CE 273
Markov Decision Processes

Lecture 7

Infinite Horizon Discounted MDPs

Previously on Markov Decision Processes

Costs are usually discounted for mathematical and practical reasons. Mathematically, they guarantee that the long-run discounted costs are bounded.

Practically, a cost of c units in one future time-step is equivalent to incurring αc now, where $0 \leq \alpha < 1$. More generally, cost c at time step n is equivalent to $\alpha^n c$ now.

One interpretation of α is that it reflects the interest rate. Another grim way to look at is to assume that time is finite, and the future may not happen with probability $(1 - \alpha)$. Define the cost over the infinite horizon as

$$C = \sum_{n=0}^{\infty} \alpha^n c(X_n)$$

C is a random variable and hence let's look at the expected total discounted cost starting from state i,

$$\phi(i) = \mathbb{E}\big[C | X_0 = i\big]$$

Previously on Markov Decision Processes

The expected costs for the discounted case **depends on the initial state**. The advantage however is that we don't need any conditions on the DTMC and the inverse exists as long as $\alpha < 1$.

Interestingly, if we average the total expected cost without discounting, the initial state does not matter!

Define the long-run expected cost per period (average cost),

$$g(i) = \lim_{N\to\infty} \frac{1}{N+1}\mathbb{E}\left\{ \sum_{n=0}^{N} c(X_n)|X_0 = i \right\}$$

For the above limit to exist, assume that the DTMC is irreducible and positive recurrent. Since πs represent the average time spent in state i, we would expect $g(i) = g = \sum_{j\in S} \pi_j c(j)$.

Previously on Markov Decision Processes

When states are countable, we can simplify the notation. Suppose indices i and j represent the states. Let

$$p_{ij}(u,k) = \mathbb{P}\left[x_{k+1} = j | x_k = i, u_k = u\right]$$

How is this different from the transition probabilities of DTMCs?

The DP algorithm can be written as

$$J_k(i) = \min_{u_k \in U_k(i)} \left\{ g_k(i,u_k) + \sum_{j \in S_{k+1}} p_{ij}(u_k,k) J_{k+1}(j) \right\}$$

where $g_k(i,u_k)$ is expected cost of choosing u_k in state i. If it depends on the disturbance, one can treat it as $g_k(i,u_k,j)$ and move it inside the summation.

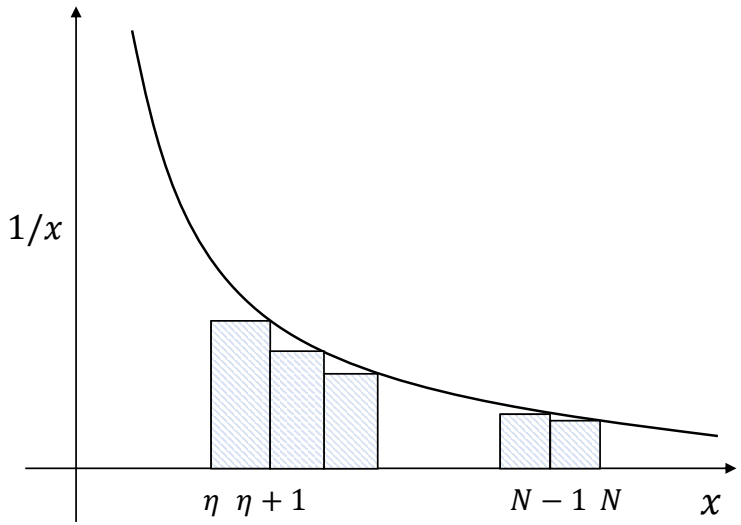
Previously on Markov Decision Processes

Thus, the number of candidates to reject before selecting the candidate with relative rank 1 is the smallest integer for which

$$\frac{1}{N-1}+\frac{1}{N-2}+\ldots+\frac{1}{\eta+1}\leq 1$$

What is the value of η for $N=5$? What if $N\to\infty$? We can approximate the above inequality as

$$\frac{1}{N}+\frac{1}{N-1}+\frac{1}{N-2}+\ldots+\frac{1}{\eta+1}\approx 1$$



which can be written as

$$\int_{\eta}^{N}\frac{1}{x}\approx 1\Rightarrow ln(N/\eta)\approx 1$$

$$\Rightarrow \eta/N\approx e^{-1}=0.3679$$

Thus, when N is large, it is optimal to reject 36.79% of the candidates and then select the top relative-ranked one!

