Appendix B

Overview in Probability and Random Processes

Po-Ning Chen, Professor

Institute of Communications Engineering

National Chiao Tung University

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

B.1 Probability space

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. *$\mathcal{F}$ is closed under complementation:* If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where $A^c = \{\omega \in \Omega \colon \omega \notin A\}$.
3. *$\mathcal{F}$ is closed under countable unions:* If $A_i \in \mathcal{F}$ for $i = 1, 2, 3, \ldots$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

- 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 $\{\Omega, \emptyset, A, A^c\}$.

B.1 Probability space

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

B.1 Probability space

- The Borel σ-field of $\mathbb{R}$, denoted by $\mathscr{B}(\mathbb{R})$, is the smallest σ-field of subsets of $\mathbb{R}$ containing all open intervals in $\mathbb{R}$.
- The elements of $\mathscr{B}(\mathbb{R})$ are called Borel sets.
- For any random variable X, we use P_X to denote the probability distribution on $\mathscr{B}(\mathbb{R})$ induced by X, given by

$$P_X(B) := \Pr[X \in B] = P\left(w \in \Omega : X(w) \in B\right), \qquad B \in \mathscr{B}(\mathbb{R}).$$

  Note that the quantities $P_X(B)$, $B \in \mathscr{B}(\mathbb{R})$, fully characterize the random variable X as they determine the probabilities of all events that concern X.

B.2 Random variables and random processes

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

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

  for each real t.

- A random process (or random source) 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.

B.2 Random variables and random processes

- The index set I can be uncountably infinite (e.g., $I = \mathbb{R}$), in which case we are dealing with a continuous-time process.
- Except for a brief interlude with the continuous-time (waveform) Gaussian channel in Chapter 5, we will consider discrete-time communication systems throughout the lectures.

  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 $(\Omega, \mathcal{F}, P)$?

Answer 1: $(\Omega, \mathcal{F}, P)$ is what truly occurs internally,

but is possibly **non-observable**.

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

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

- For example,

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

Distribution functions

- In many applications, we are perhaps more interested in the distribution functions of random variables than the underlying probability space on which they are defined.
- 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, $\Pr[X \leq x] = P_X((-\infty, x])$, 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, $F_X(x) = \Pr[X \leq x]$, without bothering to refer to its underlying probability space.
- Nevertheless, it is important to keep in mind that, formally, random variables are defined over underlying probability spaces.

Generalization of random variables

- The definition of a random variable X can be generalized by allowing it to take values that are not real numbers:

  Definition A random variable over the probability space $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathcal{X}$ satisfying the property that for every $F \in \mathcal{F}_X$,

  $$X^{-1}(F) := \{w \in \Omega : X(w) \in F\} \in \mathcal{F},$$

  where the alphabet $\mathcal{X}$ is a general set and $\mathcal{F}_X$ is a σ-field of subsets of $\mathcal{X}$ [R. M. Gray 2010, P. C. Shields 1991].

- Contrary to the standard definition of a random variable (by taking $\mathcal{X} = \mathbb{R}$), the elements in $\mathcal{X}$ may not have a pre-defined ordering; thus, the cdf,

  $$\Pr[X \leq x] = P(\{w \in \Omega : X(w) \leq x\}),$$

  needs to be explicitly defined.

- Note that this extension definition of a random variable allows $\mathcal{X}$ to be an arbitrary (often finite) set so that a random source taking values from, e.g., English alphabet, can now be regarded as a random variable.

B.3 Statistical properties of random sources

- Statistical evolution in time is an important factor for a random source.
- In particular, a "time-shift" property should be noted first.

Definition. An event E is said to be $\mathbb{T}$-*invariant* with respect to the left-shift (or shift transformation) $\mathbb{T}\colon \mathcal{X}^\infty \to \mathcal{X}^\infty$ if

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

where

$$\mathbb{T}E := \{\mathbb{T}\boldsymbol{x}\colon \boldsymbol{x} \in E\} \quad \text{and} \quad \mathbb{T}\boldsymbol{x} := \mathbb{T}(x_1, x_2, x_3, \ldots) = (x_2, x_3, \ldots).$$

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

B.3 Statistical properties of random sources

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

$$\begin{aligned} E &:= \{(x_1=1,x_2=1,x_3=1,x_4=1,\ldots),(x_1=0,x_2=1,x_3=1,x_4=1,\ldots), \\ &\qquad (x_1=0,x_2=0,x_3=1,x_4=1,\ldots)\}, \qquad \text{(B.3.1)} \\ &= \{\underbrace{1111\ldots}_{\text{all one}}, \quad \underbrace{0111\ldots}_{\text{all one but the first}}, \quad \underbrace{0011\ldots}_{\text{all one but the first two}}\} \end{aligned}$$

yields

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

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

E will get smaller and smaller (more condensed) as time evolves.

