CE 273
Markov Decision Processes

Lecture 13

Infinite Horizon Average Cost MDPs

Previously on Markov Decision Processes

Let $V_j^{(n)}$ be the number of visits to j over $\{0, 1, \ldots, n\}$. Mathematically, occupancy time of j up to time n starting from i is

$$m_{ij}^{(n)} = \mathbb{E}\left[V_j^{(n)} | X_0 = i\right], \forall\, i, j \in S, n \geq 0$$

The matrix of $m_{ij}^{(n)}$ values, is represented by

$$M^{(n)} = \left[m_{ij}^{(n)}\right]_{|S| \times |S|}$$

Intuitively, to go from i to j in n steps, we need to transition from i to some state r in k steps and from r to j in remaining $(n - k)$ steps.

Theorem (Chapman-Kolmogorov Equations)

The n-step transition probabilities satisfy

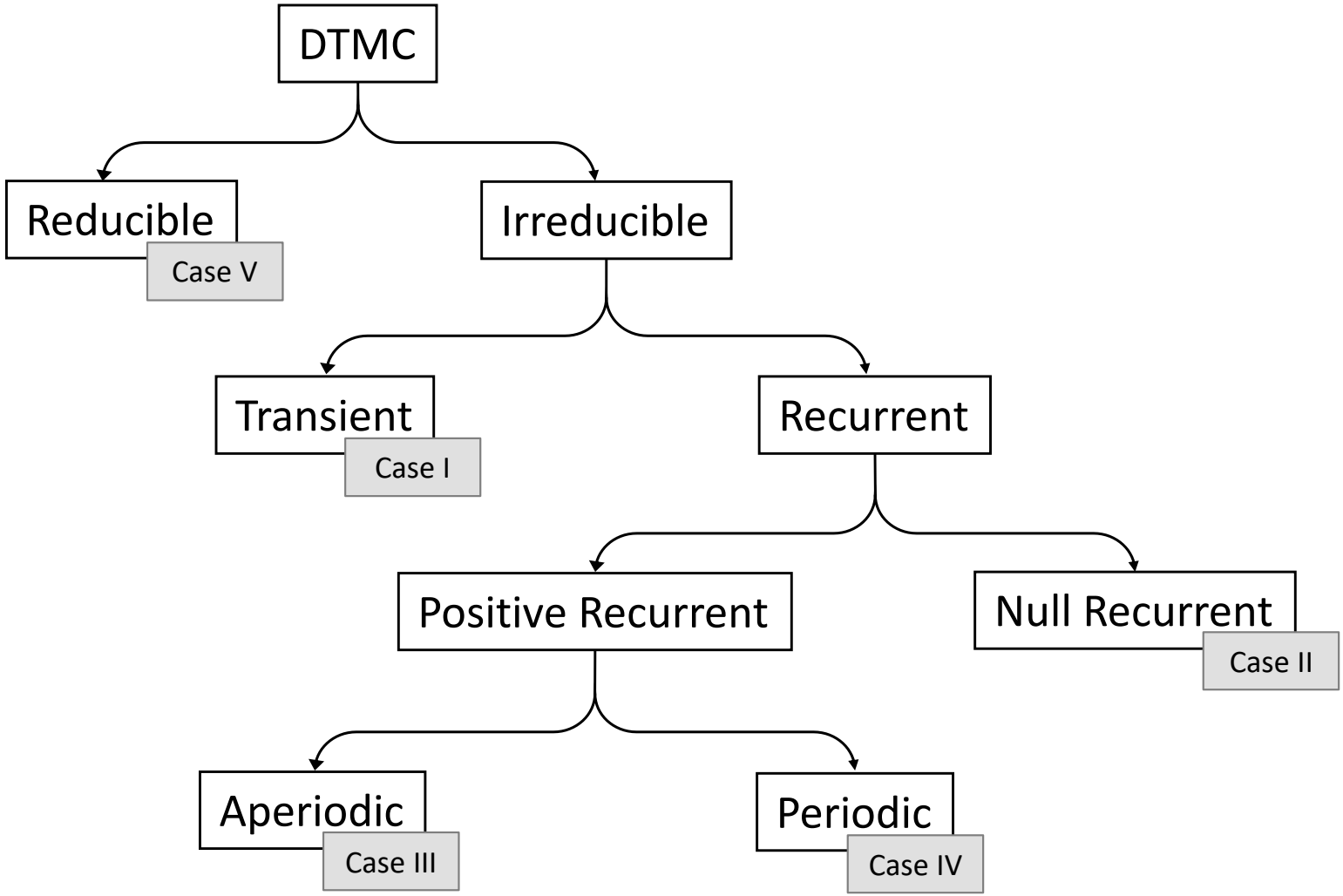
$$p_{ij}^{(n)} = \sum_{r \in S} p_{ir}^{(k)} p_{rj}^{(n-k)}, \forall\, i, j \in S, 0 \leq k \leq n$$