Lecture Outline

1. Finite Horizon MDPs Wrap-up
2. Introduction to Infinite Horizon Problems
3. Alternate Notation

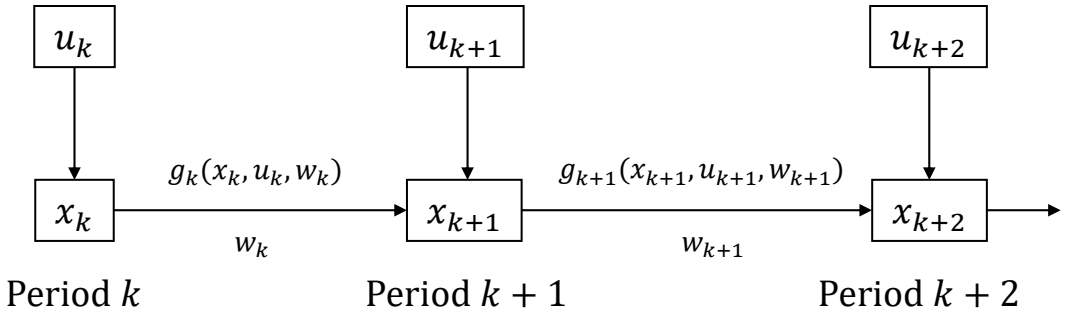
Lecture Outline

Finite Horizon MDPs Wrap-up

Finite Horizon MDPs Wrap-up

Cost Functions

There appears to be some confusion with the notation especially with regard to the cost functions. Recall that in MDPs



the objective is $\mathbb{E}\left\{ g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k, w_k) \right\}$ and the value functions for a given policy π are defined as

$$J_\pi(x_0) = \mathbb{E}\left\{ g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k) \right\}$$

The above notation is the most general form, but depending on the context $g_k(x_k, u_k, w_k)$ need not depend on the disturbance (or even the control).

Finite Horizon MDPs Wrap-up

Cost Functions

Consider the case where the cost function does not depend on w_k, i.e., $g_k(x_k, u_k, w_k) = g_k(x_k, u_k)$. Then the objective becomes

$$\mathbb{E}\left\{ g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k) \right\}$$

and the value functions for a given policy π are defined as

$$J_\pi(x_0) = \mathbb{E}\left\{ g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k)) \right\}$$

Why do we have an expectation when w_k is absent in the above expressions? Because x_{k+1}'s still depend on w_ks and hence are random variables!

Finite Horizon MDPs Wrap-up

Cost Functions

Let us revisit the disturbance and the cost/reward structure in the examples we discussed so far:

Tetris: The disturbance was the shape of the next block. One-step rewards correspond to the number of rows cleared and they do not depend on the disturbance.

Inventory Control: Disturbance was the demand in period k and the one-step cost was $cu_k + p\max(0, -x_k - u_k + w_k) + h\max(0, x_k + u_k - w_k)$ and it clearly depends on w_k.

Finite Horizon MDPs Wrap-up

Cost Functions

Secretary Problem: Disturbance was the relative rank of the next candidate (either top relative rank or not). Since the objective was to maximize the probability of selecting the top true-ranked candidate, the rewards were collected only at the end of the interview.

