CE 273
Markov Decision Processes

Lecture 15

Value Iteration for Average Cost MDPs

Previously on Markov Decision Processes

Definition

The gain J_μ of a policy μ is defined as

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

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.

Previously on Markov Decision Processes

Unlike discounted and total cost MDPs, where we could solve a system of equations for a given policy (and use this in the policy iteration algorithm), we cannot simply solve

$$J = P_\mu J$$
$$J + h = g_\mu + P_\mu h$$

to get the average cost of policy μ. (Why?)

If (J_μ, h_μ) solves the above system, then $(J_\mu, h_\mu + \text{constant})$ also satisfies the above system. Hence, there are an infinite number of solutions. We will call these policy evaluation equations for easy referencing.

In general, it can be shown that all solutions to the above system are of the form $(J_\mu, h_\mu + d)$, where $d = P_\mu d$.

Previously on Markov Decision Processes

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

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

Definition

A stationary policy μ is said to be Blackwell optimal if it is optimal for all α-discounted problems with $\alpha \in (\bar{\alpha}, 1)$, where $0 < \bar{\alpha} < 1$

A Blackwell optimal policy is optimal to the average cost problem when we restrict our attention to stationary policies.

Previously on Markov Decision Processes

Proposition

1. *All Blackwell optimal policies have the same gain and bias*
2. *Let (J^*, h^*) be the gain-bias pair of a Blackwell optimal policy, then*

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) J^*(j) \, \forall \, i = 1, \ldots, n$$

Let $\bar{U}(i)$ be the set of controls that attain the minimum in the above equation.

$$J^*(i) + h^*(i) = \min_{u \in \bar{U}(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h^*(j) \right\} \forall \, i = 1, \ldots, n$$

If μ^* is Blackwell optimal, it attains the minimum in the RHS of the above two equations.

Previously on Markov Decision Processes

The earlier proposition and discussion established that a Blackwell optimal policy is optimal to the average cost problem.

Further, optimal policies were found to satisfy some equations which are the necessary conditions for optimality. It can also be shown that they are sufficient.

Proposition

If J' and h' satisfy the following pair of optimality equations

$$J(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) J(j) \, \forall \, i = 1, \ldots, n$$

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

where $\bar{U}(i)$ is the set of controls that attain the minimum in the above equation. Then, $J' = J^*$ is the optimal average cost vector.

Further, if a stationary policy μ attains the minimum in the above equations, then it is the optimal policy μ^*.

Previously on Markov Decision Processes

In summary, if the average cost is independent of the initial state, the following proposition is true

Proposition

If a scalar λ and a vector h satisfy

$$\lambda + h(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right\} \ \forall\, i = 1, \ldots n$$

then λ is the optimal average cost $J^*(i)$ for all i, i.e.,

$$\lambda = \min_{\mu} J_\mu(i) = J^*(i) \ \forall \ i = 1, \ldots, n$$

Further, if μ^* attains the minimum in the first expression, then $J_{\mu^*}(i) = \lambda \, \forall \, i$.

In shorthand, the first equation can be rewritten as $\lambda e + h = Th$. Think of this as being analogous to $J^* = TJ^*$ in the discounted world.

Previously on Markov Decision Processes

What if the state space can be divided into $C \cup \mathcal{C}$, where C is a recurrent class and $\mathcal{C}$ is the set of transient states for **every** policy?

MDPs which satisfy this property are called **Unichain MDPs** and the simplified optimality equations can be used in this case.

MDPs in which at least one policy results in two or more closed communicating classes and a transient class (possibly empty) are called **Multichain MDPs**.

The equal costs property does not hold in this case, but it still holds within each closed communicating class.

Lecture Outline

1. Value Iteration
2. VI for Unichain MDPs
3. VI for Multichain MDPs

Lecture Outline

Preliminaries

Value Iteration

Classification of MDPs Revisited

There are a few more ways to classify MDPs. For instance,

Definition (Communicating MDP)

An MDP is said to be communicating if for every pair of states i and j there exists some stationary policy μ under which $i \leftrightarrow j$.

Definition (Weakly Communicating MDP)

An MDP is said to be weakly communicating if there exists $C \subseteq S$ such that $i \leftrightarrow j \, \forall \, i, j \in C$ under some stationary policy μ and a possibly empty set $\mathcal{C} \subset S$ which is transient under every policy

The main difference in these MDPs is that the conditions have to be satisfied for some stationary policy whereas unichain and multichain definitions hold true for all policies.

There are subtle differences in results across these different types of MDPs. We will focus only on unichain and multichain models in this course.

Consider the mapping Th. How can we interpret this function? It is the optimal cost of the undiscounted 1-stage problem with a terminal cost function h.

Likewise, $T^k h$ is the optimal cost of the undiscounted k-stage problem with a terminal cost function h.

