MS&E338 Reinforcement Learning Lecture 9 - May 4, 2020

RTVI with Δ-Satisficing State Representation

Lecturer: Ben Van Roy *Scribe: Adithya M. Devraj*

Agenda

(i) Bug fix from lecture on 27th April

(ii) RTVI with Δ-satisficing

(iii) Q-learning

1 Bug Fix

Recall the RTVI algorithm: At the k^{th} iteration, given the current history-action pair is $(h_k\,, a_k)$, we update the Q-value estimate:

$$\widetilde{Q}_{k+1} \leftarrow G_{h_k,a_k}\widetilde{Q}_k \tag{1}$$

where,

$$(G_{\overline{h},\overline{a}}\widetilde{Q})(s,a) = \begin{cases} \overline{r}_{\overline{h}\overline{a}} + \sum_{o\in\mathcal{O}} \rho(o|\overline{h},\overline{a}) \max_{a'} \widetilde{Q}(f(s,a,o)\,,a')\,, & \text{if } s = f(\overline{h}) \text{ and } a = \overline{a} \\ \widetilde{Q}(s,a)\,, & \text{otherwise} \end{cases} \tag{2}$$

Two properties of the operator $G_{\overline{h},\overline{a}}$ that was used in the convergence proof of the algorithm during lecture 7 were:

(i) Monotonicity: For each history-action pair $(\overline{h},\overline{a})$,

$$(G_{\overline{h},\overline{a}}\widetilde{Q})(s,a) \le \widetilde{Q}(s,a)$$

(ii) It has a unique fixed point, $\widetilde{Q}_*$:

$$(G_{\overline{h},\overline{a}}\widetilde{Q}_*)(s,a) = \widetilde{Q}_*(s,a)$$

Given a large enough initialization, the monotonictity property was used to argue that $\{\widetilde{Q}_n\}$ is a sequence of non-increasing functions that is bounded below by $\widetilde{Q}_*$, and therefore it has to converge to $\widetilde{Q}_*$. However there is a bug in this argument: the montonicity property does not guarantee that the sequence $\{\widetilde{Q}_n\}$ is non-increasing. This is specifically due to the fact that if $s = f(\overline{h}) = f(\overline{h}')$, $\overline{h} \neq \overline{h}'$,

$$(G_{\overline{h}'\overline{a}'}G_{\overline{h}\overline{a}})\widetilde{Q}(s,a) \le (G_{\overline{h}\overline{a}})\widetilde{Q}(s,a)$$

does not necessarily hold if $(\overline{h},\overline{a}) \neq (\overline{h}',\overline{a}')$. This is made precise in the following counter example.

1.1 Counter Example

Consider an environment with $\mathcal{A} = \{1\}$, $\mathcal{O} = \{1,2\}$, $\mathcal{S} = \{1,2,3\}$. Each state has two histories that are grouped together: $\{h_i\,, h_i'\}$, for $1 \le i \le 3$. Let's say the episode begins with one of two histories: h_1 or h_1'. For each of the two histories, there are two possible next histories (say equally likely – it doesn't matter): h_2 and h_2' starting from h_1, h_3 and h_3' starting from h_1'. The associated rewards are:

$$r_{h_1ah_2} = r_{h_1ah_2'} = 0 \quad \text{and} \quad r_{h_1'ah_3} = r_{h_1'ah_3'} = 1$$

The episode reaches the terminal state t in the next time-step, starting from any of the four histories $\{h_2, h_2', h_3, h_3'\}$, with associated rewards:

$$r_{h_2 a t} = r_{h_2' a t} = 0 \quad \text{and} \quad r_{h_3 a t} = r_{h_3' a t'} = -1$$

See Figure 1 for details.

