MS&E338 Reinforcement Learning Lecture 8 - Wednesday 29th April, 2020

Transient Behavior of Real-Time Value Iteration

Lecturer: Ben Van Roy *Scribe: Tina Diao & Abhinav Rangarajan*

1 Recap from Last Lecture

Last time we covered the algorithm *Real Time Value Iteration (RTVI)* (shown below as Algorithm 1) which involved simulating an environment and tuning, over episodes of simulation, a Q-function over state-action pairs. It argued that eventually the simulated actions become optimal. The simplifying assumptions made were

1. The termination was memoryless, i.e. at each time step, there is probability of $1-\gamma$ of termination regardless of history. It allows us to have a simple initialization of $\tilde{Q} = M$, where $M = \frac{1}{1-\gamma} \max\limits_{h,a} \overline{r}_{ah}$.

   As a result, the initial $\tilde{Q}$ dominates Q_* and the bellman operator, when applied, only decreases the value of $\tilde{Q}$.

2. There is sufficient agent state, i.e. for each agent state s, the histories compressed into s all have the same Q_* values.

Algorithm 1: Real-Time Value Iteration Pseudo-code

```
Initialization  $\tilde{Q}(s,a) = M, \ \forall s,a$;
for episode $l = 1,2,\dots$ do
    observe $o_0$;
    $h_0 \leftarrow (o_0), s_0 \leftarrow f(h_0)$;
    while State $s \neq$ terminated do
        $a_t \leftarrow arg\max_{a \in A} \tilde{Q}(s_t, a)$;
        execute $a_t$, observe $o_{t+1}$;
        $\tilde{Q}(s_t,a_t) \leftarrow \min\left\{(G_{h_t a_t}\tilde{Q})(s_t,a_t),\ \tilde{Q}(s_t,a_t)\right\}$ # ensures that $\tilde{Q}$ is monotonically decreasing ;
        $s_{t+1} \leftarrow f(s_t, a_t, o_{t+1})$, $h_{t+1} \leftarrow (h_t, a_t, o_{t+1})$;
    end
end
```

The operator G_{ha} is defined as

$$G_{\overline{h}\overline{a}}\tilde{Q}(s,a) = \begin{cases} \overline{r}_{\overline{h}\overline{a}} + \sum_{o \in \mathcal{O}} \rho(o|\overline{h}\overline{a}) \max_{a' \in \mathcal{A}} \tilde{Q}(f(s,a,o),a') & \text{if } s = f(\overline{h}), a = \overline{a} \\ Q(s,a) & \text{otherwise} \end{cases}$$

Today we will look at the transient behavior of an algorithm like this, which helps us understand whether or not and how an algorithm is efficient.

A key argument made last time was that upon convergence (of values along trajectories that are simulated) we have

$$\tilde{Q}(H_t, A_t) = F\tilde{Q}(H_t, A_t) \quad \forall H_t, A_t \text{ visited infinitely often}$$

i.e. bellman equation is satisfied. It follows that $\tilde{\pi}$ is optimal.

2 Shortfalls

Theorem 1. $\forall \pi, V\colon \mathcal{H} \to \mathbb{R}$

$$\rho V - \rho V_\pi = \mathbb{E}\big[\sum_{t=0}^{\tau-1} \underbrace{V(H_t) - (F_\pi V)(H_t)}_{\textit{error term}} \mid \mathcal{E}, \pi\big]$$

This theorem establishes an expression for the difference on the LHS and holds for any function V. ρ is the distribution over the initial history and ρV is the average value over all possible initial conditions. $\rho V_\pi = \bar{V}_\pi$ is the expected episode reward for policy π. The theorem does not assume memoryless termination.