Hence, one would expect that the average cost problem can be solved by computing the limit of $\frac{1}{k} T^k h$.

This is in fact true. To prove this, we'll first try to bound $T^k h$.

Value Iteration

Preliminaries

Proposition

Let J^* be the optimal average cost and $\hat{h}$ satisfy $J^* + \hat{h} = T\hat{h}$. Then for any h,

$$\min_j \left\{ h(j) - \hat{h}(j) \right\} + kJ^*(i) + \hat{h}(i) \leq (T^k h)(i) \leq \max_j \left\{ h(j) - \hat{h}(j) \right\} + kJ^*(i) + \hat{h}(i)$$

Proof.

Let's prove the RHS of the above inequality. Consider a policy μ.

$$\begin{aligned} T_\mu h - T_\mu \hat{h} &= g_\mu + P_\mu h - g_\mu - P_\mu \hat{h} \\ &= P_\mu (h - \hat{h}) \end{aligned}$$

In a similar manner, we can write

$$\begin{aligned} & T_\mu^k h - T_\mu^k \hat{h} = P_\mu^k (h - \hat{h}) \\ \Rightarrow & T_\mu^k h - T_\mu^k \hat{h} = P_\mu^k (h - \hat{h}) \leq \max_j \left\{ h(j) - \hat{h}(j) \right\} e \\ \Rightarrow & T^k h - T_{\mu^*}^k \hat{h} \leq \max_j \left\{ h(j) - \hat{h}(j) \right\} e \end{aligned}$$

Proof.

It is assumed that $\hat{h}$ satisfies $J^* + \hat{h} = T\hat{h} \leq T_\mu \hat{h}$. Applying T_μ on both sides and using the monotonicity lemma,

$$\begin{aligned} T_\mu^2 \hat{h} &\geq T_\mu (J^* + \hat{h}) \\ &\geq g_\mu + P_\mu (J^* + \hat{h}) \\ &\geq T_\mu \hat{h} + P_\mu J^* \\ &\geq T_\mu \hat{h} + J^* \\ &\geq J^* + \hat{h} + J^* = 2J^* + \hat{h} \end{aligned}$$

Similarly, it can be shown that $T_\mu^k \hat{h} \geq kJ^* + \hat{h}$. Equality occurs only if $\mu = \mu^*$, in which case we can write

$$T_{\mu^*}^k \hat{h} = kJ^* + \hat{h}$$

The RHS of the inequality follows. ■

Corollary

Given an optimal average cost vector J^* and $\hat{h}$ satisfying $T\hat{h} = J^* + \hat{h}$, for all k,

$$T^k \hat{h} = kJ^* + \hat{h}$$

Replace h with $\hat{h}$ in the previous proposition.

Value Iteration

Main Result

Theorem

$$J^* = \lim_{k\to\infty} \frac{1}{k} T^k h$$

Proof.

From the earlier proposition,

$$\min_j \left\{ h(j) - \hat{h}(j) \right\} + kJ^*(i) + \hat{h}(i) \leq (T^k h)(i)$$
$$\leq \max_j \left\{ h(j) - \hat{h}(j) \right\} + kJ^*(i) + \hat{h}(i)$$

The theorem follows from dividing both sides with k and taking limits as $k \to \infty$. ■

Note that no assumption on the underlying Markov chains or the type of MDP were made so far. This method works for any average cost problem.

Value Iteration

Drawbacks

While this method works for all average cost MDPs, the values of $T^k h$ keep increasing. (Why?)

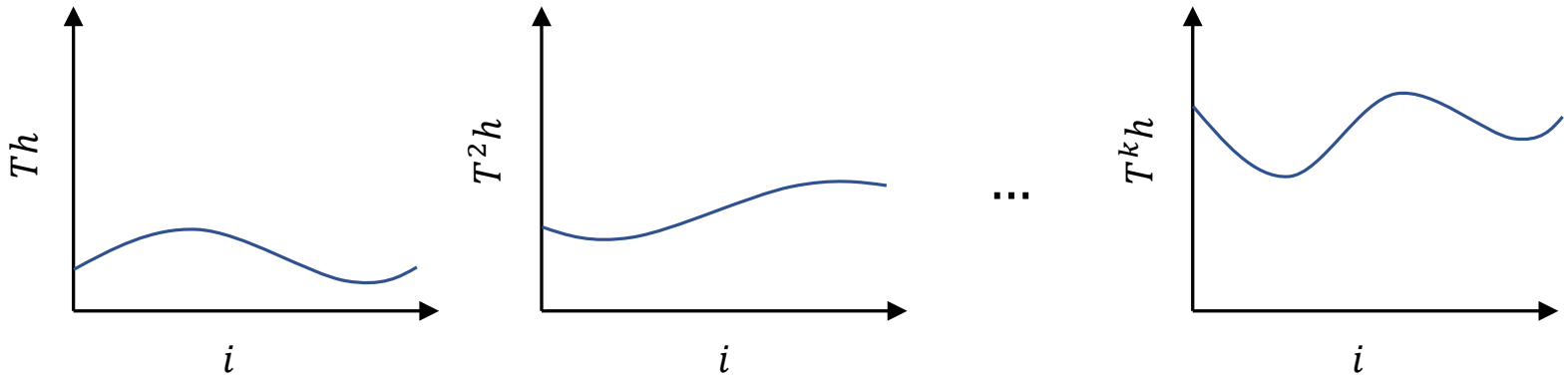
This poses a computational challenge as some components may diverge to ∞. This issue can be easily addressed for certain types of MDPs.