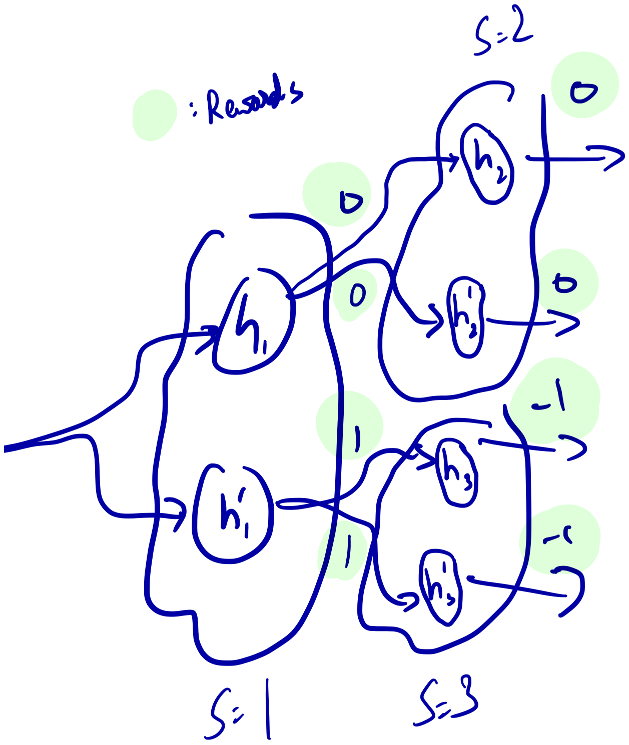

Figure 1: Counterexample for arguments on the convergence of the previous RTVI algorithm

Note that we have a sufficient state representation because it's easy to verify that:

$$Q_*(h_1, a) = Q_*(h_1', a) = \widetilde{Q}^*(1, a) = 0$$
$$Q_*(h_2, a) = Q_*(h_2', a) = \widetilde{Q}^*(2, a) = 0$$
$$Q_*(h_3, a) = Q_*(h_3', a) = \widetilde{Q}^*(3, a) = -1$$

Now, let's look at the behavior of the RTVI algorithm. Let's assume we initialize with $\widetilde{Q}_0(i, a) = 0$ for all states i. Suppose we start the episode with h_1, then we have:

$$\widetilde{Q}_1(1, a) = 0$$

and the $\widetilde{Q}$ values will remain unchanged for rest of the states. The next history will either be h_2 or h_2', and irrespective of which of these histories we end up with, we have

$$\widetilde{Q}_2(2, a) = 0$$

with the rest of the $\widetilde{Q}$ values unchanged. The episode then terminates. Suppose in the next episode, we begin with history h_1', we have:

$$\widetilde{Q}_3(1, a) = +1$$

and in the following step, we update:

$$\widetilde{Q}_4(3,a) = -1$$

and the episode terminates. In the third episode, suppose we begin with h'_1 again, note that the $\widetilde{Q}(3,a) = -1$ now (as opposed to 0 the last time h'_1 was observed), so we obtain:

$$\widetilde{Q}_5(1,a) = 0$$

From next time-step on-wards, none of the $\widetilde{Q}$-values get updated, since $\widetilde{Q} = \widetilde{Q}_*$.

The key thing to note is $\widetilde{Q}_1(1,a) = \widetilde{Q}_2(1,a) < \widetilde{Q}_3(1,a) = \widetilde{Q}_4(1,a) > \widetilde{Q}_5(1,a)$, breaking the monotonicity argument from Lecture 7.

1.2 Bug Fix

A simple modification to the RTVI algorithm will resolve this issue. Modifying the update rule in (1) to the following

$$\widetilde{Q}_{k+1}(s_k,a_k) \leftarrow \min\left((G_{h_k,a_k}\widetilde{Q}_k)(s_k,a_k), \widetilde{Q}_k(s_k,a_k)\right), \tag{3}$$

it is easy to see that the monotonicity argument will hold: $\widetilde{Q}_{k+1} \leq \widetilde{Q}_k$ for each k. Moreover, $\widetilde{Q}_*$ is the stationary point of (3). Therefore, the above algorithm converges using the same arguments of lecture 7: $\{\widetilde{Q}_n\}$ is a sequence of non-increasing functions that is bounded below by the fixed point $\widetilde{Q}_*$.