Proof.

$$\begin{aligned}
\mathbb{E}\big[\sum_{t=0}^{\infty} V(H_t) \mid \mathcal{E}, \pi\big] &= \rho \sum_{t=0}^{\infty} P_\pi^t V \\
&= \rho V + \rho \sum_{t=1}^{\infty} P_\pi^t V \\
&= \rho V + \rho \sum_{t=0}^{\infty} P_\pi^{t+1} V \quad \text{re-indexing } t \\
&= \rho V - \rho V_\pi + \rho V_\pi + \rho \sum_{t=0}^{\infty} P_\pi^{t+1} V \\
&= \rho V - \rho V_\pi + \rho \sum_{t=0}^{\infty} P_\pi^t \bar{r}_\pi + \rho \sum_{t=0}^{\infty} P_\pi^{t+1} V \quad \text{rewriting } V_\pi \text{ in } 3^{rd} \text{ term} \\
&= \rho V - \rho V_\pi + \rho \sum_{t=0}^{\infty} P_\pi^t (\bar{r}_\pi + P_\pi V) \\
&= \rho V - \rho V_\pi + \rho \sum_{t=0}^{\infty} P_\pi^t (F_\pi V) \\
&= \rho V - \rho V_\pi + \mathbb{E}\big[\sum_{t=0}^{\infty} (F_\pi V)(H_t) \mid \mathcal{E}, \pi\big]
\end{aligned}$$

□

Theorem 2. *$\forall \tilde{Q}\colon \mathcal{H} \times \mathcal{A} \to \mathbb{R}$, if $\tilde{V}(h) = \max_{a \in \mathcal{A}} \tilde{Q}(h, a)$ $\forall h$ and $\tilde{\pi}$ greedy w.r.t. $\tilde{Q}$, then*

$$\rho \tilde{V} - \rho V_{\tilde{\pi}} = \mathbb{E}\big[\sum_{t=0}^{\tau-1} \tilde{Q}(H_t, A_t) - (F\tilde{Q})(H_t, A_t) \mid \mathcal{E}, \tilde{\pi}\big]$$

The differences from Theorem 1 are that we are looking at $(\textit{history}, \textit{action})$ pairs and use $\tilde{\pi}$ now. The LHS is the "calibration error".

Proof.

$$
\begin{aligned}
\tilde{Q}(H_t, A_t) &= \max_a Q(\tilde{H_t}, a) = \tilde{V}(h) \\
\mathbb{E}\big[(F\tilde{Q})(H_t, A_t) \mid H_t\big] &= \mathbb{E}\big[\bar{r}_{A_t H_t} + \sum_{h' \in \mathcal{H}} P_{A_t H_t h'} \cdot \max_{a' \in \mathcal{A}} \tilde{Q}(h', a') \mid H_t\big] \quad \text{writing Bellman operator} \\
&= \mathbb{E}\big[\bar{r}_{A_t H_t} + \sum_{h' \in \mathcal{H}} P_{A_t H_t h'} \cdot \tilde{V}(h') \mid H_t\big] \\
&= \bar{r}_{\tilde{\pi} H_t} + \sum_{h' \in \mathcal{H}} P_{\tilde{\pi} H_t h'} V(h') \quad \text{average is taken across actions, which is given by } \tilde{\pi} \\
&= F_{\tilde{\pi}} \tilde{V} \quad \text{definition of } F_{\tilde{\pi}} \\
\rho\tilde{V} - \rho V_{\tilde{\pi}} &= \mathbb{E}\Big[\sum_{t=0}^{\tau-1} \tilde{V}(H_t) - (F_{\tilde{\pi}}\tilde{V})(H_t) \mid \mathcal{E}, \tilde{\pi}\Big] \quad \text{applying Theorem 1} \\
&= \mathbb{E}\Big[\sum_{t=0}^{\tau-1} \tilde{V}(H_t) - \mathbb{E}\big[(F\tilde{Q})(H_t, A_t) \mid H_t\big] \mid \mathcal{E}, \tilde{\pi}\Big] \quad \text{using expression of } F_{\tilde{\pi}}\tilde{V} \\
&= \mathbb{E}\big[\sum_{t=0}^{\tau-1} \tilde{Q}(H_t, A_t) - (F\tilde{Q})(H_t, A_t) \mid \mathcal{E}, \tilde{\pi}\big] \quad \text{by law of total expectation}
\end{aligned}
$$

