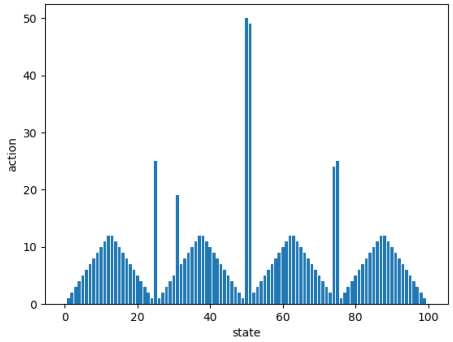
hw 3 Solution

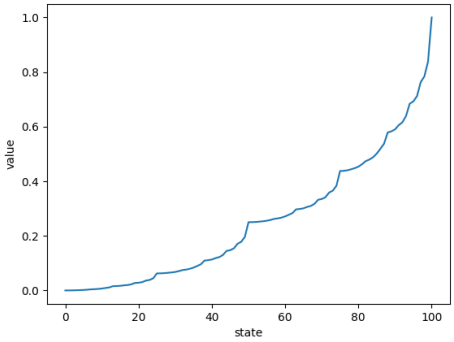
chenyu wu

Oct 2023

Exercise 4.9 (25pts/15pts): For $p_h = 0.25, 0.55$ we have

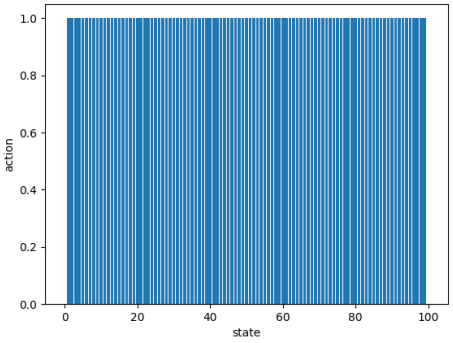


(a) Policy

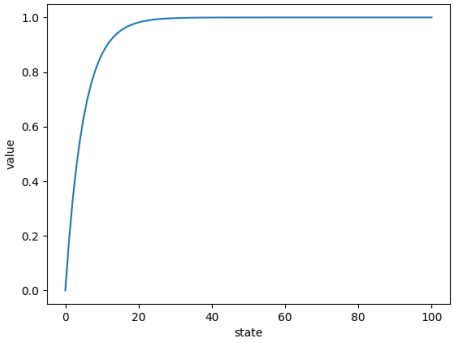


(b) Value

Figure 1: $p_h = 0.25$



(a) Policy



(b) Value

Figure 2: $p_h = 0.55$

Problem 2 (25pts/15pts)

According to the problem setting, we first write the Bellman equation for this discounted problem: For all $i \in \{500, 600, 700, 800, 1000\}$, we have

$$J^*(i) = \max_{u \in \{0,1\}} \sum_{j=1}^{5} p_j(g(i,u,j) + \alpha J^*(j)),$$

where $u = 0$ denotes the decision to reject the offer and $u = 1$ denotes the decision to accept the offer and $u = 1$ denotes the decision to accept the offer, $g(i, 0, j) = -60, \alpha = 0.97$, and when $u = 1, J(i) = i$.

(a)

The value iteration algorithm will be as follows:
for all $i \in \{500, 600, 700, 800, 1000\}$, and any initial conditions $J_0(500), \cdots, J_0(1000)$(starting with arbitrary initial condition), the VI algorithm generates the sequence $\{J_k\}$ according to

$$J_{k+1}(i) = \max_{u \in \{0,1\}} \sum_{j=1}^{5} p_j(g(i,u,j) + \alpha J_k(j)).$$

We suppose $J_0(i) = i$ for all i so that we have

$$J_1 \approx [610.92, 610.92, 700, 800, 1000]$$
$$J_2 \approx [640.46, 640.46, 700, 800, 1000]$$

(b)