B.3 Statistical properties of random sources

- 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

$$\cdots \subseteq \mathbb{T}^3E \subseteq \mathbb{T}^2E \subseteq \mathbb{T}E \subseteq E.$$

- Thus, if an element say $(1, 0, 0, 1, 0, 0, \ldots)$ is in a $\mathbb{T}$-invariant set E, then all its left-shift counterparts (i.e., $(0, 0, 1, 0, 0, 1 \ldots)$ and $(0, 1, 0, 0, 1, 0, \ldots)$) should be contained in E.
- As a result, for a $\mathbb{T}$-invariant set E, an element and all its left-shift counterparts are either all in E or all outside E, but cannot be partially inside E.
- Hence, a "$\mathbb{T}$-invariant group" such as one containing

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

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

B.3 Statistical properties of random sources

- Although we are in particular interested in these "$\mathbb{T}$-invariant indecomposable groups" (especially when defining an ergodic random process), it is possible that some single "transient" element, such as $(0, 0, 1, 1, \ldots)$ in (B.3.1), is included in a $\mathbb{T}$-invariant set, and will be excluded after applying left-shift operation $\mathbb{T}$.
- This however can be resolved by introducing the inverse operation $\mathbb{T}^{-1}$.
- 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 [P. C. Shields 1991, p. 3] allows us to "enlarge" the $\mathbb{T}$-invariant set such that all right-shift counterparts of the single "transient" element are included:

$$\mathbb{T}^{-1}E := \{\boldsymbol{x} \in \mathcal{X}^{\infty} \colon \mathbb{T}\boldsymbol{x} \in E\}\,.$$

B.3 Statistical properties of random sources

- We then notice from the above definition that if

$$\mathbb{T}^{-1}E = E, \tag{B.3.2}$$

then

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

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

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

- The sets that satisfy (B.3.2) 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.

B.3 Statistical properties of random sources

- As the textbook only deals with one-sided random processes, the discussion on $\mathbb{T}$-invariance only focuses on sets of one-sided sequences.
- When a two-sided random process $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ 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, $\mathbb{T}E = E$ iff $\mathbb{T}^{-1}E = E$. Ergodicity for two-sided sequences can therefore be directly defined using $\mathbb{T}E = E$.

B.3 Statistical properties of random sources

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

$$\boldsymbol{X} = \{X_1, X_2, \ldots\}.$$

◇ *Memoryless*: A random process or a source $\boldsymbol{X}$ is said to be *memoryless* if the sequence of random variables X_i is *independent and identically distributed* (i.i.d.).

◇ *Stationary process*: A process is said to be *stationary* (or *strictly stationary*) if the probability of every sequence or event is unchanged by a left (time) shift.

◇ *Ergodic process*: A process is said to be *ergodic* if any ergodic set (satisfying (B.3.2)) in $\mathcal{F}_{\boldsymbol{X}}$ has probability either 1 or 0. This definition is not very intuitive, but some interpretations and examples may shed some light.

- 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.
- Ergodicity implies that all convergent time averages converge to a constant (but not necessarily to the ensemble average or statistical expectation).

Below is an example that can be used to explain the idea.

Example. Suppose $X_1, X_2, \cdots, X_n, \cdots$ 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 $\alpha \in [0,1]$,

$$\bar{E}_n(\alpha) := \left\{ \boldsymbol{x} \in \{0,1\}^\infty \ : \ \alpha \le \limsup_{m\to\infty} \frac{x_1 + \cdots + x_m}{m} < \alpha + \frac{1}{n} \right\}$$

and

$$\underline{E}_n(\alpha) := \left\{ \boldsymbol{x} \in \{0,1\}^\infty \ : \ \alpha \le \liminf_{m\to\infty} \frac{x_1 + \cdots + x_m}{m} < \alpha + \frac{1}{n} \right\}.$$