Applying the CK equations recursively, $P^{(n)} = P^n$

Theorem

Let $P^0 = I$. For a fixed n, $M^{(n)} = \sum_{r=0}^{n} P^r$

Previously on Markov Decision Processes



Previously on Markov Decision Processes

Theorem (Case I)

Let $\{X_n, n \geq 0\}$ be an transient, irreducible DTMC. Then

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0 \,\forall\, i, j \in S$$

Theorem (Case II)

Let $\{X_n, n \geq 0\}$ be an null recurrent, irreducible DTMC. Then

$$\lim_{n\to\infty} p_{ij}^{(n)} = 0 \,\forall\, i, j \in S$$

Previously on Markov Decision Processes

Theorem (Case III)

Let e be a column vector of ones. For an aperiodic, positive recurrent, irreducible DTMC, there exists unique $\pi_j > 0, j \in S$ such that

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j, \forall\, i,j \in S$$

$$\pi P = \pi \text{ (Balance Equation)}$$

$$e\pi = 1 \text{ (Normalizing Equation)}$$

Theorem (Case IV)

Let e be a column vector of ones. For a periodic, positive recurrent, irreducible DTMC, there exists unique $\pi_j > 0, j \in S$ such that

$$\lim_{n\to\infty} \frac{m_{ij}^{(n)}}{n+1} = \pi_j, \forall\, i,j \in S$$

$$\pi P = \pi$$

$$e\pi = 1$$

Previously on Markov Decision Processes

Theorem (Case V)

Let $i \in T$ and $j \in C_r$.

1. *If C_r is transient or null recurrent $d_{ij}^{(n)} \to 0$*
2. *If C_r is aperiodic and positive recurrent, $d_{ij}^{(n)} \to u_i(r)\pi_j$, where π_js are derived from limiting distribution of $P(r)^{(n)}$*
3. *If C_r is periodic and positive recurrent, $d_{ij}^{(n)}$ does not have a limit. However $\sum_{m=0}^{n} d_{ij}^{(m)}/(n+1) \to u_i(r)\pi_j$, where π_js are derived from limiting distribution of $P(r)^{(n)}$*

Previously on Markov Decision Processes

From the examples, we can see that

- $\lim_{n\to\infty} P^{(n)}$ doesn't always exist
- $\lim_{n\to\infty} \frac{M^{(n)}}{n+1}$ however always exists and equals $\lim_{n\to\infty} P^{(n)}$ when the later exists. (Why is this intuitively true?)

Case	$\lim_{n\to\infty} P^{(n)}$	$\lim_{n\to\infty} \frac{M^{(n)}}{n+1}$	Identical Rows	Row Sum $= 1$
I	✓	✓	✓	X
II	✓	✓	✓	X
III	✓	✓	✓	✓
IV	X	✓	✓	✓
V	✓	✓	X	✓

Lecture Outline

1. Introduction and Motivating Examples
2. Connections With Discounted MDPs

Lecture Outline

Introduction and Motivating Examples

In situations where the dynamic program does not have an economic interpretation, total discounted cost could be used.

However, this objective can be unbounded without zero-cost terminal states. In such cases, an alternate objective, the average cost per stage can be used to find optimal decisions.

This objective is well suited for queuing applications in traffic and communications.

Introduction and Motivating Examples

Objective

Consider a problem with state space X. As before, assume that we can choose actions from the set $U(i)$ when in state i.

The probability of transitioning from i to j when action u is chosen is $p_{ij}(u)$. Consider a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, where $\mu_k(i) \in U(i)$.