1.3 Expected short-fall

Now let us look at the expected shortfall for the above algorithm. Recall that we have the following bound: In the ℓ^{th} episode, with π_ℓ denoting the greedy policy with respect to $\widetilde{Q}_{\ell,0}$,

$$\overline{V}_* - \overline{V}_{\pi_\ell} \leq \mathsf{E}\Big[\sum_{t=0}^{\tau_\ell-1}\left(\widetilde{Q}_{\ell,t}(S_t,A_t) - (G_{H_t,A_t}\widetilde{Q}_{\ell,t})(S_t,A_t)\right)\Big|\mathcal{E},\pi_\ell,\widetilde{Q}_{\ell,0}\Big]$$

The right hand side can be further upper bounded to obtain

$$\overline{V}_* - \overline{V}_{\pi_\ell} \leq \mathsf{E}\Big[\sum_{t=0}^{\tau_\ell-1}\underbrace{\Big(\widetilde{Q}_{\ell,t}(S_t,A_t) - \min\Big\{(G_{H_t,A_t}\widetilde{Q}_{\ell,t})(S_t,A_t)\,,\widetilde{Q}_{\ell,t}(S_t,A_t)\Big\}\Big)}_{\delta_{\ell,t}}\Big|\mathcal{E},\pi_\ell,\widetilde{Q}_{\ell,0}\Big] \tag{4}$$

where $\delta_{\ell,t} \geq 0$ is precisely the change in $\widetilde{Q}$-value corresponding to the state-action pair (S_t,A_t), when applying the RTVI algorithm.

Given an initial $\widetilde{Q}_{0,0} = \frac{1}{1-\gamma}r_{\max}$ at episode 0, the maximum possible *decrease*, $\delta_{\ell,t}$ that corresponds to each state-action pair (s,a) is $\frac{2}{1-\gamma}r_{\max}$ (since $\widetilde{Q}_*$ is lower bounded by $-\frac{1}{1-\gamma}r_{\max}$):

$$\widetilde{Q}_{0,0}(s,a) - \widetilde{Q}_*(s,a) \leq \frac{2}{1-\gamma}r_{\max}$$

We therefore have the following bound on the sum of short fall over L episodes:

$$\begin{aligned}\sum_{\ell=1}^{L}\left(\overline{V}_* - \overline{V}_{\pi_\ell}\right) &\leq \sum_{\ell=1}^{L}\mathsf{E}\Big[\sum_{t=0}^{\tau_\ell-1}\delta_{\ell,t}\Big|\mathcal{E},\pi_\ell,\widetilde{Q}_{\ell,0}\Big] \\ &\leq \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\Big(\widetilde{Q}_{0,0}(s,a) - \widetilde{Q}_*(s,a)\Big) \\ &\leq \frac{2|\mathcal{S}||\mathcal{A}|r_{\max}}{1-\gamma}\end{aligned} \tag{5}$$

2 RTVI with Δ-satisficing

Suppose we have a Δ-satisficing state representation. We are interested in the fundamental limits of RL algorithms in this case. We will assume memoryless termination to simplify the math.

We will first show that the optimism can break down under this representation, if we use the RTVI algorithm (3) as it is. That is, it is possible for $\widetilde{Q}(s,a) < Q^*(h,a)$, for $s = f(h)$. Consider a state-representation where there exists a state $s = f(h) = f(h')$, with $Q^*(h,a) = \Delta/2$ and $Q^*(h',a) = -\Delta/2$. This is possible under the Δ-saticficing assumption. In this case, suppose we visit h' first in the RTVI algorithm, we will update $\widetilde{Q}(s,a) = -\Delta/2$, which can be pessimistic in the perspective of history h, since the true value $Q^*(h,a) = \Delta/2 > \widetilde{Q}(s,a)$.

