Ma 661 Dynamic Programming and Reinforcement Learning
Darinka Dentcheva darinka.dentcheva@stevens.edu

Homework 2
Solutions

Problem 1. Assume that you play a chess match with a friend. If you play timid your probability of making a draw is $p = 0.9$, the probability to win is 0 and the probability to lose is 0.1. If you play bold you either win with probability $q = 0.45$, or you lose. Each win brings one point to the score of the winer. The match consists of 5 games. If the score is a tie after the fifth game, then a "sudden death" rule is adopted; that is, whoever wins the next game is a winner of the math; if it is a draw, then the game is repeated with the same rule.

Formulate a Markov decision problem to determine the optimal strategy of your play (to maximize the probability of winning the match) and solve it. Clearly describe the state space, control space, transition probabilities, and the reward function.

Solution

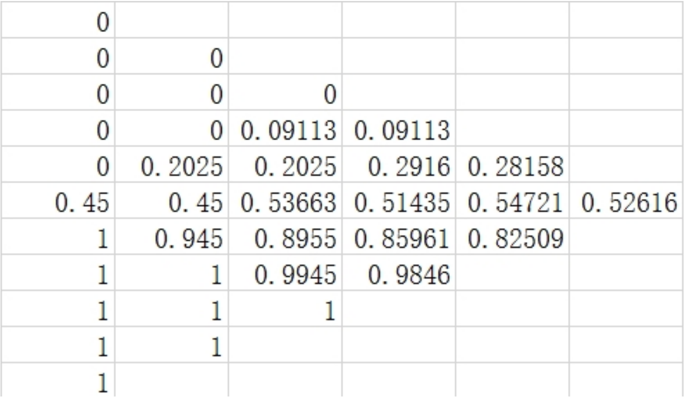
- State space: $\mathscr{X} = \{-5,-4,-3,-2,-1,0,1,2,3,4,5\}$ - your advantage at the moment.;
- Control space: $\mathscr{U} = \{0 \text{ (timid) }, 1 \text{ (bold)}\}$
- Transition probabilities:

$$\begin{gathered}
P\{x_{t+1} = x_t - 1 \,|\, x_t\,, 0\} = 0.1, \qquad P\{x_{t+1} = x_t \,|\, x_t\,, 0\} = 0.9 \\
P\{x_{t+1} = x_t - 1 \,|\, x_t\,, 1\} = 0.55, \qquad P\{x_{t+1} = x_t + 1 \,|\, x_t\,, 1\} = 0.45 \\
\text{for all other states } y \in \mathscr{X}, \text{ we have } P\{x_{t+1} = y \,|\, x_t\,, u_t\} = 0;
\end{gathered}$$

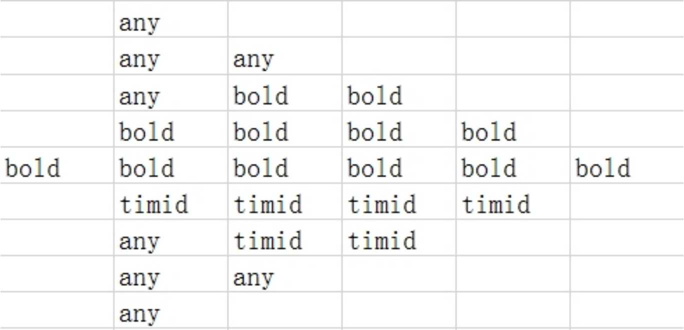
- Reward function for one period: $r_t(x,u) = 0$.
- Dynamic programming equations: The value function $v_t^*(x)$ expresses the probability to win at time t if the state (score) is x; we maximize it by devising a policy for the match by dynamic programming.

