CS6700: Reinforcement Learning — Homework 1

Avinash Kori — ED15B006

Indian Institute of Technology Madras

---

let,
$X = 1, 2, 3...n$ be all possible states
$U = a_1, a_2, a_3...a_m$ be all possible actions
$\mu_1, \mu_2, ...\mu_N$ be sequence of functions in policy π

1. QUESTION 1

Expected optimal cost without DP algorithm is given by 1

$$J^*_\pi(x_0) = min_{a_i \in U}\, \mathbb{E}_{x' \in X}(g_N(x_N) + \Sigma_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), x_{k+1})) \tag{1}$$

Expected optimal cost with DP algorithm is given by 2

$$J^*(x_k) = min_{a_i \in U}\, \mathbb{E}_{x' \in X}(g_k(x_k, \mu_k(x_k), x_{k+1}) + J^*(x_{k+1})) \tag{2}$$

1.1 Complexity analysis of Expected total cost (equation 1):

- Total number of state-stage pares: nN
- Number of actions for each state-space pare: m

Total number of operations is given by m^{nN}

1.2 Complexity analysis of equation 2:

- Total number of state-stage pares: nN
- Total number of action state pares: mn

Total number of operations is given by mn^2N

1.3 Conclusion:

As the number of operations required in calculating J using DP algorithm is less than Expected cost calculating for each policy, DP algorithm is computationally less intensive.

---

2. QUESTION 2

Alternative cost function for finite horizon MDP is given by 3

$$J^*_{\pi}(x_0) = min_{a_i \in U}\, \mathbb{E}_{x' \in X}[exp(g_N(x_N) + \Sigma_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), x_{k+1}))] \tag{3}$$

2.1 DP-variant

$$J^*(x_k) = min_{a_i \in U}\, \mathbb{E}_{x' \in X}[exp(g_k(x_k, \mu_k(x_k), x_{k+1}) + J^*(x_{k+1}))] \tag{4}$$

Proof by induction, to show optimal policy obtained by 3 is same as policy obtained by 4

Claim $J_0(x_0)$ by DP algorithm is equal to $J^*_0(x_0)$

let $\pi^* = \{\mu^*_1, \mu^*_2, ..., \mu^*_N\}$ be optimal policy for any policy $\pi = \{\mu_1, \mu_2, ..., \mu_N\}$

let $J^*_k(x_k)$ be optimal cost to-go for tail sub-problem

For k=N, $J^*_N(x_N) = g_N(x_N) = J_N(x_N)$

Assume $J^*_{k+1}(x_{k+1}) = J_{k+1}(x_{k+1}) \qquad \forall x_{k+1} \in X$

$$J^*_k(x_k) = min_{\{\mu_k, \pi_{k+1}\}}\, \mathbb{E}_{x' \in X}[exp(g_N(x_N) + g_k(x_k, \mu_k, x_{k+1}) + \Sigma_{j=k+1}^{N-1} g_j(x_j, \mu_j(x_j), x_{j+1}))]$$

$$\Rightarrow min_{\mu_k}\, \mathbb{E}_{x_k}[exp(g_N(x_N) + g_k(x_k, \mu_k, x_{k+1})) * min_{\pi_k}\, \mathbb{E}_{x_{k+1}, x_{k+2}, ..., x_{N-1}}[exp(\Sigma_{j=k+1}^{N-1} g_j(x_j, \mu_j(x_j), x_{j+1}))]]$$

$$\Rightarrow min_{\mu_k}\, \mathbb{E}_{x_k}[exp(g_k(x_k, \mu_k, x_{k+1})) * min_{\pi_k}\, \mathbb{E}_{x_{k+1}, x_{k+2}, ..., x_{N-1}}[exp(g_N(x_N) + \Sigma_{j=k+1}^{N-1} g_j(x_j, \mu_j(x_j), x_{j+1}))]]$$

$$\Rightarrow min_{\mu_k}\, \mathbb{E}_{x_k}[exp(g_k(x_k, \mu_k, x_{k+1})) * J_{k+1}(x_{k+1})$$

$$\Rightarrow J_k(x_k) \quad by \quad DP$$

2.2 .

let $V_k(x_k) = logJ_k(x_k)$

$$J_k(x_k) = min_{a_k \in A(x_k)}\, \mathbb{E}_{x_{k+1}}[exp(g_k(x_k, \mu_k, x_{k+1})) * J_{k+1}(x_{k+1})]$$

$$log(J_k(x_k)) = min_{a_k \in A(x_k)} log(\mathbb{E}_{x_{k+1}}[exp(g_k(x_k, \mu_k)) * J_{k+1}(x_{k+1}))]$$

$$V_k(x_k) = min_{a_k \in A(x_k)} log(exp(g_k(x_k, \mu_k)) * \mathbb{E}_{x_{k+1}}[J_{k+1}(x_{k+1}))]$$

$$\Rightarrow min_{a_k \in A(x_k)}(g_k(x_k, \mu_k) + log(\mathbb{E}_{x_{k+1}}[J_{k+1}(x_{k+1}))])$$

$$\Rightarrow min_{a_k \in A(x_k)}(g_k(x_k, \mu_k) + log(\mathbb{E}_{x_{k+1}}[exp(v_{k+1}(x_{k+1})))])$$

Hence Proved...