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

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

Observe that

$$\Omega = \bigcup_{k=0}^{n} \bar{E}_n\big(\tfrac{k}{n}\big) = \bigcup_{k=0}^{n} \underline{E}_n\big(\tfrac{k}{n}\big)$$

and

$$\bar{E}_n\big(\tfrac{k}{n}\big) \bigcap \bar{E}_n\big(\tfrac{\ell}{n}\big) = \underline{E}_n\big(\tfrac{k}{n}\big) \bigcap \underline{E}_n\big(\tfrac{\ell}{n}\big) = \emptyset \text{ for } k \neq \ell.$$

B.3 Statistical properties of random sources

The definition of ergodicity implies the existence of k and ℓ such that

$$\Pr\left[\boldsymbol{X} \in \bar{E}_n\left(\tfrac{k}{n}\right)\right] = \Pr\left[\boldsymbol{X} \in \underline{E}_n\left(\tfrac{\ell}{n}\right)\right] = 1.$$

If $\frac{X_1+\cdots+X_n}{n}$ converges with probability one, then $k = \ell$.

In other words,

$$\frac{k}{n} \leq \limsup_{m\to\infty} \frac{X_1 + \cdots + X_m}{m} < \frac{k+1}{n} \text{ with probability 1}$$

and

$$\frac{k}{n} \leq \liminf_{m\to\infty} \frac{X_1 + \cdots + X_m}{m} < \frac{k+1}{n} \text{ with probability 1.}$$

As a result,

$$\left|\limsup_{m\to\infty} \frac{X_1 + \cdots + X_m}{m} - \liminf_{m\to\infty} \frac{X_1 + \cdots + X_m}{m}\right| < \frac{1}{n} \text{ with probability 1.}$$

□

Ergodicity implies that all convergent time averages converge to a constant.

- It needs to be pointed out that in the above example, ergodicity does not guarantee that the ensemble average lies in $[k/n, (k+1)/n)$.
- A quick example is that

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

and

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

assure the validity of ergodicity, but

$$\frac{X_1 + \cdots + X_n}{n} \to \frac{1}{2} \text{ with probability 1.}$$

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

- In principle,
  - ergodicity implies that all convergent sample averages converge to a constant (but not necessarily to the statistical expectation), 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.

B.3 Statistical properties of random sources

Theorem B.4 (Pointwise ergodic theorem) Consider a discrete-time stationary random process, $\boldsymbol{X} = \{X_n\}_{n=1}^{\infty}$. 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

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

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

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

Operational meaning of stationary ergodic assumption

- 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.
  - 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:

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

  - 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.

Operational meaning of stationary ergodic assumption

- In communications theory, one may assume that *the source is stationary* or *the source is stationary ergodic*. But it is not common to see the assumption of *the source being ergodic but non-stationary*. Why?
  - This is perhaps because an ergodic but non-stationary source in general does not facilitate the analytical study of communications problems.
- 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.

B.3 Statistical properties of random sources

We continue to classify useful statistical properties of (one-sided) random process

$$\boldsymbol{X} = \{X_1, X_2, \ldots\}.$$

◇ Markov chain for three random variables:

Three random variables X, Y and Z are said to form a Markov chain if

$$P_{X,Y,Z}(x, y, z) = P_X(x) \cdot P_{Y|X}(y|x) \cdot P_{Z|Y}(z|y); \tag{B.3.3}$$

i.e.,

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

This is usually denoted by

$$X \to Y \to Z.$$

- "$X \to Y \to Z$" is sometimes read as "X and Z are conditionally independent given Y" because it can be shown that (B.3.3) is equivalent to

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

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

B.3 Statistical properties of random sources

◇ kth-order Markov sources:

The sequence of random variables $\{X_n\}_{n=1}^{\infty} = X_1, X_2, X_3, \ldots$ with common finite-alphabet $\mathcal{X}$ is said to form a k-th order Markov chain (or k-th order Markov source or process) if for all $n > k$, $x_i \in \mathcal{X}$, $i = 1, \ldots, n$,

$$\begin{aligned} &\Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1] \\ &= \Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}]. \end{aligned} \tag{B.3.4}$$

Each $x_{n-k}^{n-1} := (x_{n-k}, x_{n-k+1}, \ldots, x_{n-1}) \in \mathcal{X}^k$ is called the *state* of the Markov chain at time n.

