MS&E338 Reinforcement Learning Lecture 6 - 4/22/20

Planning over Agent States

Lecturer: Ben Van Roy *Scribe: Eric Frankel, Yueyang Liu, William Chong*

1 Recap from Last Lecture

In the previous lecture, we learned about the Markov Decision Process abstraction for an environment. For an environment $\mathcal{E} = (\mathcal{O}, \mathcal{A}, \mathcal{H}, \rho)$, a compression function $f : \mathcal{H}^+ \to \mathcal{S} \cup \{\text{terminal}\}$ mapping from histories to states, a reward function r, and relevance weights ν, we can write it in terms of an MDP $M = (\mathcal{S}, A, \tilde{p}, \tilde{P}, \tilde{r})$, where

$$\begin{aligned} A &= \mathcal{A} \\ \tilde{\rho}(s) &= \sum_{o \in \mathcal{O}: f(o) = s} \rho(o) \\ \tilde{P}_{ass'} &= \frac{1}{\nu(\mathcal{H}_s)} \sum_{h \in \mathcal{H}_s} \nu(h) \sum_{h' \in \mathcal{H}_{s'}} P_{ahh'} \\ \tilde{r}_{as} &= \frac{1}{\nu(\mathcal{H}_s)} \sum_{h \in \mathcal{H}_s} \nu(h) \bar{r}_{ah} \end{aligned}$$

Similarly, we learned last time the following theorem (and its corresponding corollary):

Theorem 1 (error to performance)**.** *For any environment $\mathcal{E}$ and value function $\tilde{V}$ with corresponding action-value function $\tilde{Q}$, if $\tilde{\pi}$ a policy greedy with respect to $\tilde{V}$, then*

$$\bar{V}_* - \bar{V}_{\tilde{\pi}} \leq 2\mathbb{E}[\tau | \mathcal{E}, \tilde{\pi}] \|Q_* - \tilde{Q}\|_\infty$$

Corollary 2. *If compression function f is Δ-satisficing, then there exists a $\tilde{Q}$ such that there is a a greedy policy $\tilde{\pi}$ that satisfies*

$$\bar{V}_* - \bar{V}_{\tilde{\pi}} \leq \mathbb{E}[\tau | \mathcal{E}, \tilde{\pi}] \Delta, \quad \Delta = \max_{a \in \mathcal{A}} \max_{s \in S} \Big(\max_{h \in H_s} Q_*(h, a) - \min_{h \in H_s} Q_*(h, a) \Big)$$

2 The Projection Π_ν

We want to get some notion of a projection for a given action-value function $Q(h, a)$. Define the projection operator Π_ν as

$$\Pi_\nu Q(h, a) = \frac{1}{\nu(H_s)} \sum_{h' \in H_s} \nu(h') Q(h', a),$$

where s is the state corresponding to a history h given by the compression function $s = f(h)$, the notation here being a slight abuse of the formal definition of f. Note that this projection takes Q and replaces it with a piecewise function across cells of histories H_s whose values are the weighted average of the action-values in the cell. Note that Π_ν is not well defined for cells with weight 0. This projection operator can then give us the Bellman operator for an MDP as

$$\tilde{F} = \Pi_\nu F$$

where F is the Bellman operator for the environment. This can be shown through the following observation:

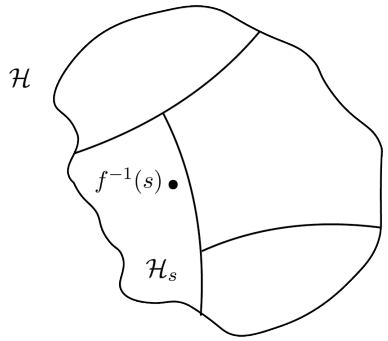

Figure 1: An illustration of $\mathcal{H}$. The intuition behind this is to divide the set of actionable histories into cells that correspond to a single given state of the MDP.

Observation 3. *Suppose $Q = \Pi_v Q$, i.e. Q is already represented in a piecewise manner. Then note that we have*

$$\begin{aligned}
\Pi_\nu FQ(h,a) &= \frac{1}{\nu(\mathcal{H}_s)} \sum_{h' \in \mathcal{H}_s} \nu(h') FQ(h', a) \\
&= \frac{1}{\nu(\mathcal{H}_s)} \sum_{h' \in H_s} \nu(h') \Big(\bar{r}_{ah'} + \sum_{h'' \in \mathcal{H}} P_{ah'h''} \max_{a' \in A} Q(h'', a') \Big) \\
&= \frac{1}{\nu(\mathcal{H}_s)} \Big(\sum_{h' \in \mathcal{H}_s} \nu(h') \bar{r}_{ah'} + \sum_{h' \in \mathcal{H}_s} \nu(h') \sum_{h'' \in \mathcal{H}} P_{ah'h''} \max_{a' \in A} Q(h'', a') \Big) \\
&= \frac{1}{\nu(\mathcal{H}_s)} \Big(\sum_{h' \in \mathcal{H}_s} \nu(h') \bar{r}_{ah'} + \sum_{h' \in \mathcal{H}_s} \nu(h') \sum_{s' \in \mathcal{S}} \sum_{h'' \in \mathcal{H}_{s'}} P_{ah'h''} \max_{a' \in A} Q(h'', a') \Big).
\end{aligned}$$