□

3 An Observation

We make an important observation arising from the result above that strikes at the heart of many theoretical results in reinforement learning. We study the difference between the optimal value of an episode $\bar{V}_*$ and the value from a policy $\tilde{\pi}$ that is greedy w.r.t. some approximation $\tilde{Q}$. We start by subtracting and adding back the same term, $\rho\tilde{V}$.

$$
\begin{aligned}
\bar{V}_* - \bar{V}_{\tilde{\pi}} &= \overbrace{(\rho V_* - \rho\tilde{V})}^{\text{pessimism}} + \overbrace{(\rho\tilde{V} - \rho V_{\tilde{\pi}})}^{\text{miscalibration}} \\
&\leq \rho\tilde{V} - \rho V_{\tilde{\pi}} \qquad \text{since the ``pessimism'' term } \leq 0 \text{ in RTVI} \\
&= \mathbb{E}\big[\sum_{t=0}^{\tau-1} \underbrace{\tilde{Q}(H_t, A_t) - (F\tilde{Q})(H_t, A_t)}_{\text{Bellman Error}} \mid \mathcal{E}, \tilde{\pi}\big]
\end{aligned}
$$

The "pessimism" term refers to that of our current prediction, generated by $\tilde{Q}$ that gives us $\tilde{V}$. V_* is the maximum possible reward and our prediction is $\tilde{V}$. Hence the difference is how pessimistic $\tilde{V}$ is relative to V_*. This difference is ≤ 0 (in RTVI) as we start with a large $\tilde{V}$ that always dominates V_*.

The "miscalibration" term refers to the extent to which we overestimate how well we do in an episode, as $\tilde{V}$ predicts how well we will do in an episode and $V_{\tilde{\pi}}$ is how well we actually do in an episode. An important element of making a reinforement learning algorithm to work well is to have some form of "optimism" such that the "pessimism" term is negative, then calibrate the "miscalibration" term to be 0. Combined, it forces the LHS to be 0 as well since it cannot be negative.

"Bellman Error" is 0 if $\tilde{Q} = Q_*$ since $Q_* = FQ_*$. If $\tilde{Q} \neq Q_*$, the difference (the "Bellman Error") can be eliminated by replacing $\tilde{Q}$ with the value iteration, $F\tilde{Q}$.

This relationship informs that if the policy is not good, "Bellman Error" has to be large. However, if the "Bellman Error" is large, we can make it small by applying value iteration. In addition, value iteration need not be applied everywhere, but only for the histories and actions we see as we only care about what is in the expectation term. This lies at the heart of many reinforcement learning analyses. There can be complicated noises in the environment that require technical treatment such as using probabilities to ensure the noises average out elegantly. The essence is to take advantage of properties like this.

The intuition is *almost* enough to characterize the transient property of RTVI, which has technical caveats that make it more complicated to analyze. They arise from the fact that RTVI is updating $\tilde{Q}$ and the policy *during the episode*. We will talk about extension of results so far to address those.

4 Specialize to RTVI

We now extend Theorem 2 to the **Real-Time Value Iteration** (RTVI) algorithm. To recap, Theorem 1 and 2 helped us bound the miscalibration error when updating the V or Q function for a single history or state-action at a time. The reason RTVI requires additional analysis is because the policy also changes intra-episodically along with the Q function; Theorem 1 and 2 had assumed the policy remained fixed throughout the episode. Specifically, there are two important aspects to notice:

- *Policy Changes*: The next action given a certain state S_t at time t is determined in the RTVI algorithm by taking the argmax, i.e. determined greedily:

$$A_t = \arg\max_{a \in A} \tilde{Q}(s_t, a) \tag{1}$$

  At each time step, we are updating the $\tilde{Q}$ function values for all histories h associated with state s_t, or $f(h) = s_t$. Therefore, the greedy policy $\tilde{\pi}$ with respect to $\tilde{Q}$ must also be changing. (NOTE: $s_t = f(h)$ where f is the compression function).

- *Intra-Episodic Changes*: It also matters that RTVI is updating state-action pairs, not just individual histories. If the specific history h_t was the only entry in the Q function updated by RTVI, since we cannot visit that same history again in the same episode, we could have applied Theorem 1 and 2 to RTVI directly without any issue. In RTVI, however, the greedy policy $\tilde{\pi}$ maps *states* to actions. We may visit a state again in the same episode, so the intra-episodic changes have material effect on the expected performance of that episode. We denote the updated $\tilde{Q}$ at time step t as $\tilde{Q}_t$ (notice the subscript) going forward.

