Module B: Random Processes

> A random process is a family/ collection of random variables indexed by a set T, stated at $\{X_t\}_{t \in T}$.

The set T is often interpreted as "time."

- When $T = \{1, 2, ...n\}$, then $\{X_t\}_{t \in T} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$ is a random vector.
- When $T = \{1, 2, 3, ...\} = \mathbb{N}$, then $\{X_t\}_{t \in T} = (X_1, X_2, X_3, ...)$ is called a discrete-time random process.
- When $T = \mathbb{R}$, $\{X_t\}_{t \in T}$ is an uncountable collection of random variables and is called a continuous-time random process.

Recall: $X_t : \Omega \to \mathbb{R}$ fix ω: $X_t(\omega)$: function of t is called the **sample path**.

Example $X_t = \cos(2\pi\ \omega t)$ where ω the random outcome $\omega = \begin{cases} 1 & w.p & \frac{1}{3} \\ 2 & w.p & \frac{1}{3} \\ 3 & w.p & \frac{1}{3} \end{cases}$

How do we specify a random process $\{X_t\}_{t \in T}$: To fully specify a random process, for any finite collection of indices $(t_1, t_2, ...t_n)$, the joint distribution $(X_{t_1}, X_{t_2}, ...X_{t_n})$ should be provided.

Deterministic vs Stochastic Dynamical Systems

- Deterministic: starting from $x_0 \in \mathbb{R}^n$ for all $t \geq 0$

$$x_{t+1} = f(t, x_t).$$

More generally: $x_{t+1} = f(t, x_t, \ldots, x_{t-m})$ where m is the memory of the system.

Example: $n = 1$, starting at $x_0 > 0$, a simple (deterministic) population growth model:

$$x_{t+1} = r_0 x_t.$$

Note that $x_t = r_0^t x_0$.

- Random process: starting from $x_0 \in \mathbb{R}^n$ for all $t \geq 0$

$$x_{t+1} = f(t, x_t, w_t).$$

More generally: $x_{t+1} = f(t, x_t, \ldots, x_{t-m}, w_t)$ where m is the memory of the system and w_t is a random variable/vector.

Example: Beginning phase of a pandemics: for some initially infected population $x_0 > 0$, the population of infected people at the beginning phase of a pandemics can be modeled by:

$$x_{t+1} = r_t x_t,$$

where r_t is a non-negative random variable independent of r_k for $k < t$ with some $\mathbb{E}[r_t] = r_0$.

Examples of Random Processes

- Averaging: suppose that $\{w_t\}$ is an independently and identically distributed random process with $\mathbb{E}[w_k] = \mu$.

  How does the running average $x_t = \frac{w_1 + \ldots + w_t}{t}$ behave as $t \to \infty$?

  In this case:

$$
\begin{aligned}
tx_t &= (t-1)x_{t-1} + w_t \\
x_t &= (1 - \frac{1}{t})x_{t-1} + \frac{1}{t}w_t \\
x_t &= f_t(x_{t-1}, w_t) \\
f_t(x, w) &= (1 - \frac{1}{t})x + \frac{1}{t}w.
\end{aligned}
$$

- What happens if we use other weights such as: $x_t = \frac{w_1 + \ldots + w_t}{\sqrt{t}}$?
- What if we don't have any weights at all, i.e., $x_t = w_1 + \ldots + w_t$? What happens?
- What can we say about asymptotic behavior of such processes in general?

Terminology

For a random process $X = \{X_t\}_{t \in T}$

(a) **Mean function**:

$$\mu_X(t) := \mathbb{E}[X_t]$$

(b) **Autocorrelation function**:

$$R_X(t_1, t_2) := \mathbb{E}[X_{t_1} X_{t_2}]$$

(c) **Autocovariance function**:

$$C_X(t_1, t_2) := R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2)$$

For an i.i.d. process

If the random process $\{X_t\}$ is i.i.d., then

(a) For the mean function:

$$\mu_X(t) = \mathbb{E}[X_t] = \mathbb{E}[X_0].$$

Therefore, we have a constant mean function.

(b) For the **Autocorrelation function**:

$$R_X(t_1, t_2) = \mathbb{E}[X_{t_1} X_{t_2}] = \begin{cases} \mathbb{E}[X_{t_1}^2] = \mathbb{E}[X_1^2] & t_1 = t_2 \\ \mathbb{E}[X_{t_1}]\mathbb{E}[X_{t_2}] = \mu_X(t_1)\mu_X(t_2) = \mu_X(0)^2 & t_1 \neq t_2 \end{cases}.$$

(c) **Autocovariance function**:

$$C_X(t_1, t_2) = R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2) = \begin{cases} \texttt{var}(X_1) & t_1 = t_2 \\ 0 & t_1 \neq t_2 \end{cases}.$$

That is, the random process is uncorrelated in time.

Plus many other properties are true.

