stationary* (or *strictly stationary*) if the probability of every sequence or event is unchanged by a left (time) shift.

Statistical properties of random sources

I: b-16

- **Ergodic process:** A process is *ergodic* if any ergodic set (satisfying (B.5.3)) in $\mathcal{F}_{\mathbf{X}}$ has probability either 1 or 0. This definition is not very intuitive, but some interpretations and examples may shed some light.
 - First, observe that the definition has nothing to do with stationarity. It simply states that events that are unaffected by time-shifting (both left- and right-shifting) must have probability either zero or one.
 - Secondly, ergodicity actually implies that all convergent sample averages² converge to a constant.
 - Below is an example that could be used to explain the idea.

Example B.3 Suppose $X_1, X_2, \dots, X_n, \dots$ is an ergodic process, where each X_n takes values in $\{0, 1\}$. Let Ω be the set of all one-sided zero-one sequences. Define for $\alpha \in [0, 1]$,

$$\bar{E}(\alpha) \triangleq \left\{ \mathbf{x} \in \{0, 1\}^\infty : \limsup_{n \rightarrow \infty} \frac{x_1 + \dots + x_n}{n} = \alpha \right\}$$

and

$$\underline{E}(\alpha) \triangleq \left\{ \mathbf{x} \in \{0, 1\}^\infty : \liminf_{n \rightarrow \infty} \frac{x_1 + \dots + x_n}{n} = \alpha \right\}.$$

²Two alternative names for *sample average* are *time average* and *Cesàro mean*. In this book, these names will be used interchangeably.

Statistical properties of random sources

I: b-17

Then it can be verified (see footnote 7 in text) that both $\bar{E}(\alpha)$ and $\underline{E}(\alpha)$ are ergodic sets, i.e.,

$$\bar{E}(\alpha) = \mathbb{T}^{-1}\bar{E}(\alpha) \quad \text{and} \quad \underline{E}(\alpha) = \mathbb{T}^{-1}\underline{E}(\alpha),$$

and hence

$$\Omega = \bigcup_{\alpha \in [0,1]} \bar{E}(\alpha) = \bigcup_{\alpha \in [0,1]} \underline{E}(\alpha)$$

and

$$\bar{E}(\alpha) \cap \bar{E}(\beta) = \underline{E}(\alpha) \cap \underline{E}(\beta) = \emptyset \text{ for } \alpha \neq \beta.$$

For any $\gamma > 0$ arbitrarily small, the definition of ergodicity then implies the existence of α_0 and β_0 such that

$$\Pr \left[\mathbf{X} \in \bigcup_{\alpha \in [\alpha_0 - \gamma, \alpha_0 + \gamma]} \bar{E}(\alpha) \right] = \Pr \left[\mathbf{X} \in \bigcup_{\alpha \in [\beta_0 - \gamma, \beta_0 + \gamma]} \bar{E}(\alpha) \right] = 1,$$

where we allow for the possibility that 0 (resp. 1) is contained in the two intervals $[\alpha_0 - \gamma, \alpha_0 + \gamma]$ and $[\beta_0 - \gamma, \beta_0 + \gamma]$.

Statistical properties of random sources

I: b-18

In other words,

$$\limsup_{n \rightarrow \infty} \frac{X_1 + \cdots + X_n}{n} \in [\alpha_0 - \gamma, \alpha_0 + \gamma] \text{ with probability 1}$$

and

$$\liminf_{n \rightarrow \infty} \frac{X_1 + \cdots + X_n}{n} \in [\beta_0 - \gamma, \beta_0 + \gamma] \text{ with probability 1.}$$

Statistical properties of random sources

I: b-19

Notably, there exists sequence \mathbf{x} contained in both $\bar{E}(\alpha_0)$ and $\underline{E}(\beta_0)$ (cf. footnote 9 in text), it is not necessary that the two intervals interest.