- **Irreducible:** A Markov chain is *irreducible* if with some (non-zero) 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

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

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

$$\begin{aligned}&\Pr[X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}] \\ &\quad = \Pr[X_{k+1} = x_{k+1} | X_k = x_k, \ldots, X_1 = x_1].\end{aligned}$$

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

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

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

- **Aperiodic:**
  - In a first-order Markov chain, the *period* $d(x)$ of state $x \in \mathcal{X}$ is defined by

$$d(x) := \gcd\{n \in \{1,2,3,\ldots\}\colon \Pr\{X_{n+1} = x | X_1 = x\} > 0\},$$

    where gcd denotes the greatest common divisor; in other words, if the Markov chain starts in state x, then the chain cannot return to state x at any time that is not a multiple of $d(x)$.
  - If $\Pr\{X_{n+1} = x | X_1 = x\} = 0$ for all n, we say that state x has an infinite period and write $d(x) = \infty$.
  - We also say that *state x is aperiodic* if $d(x) = 1$ and *periodic* if $d(x) > 1$.
  - The first-order Markov chain is called **aperiodic** if all its states are aperiodic. In other words, the first-order Markov chain is aperiodic if

$$\gcd\{n \in \{1,2,3,\ldots\}\colon \Pr\{X_{n+1} = x | X_1 = x\} > 0\} = 1 \quad \forall x \in \mathcal{X}.$$

Property. In an irreducible first-order Markov chain, all states have the same period. Hence, if one state in such a chain is aperiodic, then the entire Markov chain is aperiodic.

B.3 Statistical properties of random sources

- **Stationarity:** A distribution $\pi(\cdot)$ on $\mathcal{X}$ is said to be a *stationary* distribution for a homogeneous (i.e., time-invariant) first-order Markov chain, if for every $y \in \mathcal{X}$,

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

Properties.

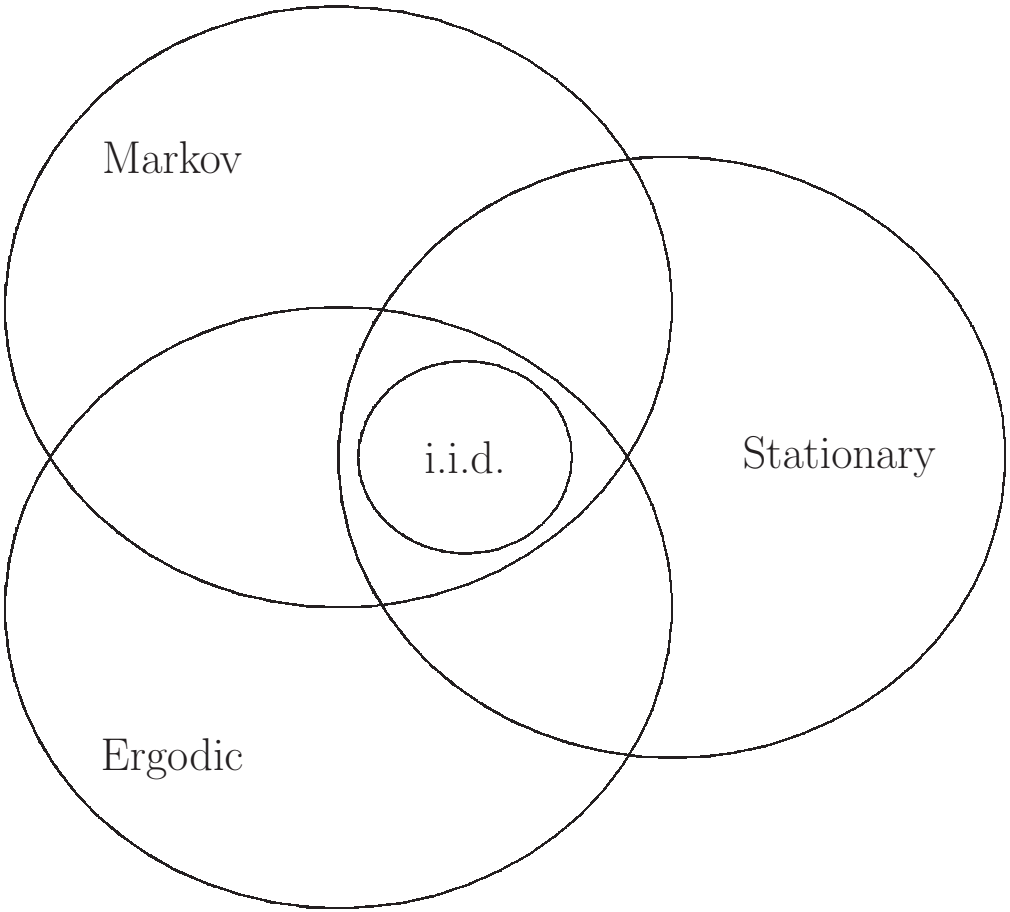
1. For a finite-alphabet homogeneous first-order Markov chain, $\pi(\cdot)$ always exists.
2. $\pi(\cdot)$ is unique if the Markov chain is irreducible.
3. For a finite-alphabet homogeneous first-order Markov chain that is both irreducible and aperiodic,