Stationary Processes

A random process is **Strict Sense Stationary** (SSS) if the (finite) joint probability distributions (CDFs) are invariant under shift, i.e., for all $t_1 < t_2 < \cdots < t_k$ and all $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$:

$$F_{X_{t_1},\ldots,X_{t_k}}(\alpha_1, \ldots, \alpha_k) = F_{X_{t_1+s},\ldots,X_{t_k+s}}(\alpha_1, \ldots, \alpha_k)$$

for all $-t_1 \le s$.
Example: i.i.d. processes as

$$F_{X_{t_1},\ldots,X_{t_k}}(\alpha_1, \ldots, \alpha_k) = F_{X_{t_1}}(\alpha_1) \cdots F_{X_{t_1}}(\alpha_k) = F_X(\alpha_1) \cdots F_X(\alpha_k).$$

A random process is **Wide Sense Stationary** (WSS) if

1. the mean function does not depend on time t, and
2. the $R_X(t_1, t_2) = f(t_1 - t_2)$, i.e., autocorrelation function is just a function of $t_1 - t_2$.

Example: i.i.d. processes

For two random processes $\{X_t\}_{t \in T}$ and $\{Y_t\}_{t \in T}$,

- Cross-correlation $R_{XY}(t_1, t_2) = \mathbb{E}[X(t_1)Y(t_2)] \neq \mathbb{E}[Y(t_1)X(t_2)]$
- Cross-covariance $C_{XY}(t_1, t_2) = \mathtt{cov}[X(t_1), Y(t_2)] = R_{XY}(t_1, t_2) - \mu_X(t_1)\mu_Y(t_2)$

$\{X_t\}_{t \in T}$ and $\{Y_t\}_{t \in T}$ are jointly WSS if

- Both $\{X_t\}_{t \in T}$ and $\{Y_t\}_{t \in T}$ are individually WSS
- $R_{XY}(t_1, t_2) = R_{XY}(t_1 - t_2)$

Example: Random Walk

Let $\{X_k\}$ be a random walk, given by $X_{k+1} = X_k + Z_k$ where $\{Z_k\}$ is i.i.d. with zero mean and variance σ^2 and $X_0 = 0$ a.s.

(a) For the mean function:

$$\mu_X(k) = \mathbb{E}[X_{k-1} + Z_{k-1}] = \mathbb{E}[X_{k-1}].$$

Therefore, $\mu_X(k) = \mu_X(k-1) = \ldots = \mu_X(0) = 0$.

(b) For the **Autocorrelation function**: Let $k_1 \leq k_2$:

$$\begin{aligned} R_X(k_1, k_2) &= \mathbb{E}[X_{k_1} X_{k_2}] = \mathbb{E}[X_{k_1}(X_{k_2} - X_{k_1} + X_{k_1})] \\ &= \mathbb{E}[X_{k_1}(X_{k_2} - X_{k_1})] + \mathbb{E}[X_{k_1}^2] \\ &= \mathbb{E}[X_{k_1}^2] = k_1 \sigma^2. \end{aligned}$$

Therefore, $R_X(k_1, k_2) = \min(k_1, k_2)\sigma^2$. Thus, such a process is not WSS and hence, not an SSS.

(c) **Autocovariance function**: since the process is zero mean $C_X = R_X$.

Continuous Time Random Processes

- Example: for a deterministic $\alpha > 0$ and frequency ω, let $X_t = \alpha \cos(\omega t + \theta)$ where $\theta \sim U([0, 2\pi])$.
  - The mean function:

$$\mu_X(t) = \mathbb{E}[\alpha \cos(\omega t + \theta)] = \frac{1}{2\pi} \int_0^{2\pi} \alpha \cos(\omega t + \theta) d\theta = 0.$$

  - The correlation function:

$$\begin{aligned}
R_X(t_1, t_2) &= \mathbb{E}[\alpha \cos(\omega t_1 + \theta) \alpha \cos(\omega t_2 + \theta)] \\
&= \frac{1}{2\pi} \int_0^{2\pi} \alpha^2 \cos(\omega t_1 + \theta) \cos(\omega t_2 + \theta) d\theta \\
&= \frac{\alpha^2}{2\pi} \int_0^{2\pi} \cos(\omega t_1 + \theta) \cos(\omega t_2 + \theta) d\theta \\
&= \frac{\alpha^2}{4\pi} \int_0^{2\pi} \cos(\omega(t_1 + t_2) + 2\theta) d\theta + \frac{\alpha^2}{4\pi} \int_0^{2\pi} \cos(\omega(t_1 - t_2)) d\theta \\
&= \frac{\alpha^2}{4\pi} \int_0^{2\pi} \cos(\omega(t_1 - t_2)) d\theta \\
&= \frac{\alpha^2}{2} \cos(\omega(t_1 - t_2)).
\end{aligned}$$

Properties of WSS Processes

- Some properties of a WSS process $\{X_t\}$:

  1. $R_X(\tau) = \mathbb{E}[X(t)X(t+\tau)]$ is an even function, i.e., $R_X(\tau) = R_X(-\tau)$.
  2. $R_X(0) \geq R_X(\tau)$ for all τ.
  3. For independent processes $\{X(t)\}$ and $\{Y(t)\}$ with zero mean, $R_{X+Y}(\tau) = R_X(\tau) + R_Y(\tau)$.

Proof on board in class.

