ECE5984: Reinforcement Learning

Assignment #2

Nima Mohammadi

nimamo@vt.edu

Problem 1.
We consider an application of MDP, where there is a car and its state i can be $0, 1, 2, \ldots$ Here the state could be an indicator of how good the car is. The state depends on various factors such as the days of operation, age of the car, etc. Everyday, we want to make a decision whether to bring the car to mechanic shops or not based on its observed states. If we decide to bring it, the car is repaired instantaneously and the state of the car returns to 0 . Repairing the car (e.g., tune up) incurs a cost R, while maintaining it causes a cost $C(i)$ per day if the car is in state $i = 0, 1, \ldots$ Moreover, $C(i), i \geq 0$, is assumed to be an increasing bounded function in i, i.e., higher maintenance costs are associated with higher state indices. Given the current state i, let P_{ij} be the transition probability that the car will be at state j the next day.
Question: Let $\gamma \in (0, 1)$ and π is a policy. Write the Markov decision model for this problem, i.e., define the state and action space, the instantaneous reward $r(s, a)$, and the transition probability matrix. Also, formulate the optimization problem to find the best policy. Note that in this case we want to minimize the discounted cumulative cost.

The Markov Decision Process (MDP) can be formulated as the quintuple $(\mathcal{S}, \mathcal{A}, P, r, \text{ and } \gamma)$. The state space $\mathcal{S} = \{0, 1, 2, \cdots\}$ is an indicator of how good the car is. The action space $\mathcal{A} = \{\text{'Repair'}, \text{'Maintain'}\}$ include the possible two actions. Futhermore, the reward can be defined as a piecewise function of the current state and the chosen action, determined by R and $C(.)$:

$$r(s = i, a) = \begin{cases} -R; & \text{if } a \text{ is 'Repair'} \\ -C(i) & \text{if } a \text{ is 'Maintain'} \end{cases}$$

$-R - C(0)$

include reward for returning to $C(0)$

$\boxed{-2.5}$

The state transition matrix can be denoted via the function below that depends on the current action and the two consecutive states:

$$\widetilde{P}_{ij} = \begin{cases} P_{ij} & \text{if } a \text{ is 'Maintain'} \\ 1 & \text{if } a \text{ is 'Repair' and } j = 1 \\ 0 & \text{if } a \text{ is 'Repair' and } j \neq 1 \end{cases}$$

Then, the optimization problem would be to find $\{a_k\}$ such that

$$\underset{\{a_k\}}{\text{maximize}} \lim_{N \to \infty} \mathbb{E}\left[\sum_{k=0}^{N} \gamma^k r(s_k, a_k) \mid s_0\right]$$

Problem 2. We recall here the policy iteration (PI) algorithm for the discounted MDP with discount factor γ. There are two main steps as follows.

1. **(Policy evaluation)** Given the current policy μ_k, estimate V_{μ_k}

$$\left(\mathbf{I} - \gamma \mathbf{P}_{\mu_k}\right) V_{\mu_k} = \mathbb{E}[r]$$

or equivalently $V_{\mu_k} = T_{\mu_k} V_{\mu_k}$, where r is the vector of rewards.

2. **(Policy improvement)** Obtain a new improved policy μ_{k+1}

$$T_{\mu_{k+1}} V_{\mu_k} = T V_{\mu_k}$$

Implementing these two steps exactly in many applications are expensive. We consider here an approximation of these two steps. In particular, given the current policy μ_k, the policy evaluation step only returns a δ-approximate of the value function V_{μ_k}, i.e., a vector V_k satisfies

$$\left\| V_k - V_{\mu_k} \right\| \leq \delta, \quad \forall k$$

In addition, using this value V_k the policy evaluation problem can only compute an ϵ-approximate of the mapping T, i.e., a new policy μ_{k+1} satisfies

$$\left\| T_{\mu_{k+1}} V_k - T V_k \right\| \leq \epsilon, \quad \forall k$$

In the sequel let $\mathbf{1}$ be the vector whose entries are 1. Given a policy μ_k and its value function V_{μ_k}, we have $V_{\mu_k} = T_{\mu_k} V_{\mu_k}$ where T_{μ_k} is given as

$$\left(T_{\mu_k} V_{\mu_k}\right)(s) = \mathbb{E}\left[r\left(s, \mu_k(s)\right)\right] + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}\left(\mu_k(s)\right) V_{\mu_k}\left(s'\right)$$