The following theorem is an extension of Theorem 2 to RTVI. It is basically Theorem 2 with a few modifications, so no proof is necessary beyond explaining the modifications.

The goal of the theorem is to provide an estimate of how much our value function $\tilde{V}$ changes over an episode when running the RTVI algorithm on the vector. That estimate is quantified by the difference $\rho\tilde{V} - \rho V_{\tilde{\pi}}$. As a reminder, the reason that our value function changes is because of miscallibration. The reason this is useful is that when we study transient behavior of RTVI in the next section, we can just sum this estimate across episodes to get the overall miscallibration error bound.

Theorem 3. ***Extension of Theorem 2 to RTVI***
$\forall \tilde{Q}_0$, if $\tilde{V}_0(h) = \max_{a \in A} \tilde{Q}_0(h, a)$, and $\tilde{\pi}_{H_t}$ is followed by RTVI, and we assume that the environment $\mathcal{E}$ does not have self-transitioning states, then

$$\rho\tilde{V}_0 - \rho V_{\tilde{\pi}} = E\left[\sum_{t=0}^{\tau-1} \left(\tilde{Q}_t(s_t, A_t) - G_{H_t}\tilde{Q}_t(s_t, A_t)\right)\middle| \mathcal{E}, \tilde{\pi}, \tilde{Q}_0\right] \tag{2}$$

A few points require explanation:

- $\tilde{\pi}_{H_t}$ is followed by RTVI (operator denoted as G_{H_t})

  Again, the RTVI-operator G_{H_t} refers to the fact that $\tilde{\pi}_{H_t}$ is greedy with respect to the latest value function $\tilde{Q}_t$.

- Environment $\mathcal{E}$ does not have self-transitioning states

  Technically, G_{H_t} should be applied to $\tilde{Q}_{t+1}(s_t, A_t)$. Referring back to the pseudo-code of RTVI in the first section above, in line number 7 and 8, we are updating the value of $\tilde{Q}$, and then in line number 9, we are stepping to the next state / history, with the updated values. So when we apply the operator G_{H_t}, as shown below, we are assessing the maximum-valued action in the next state / history, with the values already updated.

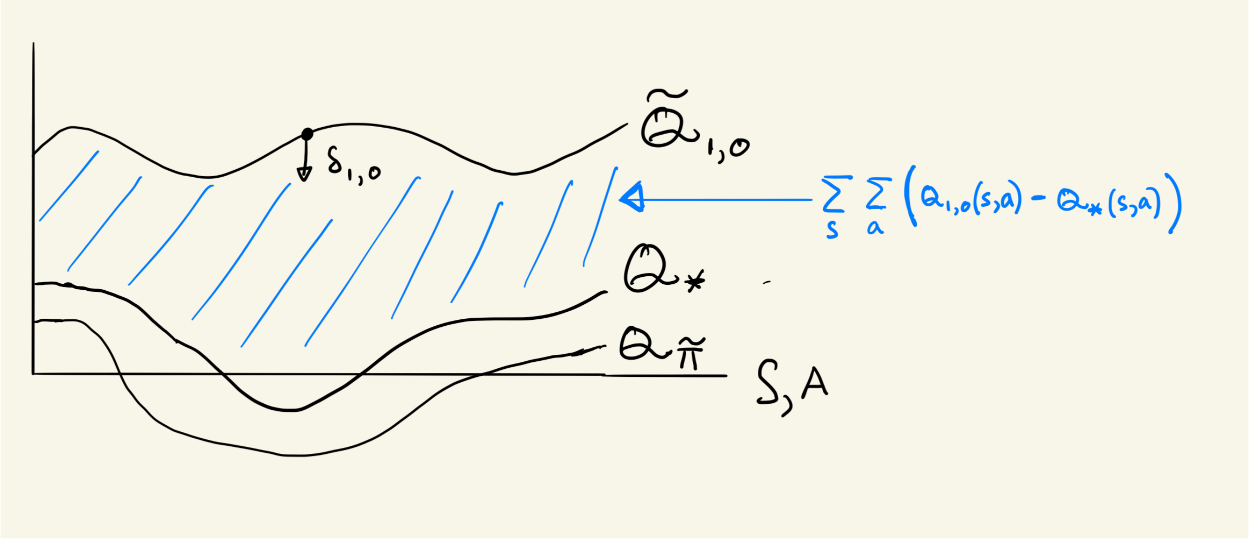
  $$G_{H_t}\tilde{Q}(s_t, A_t) = \bar{r}_{A_t H_t} + \sum_{h' \in \mathcal{H}} P_{A_t H_t h'} \cdot \max_{a' \in \mathcal{A}} \tilde{Q}(h', a') \mid H_t$$