---

3. QUESTION 3

3.1 MDP formulation

Let a_k, x_k, R_k be actions, states and reward respectively which are given by below equations:-

$g_N(x_N)$ denotes terminal reward

$$a_k = \begin{cases} a_1 & (buy) \\ a_2 & (do \quad nothing) \end{cases} \tag{5}$$

$$x_{k+1} = \begin{cases} T & if((x_k = x_N \quad \&\& \quad a_k = a_2) \quad or \quad a = a_1) \\ p & else \end{cases} \tag{6}$$

$$g_k(x_k) = \begin{cases} (N-k) & if(x_k! = T) \\ 0 & else \end{cases} \tag{7}$$

$$g_N(x_N) = \begin{cases} \frac{1}{1-p} & if(x_k! = T) \\ 0 & else \end{cases} \tag{8}$$

3.2 & 3.3 DP algorithm and policy characterization

$$J_N(x_N) = g_N(x_N)$$

$$J^*(x_k) = \begin{cases} min_{a_1,a_2}\, \mathbb{E}_{x' \in X}(g_k(x_k, \mu_k(x_k), x_{k+1}) + J^*(x_{k+1})) & if(x_k! = T) \\ 0 & else \end{cases} \tag{9}$$

considering case when $x_k! = T$:-

$$J_N(x_N) = \frac{1}{1-p}$$

$$J(x_k) = min_{a_1,a_2}\, \mathbb{E}_{x' \in X}(g_k(x_k, \mu_k(x_k), x_{k+1}) + J_{k+1}(x_{k+1}))$$

$$\Rightarrow min\{\mathbb{E}(g_k(x_k, \mu_k(x_k)), x_{k+1}), \quad \mathbb{E}_{x_{k+1} \in X}(J_{k+1}(x_{k+1}))\}$$

$$\Rightarrow min\{(N-k)x_k, \quad \mathbb{E}_{x_{k+1} \in X}(J_{k+1}(x_{k+1}))\}$$

Let's define threshold α_k such that:

$$\alpha_k = \frac{\mathbb{E}_{x_{k+1} \in X}(J_{k+1}(x_{k+1}))}{(N-k)}$$

optimal policy actions to choose:

$$a_k = \begin{cases} a_1 & if(x_k < \alpha_k) \\ a_2 & else \end{cases} \tag{10}$$

let $V_k(x_k) = \frac{J_k(x_k)}{N-k}$

Now the claim is $\alpha_k \leq \alpha_{k+1}$ To establish this claim, it is enough to show that $V_k(x) \leq V_{k+1}(x), \forall x$. For the case when $k = N-1$, we observe that

$$V_{N-1}(x_{N-1}) = min(x_{N-1}, \mathbb{E}(V_N(x_N))$$

$$\Rightarrow min(p, \frac{1}{1-p}) \leq p = V_N(x_N)$$

for $k = N-2$

$$V_{N-2}(x_{N-2}) = min(x_{N-2}, \mathbb{E}(V_{N-1}(x_{N-1}))$$

$$\Rightarrow min(p, \mathbb{E}(V_{N-1}(x_{N-1}))) \leq min(p, \mathbb{E}(J_N(x_N))) = V_{N-1}(x_{N-1})$$

then

$$V_{k+1}(x_{k+1}) = \begin{cases} p & if(\alpha_{k+1} \geq p) \\ \alpha_{k+1} & else \end{cases} \tag{11}$$

$$\begin{aligned} \alpha_k &= \mathbb{E}_{x_{k+1} \in X}(V_{k+1}(x_{k+1})) \\ &= \frac{1}{N-k} \int_0^{\alpha_{k+1}} p \times pdf(p)dp + \int_{\alpha_{k+1}}^{\infty} \alpha_{k+1} \times pdf(p)dp \\ &\leq \frac{1}{N-k} \int_0^{\infty} p \times pdf(p)dp \\ &= \frac{1}{N-k} \mathbb{E}(p) \end{aligned}$$