In addition, given a scalar c we have T_{μ_k} satisfies

$$T_{\mu_k}\left(V_{\mu_k} + c\mathbf{1}\right) = T_{\mu_k} V_{\mu_k} + \gamma c \mathbf{1}$$

1. Show that

$$\left\| T_{\mu_{k+1}} V_{\mu_k} - T_{\mu_{k+1}} V_k \right\| \leq \gamma\delta \quad \text{and} \quad \left\| T V_k - T V_{\mu_k} \right\| \leq \gamma\delta$$

$$\begin{aligned} T_{\mu_{k+1}} V_{\mu_k}(s) - T_{\mu_{k+1}} V_k(s) &= \mathbb{E}\left[r\left(s, \mu_{k+1}(s)\right)\right] + \gamma \sum_{s'} P_{ss'}\left(\mu_{k+1}(s)\right) V_{\mu_k}\left(s'\right) \\ &\quad - \mathbb{E}\left[r\left(s, \mu_{k+1}(s)\right)\right] + \gamma \sum_{s'} P_{ss'}\left(\mu_{k+1}(s)\right) V_k\left(s'\right) \\ &= \gamma \sum_{s'} P_{ss'}\left(\mu_{k+1}(s)\right)\left[V_{\mu_k}\left(s'\right) - V_k\left(s'\right)\right] \end{aligned}$$

✓

As mentioned above,

$$\left\| V_k - V_{\mu_k} \right\| \leq \delta, \quad \forall k$$

Then,

$$\begin{aligned} T_{\mu_{k+1}} V_{\mu_k}(s) - T_{\mu_{k+1}} V_k(s) &= \gamma \sum_{s'} P_{ss'}\left(\mu_{k+1}(s)\right)\left[V_{\mu_k}\left(s'\right) - V_k\left(s'\right)\right] \\ &\leq \gamma \sum_{s'} P_{ss'}\left(\mu_{k+1}(s)\right)\delta = \gamma\delta \end{aligned}$$

Then,

$$\left\| T_{\mu_{k+1}} V_{\mu_k} - T_{\mu_{k+1}} V_k \right\| \leq \gamma\delta$$

Next, to show that $\left\| T V_k - T V_{\mu_k} \right\| \leq \gamma\delta$,

$$\left\| V_k - V_{\mu_k} \right\| \leq \delta,\ \forall k \ \Rightarrow \ T V_{\mu_k}(s) - T V_k(s) \leq \gamma\delta,\ \forall s$$

$$|T V_{\mu_k}(s) - T V_k(s)| \leq \gamma\delta,\ \forall s$$

$$\Rightarrow \max_s |T V_{\mu_k}(s) - T V_k(s)| \leq \gamma\delta$$

$$\| T V_{\mu_k}(s) - T V_k(s) \| \leq \gamma\delta$$

2. Show that

$$T_{\mu_{k+1}} V_{\mu_k} \geq T V_{\mu_k} - (\epsilon + 2\gamma\delta)\mathbf{1}$$

> We refer to $\left\| V_k - V_{\mu_k} \right\| \le \delta, \quad \forall k$, then by applying $T_{\mu_{k+1}}$ we will have
>
> $$T_{\mu_{k+1}} V_k - \gamma\delta\mathbf{1} \le T_{\mu_{k+1}} V_{\mu_k}$$
>
> On the other hand, we have
>
> $$\left\| T_{\mu_{k+1}} V_k - T V_k \right\| \le \epsilon, \quad \Rightarrow T_{\mu_{k+1}} V_k \ge T V_k - \epsilon\mathbf{1}$$
>
> From two results above, we will have
>
> $$T V_k - \epsilon\mathbf{1} - \gamma\delta\mathbf{1} \le T_{\mu_{k+1}} V_{\mu_k}$$
>
> And again from applying the T operator on $\left\| V_k - V_{\mu_k} \right\| \le \delta$,
>
> $$T V_k \ge T V_{\mu_k} - \gamma\delta\mathbf{1}$$
>
> Which bring us to to the main conclusion that
>
> $$T_{\mu_{k+1}} V_{\mu_k} \ge T V_{\mu_k} - (\epsilon + 2\gamma\delta)\mathbf{1}$$