Had we interviewed everyone, in hindsight, we can assign a reward of 1 or 0 (the applet version). However, at the time of stopping, the probability with which we've chosen the top true-ranked candidate is k/N and is independent of w_k.

Imagine simulating a stopping policy. We switch to the T state and hence never know the ranks of the others.

Airline Revenue Management: The disturbance is the offer made by a customer in the next time step. The one-step reward is the offer made by the current customer if we agree to sell and is hence independent of w_k.

Finite Horizon MDPs Wrap-up

Cost Functions

One-step costs that do not depend on w_k are possible in two situations:

- Costs are realized before we see the realization of w_k and move to the new state! The index k in w_k must be carefully interpreted.
- They depend on w_k but are averaged out in the problem specification, i.e., $g_k(x_k, u_k) = \mathbb{E}_{w_k} g_k(x_k, u_k, w_k)$

You'll thus commonly see the Bellman equations in the following two formats:

$$J_k(x_k) = \min_{u_k \in U_k(x_k)} \mathbb{E}_{w_k} \left\{ g_k(x_k, u_k, w_k) + J_{k+1}(f_k(x_k, u_k, w_k)) \right\}$$

$$J_k(x_k) = \min_{u_k \in U_k(x_k)} g_k(x_k, u_k) + \mathbb{E}_{w_k} \left\{ J_{k+1}(f_k(x_k, u_k, w_k)) \right\}$$

Finite Horizon MDPs Wrap-up

Structural Results

There are many more practical problems in which structural results can be obtained. You'll see a few more examples in your assignment.

One property commonly exhibited by the value functions and policies is monotonicity. For e.g., the inventory control policy is monotone.

It is possible to derive some sufficient conditions for monotonicity of value functions and policies. Let's briefly discuss these results without delving into the proofs.

We'll use the second notation (Markov chain) as it is more ideal for this discussion.

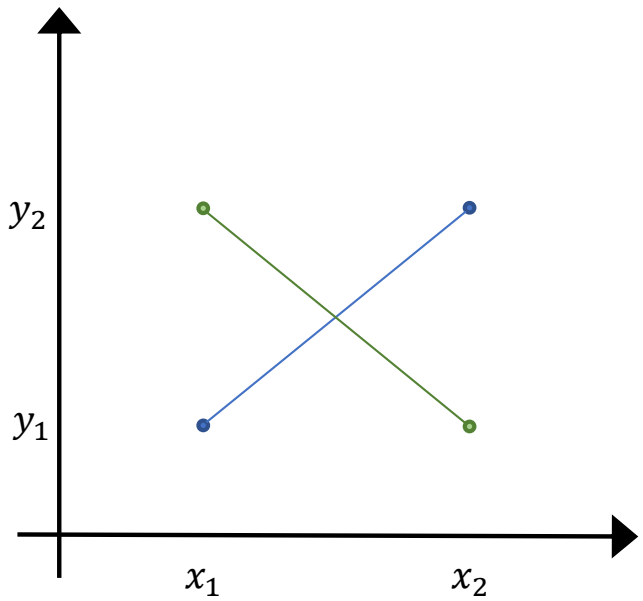
Definition (Superadditivity)

Consider a real-valued function $f : X \times Y \to \mathbb{R}$. It is superadditive or supermodular if it satisfies the quadrangle inequality

$$f(x_1, y_1) + f(x_2, y_2) \geq f(x_1, y_2) + f(x_2, y_1) \, \forall \, x_2 \geq x_1, y_2 \geq y_1$$

Examples of superadditive functions include xy, $(x + y)^2$. Similar definitions can be extended for the discrete version of this definition.

It can also be shown that for twice differentiable functions defined on $\mathbb{R}^2$, superadditivity is satisfied when $\partial^2 f / \partial x \partial y \geq 0$.