We first pick one policy, say $\forall i, \mu^0(i) = 1$ (i.e. to accept any offer), we know that $\forall i, J_{\mu^0}(i) = i$, then we proceed to do policy improvement and compute a new policy μ^1 as

$$\mu^1(i) \in \arg\max_{u \in \{0,1\}} \sum_{j=1}^{5} p_j(g(i,u,j) + \alpha J^{\mu^0}(j))$$

we get the new policy as $\mu^1 = [0, 0, 1, 1, 1]$, solve the linear equations in terms of $J_{\mu^1}(500)$ and $J_{\mu^1}(600)$, repeat the above process until $J_{\mu^{k+1}}(i) = J_{\mu^k}(i)$ for all i. Hence, we can find the optimal policy is to reject the offers of 500 and 600, and to accept the other offers.

Problem 3 (25pts/15pts)

For this problem, we know that the state $i \in \{0, 1, 2, 3\}$, let $u = 0$ denote the decision that we do nothing or replace the machine if we are at stage 3 and $u = 1$ denote the choice to repair the machine at stage 2, our policy actually only focus on the decision when we are at stage 2. Let $p_{i,j}$ be the

$(i+1, j+1)$-th element in the matrix P, we can write the cost function $g(i, u, j)$ as follows:

$$\begin{aligned} g(0,0,j) &= 0 \\ g(1,0,j) &= 1000 \\ g(2,0,j) &= 3000 \\ g(2,1,1) &= 2000 \\ g(3,0,0) &= 6000 \end{aligned}$$

Use a discount factor $\alpha = 0.95$, we do the policy iteration, first set a policy μ^k, then compute $J_{\mu^k}(i), i = 0, 1, 2, 3$, as the solution of the linear system of Bellman equations

$$J_{\mu^k}(i) = \sum_{j=0}^{3} p_{ij}(\mu^k(i))(g(i, \mu^k(i), j) + \alpha J_{\mu^k}(j)),$$

then we proceed to do policy improvement and compute a new policy μ^{k+1} as

$$\mu^{k+1} \in \arg\min_{u \in U(i)} \sum_{j=0}^{3} p_{ij}(u)(g(i, u, j) + \alpha J_{\mu^k}(j))$$

we repeat this process until $J_{\mu^{k+1}}(i) = J_{\mu^k}(i)$ for all i. We can find the optimal policy is to do nothing at stage 0 and 1, to repair the machine at stage 2, and to replace the machine if we are at stage 3.

Problem 4 (25pts/15pts)

Similarly, according to the information, we know that $i \in \{0, 1, 2\}$ and $u \in \{0, 1, 2\}$, first define the cost functions $g(i, u, j) = 10 + 5 * u$ if $u > 0$; $g(i, u, j) = 0$ if $u = 0, j = i + u$, let d denote the demand, then the holding cost is $h(j, d) = 4 * (j - d)^+$, where $x^+ = \max(0, x)$, the shortage cost is $b(j, d) = 0$ if $j \geq d = 0$; $b(j, d) = 8$ if $y - j = 1$; $b(j, d) = 32$ if $y - d = 2$. We write the Bellman equation as follows:

$$J^*(i) = \min_u \{p_{ij}(u)(g(i, u, j) + \mathbb{E}_d[h(j, d) + b(j, d) + \alpha J^*((j - d)^+)])\}$$

Following the same procedure of the policy iteration algorithm, we can find the optimal policy of this problem is to always produce nothing.

Note that if you assume the setup cost is a one-time cost, which does not affect the infinite-period policy, you may end up with the policy that starts production when the inventory is equal to 0.