Ergodic Behavior

- Statistical mean: $\mu_X(t) = \mathbb{E}[X(t)] = \int_{\omega \in \Omega} X_t(\omega)\ d\mathbb{P}_{X_t}(\omega)$.
- If we have M samples of X_{t_1}, denoted $(\widehat{x}_{t_1}^1, \widehat{x}_{t_1}^2, \dots, \widehat{x}_{t_1}^M)$, drawn from $\mathbb{P}_{X_{t_1}}$, then we can estimate the **statistical average** as $\widehat{\mu}_X(t_1) = \frac{1}{M} \sum_{i=1}^m \widehat{x}_{t_1}^i$.
- However, suppose we have a single sample path of the random process given by $x_1(\omega_0), x_2(\omega_0), \dots$. Then, we can find the **temporal mean and autocorrelation** as

$$\overline{X}(\omega_0) = \frac{1}{T} \int_0^T x_t(\omega_0) dt$$
$$\overline{R}_X(\tau) = \frac{1}{T} \int_0^T x_{t+\tau}(\omega_0) x_t(\omega_0) dt$$

- Do the temporal and statistical averages coincide? Yes, when the process is **ergodic**. The random process $\{X_t\}_{t \in T}$ is ergodic when

$$\mathbb{E}[X_t] = \lim_{T \to \infty} \frac{1}{T} \int_0^T X_t(\omega_0)\ dt.$$

  It is implicit that for ergodic process, $\mathbb{E}[X_t] = \mu_X(t) = \mu_X$ for all t.

- For a discrete-time process, we replace the integral by summation to compute temporal averages.

Mean-Square Ergodic Theorem: Let $\{X_t\}_{t \in T}$ be a wide sense stationary process with $\mathbb{E}[X_t] = \mu_X$ and auto-correlation $R_X(\tau)$, and let the Fourier transform of $R_X(\tau)$ exists. Let $\overline{X_T}(\omega) = \frac{1}{2T} \int_{-T}^T X_t(\omega)\ dt$. Then,

$$\lim_{T \to \infty} \mathbb{E}[(\overline{X_T} - \mu_X)^2] = 0.$$

In other words, $\overline{X_T}$ converges to μ_X in mean-square sense.

The implication of the above theorem is that, we can approximate mean/ Correlation by temporal average computed from a single sample path.

Random Process and LTI System

- Suppose we have a LTI system with impulse response $h(t)$. If we apply input signal $x(t)$ to this system, the output signal $y(t)$ is given as

$$y(t) = \int_{\infty}^{\infty} h(\tau)x(t-\tau)d\tau =: x(t) \circledast h(t).$$

- Now, suppose the input $X(t)$ is a random process with mean $\mu_X(t)$ and autocorrelation $R_X(t_1, t_2)$. Determine the mean and autocorrelation of Y.

- If $X(t)$ is WSS, is $Y(t)$ also WSS?

  Yes. Derivation in class.

- Are $X(t)$ and $Y(t)$ jointly WSS?

  Yes. Derivation in class. We can show that

$$R_{YX}(\tau) = h(\tau) \circledast R_X(\tau)$$

Power Spectral Density (PSD)

- From the above discussion, we have $R_{YX}(\tau) = \int_{\infty}^{\infty} h(s) R_X(\tau - s) ds$.

- For a CT WSS process $X(t)$ (that is integrable), we can find the "power spectral density" at frequency ω (rad/s):

$$S_X(\omega) := FT[R_X(\tau)] = \int_{-\infty}^{\infty} R_X(\tau) e^{-j\omega\tau} d\tau$$

- Thus, $S_{YX}(\omega) = H(\omega) S_X(\omega)$ where $H(\omega)$ is the Fourier transform of the impulse response $h(t)$.

- We can further show that

$$\begin{aligned}
R_Y(\tau) &= h(\tau) \circledast R_{XY}(\tau) \\
\Longrightarrow \; S_Y(\omega) &= H(\omega) \times S_{XY}(\omega) \\
\text{In addition,} \quad S_{YX}(\omega) &= H(\omega) \times S_X(\omega) \\
\text{Since,} \quad R_{XY}(\tau) &= R_{YX}(\tau), \quad \text{we have} \quad S_{XY}(\omega) = S_{YX}(\omega)^* \\
\Longrightarrow \; S_Y(\omega) &= |H(\omega)|^2 S_X(\omega).
\end{aligned}$$

Discrete-time WSS Processes

- A discrete-time random processs $(X_n)_{n\in\mathbb{N}}$ is a collection of random variables $(X_1, X_2, \ldots, X_n, \ldots)$.

- Mean function $\mu_X[n] = \mathbb{E}[X_n]$.

- Autocorrelation function $R_X[n_1, n_2] = \mathbb{E}[X_{n_1} X_{n_2}]$.

- Autocovariance function $C_X[n_1, n_2] = \mathtt{cov}(X_{n_1}, X_{n_2})$.

- Cross-correlation function $R_{XY}[n_1, n_2] = \mathbb{E}[X_{n_1} Y_{n_2}]$.

- For X to be W.S.S, the following properties need to be satisfied.
  1. $\mu_X[n] = \mu$ independent of n.
  2. $R_X[n_1, n_2] = R_X[n_2 - n_1]$.

- Properties such as ergodicity and output of LTI system to a WSS input continue to hold in an analogous manner.

Module B.2: Markov Chains

- Markov Process: A **random process** whose probability distribution at time $t+1$ given the past **only depends** on its value at time t. Specifically,

$$\Pr(X_{k+1} \in A \mid X_k, \ldots, X_1) = \Pr(X_{k+1} \in A \mid X_k).$$

  More generally

$$\Pr(X_{k+1} \in A \mid X_{k_i}, \ldots, X_{k_1}) = \Pr(X_{k+1} \in A \mid X_{k_i}),$$

  for any $k_1 < k_1 < \ldots < k_i < k$.