Nevertheless, we can still infer for this ergodic process that if

$$\lim_{n \rightarrow \infty} \frac{X_1 + X_2 + \cdots + X_n}{n} \text{ converges,}$$

i.e.,

$$\Pr \left[\limsup_{n \rightarrow \infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = \liminf_{n \rightarrow \infty} \frac{X_1 + X_2 + \cdots + X_n}{n} \right] = 1,$$

then

$$|\beta_0 - \alpha_0| \leq 2\gamma$$

and

$$\lim_{n \rightarrow \infty} \frac{X_1 + X_2 + \cdots + X_n}{n} \in [\alpha_0 - \gamma, \alpha_0 + \gamma] \text{ with probability 1.}$$

□

Statistical properties of random sources

I: b-20

- It needs to be pointed out that in the above example, ergodicity does not guarantee that the ensemble average lies in $[\alpha_0 - \gamma, \alpha_0 + \gamma]$.
- A quick example is that

$$\Pr\{(x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 1, \dots)\} = 0.2$$

and

$$\Pr\{(x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 0, \dots)\} = 0.8$$

assure the validity of ergodicity, but

$$\frac{X_1 + \dots + X_n}{n} \rightarrow \frac{1}{2} \text{ with probability 1.}$$

which is not equal to $E[X_i]$ for any i .

- In principle,
 - ergodicity implies that all convergent sample averages converge to a constant (not necessarily the ensemble average), and
 - stationarity assures that the time average converges to a random variable;

hence, it is reasonable to expect that they jointly imply the ultimate time average equals the ensemble average. This is validated by the well-known *ergodic theorem* by Birkhoff and Khinchin.

Statistical properties of random sources

I: b-21

Theorem B.4 (Pointwise ergodic theorem) Consider a discrete-time stationary random process, $\mathbf{X} = \{X_n\}_{n=1}^{\infty}$. For real-valued function $f(\cdot)$ on \mathbb{R} with finite mean (i.e., $|E[f(X_n)]| < \infty$), there exists a random variable Y such that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(X_k) = Y \quad \text{with probability 1.}$$

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

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(X_k) = E[f(X_1)] \quad \text{with probability 1.}$$

Physical meaning of stationary ergodic assumption

I: b-22

- Stationary ergodic random source
 - One of the important consequences that the pointwise ergodic theorem indicates is that the time average can ultimately replace the statistical average, which is a useful result in engineering.
 - Hence, with stationarity and ergodicity, one, who observes

$$X_1^{30} = 154326543334225632425644234443$$

from the experiment of rolling a dice, can draw the conclusion that the true distribution of rolling the dice can be well approximated by:

$$\begin{array}{lll} \Pr\{X_i = 1\} \approx \frac{1}{30} & \Pr\{X_i = 2\} \approx \frac{6}{30} & \Pr\{X_i = 3\} \approx \frac{7}{30} \\ \Pr\{X_i = 4\} \approx \frac{9}{30} & \Pr\{X_i = 5\} \approx \frac{4}{30} & \Pr\{X_i = 6\} \approx \frac{3}{30} \end{array}$$

- Such result is also known by the *law of large numbers*. The relation between ergodicity and the law of large numbers will be further explored later.
- 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 I: b-23

- In communications theory, one may assume that *the source is stationary* or *the source is stationary ergodic*. But it is rare to see the assumption of *the source being ergodic but non-stationary*. Why?
 - This is perhaps because an ergodic but non-stationary source not only does not facilitate the analytical study of communications problems, but may have limited application in practice.
- From this, we note that assumptions are made either to facilitate the analytical study of communications problems or to fit a specific need of applications. Without these two objectives, an assumption becomes of minor interest.
- This, to some extent, justifies that the *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.