Problem 5 (20pts)

The policy improvement theorem is stated as follows:

Let π and π' be any pair of policies such that, for all $s \in \mathcal{S}$,

$$q_\pi(s, \pi'(s)) \geq v_\pi(s). \tag{1}$$

Then the policy π' must be as good as, or better than, π. That is, it must obtain a greater or equal expected return from all states $s \in \mathcal{S}$:

$$v_{\pi'}(s) \geq v_{\pi}(s). \tag{2}$$

To adapt to the randomized policies, we need to assume

$$\mathbb{E}[q_{\pi}(s, \pi'(s))] \geq v_{\pi}(s). \tag{3}$$

Then we can follow the proof in the book:

$$\begin{aligned}
v_{\pi}(s) &\leq \mathbb{E}[q_{\pi}(s, \pi'(s))] \\
&= \sum_{a \in \mathcal{A}(s)} \pi'(a|s) \mathbb{E}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) | S_t = s, A_t = a] \\
&\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma \mathbb{E}[q_{\pi}(S_{t+1}, \pi'(S_{t+1}))] | S_t = s] \\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma \sum_{a \in \mathcal{A}(s_1)} \mathbb{E}[R_{t+2} + \gamma v_{\pi}(S_{t+2}) | S_{t+1} = s_1, A_{t+1} = a] | S_t = s] \\
&= \cdots \\
&\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s] \\
&= v_{\pi'}(s)
\end{aligned}$$

Note that in the book, its proof is already written that could adapt to the randomized policies. In particular, we can see the outer expectation is based on the new policy π'. The corresponding reward R_{t+1} inside is based on the distribution of $p_{\pi'}(a|s)$ multiply the distribution of $p(R_{t+1}, S_{t+1}|s, a)$. For deterministic policy, we just have $p_{\pi'}(a|s)$ be the indicator-type measure. The key point in the proof is mainly based on whether you recognize the assumption (3) as legitimate.

Problem 6 (20pts)

Proof by contradiction:

Suppose there exists a policy $\pi^* = \{\pi_1^*, \cdots, \pi_S^*\} \in \mathcal{P}$, where π_i^* is corresponding to the state i, such that $\rho_{\pi^*} = \max_{s \in \mathcal{S}} P_{\pi^*}(S_m \neq \Delta | S_0 = s) = 1$, then we have for some s^*

$$P_{\pi^*}(S_m = \Delta | S_0 = s^*) = 0,$$

which means for any realization realization $a_1, \cdots, a_m$ such that

$$P_{\pi^*}(S_m = \Delta, A_1 = a_1, \cdots, A_m = a_m | S_0 = s^*) \leq P_{\pi^*}(S_m = \Delta | S_0 = s^*) = 0,$$

and

$$P_{\pi^*}(A_1 = a_1, \cdots, A_m = a_m | S_0 = s^*) > 0.$$

Then we can obtain

$$\begin{aligned}
&P_{\pi^*}(S_m = \Delta, A_1 = a_1, \cdots, A_m = a_m | S_0 = s^*) \\
&= P_{\pi^*}(S_m = \Delta | A_1 = a_1, \cdots, A_m = a_m, S_0 = s^*) * P_{\pi^*}(A_1 = a_1, \cdots, A_m = a_m | S_0 = s^*) = 0.
\end{aligned}$$

Note that as we condition on $A_1, \cdots, A_m$, the policy is actually a deterministic policy so that we know from the previous result for any deterministic policy we have

$$P_{\pi^*}(S_m = \Delta | A_1 = a_1, \cdots, A_m = a_m, S_0 = s^*) > 0,$$

which leads to a contradiction.

Proof by construction (note that since $\mathbb{R}_{[0,1]}$ is infinite, we need to be careful about infinite many of $\pi \in \mathcal{P}$):

Since we have $\mathbb{R}^n_{[0,1]}$ is compact, we can see that all possible policy space $\mathcal{P}$ is compact due to finite action space and state space. Then we can simply say

$$\max_{\pi \in \mathcal{P}} \rho_\pi = \max_{\pi \in \mathcal{P}} \max_{s \in \mathcal{S}} P_\pi(S_m \neq \Delta | X_0 = s) < 1$$