CE 273
Markov Decision Processes

Lecture 24

Semi-Markov Decision Processes

Previously on Markov Decision Processes

The objective in the discounted cost MDP problem is

$$\lim_{N\to\infty} \mathbb{E}_w \sum_{k=0}^{N-1} \left\{ \alpha^k g(x_k, u_k, w_k) \right\}$$

Under most practical situations that we encounter, this limit exists and we can also exchange the limit and expectation and write

$$\mathbb{E}_w \sum_{k=0}^{\infty} \left\{ \alpha^k g(x_k, u_k, w_k) \right\}$$

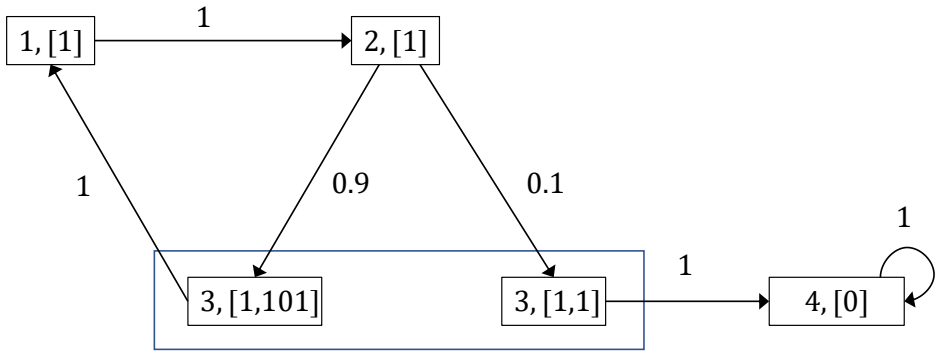
Likewise, given a particular policy $\pi = \{\mu_0, \mu_1, \ldots\}$, the value function can be written as

$$J_\pi(x_0) = \lim_{N\to\infty} \mathbb{E}_w \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}$$

We will make appropriate assumptions (such as bounded costs) that will guarantee the existence of the above limit.

Previously on Markov Decision Processes

How do the Markov chains look like when we deal with the ex ante value functions?



$$P_\mu = \begin{array}{c} \\ 4 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 4 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.1 & 0.9 & 0 & 0 \end{bmatrix} \end{array}$$

Again, the transition matrices of total cost MDP will be assumed to include only the green sub-matrix and we evaluate the cost of the policy using $(I - P_\mu)^{-1} g_\mu$.

$$\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.9 & 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \\ 0.9(1) + 0.1(1) \end{bmatrix} = \begin{bmatrix} 30 \\ 29 \\ 28 \end{bmatrix}$$

Previously on Markov Decision Processes

Let the state space be $X = \{1, 2, \ldots, n, t\}$ where t represents a termination state. Let as before, $p_{ij}(u)$ represent the probability of reaching state j when u is chosen in state i. We further assume that

- The terminal state is absorbing, i.e., $p_{tt}(u) = 1, \forall\, u \in U(t)$.
- The terminal state is cost-free, i.e., $g(t, u) = 0 \,\forall\, u \in U(t)$.

A policy μ is proper if $i \to t$ for all $i = 1, \ldots, n$ in the Markov chain associated with μ.

We make two main assumptions for the analysis of total cost MDPs:

Assumption 1: There exists at least one proper policy

Assumption 2: For all improper policies μ, $J_\mu(i)$ is ∞ for at least one i

For stochastic shortest paths, the above conditions are met if the destination is reachable from all nodes and the link travel times are positive.

Lecture Outline

Lecture Outline

Introduction

Introduction

Assumptions

So far we have looked at problems in which time was discretized into equal intervals, i.e., the time between actions and state transitions was one period.

