CE 273
Markov Decision Processes

Lecture 1

Introduction

Lecture Outline

Lecture Outline

Course Overview

What will I learn from this course?

- The primary objective of this course is to study sequential decision making in a stochastic environment.
- You'll be exposed to a variety of situations which will help you model an engineering problem as a Markov Decision Process (MDP).
- The focus of the course will also be to study the tools and techniques to solve these MDPs.

Is is course useful to my research?

- There are applications of MDPs in several fields. If you think you'll deal with situations where repeated choice making and optimization is involved, it is very likely that this course will be useful.
- We will look a few examples in this class (some from transportation) which will hopefully illustrate the point.

Course Overview

Frequently Asked Questions (FAQs)

Are MDPs same as Reinforcement learning?

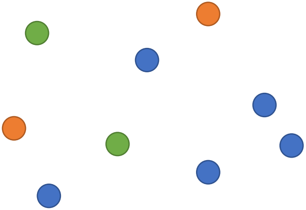
- Not exactly. MDPs assume that the inputs (system dynamics) are fully known. Think of it as a kind of an optimization framework with some uncertainty.
- Reinforcement learning (or unknown MDPs) on the other hand deals with instances in which the dynamics are inferred while simultaneously searching for the optimal solution.
- A few lectures in the latter part of the course are dedicated to introductory reinforcement learning.
- MDPs are also different from multi-stage stochastic optimization models and we'll discuss these differences later.

All the World's a State

The Big Picture

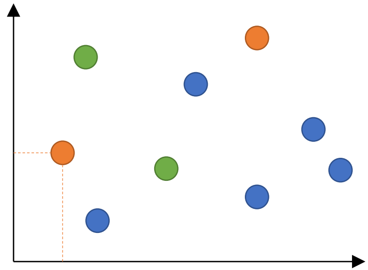
A key component of an MDP is the concept of a *state*, which is loosely a mathematical representation of the system being studied.

What is the state of the following system?



- A bunch of 9 dots
- 5 blue dots, 2 green dots, 2 orange dots

How about now?



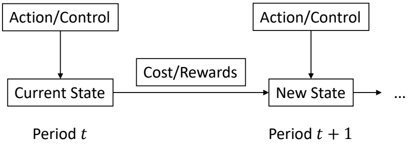
Perhaps a collection of points $(x_1, y_1), \ldots, (x_9, y_9)$.

State of a system is not necessarily unique. Its definition is context specific and depends on what actions you can take and what your objective is.

All the World's a State

The Big Picture

At a given state, the decision maker can take an action/control which takes the system to another state and the process is repeated (either for a fixed number of steps or indefinitely).



In the process, the decision maker may receive a reward or incur some cost (which depends on the current state and action).

Before choosing the action, one cannot predict the future state with certainty. State transitions are usually stochastic and are a function of the actions.

The end goal is to optimize some function of the sequence of cost/reward.

All the World's a State

The Big Picture

Let's take a tour of some problems that can be modeled using MDPs.

For each of these examples, identify the following components:

- States
- Actions
- Transitions/Dynamics
- Objective and Costs/Reward

Queuing

Adaptive signal control

Course Overview
Applications

Dynamic pricing

Course Overview

Applications

Inventory management

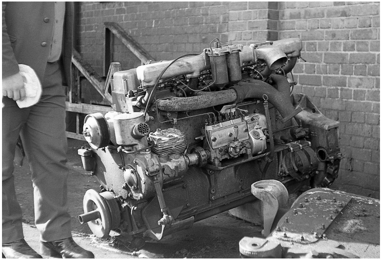
Course Overview

Applications

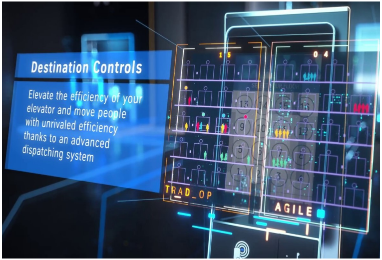
Redistribution of bicylce sharing systems