VI for Unichain MDPs

VI for Unichain MDPs

Introduction

Consider a unichain MDP. The optimal costs are equal for all starting states. The earlier discuss implies that the $T^k h$ values keep increasing.



They are also bounded above and below and the bounds are a function of k. We will try to show that the 'width' of the function or the difference between the max and min values of the function does not grow with k.

From the previous bounds on $T^k h$,

$$(T^k h)(i) \leq \max_i (T^k h)(i) \leq \max_i \left\{ \max_j \left\{ h(j) - \hat{h}(j) \right\} + kJ^*(i) + \hat{h}(i) \right\}$$

$$\leq \max_j \left\{ h(j) - \hat{h}(j) \right\} + kJ^*(i) + \max_i \hat{h}(i)$$

VI for Unichain MDPs

Introduction

Similarly, one can write

$$(T^k h)(i) \geq \min_i (T^k h)(i) \geq \min_i \left\{ \max_j \left\{ h(j) - \hat{h}(j) \right\} + kJ^*(i) + \hat{h}(i) \right\}$$
$$\geq \max_j \left\{ h(j) - \hat{h}(j) \right\} + kJ^*(i) + \min_i \hat{h}(i)$$

This implies

$$-(T^k h)(i) \leq -\min_i (T^k h)(i) \leq -\max_j \left\{ h(j) - \hat{h}(j) \right\} - kJ^*(i) - \min_i \hat{h}(i)$$

Adding the above inequalities,

$$0 \leq \max_i (T^k h)(i) - \min_i (T^k h)(i) \leq \max_i \hat{h}(i) - \min_i \hat{h}(i)$$

Thus, the width of $T^k h$ is independent of k!

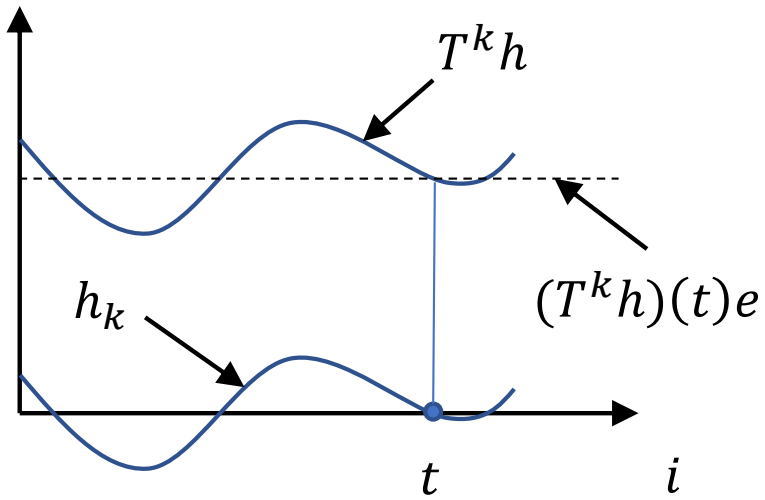
VI for Unichain MDPs

Relative Value Iteration

We can exploit this result by defining iterates of the form

$$h_k = T^k h - (T^k h)(t)e$$

where t is some arbitrary state. Effectively, we are shifting the entire function by a constant. But note that the constant varies across iterations.



From our previous discussion, the iterates h_k remain bounded and the bounds do not depend on k.

VI for Unichain MDPs

Relative Value Iteration

The h_k vector can be interpreted as the k-stage optimal cost relative to state t.

The iterates can also be written just using the T operator as

$$h_{k+1} = Th_k - (Th_k)(t)e$$

If this procedure converges to some h^*, then

$$h^* = Th^* - (Th^*)(t)e \Rightarrow (Th^*)(t)e + h^* = Th^*$$

From the optimality conditions of unichain MDPs, we know that $(Th^*)(t)$ is the optimal average cost and h^* is an associated bias vector.

It turns out that convergence is guaranteed only when each policy of the unichain MDP results in an aperiodic Markov chain.