However, there are several problems (e.g., queuing) in which

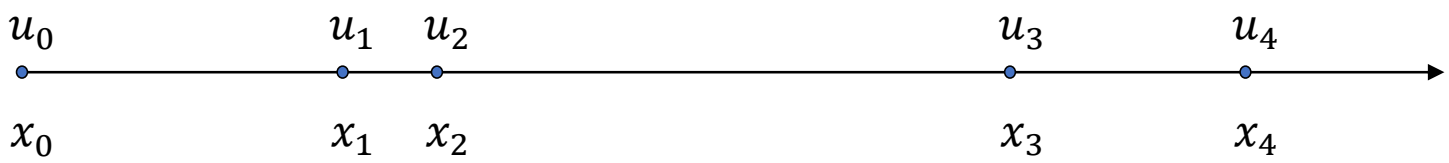
- Time between actions and state transitions is a random variable (can also depend on the current state and choice of control)
- Cost is continuously accumulated

Such problems are also called Semi-Markov Decision Processes. For this lecture, assume infinite horizon problems with finite states and controls but **random time** (could be discrete or continuous).

Introduction

Assumptions

Since time is continuous, we denote the state and control at t using $x(t)$ and $u(t)$.



Let t_k be the time of occurrence of the kth transition. Assume $t_0 = 0$. Define x_k and u_k as states and controls that satisfy

$$x(t) = x_k \text{ for } t_k \leq t < t_{k+1}$$
$$u(t) = u_k \text{ for } t_k \leq t < t_{k+1}$$

Introduction

Notation

The transition probabilities are now replaced with joint transition distributions which predict what future state we might go to and after how much time.

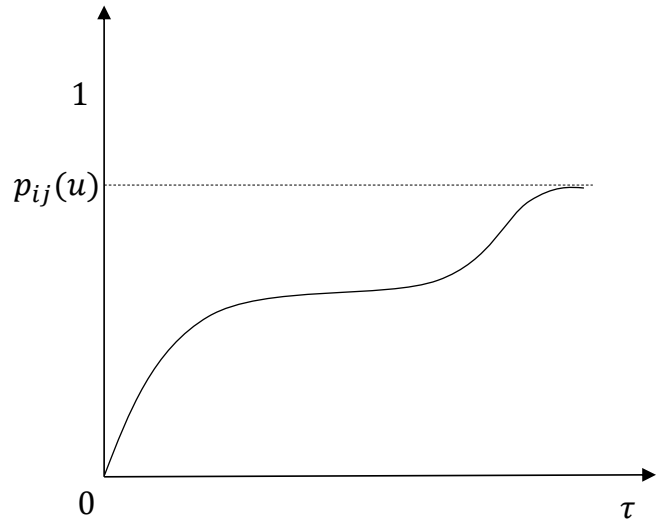
Given a state i and action u, define $Q_{ij}(\tau, u)$ the joint distribution of the transition interval and the future state, i.e.,

$$Q_{ij}(\tau, u) = \mathbb{P}\left[t_{k+1} - t_k \leq \tau, x_{k+1} = j | x_k = i, u_k = u\right]$$

What is the maximum value $Q_{ij}(\cdot, u)$ can take? Set $\tau \to \infty$.

The marginal distribution of the future state is the usual transition distribution $p_{ij}(u)$. Mathematically,

$$\begin{aligned} p_{ij}(u) &= \mathbb{P}\left[x_{k+1} = j | x_k = i, u_k = u\right] \\ &= \lim_{\tau \to \infty} Q_{ij}(\tau, u) \end{aligned}$$

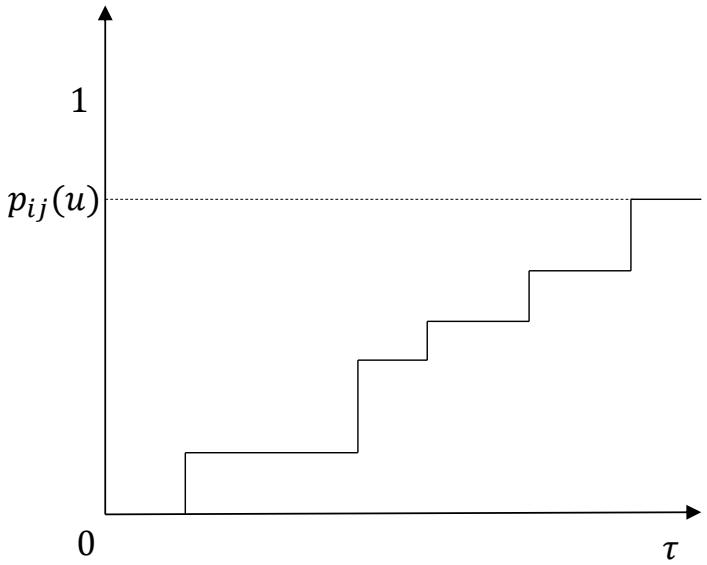


We will use the following problem as a running example in this lecture.