Physical meaning of stationary ergodic assumption I: b-24

- First-order Markov: Three random variables X , Y and Z are said to form a Markov chain or a first-order Markov chain if

$$P_{X,Y,Z}(x, y, z) = P_X(x) \cdot P_{Y|X}(y|x) \cdot P_{Z|Y}(z|y); \quad (\text{B.5.4})$$

i.e.,

$$P_{Z|X,Y}(z|x, y) = P_{Z|Y}(z|y).$$

This is usually denoted by

$$X \rightarrow Y \rightarrow Z.$$

- $X \rightarrow Y \rightarrow Z$ is sometimes read as “ X and Z are conditionally independent given Y ” because it can be shown that (B.5.4) is equivalent to

$$P_{X,Z|Y}(x, z|y) = P_{X|Y}(x|y) \cdot P_{Z|Y}(z|y).$$

- Therefore, $X \rightarrow Y \rightarrow Z$ is equivalent to $Z \rightarrow Y \rightarrow X$. Accordingly, the Markovian notation is sometimes expressed as $X \leftrightarrow Y \leftrightarrow Z$.

- k th-order Markov: The random variables X_1, X_2, X_3, \dots are said to form a k -th order Markov chain if for all $k < n$,

$$\Pr[X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_1 = x_1] = \Pr[X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_{n-k} = x_{n-k}].$$

- Each $x_{n-k}^{n-1} \in \mathcal{X}^k$ is called the *state* at time n .

Physical meaning of stationary ergodic assumption I: b-25

Irreducible A Markov chain is *irreducible* if with some probability, we can go from any state in \mathcal{X}^k to another state in a finite number of steps, i.e., for all $x^k, y^k \in \mathcal{X}^k$ there exists $j \geq 1$ such that

$$\Pr \left\{ X_j^{k+j-1} = x^k \mid X_1^k = y^k \right\} > 0.$$

Time-invariant A Markov chain is said to be *time-invariant* or *homogeneous*, if for every $n > k$,

$$\Pr[X_n = x_n \mid X_{n-1} = x_{n-1}, \dots, X_{n-k} = x_{n-k}] = \Pr[X_{k+1} = x_{k+1} \mid X_k = x_k, \dots, X_1 = x_1].$$

– Therefore, a homogeneous first-order Markov chain can be defined through its transition probability:

$$\left[\Pr\{X_2 = x_2 \mid X_1 = x_1\} \right]_{|\mathcal{X}| \times |\mathcal{X}|},$$

and its initial state distribution $P_{X_1}(x)$.

– A distribution $\pi(x)$ on \mathcal{X} is said to be a *stationary* distribution for a homogeneous first-order Markov chain, if for every $y \in \mathcal{X}$,

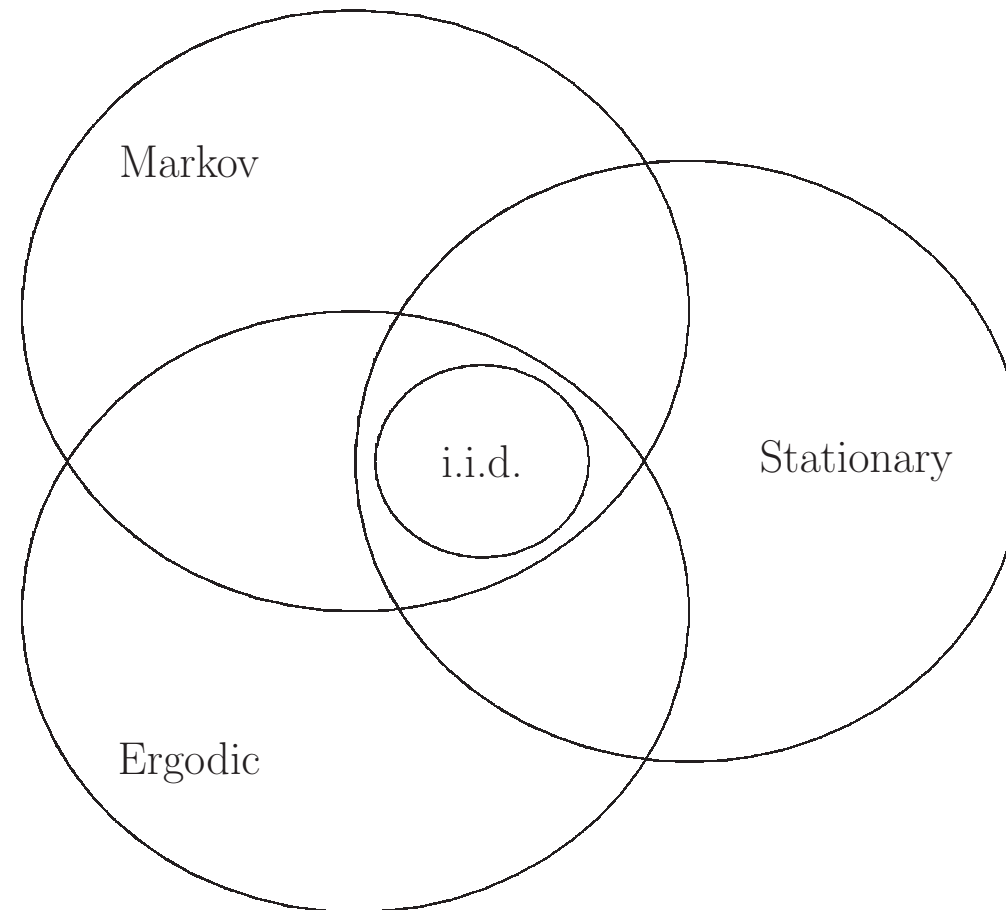
$$\pi(y) = \sum_{x \in \mathcal{X}} \pi(x) \Pr\{X_2 = y \mid X_1 = x\}.$$