$$\Rightarrow 0 \leq \alpha_k \leq \frac{\mathbb{E}(p)}{N-k}$$

---

4. QUESTION 4

N stage problem with T_i as execution time and β_i portion of T_i used for execution with probability p_i

let policy

$$L^1 = (1, \quad 2, \quad 3, ..., \quad i, \quad j, ..., \quad N-1, \quad N)$$
$$L^2 = (1, \quad 2, \quad 3, ..., \quad j, \quad i, ..., \quad N-1, \quad N)$$

reward to maximize the portion of job (maximize the residual time): $R_i = \beta_i(1-\beta_i)T_i$
Optimization fn: $J_k = max\{\Sigma R_i\}$

$$J_k^{L^1} = \Sigma_{m=1}^{i-1}(\Pi_{n=1}^{m}(p_n)R_m) + \Pi_{n=1}^{i-1}(p_n)p_iR_i + \Pi_{n=1}^{i-1}(p_n)p_ip_jR_j + \Sigma_{m=i+2}^{N}(\Pi_{n=1}^{m}(p_n)R_m)$$
$$\Rightarrow \Sigma_{m=1}^{i-1}(\Pi_{n=1}^{m}(p_n)\beta_m(1-\beta_m)T_m) + \Pi_{n=1}^{i-1}(p_n)p_i\beta_i(1-\beta_i)T_i + \Pi_{n=1}^{i-1}(p_n)p_ip_j\beta_j(1-\beta_j)T_j + \Sigma_{m=i+2}^{N}(\Pi_{n=1}^{m}(p_n)\beta_m(1-\beta_m)T_m)$$

Similarly policy on L^2 is given by:

$$J_k^{L^2} = \Sigma_{m=1}^{i-1}(\Pi_{n=1}^{m}(p_n)R_m) + \Pi_{n=1}^{i-1}(p_n)p_jR_j + \Pi_{n=1}^{i-1}(p_n)p_ip_jR_i + \Sigma_{m=i+2}^{N}(\Pi_{n=1}^{m}(p_n)R_m)$$
$$\Rightarrow \Sigma_{m=1}^{i-1}(\Pi_{n=1}^{m}(p_n)\beta_m(1-\beta_m)T_m) + \Pi_{n=1}^{i-1}(p_n)p_j\beta_j(1-\beta_j)T_j + \Pi_{n=1}^{i-1}(p_n)p_ip_j\beta_i(1-\beta_i)T_i + \Sigma_{m=i+2}^{N}(\Pi_{n=1}^{m}(p_n)\beta_m(1-\beta_m)T_m)$$

let $J_k^{L^1} \geq J_k^{L^2}$

$$\Pi_{n=1}^{i-1}(p_n)p_i\beta_i(1-\beta_i)T_i + \Pi_{n=1}^{i-1}(p_n)p_ip_j\beta_j(1-\beta_j)T_j \geq \Pi_{n=1}^{i-1}(p_n)p_j\beta_j(1-\beta_j)T_j + \Pi_{n=1}^{i-1}(p_n)p_ip_j\beta_i(1-\beta_i)T_i$$

$$\Rightarrow p_i\beta_i(1-\beta_i)T_i + p_ip_j\beta_j(1-\beta_j)T_j \geq p_j\beta_j(1-\beta_j)T_j + p_ip_j\beta_i(1-\beta_i)T_i$$

$$\Rightarrow \frac{p_i\beta_i(1-\beta_i)T_i}{1-p_i} \geq \frac{p_j\beta_j(1-\beta_j)T_j}{1-p_j}$$
$$\Rightarrow \frac{p_i\beta_i Z_i}{1-p_i} \geq \frac{p_j\beta_j Z_j}{1-p_j}$$

Index based policy....

---

5. QUESTION 5

Cost for tail sub-problem from stage k to N is given by:-

$$J_k(x_0) = \mathbb{E}_{x'\in X}(g(x_N, a_N, x') + \Sigma_{m=k}^{N-1} g(x_m, \mu_m(x_m), x_{m+1}))$$

$$J_{k+1}(x_0) = \mathbb{E}_{x'\in X}(g(x_N, a_N, x') + \Sigma_{m=k+1}^{N-1} g(x_m, \mu_m(x_m), x_{m+1}))$$