Consider a queue in which the time between successive arrivals of customers is uniform $[0, \tau_{\max}]$. The server processes customers only in batches due to a setup cost of K (which is incurred when serving a new batch of customers).

Suppose decisions are made just before a new customer joins the queue. The state is defined as the number of customers in the queue and the choices available are to serve S or idle D. What are $Q_{ij}(\tau, S)$ and $Q_{ij}(\tau, D)$?

As mentioned earlier, τ can be a discrete random variable as well. In this case, the transition distribution may be imagined to look as



By definition of conditional probability, $\mathbb{P}\big[A|B\big] = \mathbb{P}\big[A \cap B\big]/\mathbb{P}\big[B\big]$. Thus, assuming $p_{ij}(u) > 0$, the conditional cumulative distribution function of the inter-transition time is given by

$$\mathbb{P}\big[t_{k+1} - t_k \leq \tau | x_k = i, x_{k+1} = j, u_k = u\big] = \frac{Q_{ij}(\tau, u)}{p_{ij}(u)}$$

Recall that $\mathbb{E}[X] = \int x \, dF_X(x)$. Let τ represent the inter-transition time. We will abuse notation a bit and use τ for the realizations of the inter-transition times as well. Therefore, the conditional expected value of τ given i, j, and u can be written as

$$\begin{aligned}\bar{\tau} = \mathbb{E}[\tau | i, j, u] &= \int_0^\infty \tau \frac{dQ_{ij}(\tau, u)}{p_{ij}(u)} \\ &= \int_0^\infty \tau \frac{q_{ij}(\tau, u)}{p_{ij}(u)} d\tau\end{aligned}$$

where q_{ij} is like a scaled density function. What is the expected transition time when a control u is applied in state i?

$$\begin{aligned}\bar{\tau}_i(u) &= \sum_{j=1}^n p_{ij}(u) \mathbb{E}[\tau | i, j, u] \\ &= \sum_{j=1}^n \int_0^\infty \tau \, dQ_{ij}(\tau, u)\end{aligned}$$

$\bar{\tau}_i(u)$ is assumed to be finite. Notice that the controls are assumed to be chosen based on which state we are in and not based on how much time elapsed since last transition.

This assumption makes the problem tractable since we won't have to deal with continuous and infinite state spaces.

However, there is one exception where letting the controls depend on the time elapsed isn't advantageous.

If the joint transition distributions are of the form

$$Q_{ij}(\tau, u) = p_{ij}(u)(1 - e^{-\nu_i(u)\tau})$$

then from the earlier expression of the conditional cdf,

$$\mathbb{P}[\text{Transition interval} \leq \tau | i, u] = 1 - e^{-\nu_i(u)\tau}$$

which implies that the transition time interval is exponentially distributed and hence has the memoryless property, i.e.,

$$\mathbb{P}[\text{Transition interval} > a + b \,|\, \text{Transition interval} > a] = \mathbb{P}[\text{Transition interval} > b]$$

Discounted Problems

Discounted Problems

Objective

Costs are accumulating continuously in SMDPs. Hence, we view $g(i,u)$ as a "cost rate". That is, the cost incurred in dt is $g(i,u)dt$.

