ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institute for
Dynamic Systems and Control
IDSC
Institut für Dynamische Systeme
und Regelungstechnik

Dynamic Programming and Optimal Control

Fall 2021

Problem Set:
Infinite Horizon Problems, Value Iteration, Policy Iteration

Notes:

- Problem marked with BERTSEKAS are taken from the book *Dynamic Programming and Optimal Control by Dimitri P. Bertsekas, Vol. I, 3rd edition, 2005, 558 pages, hardcover.*
- The solutions were derived by teaching assistants. Please report any error that you may find on Piazza forum.

Problem Set 2

Problem 1 (BERTSEKAS, p. 445, exercise 7.1, Lecture 4)

A tennis player has a Fast serve and a Slow serve, denoted F and S respectively. The probability of F (or S) landing in bounds is p_F (or p_S, respectively). The probability of winning the point assuming the serve landed in bounds is q_F (or q_S, respectively). We assume that $p_F < p_S$ and $q_F > q_S$. The problem is to find the serve to be used at each possible scoring situation during a single game in order to maximize the probability of winning that game.

1. Formulate this as a stochastic shortest path problem, and write Bellman's equation.

2. Computer assignment: Assume that $q_F = 0.6$, $q_S = 0.4$, and $p_S = 0.95$. Use value iteration to calculate and plot (in increment of 0.05) the probability of the server winning a game with optimal serve selection as a function of p_F.

Problem 2 (BERTSEKAS, p. 446, exercise 7.3, Lecture 5)

A computer manufacturer can be in one of two states. In state 1 his product sells well, while in state 2 his product sells poorly. While in state 1 he can advertise his product in which case the one-stage reward is 4 units, and the transition probabilities are $P_{11} = 0.8$ and $P_{12} = 0.2$. If in state 1, he does not advertise, the reward is 6 units and the transition probabilities are $P_{11} = P_{12} = 0.5$. While in state 2, he can do research to improve his product, in which case the one-stage reward is -5 units, and the transition probabilities are $P_{21} = 0.7$ and $P_{22} = 0.3$. If in state 2 he does not do research, the reward is -3, and the transition probabilities are $P_{21} = 0.4$ and $P_{22} = 0.6$. Consider the infinite horizon, discounted version of the problem

1. Show that when the discount factor α is sufficiently small, the computer manufacturer should follow the "shortsighted" policy of not advertising (not doing research) while in state 1 (state 2). By contrast, when α is sufficiently close to unity, he should follow the "farsighted" policy of advertising (doing research) while in state 1 (state 2).

2. For $\alpha = 0.9$ calculate the optimal policy using policy iteration.

3. For $\alpha = 0.99$, use a computer to solve the problem by value iteration.

Problem 3 (BERTSEKAS, p. 60, exercise 1.23, Lecture 5)

An evident, yet very important property of the DP algorithm is that if the terminal cost g_N is changed to a uniformly larger cost $\overline{g}_N$ (i.e., $g_N(x_N) \leq \overline{g}_N(x_N)$ for all x_N), then the last stage cost-to-go $J_{N-1}(x_{N-1})$ will be uniformly increased. More generally, given two functions J_{k+1} and $\overline{J}_{k+1}$ with $J_{k+1}(x_{k+1}) \leq \overline{J}_{k+1}(x_{k+1})$ for all x_{k+1}, we have, for all x_k and $u_k \in \mathcal{U}_k(x_k)$

$$\underset{(w_k|x_k,u_k)}{\mathrm{E}} \left[g_k(x_k, u_k, w_k) + J_{k+1}\big(f_k(x_k, u_k, w_k)\big)\right] \leq \underset{(w_k|x_k,u_k)}{\mathrm{E}} \left[g_k(x_k, u_k, w_k) + \overline{J}_{k+1}\big(f_k(x_k, u_k, w_k)\big)\right]$$

Suppose now that in the basic problem the system and cost are time invariant; that is, $\mathcal{S}_k \equiv \mathcal{S}$, $\mathcal{U}_k(x_k) \equiv \mathcal{U}(x_k)$ for all $x_k \in \mathcal{S}$, $f_k(\,\cdot\,,\,\cdot\,,\,\cdot\,) \equiv f(\,\cdot\,,\,\cdot\,,\,\cdot\,)$, $w_k \sim p_{w|x,u}$ and $g_k(\,\cdot\,,\,\cdot\,,\,\cdot\,) \equiv g(\,\cdot\,,\,\cdot\,,\,\cdot\,)$ for some $\mathcal{S}, \mathcal{U}(\,\cdot\,)$, $f(\,\cdot\,,\,\cdot\,,\,\cdot\,)$, and g. Show that if in the DP algorithm we have $J_{N-1}(x) \leq J_N(x)$ for all $x \in \mathcal{S}$, then

$$J_k(x) \leq J_{k+1}(x) \quad \text{for all } x \in \mathcal{S} \text{ and } k.$$

Similarly, if we have $J_{N-1}(x) \geq J_N(x)$ for all $x \in \mathcal{S}$, then

$$J_k(x) \leq J_{k-1}(x) \quad \text{for all } x \in \mathcal{S} \text{ and } k.$$