Note that since Q is piecewise constant by assumption, we have

$$\begin{aligned}
\Pi_\nu FQ(h,a) &= \frac{1}{\nu(\mathcal{H}_s)} \Big(\sum_{h' \in \mathcal{H}_s} \nu(h') \bar{r}_{ah'} + \sum_{h' \in \mathcal{H}_s} \nu(h') \sum_{s' \in \mathcal{S}} \sum_{h'' \in \mathcal{H}_{s'}} P_{ah'h''} \max_{a' \in A} Q(s', a') \Big) \\
&= \frac{1}{\nu(\mathcal{H}_s)} \Big(\sum_{h' \in \mathcal{H}_s} \nu(h') \bar{r}_{ah'} + \sum_{s' \in \mathcal{S}} \sum_{h' \in \mathcal{H}_s} \nu(h') \sum_{h'' \in \mathcal{H}_{s'}} P_{ah'h''} \max_{a' \in A} Q(s', a') \Big).
\end{aligned}$$

We can think of these two terms as the weighted average of the rewards and the weighted average of transition probabilities in the current agent state. Therefore, we can reduce this to

$$\Pi_\nu FQ(h,a) = \tilde{r}_{as} + \sum_{s' \in S} \tilde{P}_{ass'} \max_{a' \in A} Q(s', a') = (\tilde{F}Q)(s,a).$$

This formulation tells us that using the environment and projecting is the same as the Bellman operator for the MDP.

3 Properties of Π_ν

We begin by stating, but not proving, the contraction property of Π_ν.

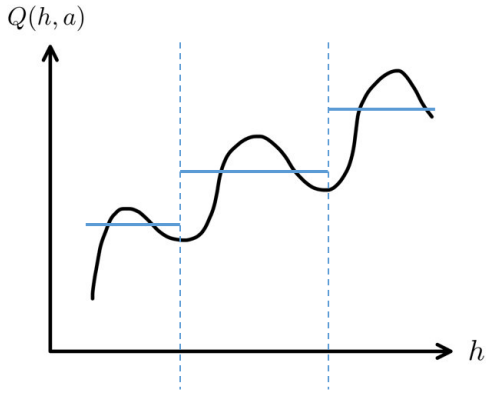

Figure 2: An illustration of weighted averages of action-values in cells

Theorem 4 (contraction)**.** *For action-value functions $Q, \bar{Q}$, we have that*

$$\|\Pi_\nu F Q - \Pi_\nu F \bar{Q}\|_{\infty,\tilde{\tau}} \leq \tilde{\gamma} \|Q - \bar{Q}\|_{\infty,\tilde{\tau}} \text{ for } \tilde{\gamma} = 1 - \frac{1}{\tilde{\tau}},\ \tilde{\tau} = \max_\pi \mathbb{E}[\tau | M, \pi, A = a]$$

for M the MDP associated with relevance weights ν.

We can use this theorem to bound the error between the optimal MDP action-value function and the optimal action-value function for an environment.

Theorem 5 (error bound)**.** *For $\tilde{Q}_*$ and Q_* satisfying the equations $\tilde{Q}_* = \Pi_\nu F \tilde{Q}_*$ and $Q_* = F Q_*$ respectively, we have the inequality*

$$\|Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}} \leq \frac{1}{1 - \tilde{\gamma}} \|Q_* - \Pi_\nu Q_*\|_{\infty,\tilde{\tau}}.$$

Proof. By the triangle inequality we have

$$\|Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}} \leq \|Q_* - \Pi_\nu Q_*\|_{\infty,\tilde{\tau}} + \|\Pi_\nu Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}}$$

Now, since Q_* and $\tilde{Q}_*$ satisfy the respective equations included above, by the contraction theorem we have

$$\begin{aligned}
\|Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}} &\leq \|Q_* - \Pi_\nu Q_*\|_{\infty,\tilde{\tau}} + \|\Pi_\nu Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}} \\
&= \|Q_* - \Pi_\nu Q_*\|_{\infty,\tilde{\tau}} + \|\Pi_\nu F Q_* - \Pi_\nu F \tilde{Q}_*\|_{\infty,\tilde{\tau}} \\
&\leq \|Q_* - \Pi_\nu Q_*\|_{\infty,\tilde{\tau}} + \tilde{\gamma} \|Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}}
\end{aligned}$$