$$\lim_{n \to \infty} \Pr\{X_{n+1} = y | X_1 = x\} = \pi(y)$$

for all states x and y in $\mathcal{X}$.

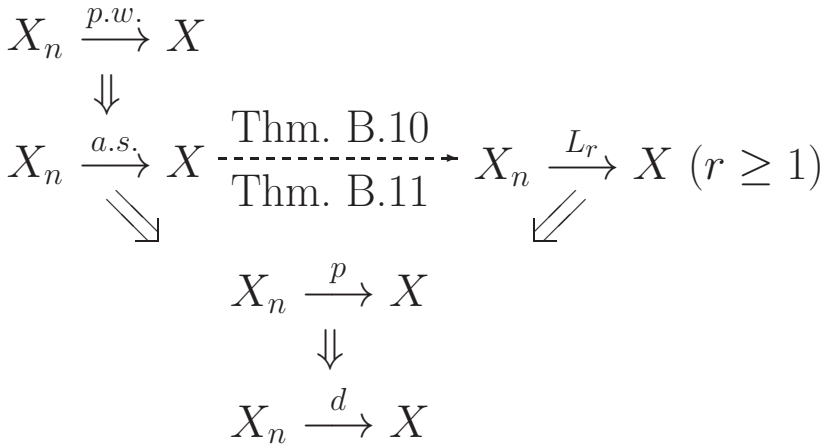
If the initial state distribution is equal to a stationary distribution, then the homogeneous first-order Markov chain becomes a stationary process.

General Relation of Random Processes



B.4 Convergence of sequences of random variables

- Relation of five modes of convergence



B.4 Convergence of sequences of random variables

- Pointwise convergence and almost surely convergence

  Example B.7 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 $\mathbb{R}$. 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

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

  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

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

B.4 Convergence 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

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

B.4 Convergence of sequences of random variables

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.

B.4 Convergence of sequences of random variables

Theorem B.9 (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.10 (Dominated convergence theorem)

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

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

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

B.5 Ergodicity and law of large numbers

Theorem B.13 (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,

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[X_i] = 0, \quad (\text{equivalently}, \frac{X_1 + \cdots + X_n}{n} \xrightarrow{\mathcal{L}_2} \mu)$$

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: $X_n \xrightarrow{\mathcal{L}_2} X$ implies $X_n \xrightarrow{p} X$.

B.5 Ergodicity and law of large numbers

Theorem B.14 (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

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

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 sure* convergence.

B.5 Ergodicity and law of large numbers

- 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.

B.5 Ergodicity and law of large numbers

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.

B.5 Ergodicity and law of large numbers

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!

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

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

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

- *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 $A_1, A_2, A_3, \ldots$ or $B_1, B_2, B_3, \ldots$, 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 !

B.5 Ergodicity and law of large numbers

- 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.

B.6 Central limit theorem

Theorem B.15 (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 $n \to \infty$) and $Z \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian distributed random variable with mean 0 and variance σ^2.

B.7 Convexity, concavity and Jensen's inequality

Definition B.16 (Convexity) Consider a convex set $\mathcal{O} \subset \mathbb{R}^m$, where m is a fixed positive integer. Then a function $f : \mathcal{O} \to \mathbb{R}$ is said to be *convex* over $\mathcal{O}$ if for every $\underline{x}$, $\underline{y}$ in $\mathcal{O}$ and $0 \leq \lambda \leq 1$,

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

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

- A set $\mathcal{O} \subset \mathbb{R}^m$ is said to be *convex* if for every $\underline{x} = (x_1, x_2, \cdots, x_m)^T$ and $\underline{y} = (y_1, y_2, \cdots, y_m)^T$ 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.17 (Concavity) A function f is *concave* if $-f$ is convex.