Then the average cost per stage starting from x_0 is defined as

$$J_\pi(x_0) = \limsup_{N\to\infty} \frac{1}{N+1} \mathbb{E}\left\{ \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) \right\}$$

We've switched to non-stationary policies and lim sup to illustrate cases in which the lim may not exist and stationary policies may not be optimal! The above lim sup exists when the one-step costs are bounded.

As before, we write g_μ and J_μ to represent the one-step costs and the average cost of using policy μ and P_μ is the one-step transition matrix.

The average cost also presents unique technical challenges that were absent in the previous MDP objectives:

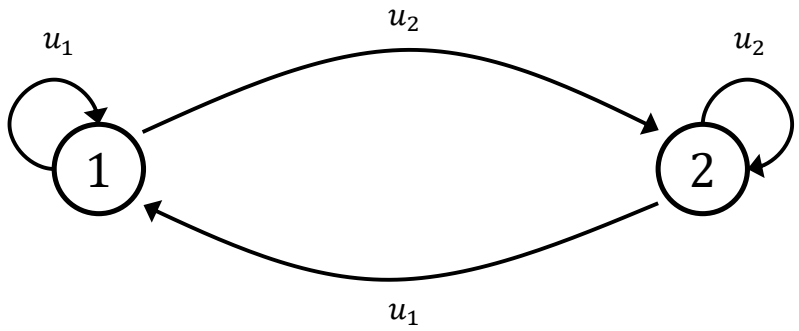
- The limit of the average cost objective may not exist for all types of policies.
- The solution methods depend heavily on the underlying stochastic processes associated with different policies.

Let us look at some of these features with a few examples. For now, we will assume that the state space can be countably infinite, but we will switch to finite state spaces later.

Introduction and Motivating Examples

Example 1

Consider the two-state system in which one can take two actions u_1 and u_2 in each state. The rewards and the transition probabilities are shown below:



- $U(1) = \{u_1, u_2\}$
- $g(1, u_1) = 2, g(1, u_2) = 2$
- $p_{1j}(u_1) = [1 \quad 0]$
- $p_{1j}(u_2) = [0 \quad 1]$

- $U(2) = \{u_1, u_2\}$
- $g(2, u_1) = -2, g(2, u_2) = -2$
- $p_{2j}(u_1) = [1 \quad 0]$
- $p_{2j}(u_2) = [0 \quad 1]$

Consider the non-stationary policy which starting in 1, remains in 1 for one period, proceeds to 2 and remains there for 3 periods, returns to 1 and remains there for 3^2 periods and proceeds to 2 and remains there for 3^3 periods.

However, if you use any stationary policy, both lim sup and lim inf exist and are equal. This is true for all stationary policies of an average cost MDP.

So why not just look at stationary policies and make our life easy? Because,

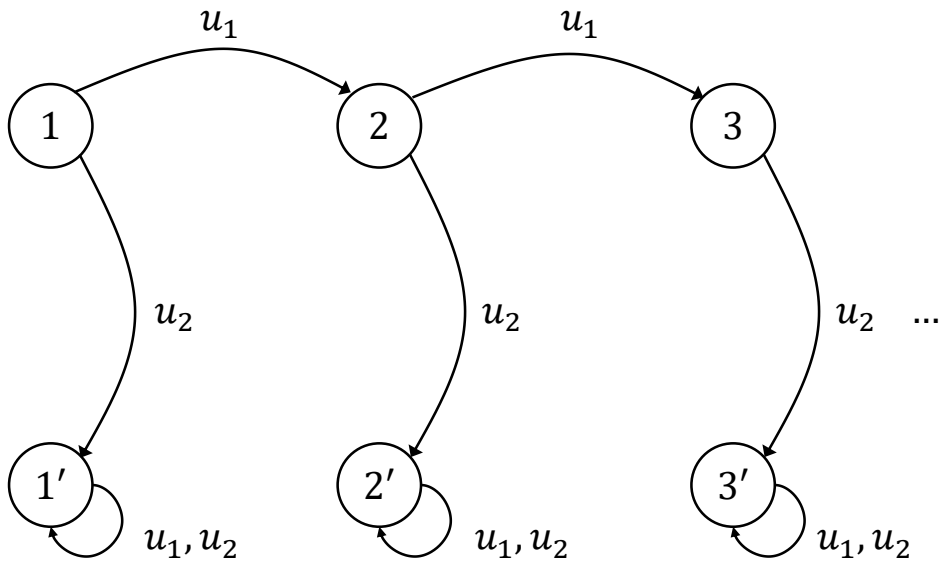
- No stationary or randomized policy may be optimal
- Randomized polices and non-stationary policies can do better than stationary policies
- The performance of stationary policies may be far from that of randomized policies

