CE 273
Markov Decision Processes

Lecture 12

Solution Methods for Total Cost MDPs

Previously on Markov Decision Processes

Proposition (Policy Improvement Property (PIP))

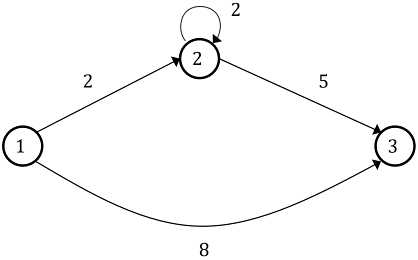
Let μ and μ' be stationary policies such that $T_{\mu'} J_\mu = TJ_\mu$. Then,

$$J_{\mu'}(i) \leq J_\mu(i) \,\forall\, i = 1, \ldots, n$$

Furthermore, if μ is not optimal, strict inequality holds for at least one i.

Previously on Markov Decision Processes

Even when costs are not involved, the notion of discounting may hold water since it is human nature to place more weight on short-term costs/rewards. But discounting doesn't always make sense.

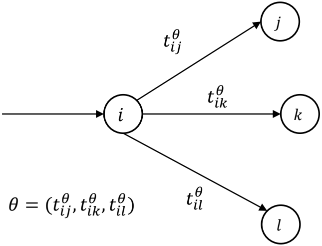


With a discount factor of say $\alpha = 0.5$, it is always optimal to cycle one more time and we'd never reach the destination!

Previously on Markov Decision Processes

Suppose the graph is represented by $G = (N, A)$, where N is the set of nodes and A is the set of links/arcs.

Upon arriving at a node i, a traveler observes a information vector $\theta \in \Theta_i$ drawn with probability q^θ informing him or her of the travel time of each link leaving node i.



Thus, the states are tuples (i, θ). Policies are functions $\mu(i, \theta)$ which tell us which node to go to next. Note that this is a problem with uncontrollable state components like Tetris.

Previously on Markov Decision Processes

We can simplify the problem by defining ex ante value functions (this the value at a node before we observe the information vector) using $\hat{J}(i) = \sum_{\theta \in \Theta_i} q^\theta J(i,\theta)$
Thus, one can hypothesize that the value iteration algorithm looks like

$$\hat{J}_{k+1}(i) = \sum_{\theta \in \Theta_i} q^\theta \min_{j \in \Gamma(i)} \left\{ t_{ij}^\theta + \hat{J}_k(j) \right\}$$

with $\hat{J}_k(t) = 0$ for all k. The problem is easier to solve because we have a smaller number of states. Let's say this algorithm converges. How do we find the optimal policies from $\hat{J}^*$?

$$\mu^*(i,\theta) = \arg\min_{j \in \Gamma(i)} \left\{ t_{ij}^\theta + \hat{J}^*(j) \right\}$$

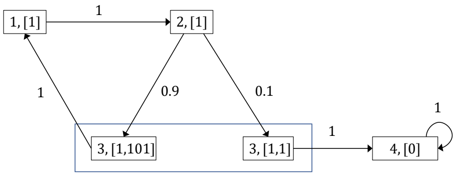
Likewise, given a policy μ, we expect the ex ante cost of the policy μ to be a solution to

$$\hat{J}_\mu(i) = \sum_{\theta \in \Theta_i} q^\theta \left\{ t_{i,\mu(i,\theta)}^\theta + \hat{J}_\mu\left(\mu(i,\theta)\right) \right\}$$

and $\hat{J}_\mu(t) = 0$.

Previously on Markov Decision Processes

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



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

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.

Suppose, J_0 represents a vector of zeros. What happens when we apply the T_μ repeatedly? We would accumulate the one-step costs and hence get the total cost of associated Markov chain.

$$J_\mu = \lim_{N \to \infty} \sum_{k=0}^{N-1} P_\mu^k g_\mu$$

We will soon extend this by proving that we get J_μ by applying T_μ repeatedly on **any** initial guess J_0.

Previously on Markov Decision Processes

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

Since, $J_\mu = \lim_{N\to\infty} \sum_{k=0}^{N-1} P_\mu^k g_\mu$, the second assumption implies that some component of $\sum_{k=0}^{N-1} P_\mu^k g_\mu$ diverges to ∞ as $N \to \infty$.