$$\begin{aligned}
&v_t^*(x) = \max\{0.9v_{t+1}^*(x) + 0.1v_{t+1}^*(x-1), 0.45v_{t+1}^*(x+1) + 0.55v_{t+1}^*(x-1)\}, \ t = 1,\dots,5; \\
&\qquad \text{for } t = 1, x = 0;\ t = 2, x = -1,0,1;\ t = 3, x = -2,-1,0,1,2; \\
&\qquad\quad t = 4, x = -3-2,-1,0,1,2,3;\ t = 5, x = \mathscr{X}; \\
&v_t^*(x) = 0 \text{ for other states when } t = 1,2,3,4,5.
\end{aligned}$$

$$v_6^*(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \\ 0.45 & \text{if } x = 0. \end{cases}$$



0					
0	0				
0	0	0			
0	0	0.09113	0.09113		
0	0.2025	0.2025	0.2916	0.28158	
0.45	0.45	0.53663	0.51435	0.54721	0.52616
1	0.945	0.8955	0.85961	0.82509	
1	1	0.9945	0.9846		
1	1	1			
1	1				
1					



	any				
	any	any			
	any	bold	bold		
	bold	bold	bold	bold	
bold	bold	bold	bold	bold	bold
	timid	timid	timid	timid	
	any	timid	timid		
	any	any			
	any				

Problem 2. A software manufacturer can be in one of two states. In state 1 their software sells well, and in state 2, the product sells poorly. While in state 1, the company can invest in development of upgraded version of the software, in which case the one-stage reward is 4 units, and the probability of degrading to state 2 is 0.2. If no investment in new development occurs, then the reward is 6 units, but the probability of transition to state 2 is 0.5. While in state 2, if the company invests in software development, then the reward is -2 units, but the probability of transition to state 1 is 0.7. Without special efforts to improve, the reward is 1 and the probability of upgrading to state 1 is 0.

Formulate a dynamic programming problem to determine an optimal reserch and development policy. Solve the problem for a time horizon of 12 time intervals.

Solution:

- The state space is $S = \{1 \text{ (good) }, 2 \text{ (bad)}\}$
- The control space is $U = \{0 \text{ (do nothing) }, 1 \text{ (research and advertise) }\}$.
- Transition probabilities:

$$\begin{aligned}
p(x_{t+1} = 1 | x_t = 1, u_t = 1) &= 0.8 \quad & p(x_{t+1} = 2 | x_t = 1, u_t = 1) &= 0.2 \\
p(x_{t+1} = 1 | x_t = 2, u_t = 1) &= 0.7 \quad & p(x_{t+1} = 2 | x_t = 2, u_t = 1) &= 0.3 \\
p(x_{t+1} = 1 | x_t = 1, u_t = 0) &= 0.5 \quad & p(x_{t+1} = 2 | x_t = 1, u_t = 0) &= 0.5 \\
p(x_{t+1} = 1 | x_t = 2, u_t = 0) &= 0.1 \quad & p(x_{t+1} = 2 | x_t = 2, u_t = 0) &= 0.9
\end{aligned}$$

Another way to write the transition probabilities is the following:

$$P(1) = \begin{pmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{pmatrix} \quad P(0) = \begin{pmatrix} 0.5 & 0.5 \\ 0.1 & 0.9 \end{pmatrix}$$

- The reward function is

$$r(x,u) = \begin{cases} 4 & \text{if } x=1, u=1 \\ 6 & \text{if } x=1, u=0 \\ 1 & \text{if } x=2, u=1 \\ 2 & \text{if } x=2, u=0. \end{cases}$$

- The dynamic programming equations are

$$v_{T+1} = 0;$$
$$v_t(x) = \max_{u \in \{0,1\}} \{r(x,u) + p(x_{t+1}=1|x,u)v_{t+1}(1) + p(x_{t+1}=2|x,u)v_{t+1}(2)\},$$
$$t = T, T-1, \dots, 1.$$

More specifically for $t = T, T-1, \dots, 1$, we have

$$v_t(1) = \max\{4 + 0.8v_{t+1}(1) + 0.2v_{t+1}(2), 6 + 0.5v_{t+1}(1) + 0.5v_{t+1}(2)\};$$
$$v_t(2) = \max\{1 + 0.7v_{t+1}(1) + 0.3v_{t+1}(2), 2 + 0.1v_{t+1}(1) + 0.9v_{t+1}(2)\}.$$