http://tiny.cc/x9qysz

Periodic maintenance problems

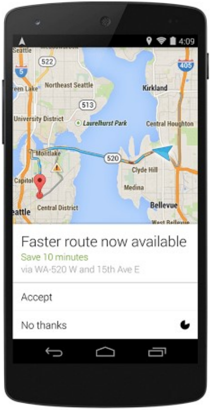


Intelligent elevators



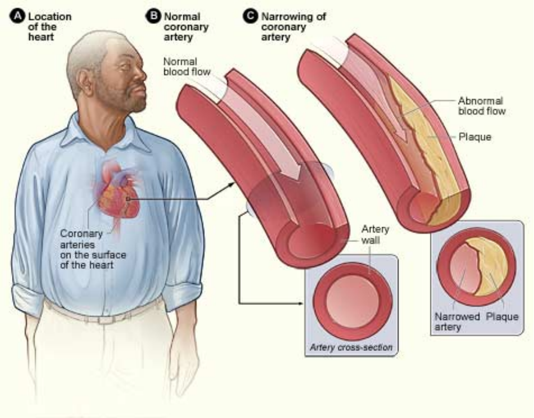
https://www.youtube.com/watch?v=T6gzm_ifzg8

Route planning under uncertainty

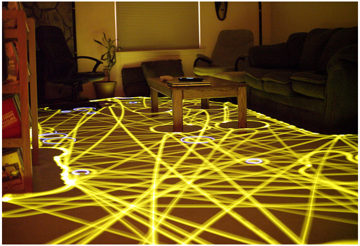


Investment management

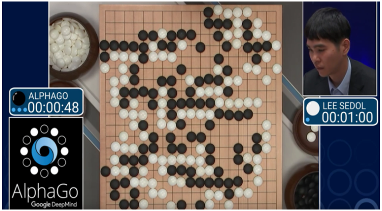
Disease diagnosis and treatment



Autonomous systems



Game playing



https://www.youtube.com/watch?v=WXuK6gekU1Y

Markov Chains

Definition

A stochastic process is a collection of random variables $\{X(\tau), \tau \in T\}$.

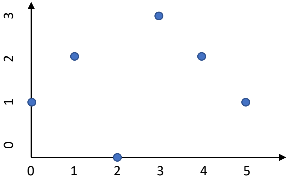
- The support of these random variables is called the state space S. (Can be discrete or continuous.)
- T usually represents time and can be discrete or continuous.

Definition

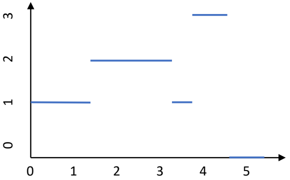
A sample path is a realization of the stochastic process $\{x(\tau), \tau \in T\}$.

Definition

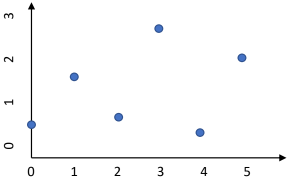
The set of all sample paths or trajectories of the stochastic process is called the sample space.



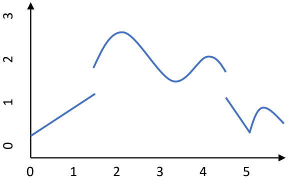
Discrete S, Discrete T



Discrete S, Continuous T



Continuous S, Discrete T



Continuous S, Continuous T

Discrete state space or time does not imply that S and T must be subsets of $\mathbb{Z}$! They must simply be countable.

- For instance, imagine a system of queues at the airport. The state could be a vector representing the number of individuals in each queue and yet the state space is countable.
- In many cases, continuous time systems can be modeled as discrete time processes via a method called *uniformization*. In this method, time is re-indexed by tracking events involving state transitions.

We will deal with discrete state and time processes for most part of the course and denote the stochastic process as $\{X_n, n \geq 0\}$.

Consider a discrete-time stochastic process $\{X_n, n \geq 0\}$ with a countable state space $S = \{0, 1, 2, \ldots\}$.

- We will focus on stochastic processes in which the present state contains all the information required to predict the future.
- In other words, the future is independent of the past, or the current state is a 'sufficient statistic' of the future.

This assumption is also called the **Markov property** and such stochastic processes are called Discrete-Time Markov Chains (DMTCs).