The following three examples give us reason to believe the above observations. Assume that, in these examples, we wish to maximize rewards instead of minimizing costs.

Introduction and Motivating Examples

Example 2

Suppose the state space is $S = \{1, 1', 2, 2', \ldots\}$ and at each state we can choose u_1 and u_2. Let the transition diagram associated with these actions be as shown below.



Suppose $g(i, u) = 0 \, \forall \, i \in \{1, 2, \ldots\}, u \in U(i)$ and $g(i, u) = 1 - 1/i \, \forall \, i \in \{1', 2', \ldots\}, u \in U(i)$. What is the average reward of the optimal policy?

Introduction and Motivating Examples

Example 2

In this example, lim sup exists (equals 1) but it is not attained by any policy. Every stationary or non-stationary policy has an average reward strictly less than 1.

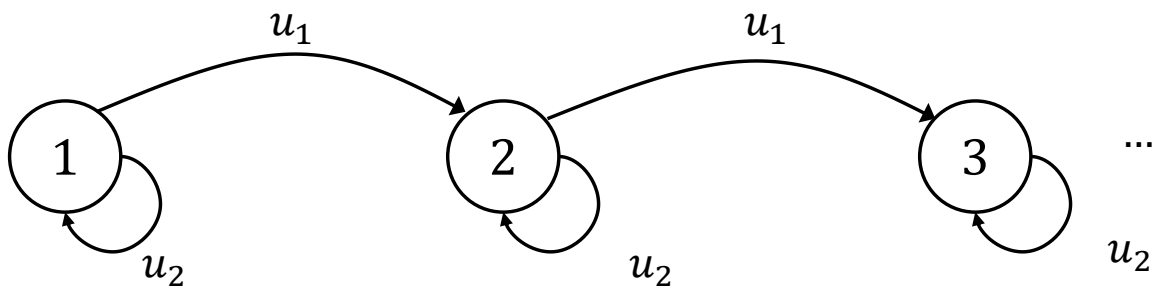
Think of this as being equivalent to minimizing a function like e^{-x}.

However, in this example, it is possible to construct stationary policies that are ϵ-optimal. That is, $J_{\mu_\epsilon} > 1 - \epsilon$.

Introduction and Motivating Examples

Example 3

Suppose the state space is $X = \{1, 2, \ldots\}$ and at each state we can choose u_1 and u_2. Let the transition diagram associated with these actions be as shown below.



Suppose $g(i, u_i) = 0 \, \forall \, i$ and $g(i, u_2) = 1 - 1/i \, \forall \, i \in \{1, 2, \ldots\}$. What is the average reward of the optimal policy?

Introduction and Motivating Examples

Example 3

The average reward of every stationary policy is strictly less than 1.