Finite Horizon MDPs Wrap-up

Monotonicity of Optimal Value Functions

Let the state space be $S = \{0, 1, \ldots\}$ for all k. Recall that $p_{ij}(u_k, k)$ is the probability of moving from state i to j when an action u_k is taken at time k. Define

$$q_{ii'}(u_k, k) = \sum_{j=i'}^{\infty} p_{ij}(u_k, k)$$

which is the probability of transitioning from i to a state numbered $\geq i'$.

Proposition

Suppose that

1. $q_{ii'}(u_k, k)$ *is nondecreasing in* i *for all* $i' \in S$, $u_k \in U_k(i)$, $k \in \{0, 1, \ldots, N-1\}$
2. $g_k(i, u_k)$ *is nondecreasing in* i *for all* $u_k \in U_k(i)$, $k \in \{0, 1, \ldots, N-1\}$
3. $g_N(i)$ *is nondecreasing in* i

then the optimal value functions $J_k^*(i)$ *are nondecreasing in* i *for all* $k \in \{0, 1, \ldots, N-1\}$

Finite Horizon MDPs Wrap-up

Monotonicity of Optimal Policies

Similar results exist for proving the monotonicity of the optimal policies.

Proposition

Suppose that

1. $q_{ii'}(u_k, k)$ *is nondecreasing in* i *for all* $i' \in S$, $u_k \in U_k(i)$, $k \in \{0, 1, \ldots, N-1\}$
2. $g_k(i, u_k)$ *is nondecreasing in* i *for all* $u_k \in U_k(i)$, $k \in \{0, 1, \ldots, N-1\}$
3. $g_N(i)$ *is nondecreasing in* i
4. $q_{ii'}(u_k, k)$ *is superadditive on* $S \times U_k(i)$, $k \in \{0, 1, \ldots, N-1\}$
5. $g_k(i, u_k)$ *is superadditive on* $S \times U_k(i)$, $k \in \{0, 1, \ldots, N-1\}$

then the optimal policy $\mu_k^*(i)$ *is nondecreasing in* i *for all* $k \in \{0, 1, \ldots, N-1\}$

One can replace nondecreasing with nonincreasing and superadditivity with subadditivity to claim that the value functions and policies are nonincreasing.

How does this help? Notice that the Bellman equations require us to minimize a function over $U_k(i)$

$$J_k(i) = \min_{u_k \in U_k(i)} \left\{ g_k(i, u_k) + \sum_{j \in S_{k+1}} p_{ij}(u_k, k) J_{k+1}(j) \right\}$$

This has to be carried out for every $i \in S$. Thus, if we know the optimal policy for some state i, $\mu_k^*(i)$, we can shrink the search space for higher states.

That is, for $i+1$, the minimization is carried out over $u_k \in U_k(i+1)$ and $u_k \geq \mu^*(i)$.

Finite Horizon MDPs Wrap-up

Deterministic vs. Randomized Policies

At any stage k, we have so far chosen a single control depending on the state. However, one can also randomize over the set of available actions.

There are results with mild requirements (such as bounded costs) which guarantee the existence of deterministic policies.

All the examples we've discussed in the last few lectures can be shown to satisfy the conditions required for these results.

Lecture Outline

Introduction to Infinite Horizon MDPs

Introduction to Infinite Horizon MDPs

Motivation

In most practical sequential decision making problems, the time horizon is finite. However, it is worthwhile to look at cases where $N \to \infty$ if

- We are optimizing a system that doesn't have a terminal stage. For e.g., maintenance of roads, water resource management, or training a robot.
- We could also be looking at systems where several a finite, but exponential number of stages are involved (e.g., Go)
- Alternately, we might be dealing with optimal stopping problems where N can theoretically be infinite (e.g., Stochastic shortest paths with cycles)

A common feature in infinite horizon models is stationarity. The transition dynamics and one-step costs/rewards, and distribution of disturbances do not depend on time.

This is very similar to the time-homogeneity assumption in Markov chains.