VI for Unichain MDPs

Example 1

Consider the situation where there is only one policy μ for which the one-step costs and transition matrices are

$$g_\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad P_\mu \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Calculate the sequences of $\frac{1}{k} T^k h$ and $T^k h - (T^k h)(t)e$.

The first sequence converges but the second does not.

VI for Unichain MDPs

Algorithm

The relative value iteration (RVI) works for unichain MDPs which induce aperiodic Markov Chains.

Define the span semi-norm of a vector h as

$$sp(h) = \max_{i \in X} h(i) - \min_{i \in X} h(i)$$

Relative Value Iteration

Fix a tolerance level $\epsilon > 0$ and select a state t
Select $h_0 \in B(X)$ and $k \leftarrow 0$
$h_1 \leftarrow Th_0 - (Th_0)(t)e$
while $sp(h_{k+1} - h_k) > \epsilon$ **do**
    $k \leftarrow k + 1$
    $h_{k+1} \leftarrow Th_k - (Th_k)(t)e$
end while

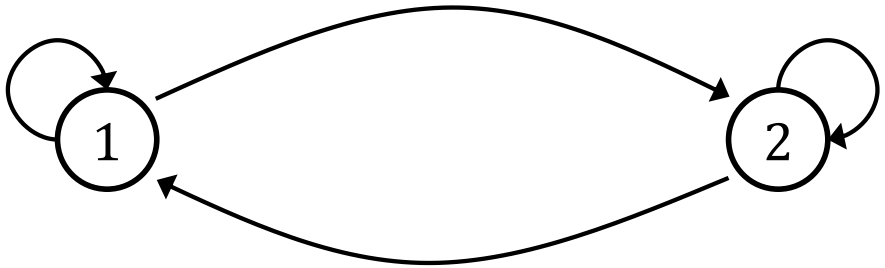
Select μ such that

$$\mu(i) \in \arg \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h_k(j) \right\}$$

VI for Unichain MDPs

Example 2

Perform three iterations of the RVI algorithm for the following example with two states 1 and 2. Use state 1 as the reference state t.



- $U(1) = \{u_1, u_2\}$
- $g(1, u_1) = 2, g(1, u_2) = 0.5$
- $p_{1j}(u_1) = [3/4 \;\; 1/4]$
- $p_{1j}(u_2) = [1/4 \;\; 3/4]$

- $U(2) = \{u_1, u_2\}$
- $g(2, u_1) = 1, g(2, u_2) = 3$
- $p_{2j}(u_1) = [3/4 \;\; 1/4]$
- $p_{2j}(u_2) = [1/4 \;\; 3/4]$

VI for Unichain MDPs

Periodic Markov Chains

When policies of a unichain MDP can induce periodic Markov chains, as seen earlier, RVI does not converge but will exhibit oscillatory behavior.

There are methods to address this issue which modify the transition matrix to wipe out the periodic nature.

VI for Multichain MDPs

VI for Multichain MDPs

Introduction

For the multi-chain case, we still have $\frac{1}{k} T^k h \to J^*$, but we cannot find an equivalent relative value iteration method.

If we find $h_k = T^k h$, the rate at which these iterates diverge is a function of the state i since there are multiple recurrent classes. For the multichain case, define a *residual sequence*

$$r_k = h_k - kJ^* = T^k h - kJ^*$$

Consider a vector $\hat{h}$ that satisfies $J^* + \hat{h} = T\hat{h}$. From the first proposition we saw today, $T^k \hat{h} = kJ^* + \hat{h}$.

Thus, r_k can be written as $\hat{h} + (T^k h - T^k \hat{h})$. It turns out that for problems in which each policy results in an aperiodic DTMC, $(T^k h - T^k \hat{h})$ converges.

Notice that $T^k h - T^k \hat{h}$ is the difference of the optimal k-stage problems with different terminal costs.

Since $r_k = h_k - kJ^*$, we can write

$$h_k = kJ^* + r_k$$
$$h_{k+1} = (k+1)J^* + r_{k+1}$$

Subtracting the above equations,

$$h_{k+1} - h_k = J^* + (r_{k+1} - r_k)$$

Since r_k converges, $(r_{k+1} - r_k) \to 0$. Thus, we can conclude that $h_{k+1} - h_k$ converges to J^*.

VI for Multichain MDPs

Main Result

Proposition

Let J^* be the optimal average cost and assume that the sequence $\{h_k\}$ is generated from the VI method $h_{k+1} = Th_k$. If every stationary policy results in an aperiodic DTMC, then

1. $J^* = \lim_{k\to\infty}(h_{k+1} - h_k)$
2. *The residual sequence $\{r_k\} \to r^*$. J^* and r^* satisfy the pair of optimality equations defined in the last class.*

Your Moment of Zen