- If the (time) index set is continuous, the corresponding random process is called *Markov Process*.

- In this course: we focus on discrete-time Markov process where each random variable X_k is a discrete random variable that takes values from a finite set.

- Example: Infectious disease with reinfection where an individual can be in one of two possible states: susceptible (S) and infected (I).

Formal Definition

- **Definition**: We say that a (DT) random process $\{X_k\}$ is a Markov chain over a discrete-space if

  1. X_k's are all discrete random variables with common support S, i.e., $\mathbf{Pr}(X_k \in S) = 1$ for all k, where S is countable, and
  2. for all $i \geq 1$, all $1 \leq k_1 < k_2 < \ldots < k_i \leq k$, and all $1, \ldots, i, s \in S$:

$$\Pr(X_{k+1} = s \mid X_{k_i} = s_i, \ldots, X_{k_1} = s_1) = \Pr(X_{k+1} = s \mid X_{k_i} = s_i). \quad (1)$$

- S is called the state space and each $s \in S$ is called a state. Relation (1) is called *Markov property*.

- If S is finite, $\{X_k\}$ is called a finite state Markov chain.

Transition Probabilities

- From this point on assume S is a finite set with elements, $S = \{1, \ldots, n\}$. Unless otherwise stated, many of the following discussions hold for $n = \infty$ but for convenience we assume that n is finite.

- For any k, let π_k be the (marginal) probability mass function X_k, i.e.,

$$\pi_k(i) = \Pr(X_k = i).$$

  Note that the vector π_k is non-negative and $\sum_{i=1}^{d} \pi_k(i) = 1$. Such a vector is called a stochastic (sometimes probability) vector. It is convenient to assume that π_k is a **row** vector.

- For any $1 \leq k_1 < k_2$, define the matrix (array)

$$P_{k_1,k_2}(i, j) = \Pr(X_{k_2} = j \mid X_{k_1} = i).$$

- $P_{k_1,k_2} \in \mathbb{R}^{n_1 \times n_2}$ is called the transition matrix of MC from time k_1 to time k_2. In other words,

$$\pi_{k_2} = \pi_{k_1} P_{k_1,k_2}.$$

- We also (naturally) define $P_{k,k} := I$, where I is the $n \times n$ identity matrix.

Properties of Transition Matrices

- Definition: We say that a $n \times n$ matrix A is a row-stochastic matrix if (i) A is non-negative, and (ii) $A\mathbf{1} = \mathbf{1}$ (or each row sums up to one).
- Properties of the transition matrices:
  - **Row-stochastic**: For any $k \le m$, $P_{k,m}$ is a row-stochastic matrix: The non-negativeness follows from the definition. Also, each row adds up to one:

$$\sum_{j=1}^{n} P_{k,m}(i,j) = \sum_{j=1}^{n} \Pr(X_m = j \mid X_k = i) = 1.$$

  - For any $k \le m$, we have:

$$\pi_m = \pi_k P_{k,m}.$$

    This follow from the fact:

$$\begin{aligned} \pi_m(j) = \Pr(X_m = j) &= \sum_{i=1}^{n} \Pr(X_m = j, X_k = i) \\ &= \sum_{i=1}^{n} \Pr(X_m = j \mid X_k = i) \Pr(X_k = i) \\ &= [\pi_k P_{k,m}]_j. \end{aligned}$$

Properties of Transition Matrices cont.

- Properties of the transition matrices cont.:
  - **Semigroup property**: For any $k \leq m \leq q$, we have:

    $$P_{k,q} = P_{k,m} P_{m,q}.$$

    To show this, let i, j being fixed. Then, we have

    $$\begin{aligned} P_{k,q}(i,j) &= \Pr(X_q = j \mid X_k = i) \\ &= \sum_{\ell=1}^{n} \Pr(X_q = j, X_m = \ell \mid X_k = i) \\ &= \sum_{\ell=1}^{n} \Pr(X_q = j \mid X_m = \ell, X_k = i) \Pr(X_m = \ell \mid X_k = i) \\ \text{(by Markov property)} &= \sum_{\ell=1}^{n} \Pr(X_q = j \mid X_m = \ell) \Pr(X_m = \ell \mid X_k = i) \\ &= \sum_{\ell=1}^{n} P_{k,m}(i,\ell) P_{m,q}(\ell,j) \\ &= [P_{k,m} P_{m,q}]_{i,j}. \end{aligned}$$

    This property is widely known as *Chapman-Kolmogorov* equation.