Problem 3. Consider the equipment replacement problem discussed in class, with the following data:

- operating cost per period $c_0 + c_1 x$, $x = 0, 1, 2, \dots$;
- revenue per period R;
- replacement cost K;
- salvage value $\gamma K e^{-\mu x}$;

where $c_0 > 0$, $c_1 > 0$, $c_2 > 0$, $0 < \gamma < 1$, and $\mu > 0$. Assume that the salvage value can be collected whenever the item is replaced. The probabilities of deterioration by j steps in one period are given by the Poisson distribution

$$p_j = \frac{\lambda^j}{j!} e^{-\lambda}, \quad j = 0, 1, \dots.$$

(3.1) Formulate the corresponding Markov decision problem. Clearly define the state space, action space, transition probabilities, and the reward function.

Solution:

- State space: $\mathscr{X} = \{0, 1, 2, ...\}$ - state of the equipment - $x = 0$ - new
- Control space: $\mathscr{U} = \{0, 1\}$ - 0 - wait, 1 - replace.
- Reward function for one period: $r(x,u) = \begin{cases} R - K(1 - \gamma e^{-\mu x}) - c_0 & u = 1 \\ R - (c_0 + c_1 x) & u = 0 \end{cases}.$
- Dynamic programming equation:

$$v_t^*(x) = \max\left\{ r(x,1) + \sum_{j=0}^{\infty} p_j v_{t+1}^*(j);\ r(x,0) + \sum_{j=0}^{\infty} p_j v_{t+1}^*(x+j) \right\}$$

(3.2) Solve the problem numerically for $c_0 = 1$, $c_1 = 1$, $R = 5$, $K = 10$, $\gamma = 0.8$, $\mu = 0.2$, $\lambda = 1$, and time horizon $T = 20$. To this end, argue that a constant value $\bar{x}$ exists such that for all $x \geq \bar{x}$ replacement is always profitable. Then you will know that the value function for all $x \geq \bar{x}$ is the same as at $\bar{x}$. This will allow you have finite tables of the value function for each time t.

Solution:

We can show that the value function is nonincreasing by induction.

We observe that $r(\cdot, u)$ is decreasing because $c_1 > 0$ in the case $u = 1$; the function e^{-x} is decreasing and is multiplied by positive constant γK in the case $u = 0$.

Further, p_j are constant and do not depend on x. The functions $v^*_{t+1}(\cdot)$ $v^*_{t+1}(\cdot + j)$ are decreasing with respect to x, by induction assumption.
Denoting

$$h(x) = r(x, 1) + \sum_{j=0}^{\infty} p_j v^*_{t+1}(j),$$

$$g(x) = r(x, 0) + \sum_{j=0}^{\infty} p_j v^*_{t+1}(x + j),$$

we conclude that $h(\cdot)$ and $g(\cdot)$ are decreasing because they are sums of decreasing functions. Thus, for $x > y$

$$\max\{h(x); g(x)\} \leq \max\{h(y); g(y)\}$$

We conclude that $v^*_t(x)$ is nonincreasing.

Now we find a point $\bar{x}$. Due to the monotonicy of $v^*(\cdot)$, we have for all x

$$\sum_{j=0}^{\infty} p_j v^*_{t+1}(j) \geq \sum_{j=0}^{\infty} p_j v^*_{t+1}(x + j).$$

Thus, we need to only compare the one-step reward. We observe that

$$R - K(1 - \gamma e^{-\mu x}) - c_0 \geq R - (c_0 + c_1 x) \quad \Leftrightarrow \quad c_1 x + \gamma K e^{-\mu x} \geq K$$

For the given parameters we have can verify numerically that $\bar{x} = 9$ is sufficiently large. Moreover, $p_9 \approx 10^{-6}$. Therefore, we can now limit our attention to $x \in \{0, 1, \ldots 9\}$. Under these assumptions, the numerical solution gives an optimal policy to replace if the state was at least $x_t = 2$, for $t = 1, \ldots N - 1$.