Mathematically,

Definition (Markov Property)

A stochastic process $\{X_n, n \geq 0\}$ with a countable state space S is called a DTMC if $\forall\, n \geq 0, i, j \in S$,

$$\mathbb{P}\big[X_{n+1} = j | X_n = i, X_{n-1}, X_{n-1}, \ldots, X_0\big] = \mathbb{P}\big[X_{n+1} = j | X_n = i\big]$$

Definition (Time Homogeneity)

A DTMC $\{X_n, n \geq 0\}$ is said to be time homogeneous if $\forall\, n \geq 0, i, j \in S$,

$$\mathbb{P}\big[X_{n+1} = j | X_n = i\big] = p_{ij}$$

i.e., RHS does not depend on n or $p_{ij}(n) = p_{ij}\, \forall\, n \geq 0$

The probability with which the system moves from i to j, p_{ij}, is called the **transition probability** and the matrix of p_{ij} values is called the **one-step transition probability matrix**.

$$P = \left[p_{ij}\right]_{|S|\times|S|}$$

Note that P can have countably infinite rows and columns.

Definition (Stochastic Matrix)

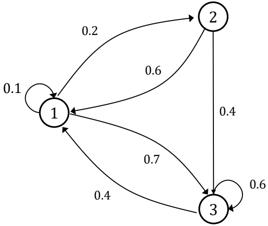
A square matrix $P = \left[p_{ij}\right]_{|S|\times|S|}$ is called right stochastic if

1. $p_{ij} \geq 0 \,\forall\, i, j \in S$
2. $\sum_{j\in S} p_{ij} = 1 \,\forall\, i \in S$

Transition matrices of a Markov chain are right stochastic matrices.

The transition probability matrix can also be visualized as a directed graph in which the states are nodes and an arc (i, j) exists only if $p_{ij} > 0$.

$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.6 & 0 & 0.4 \\ 0.4 & 0 & 0.6 \end{bmatrix} \end{array}$$



http://setosa.io/ev/markov-chains/

The P matrix alone doesn't fully describe a DTMC. We'd also need to know the initial distribution.

$$a_i = \mathbb{P}\left[X_0 = i\right] \forall\, i \in S$$

Let a be row vector of a_i's. A Markov chain can thus be fully specified using (S, P, a).

One can use the Markov property to estimate joint distributions and mass functions. For example. suppose $a = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$ and

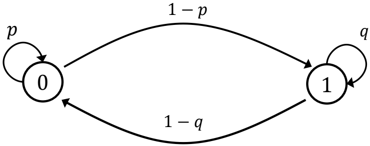
$$P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.2 & 0.3 & 0.3 \\ 0.5 & 0 & 0.5 & 0 \\ 0.6 & 0.2 & 0.1 & 0.1 \end{bmatrix} \end{array}$$

Compute

- $\mathbb{P}\left[X_3 = 4, X_2 = 1, X_1 = 3, X_0 = 1\right]$
- $\mathbb{P}\left[X_3 = 4, X_2 = 1, X_1 = 3\right]$

Consider a single parking spot. Suppose that if it's occupied in time interval n, then it will remain occupied in time interval $n+1$ with probability q.

Likewise, assume that if the spot is empty at time interval n, it remains empty in $n+1$ with probability p.



$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \end{array}$$

Now imagine two spots and parking events that are independent of each other. Let $\{X_n, n \geq 0\}$ be the number of empty spots. The state space is $S = \{0, 1, 2\}$.

The associated transition probability matrix is

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} \quad 0 \quad & \quad\quad 1 \quad\quad & \quad 2 \quad \end{array} \\ \begin{bmatrix} p^2 & 2p(1-p) & (1-p)^2 \\ p(1-q) & pq + (1-p)(1-q) & q(1-p) \\ (1-q)^2 & 2q(1-q) & q^2 \end{bmatrix} \end{array}$$

Imagine you want to shuffle a set of playing cards. For simplicity, assume that you just have 3 ace cards $\clubsuit$, $\heartsuit$, and $\spadesuit$. The state space consists of 6 permutations.