  - For DS Markov chains, the second property, and the Chapman-Kolmogorov property imply:

    $$\pi_k = \pi_1 P_{1,k} = \pi_1 P_{1,2} P_{2,k} = \cdots = \pi_1 P_{1,2} P_{2,3} \cdots P_{k-1,k}.$$

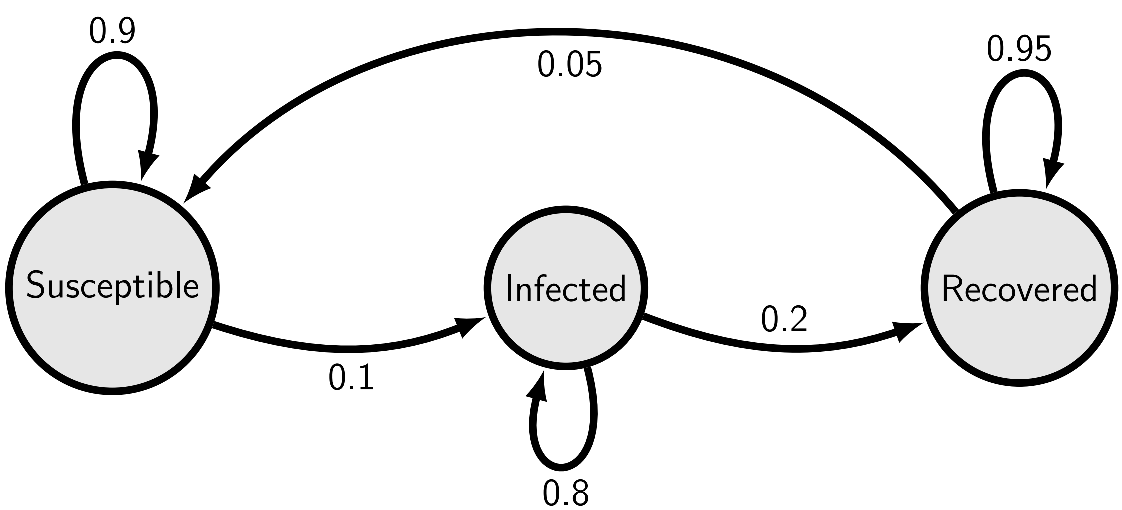
Homogeneous Markov Chains

Definition: We say that a Markov chain $\{X_k\}$ is (time-)homogeneous if $P_{1,2} = P_{m,m+1}$ does not depend on m.

- Denote $P := P_{m,m+1}$. P is called the one-step transition matrix of the underlying Homogeneous Markov chain.
- P is a row-stochastic matrix.
- For Homogeneous Markov chains, we have $P_{m,n} = P^{n-m}$.
- Distribution of X_k is given by $\pi_k = \pi_{k-1}P = \pi_0 P^k$.
- With the abuse of notation, for a Homogeneous Markov chain P is also called (one-step) transition probability matrix (TPM).
- For homogeneous markov chains, the initial distribution and the one-step TPM completely specifies the random process.

Graph-Theoretic Interpretation

- Consider a homogeneoys MC on state space S with TPM P.
- Consider a directed weighted graph $G = (V, E, P)$ where
  - $V = S = \{1, \ldots, n\}$,
  - $E = \{(i, j) \mid P_{ij} > 0\}$, and
  - P_{ij} is the weight of edge i, j.
- Then, the MC can be viewed as a random walk on this weighted graph.
- Example: infectious disease model. Determine the TPM, and simulate the MC.



Classification of States

We introduce a few basic definitions.

- An m-step walk on a graph $G = (V, E)$ is an ordered string of nodes $i_0, i_1, \ldots, i_m$ such that $(i_{k-1}, i_k) \in E$ for all $k \in \{1, 2, \ldots, m\}$.

- A path is a walk where no two nodes are repeated. A cycle is a walk where the first and last nodes are identical and no other node is repeated.

- Let $G = (V, E, P)$ be the graph associated with a MC with TPM P. A state j is accessible from state i, denoted $i \to j$ if there is a walk in the graph from node i to node j.

- In other words, there exists nodes $i_1, i_2, \ldots, i_k$ such that $(i, i_1) \in E, (i_1, i_2) \in E, \ldots, (i_k, j) \in E$. The length of this walk is $k + 1$.

- Equivalently, $P_{i,i_1} > 0, P_{i_1,i_2} > 0, \ldots, P_{i_k,j} > 0$. Thus, $[P^{k+1}]_{i,j} > 0$.

- Two states i and j communicate if $i \to j$ and $j \to i$. This is denoted by $i \leftrightarrow j$.

- Naturally, if $i \leftrightarrow j, j \leftrightarrow k$, then $i \leftrightarrow k$.

- A subset of states $C \subseteq V$ is a communicating class if

  1. $i \in C, j \in C \implies i \leftrightarrow j$, and
  2. $i \in C, j \notin C \implies i \not\leftrightarrow j$.

  The set of states can be partitioned into distinct communicating classes. Each state belongs to exactly one communicating class.

Definition: A state i is called recurrent if $i \to j \implies j \to i$. A state is transient if it is not recurrent.

If a state is recurrent, there is no path to a state from which there is no return.

Classification of States Cont.

Theorem: In a given communicating class, either all states are recurrent or all states are transient. Furthermore, in a finite-state MC, there is at least one recurrent communicating class.

- A matrix P is **irreducible** if for any i, j, $[P^{k_{ij}}]_{ij} > 0$ for some $k_{ij} \geq 1$. In other words, $i \leftrightarrow j$ for every pair of states i, j.
- Graph theoretic interpretation: P is irreducible if there is a directed path between any two nodes on the graph.
- In this case, there is a single communicating class which is recurrent.

Definition: The period γ_i of a state i, to be greatest common divisor (gcd) of $\texttt{gcd}(k \mid [P^k]_{ii} > 0)$.

- Graph theoretic interpretation: gcd of lengths of all paths from i to itself.
- Example: for $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, determine the period of its states.
- All states in the same communicating class have the same period.
- We say that a non-negative matrix P is aperiodic if $\gamma_i = 1$ for all i.
- A (homoegeneous) Markov chain with the transition matrix P is said to be irreducible (aperiodic) if P is irreducible (aperiodic).

