6.7950 Fall 2022: - Recitation 3 Handout

1 Warmup

Let's look at certain transformation that can be applied to an MDP without affecting the underling problem. Consider an infinite horizon discounted MDP with

- States $s \in S$
- Actions $a \in A$
- Policies $\pi \in \Pi$
- Discount factor γ

1. Assume we have an upper bound r^{max} for the reward function such that $r(s,a) \leq r^{max}, \forall s \in S, a \in A$. Prove that $\forall s \in S, \pi \in \Pi$, we have that

$$V^{\pi}(s) \leq \frac{r^{max}}{1-\gamma} \tag{1}$$

2. Assume that, besides the aforementioned r^{max}, you also have a lower bound $r^{min} \neq r^{max}$ such that $r^{min} \leq r(s,a)$. Use these values to create a modified MDP with rewards $\overline{r}$ such that $\forall \pi in \Pi$, we have the modified value function $\overline{V}$ satisfying $0 \leq \overline{V}^{\pi}(s) \leq 1$ and that both MDP share the same optimal policy π^*.

Solution:

1. We can expand the expression for V^{π} as

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s, \pi(a))\right] \leq \sum_{t=0}^{\infty} \gamma^t r^{max} = \frac{r^{max}}{1-\gamma} \tag{2}$$

2. We can take the modified reward

$$\overline{r}(s,a) = \frac{r(s,a) - r^{min}}{r^{max} - r^{min}}(1-\gamma) \tag{3}$$

And by an analogous procedure as the previous item we get $\overline{V} \leq 1$. The lower bound is similarly obtained as $\overline{r}(s,a) \geq 0$ and thus $\overline{V}(s,a) \geq 0$. As all rewards are just scaled by a positive constant and shifted by the same amount, so are all the values $\overline{V}$ in comparison to V and the same actions that maximize one MDP also maximize the other.

2 Modified policy iteration

The are many ways to modify the policy iteration algorithm while still guaranteeing convergence. In homework 2, you are going to provide a general proof for one such modification. In this recitation, let's make a brief analysis of another version.

In this variant, the evaluation step for a new policy $\bar{\pi}$ is carried out iteratively for 2 steps only and averaged. In particular, the algorithm is (assuming finite-state infinite-horizon discounted problem, with finite action space)

- Let V_0 be an arbitrary n-dimensional vector.
- The algorithm generates a sequence of vectors $V_1, V_2, \ldots$ and stationary policies $\pi_0, \pi_1, \ldots$.
- Each policy π_t is chosen to satisfy

$$\mathcal{T}_{\pi_t} V_t = \mathcal{T} V_t$$

- The next vector V_{t+1} is computed according to

$$V_{t+1} = \frac{\mathcal{T}_{\pi_t} V_t + \mathcal{T}_{\pi_t}^2 V_t}{2}$$

Assuming $\mathcal{T} V_0 \geq V_0$, prove $\lim_{t\to\infty} V_t = V^*$. Hint: first prove that $V_{t+1} \geq \mathcal{T} V_t$

Solution: We first prove that $V_{t+1} \geq \mathcal{T} V_t$ by induction for all t. Let's assume this property holds for $V_t \geq \mathcal{T} V_{t-1}$, we then have that

$$\begin{aligned}
V_{t+1} &= \frac{\mathcal{T}_{\pi_t} V_t + \mathcal{T}_{\pi_t}^2 V_t}{2} \\
&= \frac{\mathcal{T}_{\pi_t} V_t + \mathcal{T}_{\pi_t} \mathcal{T} V_t}{2} \\
&\geq \frac{\mathcal{T}_{\pi_t} \mathcal{T} V_{t-1} + \mathcal{T}_{\pi_t} \mathcal{T}^2 V_{t-1}}{2} \\
&\geq \frac{\mathcal{T}_{\pi_t} \mathcal{T}_{\pi_{t-1}} V_{t-1} + \mathcal{T}_{\pi_t} \mathcal{T}_{\pi_{t-1}}^2 V_{t-1}}{2} \\
&= \mathcal{T}_{\pi_t} \frac{\mathcal{T}_{\pi_{t-1}} V_{t-1} + \mathcal{T}_{\pi_{t-1}}^2 V_{t-1}}{2} = \mathcal{T}_{\pi_t} V_t = \mathcal{T} V_t
\end{aligned}$$

That concludes the induction step. The base case follows from the assumption as:

$$V_1 = \frac{\mathcal{T}_{\pi_0} V_0 + \mathcal{T}_{\pi_0}^2 V_0}{2} = \frac{\mathcal{T} V_0 + \mathcal{T}_{\pi_0} \mathcal{T} V_0}{2} \geq \frac{V_0 + \mathcal{T}_{\pi_0} V_0}{2} = \frac{V_0 + \mathcal{T} V_0}{2} \geq= \frac{V_0 + V_0}{2} = V_0$$

Thus, we proved the auxiliary property that $V_{t+1} \geq \mathcal{T} V_t$. We can repeatedly apply this result to V_t in order to obtain

$$V_t \geq \mathcal{T} V_{t-1} \geq \mathcal{T}^2 V_{t-1} \geq \ldots \geq \mathcal{T}^{t-1} V_1 \geq \mathcal{T}^t V_0$$

and therefore we have that $V_t \geq \mathcal{T}^t V_0$. Since $\mathcal{T}^t V_0 \to V^*$ as $t \to \infty$, then $\lim_{t\to\infty} V_t \geq \lim_{t\to\infty} \mathcal{T}^t V_0 = V^*$. But V^* is optimal, so $V^* \geq \lim_{t\to\infty} V_t \geq V^*$, which means that $V_t \to V^*$ as $t \to \infty$.