5.1 i

$$J_{N-1}(x) \leq J_N(x)$$
$$\Rightarrow \mathbb{E}(g(x_N, a_N, x') + g(x_{N-1}, a_{N-1}, x')) \leq \mathbb{E}(g(x_N, a_N, x'))$$
$$\Rightarrow \mathbb{E}(g(x, a, x')) \leq 0$$

$$\begin{aligned} J_k(x_0) &= \mathbb{E}(g(x_N, a_N, x') + g(x_k, a_k, x') + \Sigma_{m=k+1}^{N-1} g(x_m, \mu_m(x_m), x_{m+1})) \\ &= \mathbb{E}(g(x_k, a_k, x') + J_{k+1}(x_{k+1})) \\ &= \mathbb{E}(g(x_k, a_k, x')) + \mathbb{E}(J_{k+1}(x_{k+1})) \end{aligned}$$

$$\Rightarrow J_k(x) \leq J_{k+1}(x)$$

proved

5.2 ii

$$J_{N-1}(x) \geq J_N(x)$$
$$\Rightarrow \mathbb{E}(g(x_N, a_N, x') + g(x_{N-1}, a_{N-1}, x')) \geq \mathbb{E}(g(x_N, a_N, x'))$$
$$\Rightarrow \mathbb{E}(g(x, a, x')) \geq 0$$

$$\begin{aligned} J_k(x_0) &= \mathbb{E}(g(x_N, a_N, x') + g(x_k, a_k, x') + \Sigma_{m=k+1}^{N-1} g(x_m, \mu_m(x_m), x_{m+1})) \\ &= \mathbb{E}(g(x_k, a_k, x') + J_{k+1}(x_{k+1})) \\ &= \mathbb{E}(g(x_k, a_k, x')) + \mathbb{E}(J_{k+1}(x_{k+1})) \\ &\Rightarrow J_k(x) \geq J_{k+1}(x) \end{aligned}$$

proved

---

6. QUESTION 6

6.1 MDP formulation

Let a_k, x_k, R_k be actions, states and reward respectively which are given by below equations:-

x_k denotes total number of uncorrected errors at that stage which is given by an expectation over binomial distribution in unordered pairs which is given by: $p_k(1-p_k)^{n-1} + 2p_k^2(1-p_k)^{n-2} + 3p_k^3(1-p_k)^{n-3} + ... + np_k^n$

where n is determined by x_{k-1}
$g_N(x_N)$ denotes terminal reward

$$a_k = \begin{cases} a_1 & (publish) \\ a_2 & (continue \quad proofreading) \end{cases} \tag{12}$$

$$x_k = \begin{cases} T & if((x_{k-1} = x_N \quad \&\& \quad a_{k-1} = a_2) \quad or \quad a = a_1) \\ x_{k-1} - \Sigma_{m=0}^{x_{k-1}}(mp_k^m(1-p_k)^{x_{k-1}-1}) & else \end{cases} \tag{13}$$

$$g_k(x_k) = \begin{cases} c_2 x_k & if(x_k == T) \\ c_1 & else \end{cases} \tag{14}$$

$$g_N(x_N) = \begin{cases} 0 & if(x_N = T) \\ c_2 x_N & else \end{cases} \tag{15}$$

6.2 & 6.3 DP algorithm and policy characterization

$$J^*(x_k) = \begin{cases} min_{a_1,a_2}\, \mathbb{E}_{x' \in X}(g_k(x_k, \mu_k(x_k), x_{k+1}) + J^*(x_{k+1})) & if(x_k! = T) \\ 0 & else \end{cases} \tag{16}$$

considering case when $x_k! = T$:-

$$J(x_k) = min_{a_1,a_2}\, \mathbb{E}_{x' \in X}(g_k(x_k, \mu_k(x_k), x_{k+1}) + J(x_{k+1}))$$
$$\Rightarrow min\{g_k(x_k, \mu_k(x_k), x_{k+1}), \quad \mathbb{E}_{x_{k+1} \in X}(J(x_{k+1}))\}$$
$$\Rightarrow min\{c_2 x_k, \quad c_1 + \mathbb{E}_{x_{k+1} \in X}(J(x_{k+1}))\}$$

Let's define threshold α_k such that:

$$\alpha_k = \frac{c_1 + \mathbb{E}_{x_{k+1} \in X}(J(x_{k+1}))}{c_2}$$

optimal policy actions to choose:

$$a_k = \begin{cases} a_1 & if(x_k < \alpha_k) \\ a_2 & else \end{cases} \tag{17}$$

Simplification of α_k

$$\alpha_k = \frac{c_1 + \mathbb{E}_{x_{k+1} \in X}(J(x_k - \Sigma_{m=0}^{x_k}(mp_{k+1}^m(1-p_{k+1})^{x_k-1})))}{c_2}$$

---

7. REFERENCES

Discussed with ED15B021

Prashanth L. A. CS6700: Reinforcement learning Course notes, 2018

Dimitri P. Bertsekas. Dynamic Programming and Optimal Control, vol. I. Athena Scientific, 2017.