Then by rearrangement we quickly have the inequality

$$\|Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}} \leq \frac{1}{1 - \tilde{\gamma}} \|Q_* - \Pi_\nu Q_*\|_{\infty,\tilde{\tau}}$$

Note that $\frac{1}{1-\tilde{\gamma}} = \tilde{\tau}$, so we equivalently have

$$\|Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}} \leq \tilde{\tau} \|Q_* - \Pi_\nu Q_*\|_{\infty,\tilde{\tau}}.$$

□

Observation 6. *Note that if our compression is Δ-satisficing, we have*

$$\|Q_* - \tilde{Q}_*\|_{\infty,\tilde{\tau}} \leq \frac{1}{1-\tilde{\gamma}} \|Q_* - \Pi_\nu Q_*\|_{\infty,\tilde{\tau}} \leq \frac{\tilde{\tau}}{\underline{\tilde{\tau}}} \Delta$$

where we define $\underline{\tilde{\tau}} := \min_{s \in \mathcal{S}, a \in \mathcal{A}} \tilde{\tau}(s,a)$. Combining the above properties of Π_ν, we can get a bound on the performance of value functions:

Theorem 7 (performance)**.** $\bar{V} - \bar{V}_{\tilde{\pi}_*} \leq \frac{2\tilde{\tau}^3}{\underline{\tilde{\tau}}} \Delta$ *where $\tilde{\pi}_*$ is a greedy policy with respect to $\tilde{Q}_*$.*

However, this notation is rather messy. For ease of use, we are going to assume **memory-less termination**, or where ther is a $(1-\tilde{\gamma})$ probability of termination at any time. Indeed, since the termination time is a random variable, we will have that the expected time of termination to be $\tilde{\tau} = \frac{1}{1-\tilde{\gamma}}$. Therefore, in the case of memory-less termination, our performance bound becomes

$$\bar{V} - \bar{V}_{\tilde{\pi}_*} \leq \frac{2\tilde{\tau}^3}{\underline{\tilde{\tau}}} = 2\frac{1}{(1-\tilde{\gamma})^2} \Delta$$

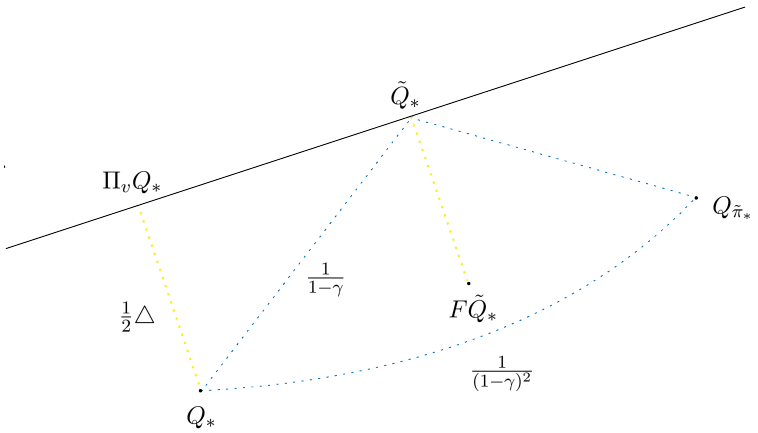

Figure 3: An illustration of the shortfall

Interestingly, in true reinforcement learning rather than MDP/planning problems, we will generally get the bounds $\bar{V} - \bar{V}_{\tilde{\pi}_*} \leq 2\frac{1}{(1-\tilde{\gamma})}\Delta$. This is the relevance weights are dynamic - we *learn* what states should have greater weight or importance. As such, suppose we define a set of relevance weights by

$$\nu_\pi(h) = \Big(\sum_{t=0}^{\infty} \rho P_\pi^t \Big)_h = \sum_{t=0}^{\infty} \mathbb{P}(H_t = h | \mathcal{E}, \pi)$$

Remark. $v_\pi(h)$ would not blow up to ∞, since the system has finite expected duration.

In other words, the relevance weights are tied to how often we see these histories. Now, recall that we have $\tilde{Q}_* = \Pi_\nu F \tilde{Q}_*$ as the optimal action-value function for the MDP. Alternatively, consider $\tilde{Q}_* = \Pi_{\nu_{\tilde{\pi}^*}} F \tilde{Q}_*$, where $\nu_{\tilde{\pi}^*}$ are the relevance weights given by the greedy policy with respect to $\tilde{Q}_*$. Then we have

$$\tilde{Q}_* = \Pi_\nu F \tilde{Q}_* \Rightarrow \bar{V}_* - \bar{V}_{\tilde{\pi}_*} \leq 2\tilde{\tau}^2 \Delta$$
$$\tilde{Q}_* = \Pi_{\nu_{\tilde{\pi}^*}} F \tilde{Q}_* \Rightarrow \bar{V}_* - \bar{V}_{\tilde{\pi}_*} \leq \tilde{\tau} \Delta$$