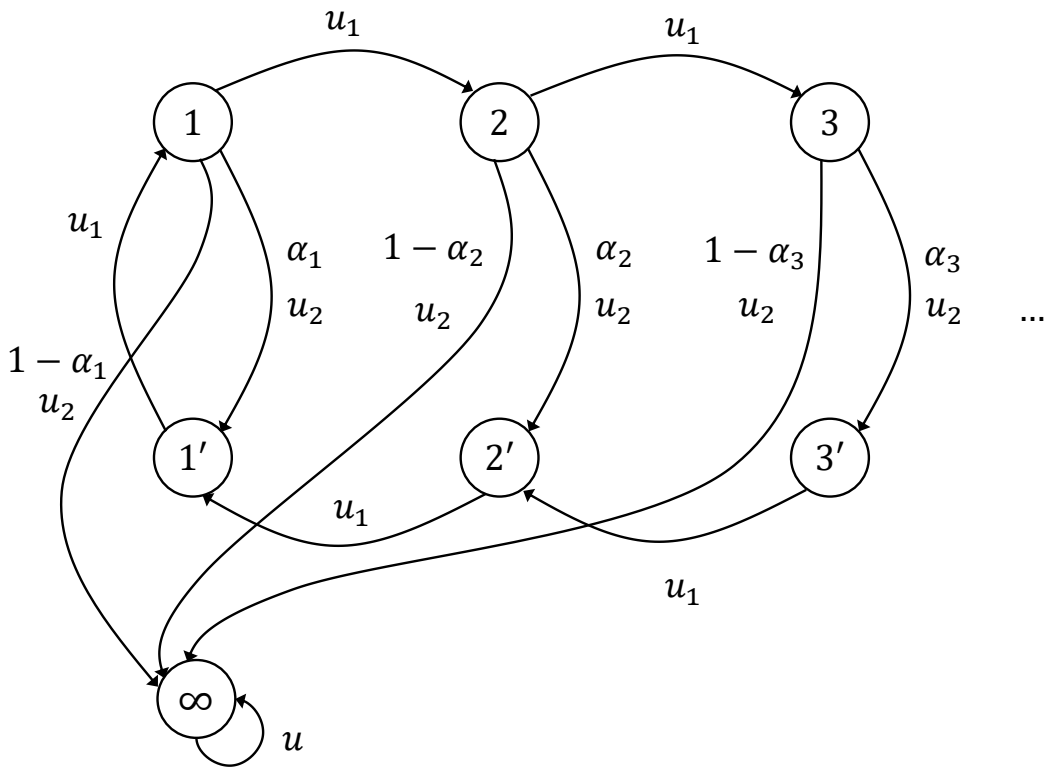
Consider the non-stationary policy that chooses u_2 i consecutive times when in state i.

The average rewards follows the sequence $0, 0, \frac{1}{2}, \frac{1}{2}, 0, \frac{2}{3}, \frac{2}{3}, \frac{2}{3}, 0, \ldots$. The limit of this sequence is 1. Hence, the average reward of this policy is 1.

Introduction and Motivating Examples

Example 4

Suppose the state space is $S = \{1, 1', 2, 2', \ldots\}$. Let the transition diagram associated with these actions be



The result of action u_1 is deterministic but taking u_2 in a state i can move the system to i' or ∞ with probabilities α_i and $1 - \alpha_i$.

Suppose $g(i, u) = 0 \, \forall \, i \in \{1, 2, \ldots\}$ and $g(i, u) = 2 \, \forall \, i \in \{1', 2', \ldots\}$ and $g(\infty, u) = 0$. What is the average reward of the optimal policy?

Introduction and Motivating Examples

Example 4

Average reward of any stationary policy is 0.

Consider the non-stationary policy that chooses u_2 on the nth return to state 1, chooses u_1 n times and then chooses u_2.

The average reward of this policy is $\prod_{i=1}^{\infty} \alpha_i$ which can be > 0 for a particular choice of αs. For example, choose α_i as $1 - \frac{1}{(i-1)^2}$ or $\frac{4i^2-1}{4i^2}$ (Wallis Product).

The infinite product is $1/2$ and $2/\pi$ for the above examples respectively and no stationary policy is ϵ-optimal.

Connections With Discounted MDPs

Connections With Discounted MDPs

Gain and Bias

Two useful functions that will help us study the theory of average cost MDPs are gain and bias.

Let us focus only on *stationary* deterministic and randomized policies. Suppose μ is such a policy.

Denote using P^*_μ the limit of the average occupancy matrix. Mathematically,

$$P^*_\mu = \lim_{N\to\infty} \frac{M^{(N)}_\mu}{N+1} = \lim_{N\to\infty} \frac{1}{N+1} \sum_{r=0}^{N} P^r_\mu$$

The above limit always exists as seen in Lecture 3.

Connections With Discounted MDPs

Gain and Bias

Now consider the cases where P_μ^* is a stochastic matrix (Cases III, IV, V).

Recall from the analysis of total cost MDPs, $\sum_{k=0}^{N-1} P_\mu^k g_\mu$ represents the cost accumulated after N stages. (We started with the zero cost vector and used the T_μ operator.)