To continuously discount the cost rate, we use a exponentially decaying discount parameter (continuous analogue of a geometric progression) $e^{-\beta\tau}$ and write the objective of the SMDP as

$$\lim_{T\to\infty} \mathbb{E}\left\{ \int_0^T e^{-\beta t} g(x(t),u(t))dt \right\}$$

Thus, given a policy $\pi = \{\mu_0, \mu_1, \ldots, \}$, we can define the value functions as

$$J_\pi(i) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \mathbb{E}\left\{ \int_{t_k}^{t_{k+1}} e^{-\beta t} g(x_k,\mu_k(x_k))dt | x_0 = i \right\}$$

Discounted Problems

Transition Costs

Proposition

Let $G(i,u)$ be the expected transition cost when action u is taken in state i. Then,

$$G(i,u) = g(i,u)\sum_{j=1}^{n}\int_0^\infty \frac{1-e^{-\beta\tau}}{\beta}\,dQ_{ij}(\tau,u)$$

Why doesn't it simply equal $g(i,u)\bar{\tau}_i(u)$?

Proof.

$$\begin{aligned}
G(i,u) &= \mathbb{E}\left\{\int_0^\tau e^{-\beta t}g(i,u)dt\right\}\\
&= g(i,u)\mathbb{E}\left\{\int_0^\tau e^{-\beta t}dt\right\}\\
&= g(i,u)\mathbb{E}_j\left\{\mathbb{E}_\tau\left\{\int_0^\tau e^{-\beta t}dt\,|\,j\right\}\right\}
\end{aligned}$$

Proof.

$$= g(i,u)\sum_{j=1}^{n} p_{ij}(u)\left\{\mathbb{E}_\tau\left\{\int_0^\tau e^{-\beta t}dt \,|\, j\right\}\right\}$$

$$= g(i,u)\sum_{j=1}^{n} p_{ij}(u)\int_0^\infty \left(\int_0^\tau e^{-\beta t}dt\right)\frac{dQ_{ij}(\tau,u)}{p_{ij}(u)}$$

$$= g(i,u)\sum_{j=1}^{n} p_{ij}(u)\int_0^\infty \left(\frac{1-e^{-\beta\tau}}{\beta}\right)\frac{dQ_{ij}(\tau,u)}{p_{ij}(u)}$$

$$= g(i,u)\sum_{j=1}^{n} \int_0^\infty \frac{1-e^{-\beta\tau}}{\beta}dQ_{ij}(\tau,u)$$

■

Thus, using $Q_{ij}(\tau,u)$, which is problem data, we can compute $G(i,u)$ for all state-action pairs.

Discounted Problems

Bellman Equations

To derive Bellman equations informally, we breakdown the cost into sum of the expected cost from the first transition and the value function from the next state which is appropriately discounted.

$$J_\pi(i) = G(i, \mu_0(i)) + \mathbb{E}[e^{-\beta\tau} J_{\pi_1}(j) | i, \mu_0(i)]$$

where $\pi_1 = \{\mu_1, \mu_2, \ldots\}$. Using a similar logic to calculate the above expectation,

$$\begin{aligned}
\mathbb{E}[e^{-\beta\tau} J_{\pi_1}(j) | i, \mu_0(i)] &= \mathbb{E}_j \Big[\mathbb{E}_\tau [e^{-\beta\tau} | j] J_{\pi_1}(j) | i, \mu_0(i) \Big] \\
&= \sum_{j=1}^{n} p_{ij}(\mu_0(i)) \Big[\mathbb{E}_\tau [e^{-\beta\tau} | j] J_{\pi_1}(j) | i, \mu_0(i) \Big] \\
&= \sum_{j=1}^{n} p_{ij}(\mu_0(i)) \left(\int_0^\infty e^{-\beta\tau} \frac{dQ_{ij}(\tau, \mu_0(i))}{p_{ij}(\mu_0(i))} \right) J_{\pi_1}(j) \\
&= \sum_{j=1}^{n} \left(\int_0^\infty e^{-\beta\tau} dQ_{ij}(\tau, \mu_0(i)) \right) J_{\pi_1}(j)
\end{aligned}$$

Discounted Problems

Bellman Equations

Let $m_{ij}(u)$ be defined as

$$m_{ij}(u) = \int_0^\infty e^{-\beta\tau}\, dQ_{ij}(\tau, u)$$

Thus, the Bellman equations can be written in a more familiar form as

$$J_\pi(i) = G(i, \mu_0(i)) + \sum_{j=1}^{n} m_{ij}(u) J_{\pi_1}(j)$$

How is this different from the Bellman equations of regular MDPs? There is no explicit discount factor in the above expression.