Here is a deterministic shuffling strategy: Take the last card and place it at the top of the deck. The transition matrix for this method is

$$\begin{array}{c} \\ \clubsuit\heartsuit\spadesuit \\ \clubsuit\spadesuit\heartsuit \\ \heartsuit\spadesuit\clubsuit \\ \heartsuit\clubsuit\spadesuit \\ \spadesuit\clubsuit\heartsuit \\ \spadesuit\heartsuit\clubsuit \end{array} \begin{array}{c} \begin{array}{cccccc} \clubsuit\heartsuit\spadesuit & \clubsuit\spadesuit\heartsuit & \heartsuit\spadesuit\clubsuit & \heartsuit\clubsuit\spadesuit & \spadesuit\clubsuit\heartsuit & \spadesuit\heartsuit\clubsuit \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

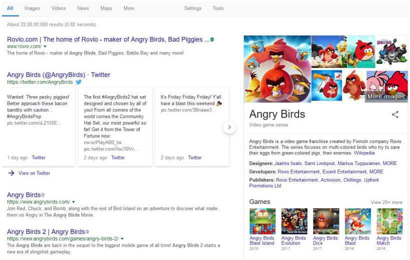
Is this a good shuffling strategy?

Instead, suppose that the last card is at the top of the deck or inserted between the first and second cards with equal probability. The corresponding transition matrix is

	♣♡♠	♣♠♡	♡♠♣	♡♣♠	♠♣♡	♠♡♣
♣♡♠	0	0.5	0	0	0.5	0
♣♠♡	0.5	0	0	0.5	0	0
♡♠♣	0.5	0	0	0.5	0	0
♡♣♠	0	0	0.5	0	0	0.5
♠♣♡	0	0	0.5	0	0	0.5
♠♡♣	0	0.5	0	0	0.5	0

Is this a better shuffling strategy?

Perhaps the most popular application of Markov chains is Google's approach to ranking web pages. Imagine there are N webpages.



Let c_{ij} be 1 if page i has a link pointing to page j and 0 otherwise. Let c_i be the total number of other pages i links to.

Suppose time step n represents nth click. Then, a user on page i will visit page j with probability c_{ij}/c_i.

However, if $c_i = 0$, then the user is assumed to choose one of the N pages with equal probability $1/N$.

As an example, suppose there are four web pages with connections as shown in the C matrix.

$$C = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \qquad P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} \end{array}$$

The average amount of time spent by the Markov chain in each of these states is used to determine the page ranking. We'll revisit this example in greater detail in Lecture 3.

Melodies in music are usually accompanied by groups of notes called chords. For example, a piece in a major scale would typically use the following seven chords: I, ii, iii, IV, V, vi and vii^o.

One can use the chord progressions to find patterns across eras or composers. For example, the next few slides show the frequencies of chord transitions for four composers spanning different eras.

---

*The data set used is very small but is still informative. *Source:* http://lib.bsu.edu/beneficencepress/mathexchange/10-01/index.html

Palestrina (1547-1580)

	I	ii	iii	IV	V	vi	vii^O
I	0	0.15	0.13	0.28	0.14	0.22	0.08
ii	0.08	0	0.15	0.13	0.28	0.14	0.22
iii	0.22	0.08	0	0.15	0.13	0.28	0.14
IV	0.14	0.22	0.08	0	0.15	0.13	0.28
V	0.28	0.14	0.22	0.08	0	0.15	0.13
vi	0.13	0.28	0.14	0.22	0.08	0	0.15
vii^O	0.15	0.13	0.28	0.14	0.22	0.08	0

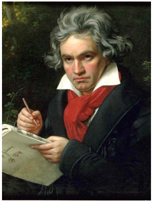
Bach (1685-1750)

	I	ii	iii	IV	V	vi	vii^O
I	0	0.15	0.01	0.28	0.41	0.09	0.06
ii	0.01	0	0	0	0.71	0.01	0.25
iii	0.03	0.03	0	0.52	0.06	0.32	0.03
IV	0.22	0.13	0	0	0.39	0.02	0.23
V	0.82	0.01	0	0.07	0	0.09	0
vi	0.15	0.29	0.05	0.11	0.32	0	0.09
vii^O	0.91	0	0.01	0.02	0.04	0.03	0

	I	ii	iii	IV	V	vi	vii^O
I	0	0.13	0	0.15	0.62	0.05	0.05
ii	0.49	0	0.01	0	0.40	0.01	0.09
iii	0.67	0	0	0	0	0.33	0
IV	0.64	0.14	0	0	0.15	0	0.07
V	0.94	0	0	0.01	0	0.04	0.01
vi	0.11	0.51	0	0.14	0.20	0	0.04
vii^O	0.82	0	0.01	0.01	0.16	0	0