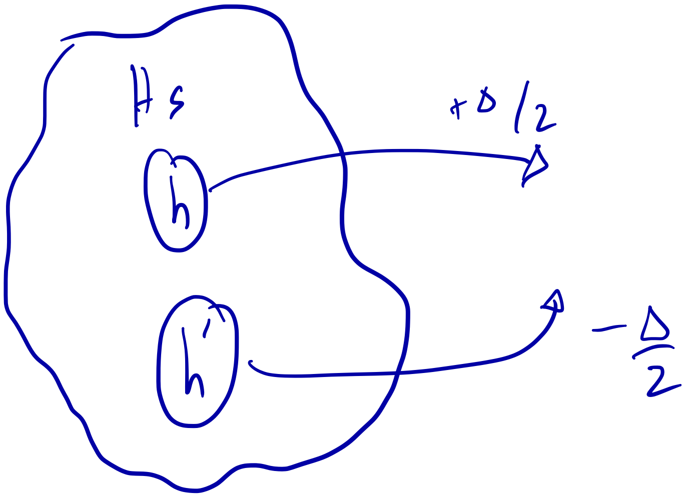

Figure 2: An example of Δ-satisficing state representation.

It is easy to see that a simple modification to the RTVI algorithm will resolve the optimism issue:

$$\widetilde{Q}_{k+1}(s_k,a_k) \leftarrow \min\big((G_{h_k,a_k}\widetilde{Q}_k)(s_k,a_k) + \Delta, \widetilde{Q}_k(s_k,a_k)\big) \tag{6}$$

Of-course, in this case, $\widetilde{Q}(s,a)$ can be off by Δ from $Q^*(h,a)$ for some histories $h = f^{-1}(s)$.

Now we consider the expected short-fall of the above algorithm over L episodes. The expected shortfall in episode ℓ is bounded by:

$$\overline{V}_* - \overline{V}_{\pi_\ell} \leq \mathsf{E}\Big[\sum_{t=0}^{\tau_\ell - 1} \delta'_{\ell,t} \Big| \mathcal{E}, \pi_\ell, \widetilde{Q}_{\ell,0}\Big] \tag{7}$$

where, $\delta'_{\ell,t}$ is the *net decrease* of the $\widetilde{Q}$ estimate in time-step t of episode ℓ. This is upper bounded by sum of two terms: $\delta_{\ell,t}$ that is defined in (4), plus an additional Δ that appears due to the modification in update rule (6):

$$\delta'_{\ell,t} \leq \delta_{\ell,t} + \Delta \tag{8}$$

Using (7) and (8), we can obtain a bound similar to (5) for the algorithm (6):

$$\begin{aligned}\sum_{\ell=1}^{L}\left(\overline{V}_* - \overline{V}_{\pi_\ell}\right) &\leq \sum_{\ell=1}^{L}\mathsf{E}\Big[\sum_{t=0}^{\tau_\ell-1}\delta'_{\ell,t}\Big|\mathcal{E},\pi_\ell,\widetilde{Q}_{\ell,0}\Big] \\ &\leq \frac{2|\mathcal{S}||\mathcal{A}|r_{\max}}{1-\gamma} + \sum_{\ell=1}^{L}\mathsf{E}\Big[\tau_\ell\Delta\Big|\mathcal{E},\pi_\ell,\widetilde{Q}_{\ell,0}\Big] \\ &= \frac{2|\mathcal{S}||\mathcal{A}|r_{\max}}{1-\gamma} + \frac{\Delta L}{1-\gamma}\end{aligned} \tag{9}$$

where we have used the memoryless property to obtain $\mathsf{E}[\tau_\ell|\mathcal{E},\pi_\ell,\widetilde{Q}_{\ell,0}] = 1/(1-\gamma)$ for all ℓ. This implies

$$\lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\left(\overline{V}_* - \overline{V}_{\pi_\ell}\right) \leq \frac{\Delta}{1-\gamma} \tag{10}$$

The arguments are illustrated in Figure 3 below.