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

1. Main Results
2. VI, PI, and LP Methods

Main Results

Main Results

Constant Shift Lemma

Suppose $e : X \to \mathbb{R}$ denotes the unit function that takes a value 1 for all i and let r be a **positive** scalar.

$$\begin{aligned}\big(T(J - re)\big)(i) &= \min_{u \in U(i)} \mathbb{E}\Big\{g(i,u) + \sum_{j=1}^{n} p_{ij}(u)(J - re)(j)\Big\} \\ &= \min_{u \in U(x)} \mathbb{E}\Big\{g(i,u) + \sum_{j=1}^{n} p_{ij}(u)J(j) - r\sum_{j=1}^{n} p_{ij}(u)\Big\} \\ &\geq \big(TJ\big)(i) - r\end{aligned}$$

Similarly, we can show $\big(T_\mu(J - re)\big)(i) \geq (T_\mu J)(i) - r$.

Main Results

Properties of Proper Policies

Proposition

For a proper policy μ, J_μ satisfies

$$\lim_{k\to\infty} (T_\mu^k J)(i) = J_\mu(i) \,\forall\, i = 1, \ldots, n$$

for every J. Further, J_μ is the unique fixed point of T_μ

Proof.

By definition of T_μ, $T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu$,

Taking limits as $k \to \infty$,

$$\begin{aligned}\lim_{k\to\infty} T_\mu^k J &= \lim_{k\to\infty} P_\mu^k J + \lim_{k\to\infty} \sum_{m=0}^{k-1} P_\mu^m g_\mu \\ &= 0 + J_\mu\end{aligned}$$

Also by definition of T_μ, $T_\mu^{k+1} J = g_\mu + P_\mu T_\mu^k J$. Taking limits as $k \to \infty$, $J_\mu = T_\mu J_\mu$. Proof of uniqueness is left as an exercise. ■

Main Results

Properties of Proper Policies

Proposition

A policy μ is proper $\Leftrightarrow J \geq T_\mu J$ for some vector J.

Proof.

($\Rightarrow$) Set $J = J_\mu$ and use the previous proposition.

($\Leftarrow$) Suppose not. If $J \geq T_\mu J$, using the monotonicity lemma,

$$J \geq T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m g_\mu$$

Since μ is not proper, taking limits as $k \to \infty$, at least one component of the RHS diverges (Assumption 2) but the LHS is J. A contradiction. ∎

We will now prove results that will allow us to use VI and PI.

Main Results

Properties of Proper Policies

Proposition (Proper Policy Improvement Property (PPIP))

Let μ be a proper policy with total cost J_μ. Choose μ' such that $T_{\mu'} J_\mu = TJ_\mu$. Then μ' is proper and

$$J_{\mu'} \leq J_\mu$$

Proof.

By definition of T and T_μ,

$$TJ_\mu \leq T_\mu J_\mu$$

J_μ is a fixed point of T_μ i.e., $T_\mu J_\mu = J_\mu$. Therefore,

$$T_{\mu'} J_\mu = TJ_\mu \leq T_\mu J_\mu = J_\mu$$

Since $J_\mu \geq T_{\mu'} J_\mu$, previous proposition implies μ' is proper. Since $T_{\mu'}$ mapping is monotonic,

$$J_\mu \geq T_{\mu'} J_\mu \geq T^2_{\mu'} J_\mu \geq \ldots \geq J_{\mu'}$$

■

Main Results

Properties of Proper Policies

Corollary

T has a unique fixed point

Proof.

Suppose we construct a sequence of policies $\{\mu_k\}$ using PPIP. Then,

$$J_{\mu_{k+1}} \leq TJ_{\mu_k} \leq J_{\mu_k}$$

Since the number of policies are finite, for some k', $\mu_{k'+1} = \mu_{k'}$ and thus $TJ_{\mu_{k'}} = J_{\mu_{k'}}$. Therefore, T has a fixed point.

Proof of uniqueness (Exercise). ■

Main Results

Bellman Equations

Theorem

1. *The optimal cost J^* is the unique solution to $J^* = TJ^*$*
2. *For any vector J, $\lim_{k\to\infty}(T^k J)(i) = J^*(i)\,\forall\, i = 1, \dots, n$*

Proof.

The earlier proposition established that T has a unique fixed point, which is also the cost of a proper policy. Let's call it J_μ. We will show that $J_\mu = J^*$ and $T^k J \to J^*$.