Physical meaning of stationary ergodic assumption I: b-26

- If the initial state distribution is equal to the stationary distribution, then the homogeneous first-order Markov chain is a stationary process.

General Relation of Random Processes

I: b-27



Convergence of sequences of random variables

I: b-28

- Relation of five modes of convergence

$$\begin{array}{ccc} X_n \xrightarrow{p.w.} X & & \\ \Downarrow & & \\ X_n \xrightarrow{a.s.} X & \xrightarrow[\text{Thm. B.11}]{\text{Thm. B.10}} & X_n \xrightarrow{L_r} X \quad (r \geq 1) \\ \Downarrow & & \Downarrow \\ & & X_n \xrightarrow{p} X \\ & & \Downarrow \\ & & X_n \xrightarrow{d} X \end{array}$$

Convergence of sequences of random variables

I: b-29

- Pointwise convergence and almost surely convergence

E.g. Give a probability space

($\Omega = \{0, 1, 2, 3\}$, 2^Ω , $P(0) = P(1) = P(2) = 1/3$).

- A random variable X_n is a mapping from a probability space to \mathbb{R} . 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

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

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

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

Convergence of sequences of random variables

I: b-30

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

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

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

- Convergence in r th mean

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

- Convergence in distribution

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

Convergence of sequences of random variables

I: b-31

The next observation facilitates the finding of limiting random variable.

Observation B.8 (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 all x .

Convergence of sequences of random variables

I: b-32

Theorem B.9 (Monotone convergence theorem)

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

Theorem B.10 (Dominated convergence theorem)

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

The implication of $X_n \xrightarrow{L_1} X$ to $E[X_n] \rightarrow E[X]$ can be easily seen from

$$|E[X_n] - E[X]| = |E[X_n - X]| \leq E[|X_n - X|].$$

Law of Large Numbers

I: b-33

Theorem B.14 (Weak law of large numbers) 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,

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

then

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

proof: By Chebyshev's inequality,

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

□

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

Law of Large Numbers

I: b-34

Theorem B.15 (Kolmogorov's strong law of large numbers) 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

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

Then

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

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 numbers

I: b-35

- After the introduction of Kolmogorov's strong law of large numbers, one may find that the pointwise ergodic theorem (Theorem B.4) actually indicates a similar result.
 - In fact, the pointwise ergodic theorem can be viewed as another version of strong law of large numbers, which states that *for stationary and ergodic processes, time averages converge with probability 1 to the ensemble expectation.*
- The notion of ergodicity is often misinterpreted, since the definition is not very intuitive. Some engineering texts may provide a definition that a stationary process satisfying the ergodic theorem is also ergodic.

Ergodicity and law of large numbers

I: b-36

Let us try to clarify the notion of ergodicity by the following remarks.