Introduction to Infinite Horizon MDPs

Types

In this course, we will study three variants of the problem:

- Discounted Cost Problems
- Undiscounted/Total Cost Problems
- Average Cost Problems

In discounted cost problems the discount factor $\alpha < 1$. This makes it mathematically elegant and relatively easy to solve.

The total cost model is similar in objective but $\alpha = 1$. Due to this assumption, the objective function can be unbounded. Hence, we'll study a special case in which there is a zero cost termination state. (Recall recurrent DTMCs.)

The average cost model is analogous to the one we saw in DTMCs and is useful when the total cost is unbounded and when discounting doesn't make sense.

Introduction to Infinite Horizon MDPs

Discounted Problems

As before, at stage k let x_k, u_k, and w_k be the state, control, and disturbance. Suppose $g(x_k, u_k, w_k)$ and $f(x_k, u_k, w_k)$ represent the one-step costs and the system dynamics.

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.

Introduction to Infinite Horizon MDPs

Stationarity

As before, if Π denotes the set of admissible policies, the optimal cost function is given by

$$J^*(x_0) = J_{\pi^*}(x_0) = \min_{\pi \in \Pi} J_\pi(x_0)$$

Note that when writing the value functions, we can drop k and think of J as a function of x alone because no matter where we are, we have an infinite number of stages over which our objective is computed.

For most problems, it turns out that the optimal policy is also stationary! That is, $\pi = \{\mu, \mu, \ldots\}$. So we can simply write $J_\mu(x)$ as the cost of the policy instead of $J_\pi(x)$.

Thus, unlike the finite horizon case, we need not find an infinite number of functions $J_k^*(x_k)$ and $\mu_k^*(x_k)$ but just compute $J^*(x)$ and $\mu^*(x)$.

Introduction to Infinite Horizon MDPs

A New DP Algorithm

Earlier, we used the DP algorithm to find the optimal values and policies. But now, we have a problem! There is no terminal state.

We'll still try to do something similar and develop a recursive set of equations which look like

$$J_0(x) = 0 \,\forall\, x \in X$$

$$J_{k+1}(x) = \min_{u \in U(x)} \mathbb{E}\Big\{ g(x,u,w) + \alpha J_k(f(x,u,w)) \Big\}$$

The key idea will be to solve this recursively and after a large number of iterations, we'll get the optimal value functions. (In fact, we'll show that the initial conditions don't matter!)

Note that we've completely dropped subscripts for several terms since unlike the objective, which has an infinite summation, we are just dealing with one transition.

What is odd about the above algorithm? How is it different from the finite horizon version?

Introduction to Infinite Horizon MDPs

A New DP Algorithm

The new DP recursive equations have k and $k+1$ switched and they seem to imply that it is forward recursive. But it is not. The subscripts have a different meaning.

Consider a N-stage sub-problem of the infinite horizon case. We'll assume that the terminal cost is $\alpha^N J(x_N)$ since it has to be discounted at the start of the process, where $J(x)$ is some known function. The objective of this finite horizon model is

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

and the Bellman equations can be written as

$$\hat{J}_N(x) = \alpha^N J(x)$$

$$\hat{J}_k(x) = \min_{u \in U(x)} \mathbb{E}_w \left\{ \alpha^k g(x, u, w) + \hat{J}_{k+1}(f(x, u, w)) \right\}$$

$$\forall\, k = N-1, \ldots, 1, 0$$

Substitute $N-k$ for k in the second equation.

$$\hat{J}_{N-k}(x) = \min_{u \in U(x)} \mathbb{E}_w \left\{ \alpha^{N-k} g(x,u,w) + \hat{J}_{N-k+1}(f(x,u,w)) \right\}$$