Mozart (1756-1791)

	I	ii	iii	IV	V	vi	vii^O
I	0	0.10	0.01	0.13	0.52	0.02	0.22
ii	0.06	0	0.02	0	0.87	0	0.05
iii	0	0	0	0	0.67	0.33	0
IV	0.33	0.03	0.07	0	0.40	0.03	0.13
V	0.56	0.22	0.01	0.04	0	0.07	0.11
vi	0.06	0.44	0	0.06	0.11	0	0.33
vii^O	0.80	0	0	0.03	0.17	0	0


Beethoven (1770-1827)

Let $\{Z_n, n \geq 1\}$ be a sequence of iid random variables with pmf

$$\alpha_k = \mathbb{P}\big[Z_n = k\big], k \in \mathbb{Z}$$

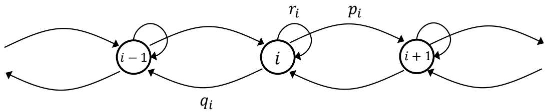
Define a stochastic process $\{X_n, n \geq 0\}$ as follows:

$$X_0 = 0, X_n = \sum_{k=1}^{n} Z_k, n \geq 1$$

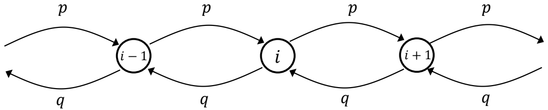
The above process is a DTMC with transition probabilities $p_{ij} = \alpha_{j-i}, \forall\, i, j \in S$ (Why?) and is called a random walk.

Most common variant of the random walk allows steps of size $0, 1$, and -1 on a 1D lattice.

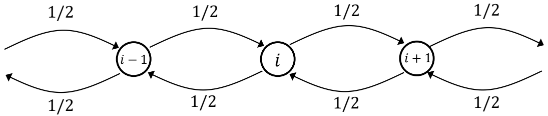
State-Dependent Random Walk:



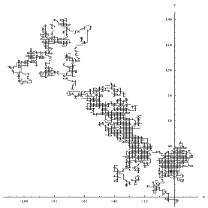
Simple Random Walk:

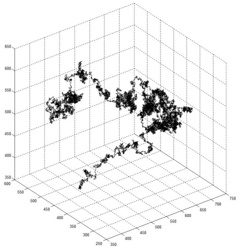


Consider the following symmetric simple random walk. Is probability of returning to 0 equal to 1?

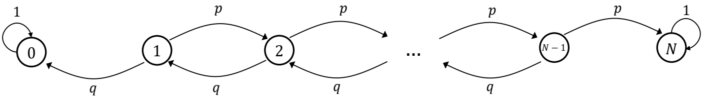


How about the 2D and 3D version?





Imagine two gamblers A and B who have ₹N. Assume that they repeatedly toss a coin and if it turns heads (whose probability is p), A wins a rupee from B and if it is tails (whose probability is q), A gives B a rupee.

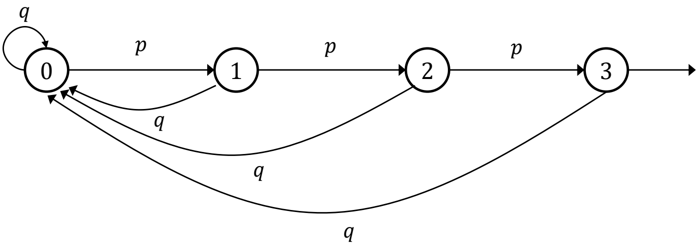


This is also called as simple random walk with absorbing barriers.

What is the probability of being broke (i.e., reaching state 0)? What if the second person had infinite amount?

Suppose a coin is tossed repeatedly and the probability of seeing heads and tails is p and q. Assume a player wins ₹1 every time it is heads and looses their entire winnings if it is tails.

Let X_n represent the player's cash after n tosses. The DTMC can be represented using the following transition diagram.