Jensen's inequality

Theorem B.18 (Jensen's inequality) If $f : \mathcal{O} \to \mathbb{R}$ is convex over a convex set $\mathcal{O} \subset \mathbb{R}^m$, and $\underline{X} = (X_1, X_2, \cdots, X_m)^T$ 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.

B.8 Lagrange multipliers tech. & KKT conditions

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

subject to

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

is a center technique to problems in information theory.

Mathematically, the problem can be formulated as:

$$\min_{\boldsymbol{x} \in \mathcal{Q}} f(\boldsymbol{x}), \tag{B.8.1}$$

where

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

B.8 Lagrange multipliers tech. & KKT conditions

- In most cases, solving the constrained optimization problem defined in (B.8.1) is hard due to the constraints.
- Instead, one may introduce a **dual** optimization problem without constraints:

$$L(\boldsymbol{\lambda}, \boldsymbol{\nu}) := \min_{\boldsymbol{x} \in \mathcal{X}} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{\ell} \nu_j h_j(\boldsymbol{x}) \right) = \min_{\boldsymbol{x} \in \mathcal{X}} L(\boldsymbol{x}; \boldsymbol{\lambda}, \boldsymbol{\nu}). \tag{B.8.2}$$

- In the literature, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_\ell)$ 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 $\boldsymbol{\nu}$.

B.8 Lagrange multipliers tech. & KKT conditions

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

$$L(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \min_{\boldsymbol{x} \in \mathcal{Q}} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{\ell} \nu_j h_j(\boldsymbol{x}) \right) \leq \min_{\boldsymbol{x} \in \mathcal{Q}} f(\boldsymbol{x}). \quad \text{(B.8.3)}$$

- 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.8.3), then

$$\begin{aligned} f(\boldsymbol{x}^*) &= \min_{\boldsymbol{x} \in \mathcal{Q}} f(\boldsymbol{x}) \\ &= L(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) = \min_{\boldsymbol{x} \in \mathcal{X}} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i g_i(\boldsymbol{x}) + \sum_{j=1}^{\ell} \tilde{\nu}_j h_j(\boldsymbol{x}) \right) \\ &\leq f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \tilde{\lambda}_i g_i(\boldsymbol{x}^*) + \sum_{j=1}^{\ell} \tilde{\nu}_j h_j(\boldsymbol{x}^*) \\ &\leq f(\boldsymbol{x}^*), \qquad \text{(B.8.4)} \end{aligned}$$

where (B.8.4) follows because the minimizer $\boldsymbol{x}^*$ of (B.8.1) lies in $\mathcal{Q}$.

B.8 Lagrange multipliers tech. & KKT conditions

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

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

and

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

and $\tilde{\lambda}_i g_i(\boldsymbol{x}^*) = 0$ for $1 \le i \le m$.[1]

- The strong duality does not in general hold.
- A situation that guarantees the validity of the strong duality has been determined by William Karush [212] (1936), and separately Harold W. Kuhn and Albert W. Tucker [235] (1951).
- 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 [56, p. 258].
  - Again, we are free to choose $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ that satisfy the KKT condition (cf. Definition B.19).

[1] Equating (B.8.4) implies $\sum_{i=1}^m \tilde{\lambda}_i g_i(\boldsymbol{x}^*) = 0$. It can then be easily verified from $\tilde{\lambda}_i g_i(\boldsymbol{x}^*) \le 0$ for every $1 \le i \le m$ that $\tilde{\lambda}_i g_i(\boldsymbol{x}^*) = 0$ for $1 \le i \le m$.

B.8 Lagrange multipliers tech. & KKT conditions

Definition B.19 (Karush-Kuhn-Tucker (KKT) condition) Point $\boldsymbol{x} = (x_1, \ldots, x_n)$ and multipliers $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_\ell)$ are said to satisfy the KKT condition if

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

- 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.

B.8 Lagrange multipliers tech. & KKT conditions

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

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

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

Example B.20 Suppose for non-negative $\{q_{i,j}\}_{1\le i\le n, 1\le j\le n'}$ with $\sum_{j=1}^{n'} q_{i,j} = 1$,

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

Then

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

Then the KKT condition implies

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

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

- 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 conditions