Divide both sides by α^{N-k} and set

$$J_k(x) = \frac{\hat{J}_{N-k}(x)}{\alpha^{N-k}}$$

The new Bellman equations thus become,

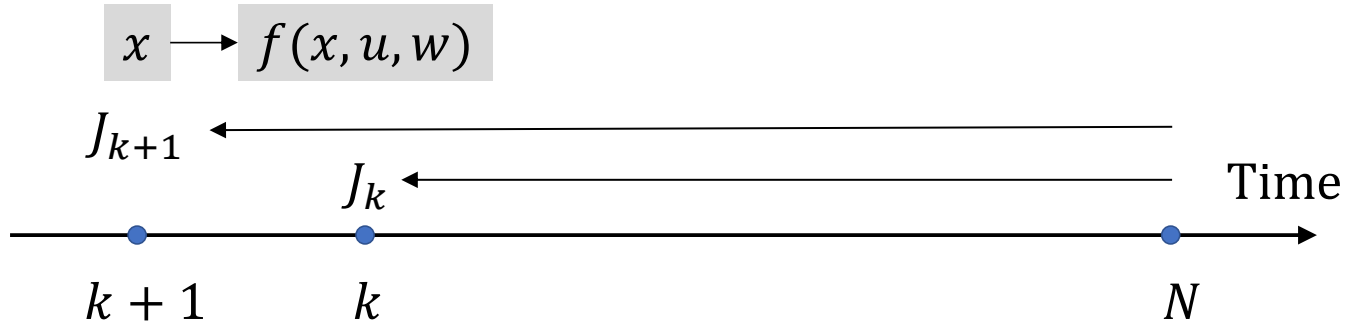
$$J_0(x) = J(x) \, \forall \, x \in X$$

$$J_{k+1}(x) = \min_{u \in U(x)} \mathbb{E} \left\{ g(x,u,w) + \alpha J_k(f(x,u,w)) \right\}$$

Introduction to Infinite Horizon MDPs

A New DP Algorithm

Time is now measured backward from some N which tends to ∞.



Thus, after N iterations, we would have found the optimal cost for the N-stage discounted problem with terminal cost function $\alpha^N J$.

If we stop the algorithm after k iterations, we would have found the optimal cost for the k-stage discounted problem with terminal cost function $\alpha^k J$.

Life can only be understood backwards; but it must be lived forwards – Søren Kierkegaard

Introduction to Infinite Horizon MDPs

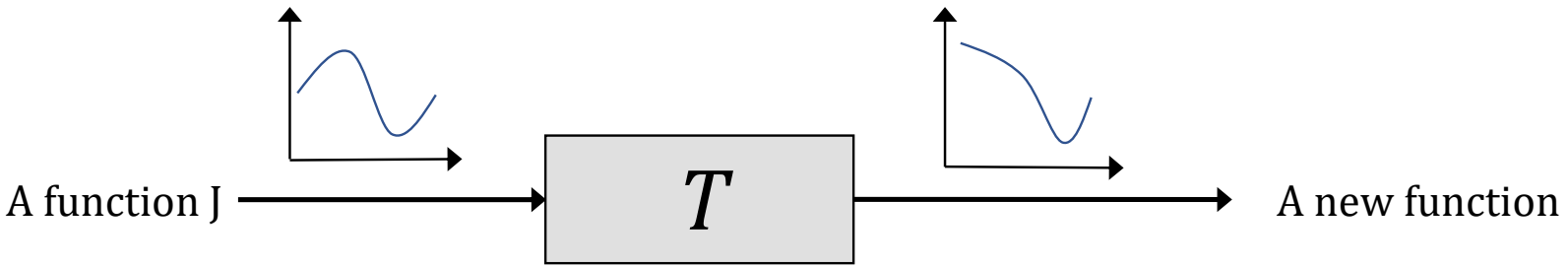
T Operator

We will define a couple of mappings that allow a compact representation of the algorithms and proofs for infinite horizon problems.