Thus, the average cost of policy μ is

$$J_\mu = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k g_\mu = \left(\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k \right) g_\mu = P_\mu^* g_\mu$$

Definition

The gain J_μ of a policy μ is defined as

$$J_\mu = P_\mu^* g_\mu$$

Connections With Discounted MDPs

Gain and Bias

Definition

The bias h_μ of a policy μ is defined as

$$h_\mu = H_\mu g_\mu$$

where $H_\mu = (I - P_\mu + P_\mu^*)^{-1} - P_\mu^*$ and is called the fundamental matrix.

In addition, suppose the associated Markov chain is aperiodic, i.e., if $P_\mu^* = \lim_{N\to\infty} P_\mu^N$ (Case III), then we can interpret h_μ as

$$h_\mu = \lim_{N\to\infty} \sum_{k=0}^{N} P_\mu^k (g_\mu - J_\mu)$$

a relative cost vector, i.e., the difference of the total cost of μ and the total cost if one-stage costs were set to J_μ.

Connections With Discounted MDPs

Road Map

It turns out that the average cost can be derived from a series expansion of the α-discounted cost model.

Let $J_{\alpha,\mu}$ represent the value functions of policy μ in a discounted MDP. We can write,

$$J_{\alpha,\mu} = \sum_{k=0}^{\infty} \alpha^k P_\mu^k g_\mu = \left(\sum_{k=0}^{\infty} \alpha^k P_\mu^k \right) g_\mu = (I - \alpha P_\mu)^{-1} g_\mu$$

Connections With Discounted MDPs

Wishful Thinking

The following set of equations is guesswork that could relate average cost MDPs and α-discounted MDPs.

$$
\begin{aligned}
J_\mu(i) &= \limsup_{N\to\infty} \frac{1}{N}\mathbb{E}\left\{ \sum_{k=0}^{N-1} g(x_k, \mu(x_k)) \right\} \\
&= \limsup_{N\to\infty} \lim_{\alpha\to 1} \frac{\mathbb{E}\left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k)) \right\}}{\sum_{k=0}^{N} \alpha^k} \\
&= \lim_{\alpha\to 1} \limsup_{N\to\infty} \frac{\mathbb{E}\left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k)) \right\}}{\sum_{k=0}^{N} \alpha^k} \\
&= \lim_{\alpha\to 1} \frac{\lim_{N\to\infty} \mathbb{E}\left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k)) \right\}}{\lim_{N\to\infty} \sum_{k=0}^{N} \alpha^k} \\
&= \lim_{\alpha\to 1} (1-\alpha) J_{\alpha,\mu}(i)
\end{aligned}
$$

Connections With Discounted MDPs

Road Map

Theorem

For any transition matrix P and $\alpha \in (0,1)$,

$$(I - \alpha P)^{-1} = (1-\alpha)^{-1} P^* + H + O(|1-\alpha|)$$

where $O(|1-\alpha|)$ is an α-dependent matrix such that $\lim_{\alpha \to 1} O(|1-\alpha|) = 0$ and P^* and H are given by

$$P^* = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k$$

$$H = (I - P + P^*)^{-1} - P^*$$

Connections With Discounted MDPs

Road Map

Recall that $J_\mu = P_\mu^* g_\mu$ and $h_\mu = H_\mu g_\mu$. Multiplying both sides of the Laurent series expansion with g_μ,

Theorem (Laurent Series Expansion)

For a given stationary policy μ with transition matrix P_μ and $\alpha \in (0,1)$,

$$J_{\alpha,\mu} = (1-\alpha)^{-1} J_\mu + h_\mu + O(|1-\alpha|)$$

where $O(|1-\alpha|)$ is an α-dependent matrix such that $\lim_{\alpha \to 1} O(|1-\alpha|) = 0$ and J_μ an h_μ represent gain and bias of the policy μ respectively.

Hence, we can write

$$J_\mu = (1-\alpha) J_{\alpha,\mu} - (1-\alpha) h_\mu + O(|1-\alpha|^2)$$

Thus, we expect that a policy minimizing $J_{\alpha,\mu}$ for α close to 1 will also minimize the average cost J_μ!

Your Moment of Zen