Stationary and Limiting Distribution of a Markov Chain

- Let $P \in \mathbb{R}^{n \times n}$ be the single-step transition probability matrix of a homogeneous markov chain.
- Let π_0 be the distribution of initial state X_0. It follows that $\pi_n = \pi_0 P^n$.

A vector $\pi^\star \in \mathbb{R}^{1 \times n}$ is called invariant/stationary/steady-state distribution of the markov chain with TPM P if

- $\pi^\star$ is a probability vector, i.e., $\pi^\star(i) \geq 0, \sum_{i=1}^{n} \pi^\star(i) = 1$, and
- $\pi^\star = \pi^\star P$.

If $\pi_k = \pi^\star$ for some k, then $\pi_m = \pi^\star$ for all $m \geq k$.

Fundamental questions in the theory of (homogeneous) Markov chains:

- *Existence and Uniqueness*: When does $\pi^\star$ exist? Is it unique?
- *Ergodicity*: When unique, under what conditions, $\pi_k \to \pi^*$?
- *Mixing time*: How fast does it converge to π^*?
- *Occupation Probability*: How often do we spend time on a given state?

We know that the TPM P satisfies the following properties.

- P is non-negative.
- P is row-stochastic, which implies that all eigenvalues reisde on or within the unit circle, and 1 is an eigenvalue.
- Note: $\pi^\star$ is nothing but the left eigenvector of eigenvalue 1.
- Thus, existence and uniqueness of stationary distribution is equivalent to showing existence and uniqueness of a non-negative left eigenvector of the TPM.

Example

- Let $P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$.
- Solving for $(u, v)P = (u, v)$ with $v = 1 - u$, we get $u = \frac{2}{5}$ and $v = \frac{3}{5}$. Is this unique?
- What about $P = I$?

Linear Algebra Viewpoint

- Give a matrix $A \in \mathbb{R}^{n\times n}$, we define its spectral radius as

$$\rho(A) := \{|\lambda| : \lambda \quad \text{is an eigenvalue of A}\}.$$

- An eigenvalue of A is called semi-simple if its algebraic multiplicity $=$ its geometric multiplicity.
  - algebraic multiplicity: number of times the eigenvalue appears as root of the characteristic equation
  - geometric multiplicity: number of linearly independent eigenvectors associated with this eigenvalue

  It is called simple when both multiplicities are equal to 1.

- The matrix A is called
  - semi-convergent if $\lim_{k\to\infty} A^k$ exists, and
  - convergent if it is semi-convergent and $\lim_{k\to\infty} A^k = 0_{n\times n}$.

Theorem 1. *A matrix $A \in \mathbb{R}^{n\times n}$ is*

- *convergent if and only if $\rho(A) < 1$, and*
- *is semi-convergent if and only if either (i) $\rho(A) < 1$ or (ii) 1 is a semi-simple eigenvalue and all other eigenvalues have magnitude strictly less than 1.*

Perron-Frobenius Theorem

A matrix $A \in \mathbb{R}^{n \times n}$ is

- **non-negative** if $A_{ij} \geq 0$ for all i, j.
- **irreducible** if $\sum_{k=0}^{n-1} A^k$ is positive, i.e., all entries are strictly larger than 0.
- **primitive** if there exists some $\bar{k}$ such that $A^{\bar{k}} > 0$.
- **positive** if $A_{ij} > 0$ for all i, j.

Theorem 2. *A matrix $A \in \mathbb{R}^{n \times n}$ be a non-negative matrix.*

- *Then, there exists a real eigenvalue $\lambda \geq |\mu| \geq 0$ where μ is any other eigenvalue. The left and right eigenvectors associated with A are non-negative.*
- *If A is irreducible, $\lambda \geq |\mu|$ is strictly positive and simple. The left and right eigenvectors associated with A are unique and positive.*
- *If A is primitive, $\lambda > |\mu|$. The left and right eigenvectors associated with A are unique and positive.*

Let $P \in \mathbb{R}^{n \times n}$ be the single-step transition probability matrix of a homogeneous markov chain.

- Is P **non-negative**?
- When is P **irreducible**? Does it imply P is semi-convergent?
- When is P **primitive**? Does it imply P is semi-convergent?
- When is P **positive**? Does it imply P is semi-convergent?

Case 1: MC with Single Recurrent Class

- In this case, TPM P is irreducible. (why?)
- From PF Theorem, largest eigenvalue 1 is simple, and left eigenvector is unique and positive. In other words, $\pi^\star$ exists and is unique.
- However, if the states have period $d > 1$, then there are d eigenvalues on the unit circle that are equally spaced. Such a matrix is not primitive, and hence not semi-convergent.
- When the states are aperiodic (i.e., period $d = 1$), then it is primitive, and P is semi-convergent.
- A MC which is both irreducible and aperiodic is called **ergodic**.
- We can show that

$$\lim_{k\to\infty} P^k = (1)^k v w^\top = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} w_1 & w_2 \ldots w_n \end{bmatrix} = \begin{bmatrix} w_1 & w_2 \ldots w_n \\ w_1 & w_2 \ldots w_n \\ & \vdots \\ w_1 & w_2 \ldots w_n \end{bmatrix} =: P_\infty,$$

  where v is the right eigenvector and w is the left eigenvector of 1. Note that $w = \pi^\star$.
- In addition, for any initial distribution π_0, we have

$$\lim_{k\to\infty} \pi_k = \lim_{k\to\infty} \pi_0 P^k = \pi_0 P_\infty = \pi^\star.$$