Definition

Given a function $J : X \to \mathbb{R}$, define $(TJ)(x)$ as

$$(TJ)(x) = \min_{u \in U(x)} \mathbb{E}\Big\{ g(x,u,w) + \alpha J(f(x,u,w)) \Big\}$$



This mapping is equivalent to one iteration of the new DP algorithm.

Introduction to Infinite Horizon MDPs

T Operator

We will also define an analogous operator for a given policy without minimization

Definition

Given a function $J : X \to \mathbb{R}$, define $(T_\mu J)(x)$ as

$$(T_\mu J)(x) = \mathbb{E}\Big\{ g(x, \mu(x), w) + \alpha J(f(x, \mu(x), w)) \Big\}$$

Introduction to Infinite Horizon MDPs

T Operator

We can also define composition mappings

$$(T^0 J)(x) = J(x) \, \forall \, x \in X$$
$$(T^k J)(x) = (T(T^{k-1} J))(x) \, \forall \, x \in X$$

How can we interpret $(T^k J)(x)$? It is equivalent to k iterations of the new DP algorithm and is hence the optimal cost of the k-stage discounted problem with terminal costs $\alpha^k J$.

Likewise,

$$(T_\mu^0 J)(x) = J(x) \, \forall \, x \in X$$
$$(T_\mu^k J)(x) = (T_\mu(T_\mu^{k-1} J))(x) \, \forall \, x \in X$$

$(T_\mu^k J)(x)$ is the cost of stationary policy μ for the k-stage discounted problem with terminal costs $\alpha^k J$.

Alternate Notation

Alternate Notation

Countable State Spaces

We will mostly deal with countable state, control, and disturbance spaces. In such cases, we can write the DP equations and the T operators in more compact form.

Suppose the state space is $X = \{1, \ldots, n\}$. The transitions no longer are a function of k and hence we can write

$$p_{ij}(u) = \mathbb{P}[x_{k+1} = j | x_k = i, u_k = u] \, \forall \, i, j \in X, u \in U(i)$$

The two T mappings take the form

$$(TJ)(i) = \min_{u \in U(i)} \left\{ g(i,u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j) \right\} \forall \, i \in X$$

$$(T_\mu J)(i) = \left\{ g(i,\mu(i)) + \alpha \sum_{j=1}^{n} p_{ij}(\mu(i)) J(j) \right\} \forall \, i \in X$$

Note that it has been implicitly assumed that g does not depend on the disturbance. How can we relax that?

Alternate Notation

Countable State Spaces

One can also write vector forms of these equations.

$$J = \begin{pmatrix} J(1) \\ \vdots \\ J(n) \end{pmatrix} \qquad TJ = \begin{pmatrix} (TJ)(1) \\ \vdots \\ (TJ)(n) \end{pmatrix} \qquad T_\mu J = \begin{pmatrix} (T_\mu J)(1) \\ \vdots \\ (T_\mu J)(n) \end{pmatrix}$$

For a given policy μ, we can also write the one-step transition probability matrix as

$$P_\mu = \begin{pmatrix} p_{11}(\mu(1)) & \dots & p_{1n}(\mu(1)) \\ \vdots & \ddots & \vdots \\ p_{n1}(\mu(n)) & \dots & p_{nn}(\mu(n)) \end{pmatrix}$$

and the cost vector for a fixed policy μ as

$$g_\mu = \begin{pmatrix} g(1, \mu(1)) \\ \vdots \\ g(n, \mu(n)) \end{pmatrix}$$

Thus, the T-mu operator in matrix form can be written as

$$T_\mu J = g_\mu + \alpha P_\mu J$$

Alternate Notation

Road Map

With these definitions, we would like to prove the following wish list:

1. $J^*(x) = \lim_{k\to\infty}(T^k J)(x) \,\forall\, x \in X$ for any bounded function J.
2. $J^* = TJ^*$, i.e., J^* is a fixed point of the mapping T.
3. If $\mu(x)$ attains the minimum in the RHS of the above equation, then it is optimal.

Your Moment of Zen