  So we should be applying the operator G_{H_t} on $\tilde{Q}_{t+1}$, but this would not allow us to extend Theorem 2. To circumvent this, assume we can rewrite the environment $\mathcal{E} \to \mathcal{E}'$, where the expected value is the same, but $\mathcal{E}'$ does not have any self-transitions; therefore, the possible states / histories we can transition to in the next time step differ from the current state / history, so we can extend Theorem 2. Removing self-transitions in an environment involves re-scaling the probabilities and rewards.

5 Transient Behavior of RTVI

The "transient" behavior refers to analyzing the behavior across episodes. In the previous section, an estimate of the intra-eposoidal change in the value function was given. Now we take the sum across episodes.

Let π_l be the policy followed in episode l (this episode is changing within the episode). Over several episodes, we expect the quantity $\overline{V}_* - \overline{V}_{\pi_l}$ to go to zero, as the policy begins to converge to optimal.



$$\begin{aligned}
\overline{V}_* - \overline{V}_{\pi_l} &= \underbrace{(\rho V_* - \rho\tilde{V})}_{\leq 0} + (\rho\tilde{V} - \rho V_{\pi_l}) \\
&\leq (\rho\tilde{V} - \rho V_{\pi_l}) \\
&= E\left[\sum_{t=0}^{\tau-1}\left(\underbrace{\tilde{Q}_{l,t}(s_t, A_t) - G_{H_t}\tilde{Q}_{l,t}(s_t, A_t)}_{\text{denote this } \delta_{l,t}}\right)\middle|\, \epsilon, \tilde{\pi}, \tilde{Q}_{l,0}\right] \text{using Extension to Theorem 2} \\
&= E\left[\sum_{t=0}^{\tau-1}\delta_{l,t} \,\middle|\, \epsilon, \tilde{\pi}, \tilde{Q}_{l,0}\right]
\end{aligned}$$

Notice that the Q functions are now indexed by l for episode and t for time step (as before).

Now we sum across all episodes as follows:

$$\begin{aligned}
\sum_{l=0}^{L} \overline{V}_* - \overline{V}_{\pi_l} &\leq \sum_{l=0}^{L} E\left[\sum_{t=0}^{\tau-1} \delta_{l,t} \middle| \epsilon, \tilde{\pi}, \tilde{Q}_{l,0}\right] \\
&= E\left[\sum_{l=0}^{L}\sum_{t=0}^{\tau-1} \delta_{l,t} \middle| \epsilon, \tilde{\pi}, \tilde{Q}_{l,0}\right] \\
&\leq \sum_{s \in S}\sum_{a \in A}\left(\tilde{Q}_{1,0}(s,a) - Q_*(s,a)\right) \text{ Worst case, visit all (state, action)-pairs many times} \\
&\leq |S| \times |A| \times \max_{s,a}\left(\tilde{Q}_{1,0}(s,a) - Q_*(s,a)\right)
\end{aligned}$$

In the derivation above, we imagine that in the worst case, we visit all (state, action)-pairs many times across episodes and drive their Q function value to towards Q_*. We do not need the expectation operator, because we do not consider transition probabilities; we are assuming the worst case behavior. The most that the function Q can change is the difference between the starting point for Q, which we denote $\tilde{Q}_{1,0}$ (episode 1, time step 0), and Q_*.