Thus, one could view this as a total cost problem in which choosing u in state i can send us to j with probability $m_{ij}(u)$ and to a fictitious terminal state with probability $\left(1 - \sum_{j=1}^{n} m_{ij}(u)\right)$!

Discounted Problems

Bellman Equations

We can thus use results from total cost MDPs to show existence of solutions to the SMDP. What assumptions do we need for this exercise?

- We can suppose that the fictitious destination or terminal state is cost free.
- The terminal state can be reached w.p. 1 from any state i.

We can be assured that the second condition is applicable if $\sum_{j=1}^{n} m_{ij}(u) < 1$. It easy to show that this holds because of the assumption $\bar{\tau}_i(u) < \infty$.

Thus, a unique solution to the following system of equations exists and solves the SMDP

$$J^*(i) = \min_{u \in U(i)} \left\{ G(i,u) + \sum_{j=1}^{n} m_{ij}(u) J^*(j) \right\}$$

We can use any of the three methods: VI, PI or LP that were discussed in the total cost MDP case.

Discounted Cost Problems

Example

In the queuing example, suppose the one-stage expected cost $G(i, S) = K$. If the cost per unit time of a customer who hasn't been served is c, what is $G(i, D)$?

$$G(i, D) = ci \sum_{j=1}^{n} \int_0^{\tau_{\max}} \frac{1 - e^{-\beta\tau}}{\beta} dQ_{ij}(u, D)$$

$$= ci \int_0^{\tau_{\max}} \frac{1 - e^{-\beta\tau}}{\beta\tau_{\max}} d\tau$$

Write the Bellman equations. Using the definition of $m_{ij}(u)$,

$$m_{i1}(S) = m_{i,i+1}(D) = \int_0^{\tau_{\max}} \frac{e^{-\beta\tau}}{\tau_{\max}} d\tau = \frac{1 - e^{-\beta\tau_{\max}}}{\beta\tau_{\max}}$$

$$J(i) = \min \left\{ K + m_{i1}(S)J(1),\ G(i, D) + m_{i,i+1}(D)J(i + 1) \right\}$$

Average Cost Problems

Average Cost Problems

Objective

One can formulate SMDPs for average cost problems as well. The objective in these problems can be written as

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{ \int_0^T g(x(t),u(t))dt \right\}$$

It helps to reformulate the above objective into the following equivalent version:

$$\lim_{N\to\infty} \frac{1}{\mathbb{E}[t_N]}\mathbb{E}\left\{ \int_0^{t_N} g(x(t),u(t))dt \right\}$$

Since we are interested in minimizing the average costs, the one-stage expected cost from choosing u in state i can be written as

$$G(i,u) = g(i,u)\bar{\tau}_i(u)$$

Average Cost Problems

Value Functions

The value function of starting from state i and following policy $\pi = \{\mu_0, \mu_1, \ldots\}$ is

$$J_\pi(i) = \lim_{N\to\infty} \frac{1}{\mathbb{E}[t_N | x_0 = i, \pi]} \mathbb{E}\left\{ \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} g(x_k, \mu_k(x_k)) dt \,|\, x_0 = i \right\}$$

We can construct what is referred to as an *embedded Markov chain* with transition probabilities $p_{ij}(u) = \lim_{\tau\to\infty} Q_{ij}(\tau, u)$.

If the embedded Markov chain satisfies a unichain-like conditions, it can be shown that $J^*(i)$ is independent of the starting state i!

Average Cost Problems

Bellman Equations

Under such conditions Bellman equations of the average cost SMDP takes the following form

$$h(i) = \min_{u \in U(i)} \left\{ G(i,u) - \lambda \bar{\tau}_i(u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right\}$$

Setting $\bar{\tau}_i(u)$ gives us the original average cost MDP.

Your Moment of Zen