- 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. 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.
- I.i.d. processes are ergodic; hence, ergodicity can be thought of as a (kind of) generalization of i.i.d.
- As mentioned earlier, stationarity and ergodicity imply 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.

Ergodicity and law of large numbers

I: b-37

Example. Let $\{A_n\}_{n=-\infty}^{\infty}$ and $\{B_n\}_{n=-\infty}^{\infty}$ be two i.i.d. binary 0-1 random variables with

$$\Pr\{A_n = 0\} = \Pr\{B_n = 1\} = 1/4.$$

Suppose that

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

where U is equiprobable binary random variable, and $\{A_n\}_{n=1}^{\infty}$, $\{B_n\}_{n=1}^{\infty}$ and U are independent.

Then $\{X_n\}_{n=1}^{\infty}$ is stationary.

Is the process ergodic? The answer is negative.

If the stationary process were ergodic, then from the pointwise ergodic theorem (Theorem B.4), its relative frequency would converge to a constant!

Ergodicity and law of large numbers

I: b-38

However, one should observe that the outputs of (X_1, \dots, X_n) form a Bernoulli process with relative frequency of 1's being either $3/4$ or $1/4$, depending on the value of U . Therefore,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_n \xrightarrow{a.s.} Y,$$

where $\Pr(Y = 1/4) = \Pr(Y = 3/4) = 1/2$, which contradicts to the ergodic theorem. \square

- *Ergodic decomposition theorem:* Under fairly general assumptions, any (not necessarily ergodic) stationary process is a mixture of stationary ergodic processes, and hence one always observes a stationary ergodic outcome. As in the above example, one always observe either A_1, A_2, A_3, \dots or B_1, B_2, B_3, \dots , depending on the value of U , for which both sequences are stationary ergodic (i.e., the time-stationary observation X_n satisfies

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

- The previous remark implies that ergodicity is not required for the strong law of large numbers to be useful.
- The next question is whether or not stationarity is required. Again the answer is negative!

