B9140 Dynamic Programming and Reinforcement Learning Fall 2017

Prof. Daniel Russo September 27, 2017

Homework 2, Due in class Monday October 9

1 Discounting Is Equivalent To A Geometrically Distributed Horizon

Consider a special case of the (undiscounted) MDPs with indefinite horizon we discussed in Class 2. The state space is $X \cup \{t\}$ where t is a terminal state. In particular, t is costless ($g(t,u) = 0$) and absorbing ($P(t,u,t) = 1$). Assume that for all $x \in X$, and controls u $P(x,u,t) = 1-\gamma$. That is, the time of termination is geometrically distributed with parameter $1-\gamma$ and is independent of the choice of policy. The agent's goal is to minimize cumulative (undiscounted) expected costs incurred prior to termination.

Since t is costless and absorbing, it natural to only track the cost to go for non-terminal states $x \in X$. For any $J \in \mathbb{R}^{|X|}$ define

$$(T_\mu J)(x) = g(x,\mu(x)) + \sum_{x' \in X} P(x,\mu(x),x')J(x')$$

and

$$(TJ)(x) = \min_{u \in U(x)} g(x,u) + \sum_{x' \in X} P(x,u,x')J(x')$$

Argue that these Bellman operators are equivalent to those in a discounted MDP with state space X (and no terminal state), but with transition probabilities given by $\tilde{P}(x,u,x') = P(x,u,x')/(1-\gamma)$. Using this, argue that any policy has the same cost-to-go function in the MDP with geometric as in the discounted MDP with infinite horizon.

2 Planning With An Approximate Cost-To-Go Function

Consider an MDP with discount factor $\gamma < 1$ and n states (i.e. $|X| = n$). Let $J \in \mathbb{R}^n$ be an approximate cost to go function. Consider a policy μ that satisfies $T_\mu J = TJ$. Show that

$$\|J_\mu - J\|_\infty \leq \frac{2\gamma \|J - J^*\|_\infty}{1-\gamma}$$

Interpretation: The policy μ is the result of optimizing performance as if J were the true cost-to-go function J^*. The result here argues that if J^* is close to J then μ must be near-optimal.

Hint: it is possible to prove this result using only that T and T_μ are contractions with respect to the same norm $\|\cdot\|_\infty$. You may wish to look back at how we proved a similar error bound in terms of $\|J - TJ\|_\infty$.

$$M^k J = J^* + \sum_{i=1} \xi_i \lambda_i \qquad \text{Re}_i,$$

so that if λ is a dominant eigenvalue and $\lambda_1, \ldots, \lambda_{n-1}$ lie within the unit circle, $M^k J$ converges to J^* at a rate governed by the subdominant eigenvalue. *Note*: This result can be generalized for the case where Q does not have a full set of linearly independent eigenvectors, and for the case where F is modified through multiple-rank corrections [Ber95a].

2.5 (A Bad Example for Policy Iteration)

This problem provides an example (privately communicated by J. Tsitsiklis) where PI is very slow because it switches the control of only one state at each iteration. Consider a deterministic discounted problem with states $1, \ldots, n$. States 1 and n are absorbing under any control. At states $i = 2, \ldots, n-2$, there are two controls: one that moves to $i-1$ at cost 1 and another that moves to $i+1$ at cost 2. At state $n-1$, there are two controls: one that moves to $n-2$ at cost 1 and another that moves to n at cost $-2n$. Let the initial policy be to move from each state $i = 2, \ldots, n-1$ to $i-1$. Show that for a discount factor α that is sufficiently close to 1 the optimal policy is to move from i to $i+1$ for all states $i = 2, \ldots, n-1$, and the PI method will require $n-1$ iterations to find it.

2.6 (Error Bound for One-Step Lookahead)

3 A Bad Example For Policy Iteration

Do question 2.5 of Bertsekas Volume II, which is reproduced below.

Bonus question: This paper suggests policy iterations stops after a number of iterations that scales at most linearly with the number of actions and logarithmically with the number of states. How do we reconcile this with the example in problem 2.5? (I have not thought about this much, but I do not know the answer...the paper could have a mistake)

4 Policy Iteration is Newton's Method

Do question 2.8 of Bertsekas Volume II, which is reproduced below.

$j=1$

where $h : \Re \mapsto \Re$ is a scalar function. Derive conditions on h and interesting special cases under which $\hat{H}$ satisfies the Contraction and Monotonicity Assumptions of Section 2.5, and we also have $\hat{J}^* = J^*$ [cf. part (b)].

2.8 (Policy Iteration and Newton's Method)

The purpose of this exercise is to demonstrate a relation between PI and Newton's method for solving nonlinear equations. Consider an equation of the form $F(J) = 0$, where $F : \Re^n \mapsto \Re^n$. Given a vector $J_k \in \Re^n$, Newton's method determines J_{k+1} by solving the linear system of equations

$$F(J_k) + \frac{\partial F(J_k)}{\partial J}(J_{k+1} - J_k) = 0,$$

where $\partial F(J_k)/\partial J$ is the Jacobian matrix of F evaluated at J_k.

(a) Consider the n-state discounted MDP of Section 2.1 and define

$$F(J) = TJ - J.$$

Show that if there is a unique μ such that

$$T_\mu J = TJ,$$

then the Jacobian matrix of F at J is

$$\frac{\partial F(J)}{\partial J} = \alpha P_\mu - I,$$

where I is the $n \times n$ identity.

(b) Show that the PI algorithm can be identified with Newton's method for solving $F(J) = 0$ (assuming it gives a unique policy at each step).