Case 2: MC with one Recurrent Class and some Transient States

- Such a markov chain is called a **unichain**. The TPM P is no longer irreducible and can be partitioned as

$$P = \begin{array}{c} m_1 \\ n-m_1 \end{array} \left(\begin{array}{c|c} \overset{m_1}{P_{RR}} & \overset{n-m_1}{0} \\ \hline P_{TR} & P_{TT} \end{array} \right)$$

  where the first m_1 states belong to the recurrent class, and the remaining states being transient.

- Though P is not irreducible, the submatrix P_{RR} is irreducible which has a unique stationary distribution $\pi_R^\star \in \mathbb{R}^{1\times m_1}$.

- Then, the vector $\pi^\star = [\pi_R^\star \quad 0_{1\times n-m_1}]$ is the unique stationary distribution of P.

- If the states in the recurrent class is apeiodic, then P is semi-convergent. Such a MC is called an **ergodic unichain**.

The following result characterizes the uniqueness and limiting behavior of the stationary distribution.

> **Theorem 3.** *Consider a finite-state homogeneous MC.*
>
> - *A MC has a unique stationary distribution $\pi^\star$ if and only if it is a unichain (i.e., it has a single recurrent class)*
> - *Let $\lim_{k\to\infty} P^k = P_\infty$. Each row of P_∞ is identical and equal to $\pi^\star$ if and only if MC is an ergodic unichain (unichain with an aperiodic recurrent class).*

Case 3: MC with Multiple Recurrent Classes

- The TPM P can be partitioned as

$$P = \left(\begin{array}{c:c:c:c} \overset{m_1}{P_{R_1}} & \overset{m_2}{0} & \overset{m_3}{0} & \overset{n-\sum m_i}{0} \\ \hdashline 0 & P_{R_2} & 0 & 0 \\ \hdashline 0 & 0 & P_{R_3} & 0 \\ \hdashline P_{TR1} & P_{TR2} & P_{TR3} & P_{TT} \end{array}\right)$$

  where the first m_1 states belong to the first recurrent class, and so on.

- For each recurrent class, the corresponding submatrix P_{R_i} is irreducible which has a unique stationary distribution $\pi_i^\star \in \mathbb{R}^{1\times m_i}$.

- Then, the vector $[0 \quad \pi_i^\star \quad \ldots 0]$ is a stationary distribution of P. Thus, stationary distribution is not unique.

- Every recurrent class adds one multiplicity to the eigenvalue 1.

- P is semi-convergent only when every recurrent class is aperiodic. In this case, $\lim_{k\to\infty} P^k = P_\infty$, but P_∞ has non-identical rows. However, rows corresponding to states in the same recurrent class are identical.

Ergodic Property

- Let the initial state $X_0 = i$.
- $T_i := \inf\{k \geq 1 \mid X_k = i\}$ (first passage time): smallest time index at which the state takes value i
- $f_i := \mathbb{P}(T_i < \infty)$: return probability
- $m_i := \mathbb{E}[T_i]$: mean return time
- $\nu_i := \sum_{k=0}^{\infty} \mathbf{1}_{\{X_k=i\}}$ number of visits to i starting from i.
- State i is recurrent if and only if $f_i = 1$. State i is transient if and only if $f_i < 1$.

Theorem: If state i is recurrent, then $\mathbb{E}[\nu_i] = \infty$. If state i is transient, then $\mathbb{E}[\nu_i] < \infty$.

Theorem: Suppose the TPM is irreducible and let $\pi^\star$ be the unique stationary distribution. Then, $m_i = \frac{1}{\pi^\star(i)}$ for all states i.

Theorem: Suppose the TPM is irreducible and aperiodic (i.e., ergodic) with the stationary distribution $\pi^\star$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{X_k=i\}} = \pi^\star(i) \quad \text{almost surely.}$$

Application: Page-Rank Algorithm

- Original idea of Google search ranking: Model a browsing person as a random walker over the graph of internet!
- Let $G = (V, E)$ where $d =$ number of webpages and there is a node for each webpage.
- $(i, j) \in E$ if i has a link to j.
- Then a person can be *modeled* as a random walker on G where

$$P_{ij} = \begin{cases} \frac{1}{d_i} & j \in \mathcal{N}_i \\ 0 & \text{otherwise.} \end{cases}$$

- Problem with this? Corresponding Markov chain is not irreducible.
- Now let us add a small reset probability, i.e., consider a Markov chain with one-step transition matrix

$$\hat{P} = (1 - a)P + aJ,$$

  where $a \in (0, 1)$ is a small reset parameter and J is the $d \times d$ matrix with all elements being $1/d$.
- Then a Markov chain with the transition matrix $\hat{P}$ is irreducible and aperiodic (why?).
- Therefore, it is ergodic, has a unique stationary distribution π^*, and $\pi_k \to \pi^*$ as $k \to \infty$.
- More importantly *average visit percentage of state (webpage) i by time k*$\to \pi_i^*$!
- Therefore, webpage i is superios to j if $\pi_i^* > \pi_j^*$.
- How does Google find π^*?