Consider a Tesla manufacturing plant which produces one car each day. Demand for cars, however, can occur in batches. Let $\{Y_n, n \geq 1\}$ be the sequence of iid demands on different days with a pmf

$$\alpha_k = \mathbb{P}\big[Y_n = k\big], k \in \mathbb{Z}^+$$

Let X_n represent the number of cars in the warehouse after the demand for a particular day is met. Assuming no back orders and that the production occurs before sales,

- Show that the Markov property holds.
- Construct the transition probability matrix.

Suppose $X_n = i$. Then for $0 < j \leq i{+}1$, $\mathbb{P}\big[X_{n+1} = j | X_n = i, X_{n-1}, \ldots, X_0\big]$

$$
\begin{aligned}
&= \mathbb{P}\big[\max\{X_n + 1 - Y_{n+1}, 0\} = j | X_n = i, X_{n-1}, \ldots, X_0\big] \\
&= \mathbb{P}\big[X_n + 1 - Y_{n+1} = j | X_n = i, X_{n-1}, \ldots, X_0\big] \\
&= \alpha_{i-j+1}
\end{aligned}
$$

If $j = 0$, $\mathbb{P}\big[X_{n+1} = 0 | X_n = i, X_{n-1}, \ldots, X_0\big]$

$$
\begin{aligned}
&= \mathbb{P}\big[\max\{X_n + 1 - Y_{n+1}, 0\} = 0 | X_n = i, X_{n-1}, \ldots, X_0\big] \\
&= \mathbb{P}\big[Y_{n+1} \geq i + 1 | X_n = i, X_{n-1}, \ldots, X_0\big] \\
&= \sum_{k=i+1}^{\infty} \alpha_k = \beta_i
\end{aligned}
$$

Markov Chains

Example 7: Inventory Modeling

The transition probability matrix can thus be written as

$$\begin{array}{c c} & \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & \dots \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} & \begin{bmatrix} \beta_0 & \alpha_0 & 0 & 0 & 0 & \dots \\ \beta_1 & \alpha_1 & \alpha_0 & 0 & 0 & \dots \\ \beta_2 & \alpha_2 & \alpha_1 & \alpha_0 & 0 & \dots \\ \beta_3 & \alpha_3 & \alpha_2 & \alpha_1 & \alpha_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

Matrices with such structure are also called lower Hessenberg matrices.

In some scenarios, Markov property may not hold. For example, say a stochastic process $\{X_n, n \geq 0\}$ satisfies

$$\mathbb{P}\big[X_{n+1} = k | X_n = j, X_{n-1} = i, X_{n-1}, \ldots, X_0\big] \\ = \mathbb{P}\big[X_{n+1} = k | X_n = j, X_{n-1} = i\big]$$

In other words, the future depends on the current state as well as the previous state.

Such a process is called second-order DTMC. It is easy to model this as a regular DTMC. (How?) Simply define $Z_n = (X_{n-1}, X_n)$. Then $\{Z_n, n \geq 1\}$ is a DTMC on $S \times S$.

There are two main goals in studying stochastic processes:

- **Transient Behavior**
  Here, we are interested in a snapshot or the distribution of X_n. We'll also look at occupancy times, which is the expected amount of time spent in various states up to n.

- **Limiting Behavior**
  We'll then look at the process in the long run. Do the sequence of random variables converge (in distribution)? Is the limit unique and how does one compute it?

Understanding these aspects will let us address several practical questions specific to the problem of interest.

This analysis will also help develop intuition and background for studying MDPs.

Coming Soon

The rest of the course will be grouped into the following parts:

1. **Stochastic Processes and Finite Horizon MDPs**:
   DTMCs, classification of DTMCs, transient and limiting behavior, finite horizon MDPs, backward induction, structural results, and applications.
2. **Infinite Horizon Discounted MDPs**:
   Banach spaces and contraction mappings, value iteration, policy iteration, linear programming methods, and applications.
3. **Infinite Horizon Total and Average Cost MDPs**:
   Existence of optimal policies, solution methods (value iteration and policy iteration), unichain and multichain models, and applications.
4. **Approximate Dynamic Programming and RL**:
   Roll-out methods, lookahead and Monte-Carlo Tree Search, model-free methods, function approximation, and policy gradient.
5. **Additional topics**:
   Dynamic discrete choice models, risk-sensitive MDPs, partially observable MDPs (POMDP), continuous-time models.

Your Moment of Zen