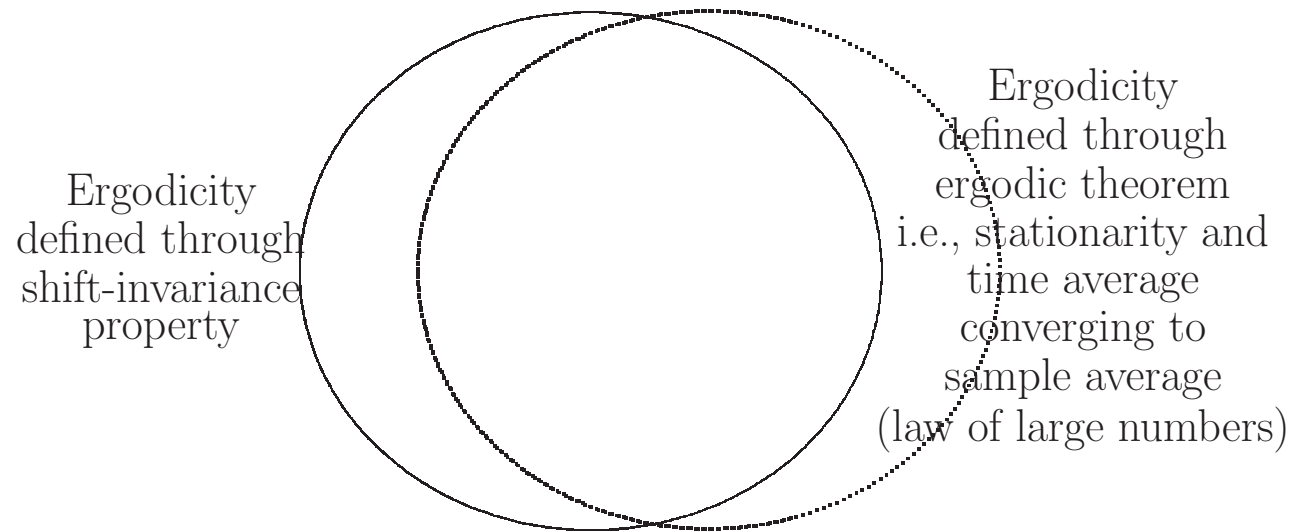
Ergodicity and law of large numbers

I: b-39

- In fact, what is needed in this course is the **law of large numbers**, which results 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 achieve the law of large numbers.
- So in my opinion, one should not take the notions of stationarity and ergodicity too seriously (if the main concern is the law of large numbers) since they can be significantly weakened and still have laws of large numbers holding (i.e., time averages and relative frequencies have desired and well-defined limits).
- Since stationary ergodicity assumption is occasionally seen in theoretical papers, it is nonetheless better to have a basic understanding of it.

Ergodicity and law of large numbers

I: b-40



Central limit theorem

I: b-41

Theorem B.16 (Central limit theorem) If $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with finite common marginal mean μ and variance σ^2 , then

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

where the convergence is in distribution (as $n \rightarrow \infty$) and $Z \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian distributed random variable with mean 0 and variance σ^2 .

Convexity, concavity and Jensen's inequality

I: b-42

Definition B.17 (Convexity) Consider a convex set $\mathcal{O} \in \mathbb{R}^m$, where m is a fixed positive integer. Then a function $f : \mathcal{O} \rightarrow \mathbb{R}$ is said to be *convex* over \mathcal{O} if for every $\underline{x}, \underline{y}$ in \mathcal{O} and $0 \leq \lambda \leq 1$,

$$f(\lambda \underline{x} + (1 - \lambda) \underline{y}) \leq \lambda f(\underline{x}) + (1 - \lambda) f(\underline{y}).$$

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

- A set $\mathcal{O} \in \mathbb{R}^m$ is said to be *convex* if for every $\underline{x} = (x_1, x_2, \dots, x_m)^T$ and $\underline{y} = (y_1, y_2, \dots, y_m)^T$ in \mathcal{O} (where T denotes transposition), and every $0 \leq \lambda \leq 1$, $\lambda \underline{x} + (1 - \lambda) \underline{y} \in \mathcal{O}$; in other words, the “convex combination” of any two “points” \underline{x} and \underline{y} in \mathcal{O} also belongs to \mathcal{O} .

Definition B.18 (Concavity) A function f is *concave* if $-f$ is convex.

Jensen's inequality

I: b-43

Theorem B.19 (Jensen's inequality) If $f : \mathcal{O} \rightarrow \mathbb{R}$ is convex over a convex set $\mathcal{O} \subset \mathbb{R}^m$, and $\underline{X} = (X_1, X_2, \dots, X_m)^T$ is an m -dimensional random vector with alphabet $\mathcal{X} \subset \mathcal{O}$, then

$$E[f(\underline{X})] \geq f(E[\underline{X}]).$$

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

Lagrange multipliers technique and KKT condition

I: b-44

Optimization of a function $f(\mathbf{x})$ over $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X} \subseteq \mathbb{R}^n$

subject to

$$\begin{cases} \text{inequality constraints } g_i(\mathbf{x}) \leq 0 \text{ for } 1 \leq i \leq m, \text{ and} \\ \text{equality constraints } h_j(\mathbf{x}) = 0 \text{ for } 1 \leq j \leq \ell \end{cases}$$

is a center technique to problems in information theory.

Mathematically, the problem can be formulated as:

$$\min_{\mathbf{x} \in \mathcal{Q}} f(\mathbf{x}), \tag{B.10.1}$$

where

$$\mathcal{Q} \triangleq \{\mathbf{x} \in \mathcal{X} : g_i(\mathbf{x}) \leq 0 \text{ for } 1 \leq i \leq m \text{ and } h_j(\mathbf{x}) = 0 \text{ for } 1 \leq j \leq \ell\}.$$

Lagrange multipliers technique and KKT condition I: b-45

- In most cases, solving the constrained optimization problem defined in (B.10.1) is hard (due to the constraints).
- Hence, one may introduce a **dual** optimization problem without constraints as

$$L(\boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq \min_{\boldsymbol{x} \in \mathcal{X}} \left(f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{\ell} \nu_j h_j(\boldsymbol{x}) \right) = \min_{\boldsymbol{x} \in \mathcal{X}} L(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\nu}). \quad (\text{B.10.2})$$

- In the literature, $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_{\ell})$ are usually referred to as **Lagrange multipliers**, and $L(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is called the **Lagrange dual function**.
 - Note that $L(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is a concave function of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ since it is the minimization of affine functions of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$.