Let e be a vector of 1's and r be a positive scalar. Imagine a function $\hat{J}$ which satisfies,

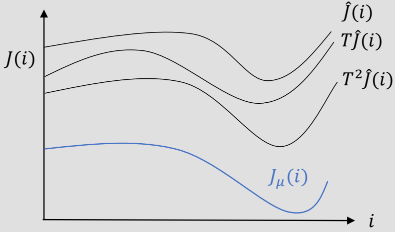
$$T_\mu \hat{J} = \hat{J} - re$$

A solution to the above equation must be unique and $J_\mu \leq \hat{J}$ (Why?). Thus, we may write

$$J_\mu = TJ_\mu \leq T\hat{J} \leq T_\mu \hat{J} = \hat{J} - re \leq \hat{J}$$
$$\Rightarrow J_\mu = T^k J_\mu \leq T^k \hat{J} \leq T^{k-1}\hat{J} \leq \hat{J}$$

Proof.

Therefore $T^k \hat{J} \to \tilde{J}$, but will it go all the way to J_μ?



Since $\lim_{k\to\infty} T^k \hat{J} = \tilde{J}$, we can write $T\tilde{J}$ as $T(\lim_{k\to\infty} T^k \hat{J})$.

T is still a piecewise concave and continuous function, so we can interchange the limit. Hence, $T\tilde{J} = \tilde{J} \Rightarrow \tilde{J} = J_\mu$ (Why?).

Using this result and constant shift lemma,

$$J_\mu - re = TJ_\mu - re \leq T(J_\mu - re) \leq TJ_\mu = J_\mu$$

Thus $T^k(J_\mu - re)$ monotonically increases and is bounded above by J_μ. We can show as before that it will converge to J_μ.

Main Results

Bellman Equations

Proof.

In summary, we saw that the functions $\hat{J}$, $T\hat{J}, \ldots$ converge to J_μ from above and $J_\mu - re$, $T(J_\mu - re), \ldots$ converge to J_μ from below.

From one of the previous inequalities,

$$J_\mu = TJ_\mu \leq T\hat{J} \leq T_\mu \hat{J} = \hat{J} - re \leq \hat{J}$$

$\Rightarrow \hat{J} \geq J_\mu + re$. Therefore, given any J, pick an $r > 0$, such that

$$J_\mu - re \leq J \leq J_\mu + re \leq \hat{J}$$

and apply the T mapping recursively and use the sandwich theorem to conclude

$$\lim_{k\to\infty} T^k J = J_\mu$$

So far, we've shown that the limit exists from any guess J and is equal to J_μ, the fixed point of T. We are yet to show $J_\mu = J^*$.

Proof.

Choose any arbitrary vector, say the zero vector J_0 and some proper policy μ'. By definition of T and $T_{\mu'}$,

$$TJ_0 \leq T_{\mu'} J_0$$

Using monotonicity lemma and taking limits as $k \to \infty$,

$$J_\mu \leq J_{\mu'}$$

Thus, μ must be optimal and $J_\mu = J^*$ ∎

It is easy to also show that

Theorem

A stationary policy μ is optimal if and only if

$$T_\mu J^* = TJ^*$$

Lecture Outline

VI, PI, and LP Methods

The previous set of results ensure that the value iteration algorithm converges from **any** initial guess J.

In general, VI converges in the limit. Under special circumstances, it is possible to guarantee convergence after a finite number of steps. (Think of deterministic shortest paths.)

Proposition

If the transition diagram associated with the optima policy is acyclic, then VI will converge to J^* after at most n iterations when initialized with $J(i) = \infty \,\forall\, i = 1, \ldots, n$.

One can also formally prove that asynchronous VI also converges with just the assumptions made so far.

VI, PI, and LP Methods

Policy Iteration

As seen earlier, PPIP can be used to develop a PI method. Start with a proper policy and repeatedly evaluate and improve it till policies obtained in successive iterations are the same.

Note that unlike the discounted case, **we must start with a proper policy** when using PI.

Unfortunately, the PPIP result does not extend to the modified policy iteration method. This method can result in improper policies even when we start with a proper policy.

VI, PI, and LP Methods

Linear Programming

Recall that the LP approach for the discounted case was based on the observation that:

Among all functions J that satisfy $J \leq TJ$, J^* is the "largest"

This holds true for the total cost problem as well because of the monotonicity of the T mapping. Hence we can find the optimal value functions using

$$\max_{i} \sum_{i=1}^{n} a_i y(i)$$

$$\text{s.t. } y(i) \leq g(i,u) + \sum_{j=1}^{n} p_{ij}(u) y(j) \qquad \forall\, i = 1, \ldots, n, u \in U(i)$$

Your Moment of Zen