✓

3. Show that

$$T_{\mu_{k+1}} V_{\mu_k} \ge V_{\mu_k} - (\epsilon + 2\gamma\delta)\mathbf{1}$$

> From the last part we can obtain the following inequality:
>
> $$T_{\mu_{k+1}} V_{\mu_k} \ge T_{\mu_k} V_{\mu_k} - (\epsilon + 2\gamma\delta)\mathbf{1}$$
>
> As $T_{\mu_k} V_{\mu_k} = V_{\mu_k}$, then
>
> $$T_{\mu_{k+1}} V_{\mu_k} \ge V_{\mu_k} - (\epsilon + 2\gamma\delta)\mathbf{1}$$

4. Show that given V_k if $T_{\mu_k} V_k \le V_k + r\mathbf{1}$, then

$$V_{\mu_k} \le V_k + \frac{r}{1-\gamma}\mathbf{1}$$

Note that $T^0 V_k = V_k$ where $T^0 = \mathbf{I}$, an identity matrix.

> Applying the T operator to both sides gives us
>
> $$T_{\mu_k} V_k \le V_k + r\mathbf{1} \;\Rightarrow\; T_{\mu_k}(T_{\mu_k} V_k) \le T_{\mu_k} V_k + \gamma r\mathbf{1}$$
>
> which along the original inequality gives us
>
> $$T_{\mu_k} V_k \le V_k + (\gamma r + r)$$
>
> Repeating this process gives us
>
> $$T_{\mu_k}(T^2_{\mu_k} V_k) \le T_{\mu_k} V_k + (\gamma^2 r + \gamma r)$$
>
> $$\Rightarrow T^3_{\mu_k} V_k \le V_k + \gamma^2 r + \gamma r + r$$
>
> Where the inequality below becomes evident:
>
> $$T^n_{\mu_k} V_k \le V_k + \gamma^{n-1} r + \cdots + \gamma r + r$$
>
> And as $n \to \infty$, we will have
>
> $$V_{\mu_k} \le V_k + \frac{r}{1-\gamma}\mathbf{1}$$

✓

Problem 3. We consider here the deterministic version of almost supermartingale convergence theorem study in the class. In particular, let $\{y_k\}$, $\{z_k\}$, and $\{w_k\}$ be nonnegative sequence satisfying

$$y_{k+1} \leq y_k - z_k + w_k$$

where w_k satisfies

$$\sum_{k=0}^{\infty} w_k < \infty$$

Show that

1. y_k converges. [Hint: Show that the sequence $V_k = y_k + \sum_{t=k}^{\infty} w_t$ is nonincreasing]

We have

$$y_{k+1} \leq y_k - z_k + w_k \leq y_k + w_k$$

where the second inquality comes from $\{z_k\}$ being nonnegative. We know that for the nonnegative sequence, if the whole sum is bounded by some positive value c, then the partial sums are also bounded, that is

$$\sum_{t=0}^{\infty} w_t \leq c < \infty \rightarrow \sum_{t=k}^{\infty} w_t \leq c$$

But we can tighten this even more. From part 2 of this problem, we know that $\{w_k\}$ converges to zero. That is, there exists a τ, such that

$$\exists \tau \ \text{ s.t. } \ w_k = 0 \ \forall k > \tau$$

Then

$$\begin{aligned} y_{k+1} &\leq y_k + w_k \leq y_k + c \\ y_{k+1} &\leq y_k \ \ \forall k > \tau \end{aligned}$$ ✓

Since we have both lower-boundedness and non-increasing properties for $\{y_k\}$, we conclude that it converges. This comes from HW1 where we showed that for such non-increasing lower bounded (non-decreasing upper bounded) sequence, $\lim_{k\to\infty} y_k$ exist.

2. $\lim_{k\to\infty} w_k = 0$

We define partial sum $S_n = \sum_{k=1}^{n} w_k$ as the summation of the first nth terms of $\{w_k\}$. This partial sum is bounded (as stated in the problem) and monotone non-decreasing. From Q2.1 of HW1, we know that a non-decreasing upper bounded sequence will converge. Assume $\lim_{n\to\infty}\{S_n\} = u$, then

$$S_{n+1} - S_n = w_{n+1} \Rightarrow u - u = 0 \Rightarrow \lim_{n\to\infty}\{w_n\} = 0$$ ✓