Lagrange multipliers technique and KKT condition I: b-46

- It can be verified that when $\lambda_i \geq 0$ for $1 \leq i \leq m$,

$$L(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \min_{\mathbf{x} \in \mathcal{Q}} \left(f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{\ell} \nu_j h_j(\mathbf{x}) \right) \leq \min_{\mathbf{x} \in \mathcal{Q}} f(\mathbf{x}). \quad (\text{B.10.3})$$

- We are however interested in when the above inequality becomes equality (i.e., when the so-called *strong duality* holds) because if there exist non-negative $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\nu}}$ that equate (B.10.3), then

$$\begin{aligned} f(\mathbf{x}^*) &= \min_{\mathbf{x} \in \mathcal{Q}} f(\mathbf{x}) \\ &= L(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \\ &= \min_{\mathbf{x} \in \mathcal{X}} \left(f(\mathbf{x}) + \sum_{i=1}^m \tilde{\lambda}_i g_i(\mathbf{x}) + \sum_{j=1}^{\ell} \tilde{\nu}_j h_j(\mathbf{x}) \right) \\ &\leq f(\mathbf{x}^*) + \sum_{i=1}^m \tilde{\lambda}_i g_i(\mathbf{x}^*) + \sum_{j=1}^{\ell} \tilde{\nu}_j h_j(\mathbf{x}^*) \\ &\leq f(\mathbf{x}^*), \end{aligned} \quad (\text{B.10.4})$$

where (B.10.4) follows because the minimizer \mathbf{x}^* of (B.10.1) lies in \mathcal{Q} .

Lagrange multipliers technique and KKT condition

I: b-47

- Hence, if the strong duality holds, the same \mathbf{x}^* achieves both

$$\min_{\mathbf{x} \in \mathcal{Q}} f(\mathbf{x})$$

and

$$L(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}),$$

and $\tilde{\lambda}_i g_i(\mathbf{x}^*) = 0$ for $1 \leq i \leq m$.³

- The strong duality does not in general hold.
- A situation that guarantees the validity of the strong duality has been determined by William Karush [26], and separately Harold W. Kuhn and Albert W. Tucker [29].
- In particular, when $f(\cdot)$ and $\{g_i(\cdot)\}_{i=1}^m$ are both convex, and $\{h_j(\cdot)\}_{j=1}^{\ell}$ are affine, and these functions are all differentiable, they found that the strong duality holds if, and only if, the KKT condition is satisfied [11, p. 258].
 - Again, we are free to choose $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ that satisfy the KKT condition.

³Equating (B.10.4) implies $\sum_{i=1}^m \tilde{\lambda}_i g_i(\mathbf{x}^*) = 0$. It can then be easily verified from $\tilde{\lambda}_i g_i(\mathbf{x}^*) \leq 0$ for every $1 \leq i \leq m$ that $\tilde{\lambda}_i g_i(\mathbf{x}^*) = 0$ for $1 \leq i \leq m$.

Lagrange multipliers technique and KKT condition I: b-48

Definition B.20 (Karush-Kuhn-Tucker (KKT) condition) Point $\mathbf{x} = (x_1, \dots, x_n)$ and multipliers $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_\ell)$ are said to satisfy the KKT condition if

$$\begin{cases} g_i(\mathbf{x}) \leq 0, & \lambda_i \geq 0, & \lambda_i g_i(\mathbf{x}) = 0 & i = 1, \dots, m \\ h_j(\mathbf{x}) = 0 & & & j = 1, \dots, \ell \\ \frac{\partial L}{\partial x_k}(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\nu}) = \frac{\partial f}{\partial x_k}(\mathbf{x}) + \sum_{i=1}^m \lambda_i \frac{\partial g_i}{\partial x_k}(\mathbf{x}) + \sum_{j=1}^{\ell} \nu_j \frac{\partial h_j}{\partial x_k}(\mathbf{x}) = 0 & & & k = 1, \dots, n \end{cases}$$