Vector-valued Random Process

A random process $X = \{X_t\}_{t \in T}$ may be such that each X_t is a random vector taking values in $\mathbb{R}^n$. Then,

(a) **Mean function**:

$$\mu_X(t) := \mathbb{E}[X_t] \in \mathbb{R}^n$$

(b) **Autocorrelation function**:

$$R_X(t_1, t_2) := \mathbb{E}[X_{t_1} X_{t_2}^\top] \in \mathbb{R}^{n \times n}$$

(c) **Autocovariance function**:

$$C_X(t_1, t_2) := \mathsf{cov}(X_{t_1}, X_{t_2}) \in \mathbb{R}^{n \times n}.$$

For WSS, every element of $C_X(t_1, t_2)$ should only depend on $t_2 - t_1$.

Other Class of Processes

- A stochastic process $\{X_t\}_{t \in T}$ is called a Gaussian Process if for every finite set of indices $t_1, t_2, \ldots, t_k$, the collection of random variables $X_{t_1}, X_{t_1}, \ldots, X_{t_k}$ is jointly Gaussian.

- A stochastic process which is both Gaussian and Markov is called Gauss-Markov Process.

- A stochastic process $\{X_t\}_{t \in T}$ is said to have **independent increments** if for every finite set of indices $t_1, t_2, \ldots, t_k$, the collection of random variables $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_k} - X_{t_{k-1}}$ are mutually indepdenent.

- The increments are stationary if $X_{t_2} - X_{t_1}$ and $X_{t_2+s} - X_{t_1+s}$ have the same distribution irrespective of the value of s.

- **Brownian Motion/Wiener Process:** A stochastic process $\{X_t\}_{t \in T}$ is a **Wiener Process** if

  1. $X_0 = 0$,
  2. the process has stationary and independent increments,
  3. $X_t - X_s \sim \mathcal{N}(0, \sigma^2(t-s))$,
  4. the sample paths are continuous with probability 1.

  For a Wiener process, one can show that the sample paths are not differentiable by showing

$$\lim_{\Delta \to 0} \texttt{var}\left[\frac{X(t+\Delta) - X(t)}{\Delta}\right] = \frac{\sigma^2}{\Delta} \to \infty.$$

Dynamical System

- Deterministic discrete-time dynamical system in state-space form is given by:
$$x_{k+1} = f_k(x_k, u_k), \quad k = 0, 1, \ldots,$$
where $x_k \in \mathbb{R}^n$ is the state at time k and $u_k \in \mathbb{R}^m$ is the input at time k.
- State variable: summarizes past information such that if we know the state at time k and the input for all $t \geq k$, then we can completely determine the future states.
- In other words, if we know the current state, we do not need to store past states and inputs to predict the future.
- If $f_k = f$ for all k, the system is time-invariant.

Stochastic Dynamical System

- **Stochastic Model:** the future state is uncertain even if the current state and input are known. There are two ways of representing such a system. Both are equivalent under reasonable assumptions.

- State-space form:

$$x_{k+1} = f_k(x_k, u_k, w_k), \quad k = 0, 1, \dots,$$

  where $w_k \in \mathbb{R}^w$ is a random variable/noise/disturbace which is not under our control (unlike input u).

- Note that $\{w_1, w_2, \dots,\}$ is a discrete-time random process, as is $\{x_1, x_2, \dots,\}$.

- Example: $x_{k+1} = ax_k + w_k$ where $w_k \in \mathcal{N}(c, 1)$ and $x_0 = 5$. What will the trajectories look like for different values of a and c? What is the distribution of x_k as $k \to \infty$? Is this process Markovian?

Stochastic Linear System

A stochastic linear system is formally defined as

$$x_{k+1} = A_k x_k + B_k u_k + w_k.$$

Problem: recursively determine the mean and variance of x_k given that $\mathbb{E}[w_k] = 0$, $\texttt{var}(w_k) = Q$ and x_0 is known.

Representation via Transition Kernel

- Recall the state-space form: $x_{k+1} = f_k(x_k, u_k, w_k), \quad k = 0, 1, \ldots.$
- Here, the distribution of x_{k+1} can be found in terms of the function f_k and indirectly, as a function of *basic random variables* $(x_0, w_0, \ldots, w_k)$.
- The alternative approach is to directly specify the distribution of x_{k+1} instead of relying on the function f_k. In particular, the **conditional distribution** of X_{k+1} given x_k and u_k is specified for all values of x_k and u_k.
- For the dynamical system to be Markovian, we need to show that for every Borel subset A and for all k,

$$\mathbb{P}(X_{k+1} \in A | x_0, u_0, x_1, u_1, \ldots, x_k, u_k) = \mathbb{P}(X_{k+1} \in A | x_k, u_k).$$

- Is the above property always true?

Observation Model

- In many instances, the states can not be directly measured.
- Instead, we observe "output" quantities that depend on the state as

$$y_k = g_k(x_k, v_k),$$

  where v_k is a random variable termed "measurement noise."
- Alternatively, the conditional distribution of y_k given x_k is specified.
- In case of a linear system, $y_k = C_k x_k + v_k$.
- One problem of significant interest is to infer or estimate the state x_k given the measured / output quantities y_k in an online and recursive manner.
- Module C will tackle this issue.