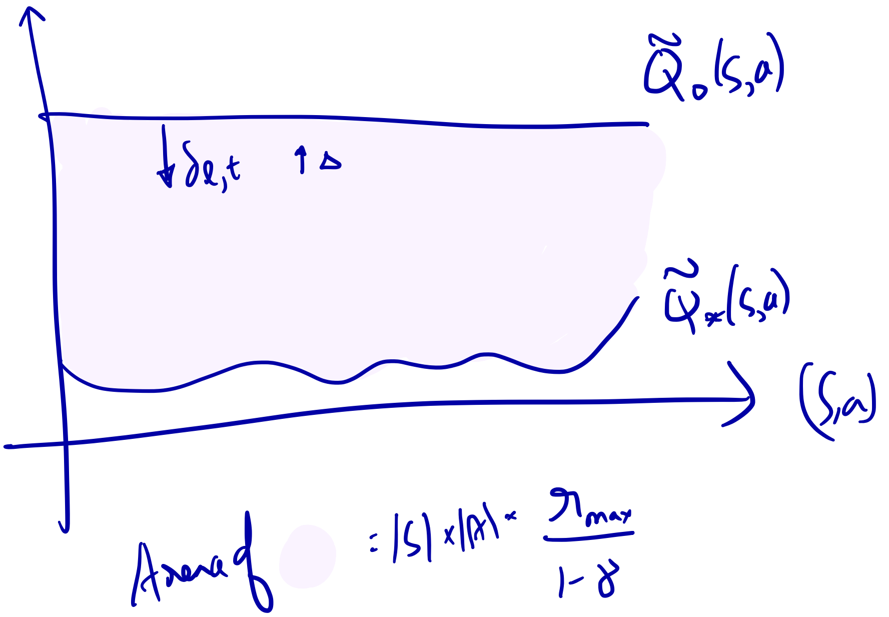
$\tilde{Q}_0(s,a)$
$\downarrow\delta_{\ell,t}$ $\uparrow\Delta$
$\tilde{Q}_*(s,a)$
(s,a)
Area of
$:|S|\times|A|\times\frac{r_{max}}{1-\gamma}$

Figure 3: Expected shortfall for Δ-satisficing state representation.

This elegant bound is quite intuitive: at each time-step, the value function V_{π_ℓ} can be off by Δ from V_*, due to the Δ-satisficing assumption. Therefore, the average per-episode shortfall scales Δ by the average episode length $1/(1-\gamma)$ to obtain (10).

Recall that the worst case bound (in particular if the relevance weights for the state-representation is not chosen properly) can look like:

$$\lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\left(\overline{V}_* - \overline{V}_{\pi_\ell}\right) \leq \frac{\Delta}{(1-\gamma)^2} \tag{11}$$

It is interesting to note that the RTVI algorithm obtains a better dependency on the factor $1/(1-\gamma)$ than this worst case bound.

3 Q-learning

The Q-learning algorithm can be thought of as a "stochastic" version of the RTVI algorithm (1,2). Given the current state s_k, and action a_k, we observe the next history h_{k+1}, and update:

$$\widetilde{Q}_{k+1}(s_k, a_k) \leftarrow \widetilde{Q}_k(s_k, a_k) + \alpha_k \Big(r_{h_k a_k h_{k+1}} + \max_{a'} \widetilde{Q}(f(s_k, a_k, o_{k+1}), a') - \widetilde{Q}_k(s_k, a_k) \Big) \tag{12}$$

where α_k is a "step-size" sequence. With $\alpha_k \equiv 1$, note that the recursion becomes

$$\widetilde{Q}_{k+1}(s_k, a_k) \leftarrow \Big(r_{h_k a_k h_{k+1}} + \max_{a'} \widetilde{Q}(f(s_k, a_k, o_{k+1}), a') \Big) \tag{13}$$

which looks exactly like the RTVI algorithm (1,2), except that we have replaced the expectations with samples.