- Note that when $f(\cdot)$ and constraints $\{g_i(\cdot)\}_{i=1}^m$ and $\{h_j(\cdot)\}_{j=1}^{\ell}$ are arbitrary functions, the KKT condition is only a necessary condition for the validity of the strong duality.
- In other words, for a non-convex optimization, we can only claim that if the strong duality holds, then the KKT condition is satisfied but not vice versa.

Lagrange multipliers technique and KKT condition I: b-49

- A case that is particularly useful in information theory is when \mathbf{x} is restricted to be a probability distribution.
- In such case, apart from other problem-specific constraints, we have additionally

$$\begin{cases} n \text{ inequality constraints } g_{m+i}(\mathbf{x}) = -x_i \leq 0 \text{ for } 1 \leq i \leq n, \text{ and} \\ \text{one equality constraint } h_{\ell+1}(\mathbf{x}) = \sum_{k=1}^n x_k - 1 = 0. \end{cases}$$

The above relation is the mostly seen form of the KKT condition when it is used in problems of information theory.

Lagrange multipliers technique and KKT condition I: b-50

Example B.21 Suppose for non-negative $\{q_{i,j}\}_{1 \leq i \leq n, 1 \leq j \leq n'}$ with $\sum_{j=1}^{n'} q_{i,j} = 1$,

$$\begin{cases} f(\mathbf{x}) = - \sum_{i=1}^n \sum_{j=1}^{n'} x_i q_{i,j} \log \frac{q_{i,j}}{\sum_{i'=1}^n x_{i'} q_{i',j}} \\ g_i(\mathbf{x}) = -x_i \leq 0 & i = 1, \dots, n \\ h(\mathbf{x}) = \sum_{i=1}^n x_i - 1 = 0 \end{cases}$$

Then

$$L(\mathbf{x}; \boldsymbol{\lambda}, \nu) \triangleq \left(f(\mathbf{x}) + \sum_{i=1}^n \lambda_i g_i(\mathbf{x}) + \nu h(\mathbf{x}) \right).$$

Then the KKT condition implies

$$\begin{cases} x_i \geq 0, \quad \lambda_i \geq 0, \quad \lambda_i x_i = 0 & i = 1, \dots, n \\ \sum_{i=1}^n x_i = 1 \\ \frac{\partial L}{\partial x_k}(\mathbf{x}; \boldsymbol{\lambda}, \nu) = \left(1 - \sum_{j=1}^{n'} q_{k,j} \log \frac{q_{k,j}}{\sum_{i'=1}^n x_{i'} q_{i',j}} \right) - \lambda_k + \nu = 0 & k = 1, \dots, n \end{cases}$$

Lagrange multipliers technique and KKT condition

I: b-51

which further implies that (we can choose)

$$\lambda_k = \begin{cases} 1 - \sum_{j=1}^{n'} q_{k,j} \log \frac{q_{k,j}}{\sum_{i'=1}^n x_{i'} q_{i',j}} + \nu = 0 & x_k > 0 \\ 1 - \sum_{j=1}^{n'} q_{k,j} \log \frac{q_{k,j}}{\sum_{i'=1}^n x_{i'} q_{i',j}} + \nu \geq 0 & x_k = 0 \end{cases}$$

□

- By this, the input distributions that achieve the channel capacities of some channels such as BSC and BEC can be identified.

Key Notes

I: b-52

- Definitions of (weakly, strictly) stationarity, ergodicity and Markovian (irreducible, homogeneous)
- Mode of convergences (almost surely or with probability 1, in probability, in distribution, in L_r mean)
- Laws of large numbers
- Central limit theorem
- Jensen's inequality (convexity and concavity)
- KKT condition