MS&E338 Reinforcement Learning	Lecture 9 - April 30, 2018
Aggregation Continued, Real-Time Q-Learning	
Lecturer: Ben Van Roy	*Scribe: Sheng Li, Jie Wu*

1 Approximate Value Iteration with Q-Functions Continued

1.1 Recap

Last time, we talked about Approximate Value Iteration with Q-Functions or μ-sampled Q-learning in the context of the aggregation case. We have defined following parameters.
State-action weights:

$$\mu(s, a). \tag{1}$$

Projections:

$$\Pi_\mu Q = \underset{Q_\theta}{\operatorname{argmin}} \|Q - Q_\theta\|_{2,\mu}. \tag{2}$$

Fixed point:

$$\tilde{Q} = \Pi_\mu F \tilde{Q}. \tag{3}$$

Notice that we use piece-wise constant approximation here. It is not recommended to use in real applications but it is one of the simplest representations. We have $Q_\theta = \Phi\theta$ where

$$\Phi = \begin{bmatrix} 1 & 0 & \\ 1 & 0 & \\ \vdots & \vdots & \cdots \\ 0 & 1 & \\ 0 & 1 & \\ \vdots & \vdots & \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|\times K} \tag{4}$$

is a matrix with K columns being indicator vectors of the K partitions of space $\mathcal{X}$ ($\mathcal{X} = \mathcal{S} \times \mathcal{A}$).

1.2 Error and performance bound intuition

We have also analyzed error bound and performance bound.
Error bound:

$$\|Q^* - \tilde{Q}\|_\infty \leqslant \frac{1}{1-\alpha}\|Q^* - \Pi Q^*\|_\infty. \tag{5}$$

Performance bound:

$$\|Q^* - Q_{\tilde{\pi}}\|_\infty \leqslant \frac{2\alpha}{(1-\alpha)^2}\|Q^* - \Pi Q^*\|_\infty, \tag{6}$$

Generally, approximate VI with Q-functions does not guarantee convergence. Figure 1 Shows a one iteration of approximate VI that leads to divergence. The Bellman operator is a contraction in the infinity norm (the square denotes value functions with infinity norm equivalent to Q_{θ_t}), while the projection is in the μ-weighted 2-norm. The dotted line is the subspace of Q functions parameterized by θ. However, with specific settings (states aggregation), we can avoid this problem as shown in Figure 2. In two dimensional case, the only possible aggregate approximation is to have a single partition as shown with the dotted diagonal line for Q_θ which leads to the contraction of ΠF.

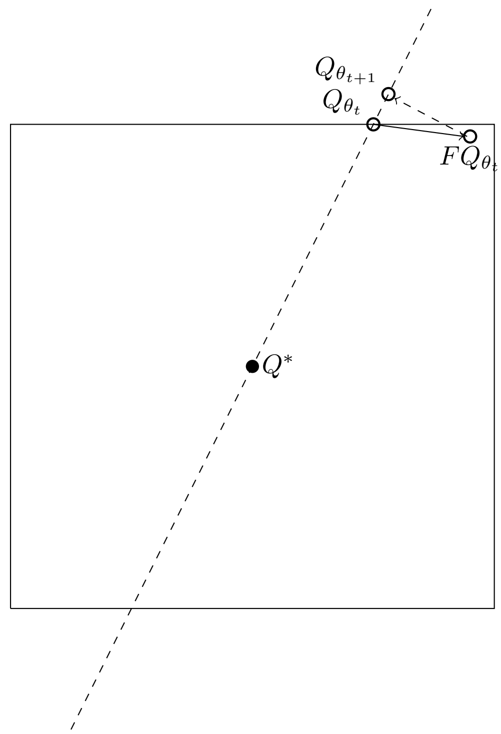

Figure 1: Possible divergence behavior of approximate VI with Q-functions

As for performance bound shown in Figure 3, Q^* is the fixed point for F. F is a contraction mapping. Notice that $\tilde{\pi}$ is the greedy policy w.r.t. $\tilde{Q}$, which means we have $F_{\tilde{\pi}}\tilde{Q} = F\tilde{Q}$ (since operator F takes actions that maximize Q-values anyway). So at each iteration, $\tilde{Q}$ needs to be closer to Q^* and $Q_{\tilde{\pi}}$ at the same time. This indicates that Q^* and $Q_{\tilde{\pi}}$ need to be relatively close to each other.

2 Real-Time Q-Learning

2.1 "Discounted" Case v.s. Discounted Case

Previously, we were working on "discounted" case. For each state in $\mathcal{S}$, the system terminates with probability $1-\alpha$ and we define a stochastic matrix

$$\overline{P} = \frac{P}{\alpha}, \tag{7}$$

The intuition is that, it is equivalent to the discounted case in which future rewards are discounted by factor α. Now we are moving to the real discount case with time discount factor α. That is:

$$(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \rho) \Rightarrow (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \alpha, \rho), \tag{8}$$

where previously it is sub-stochastic and now it is stochastic without termination. Then the objective becomes:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t r_{t+1}\right] = \max_{\pi} \rho^\top \sum_{t=0}^{\infty} \alpha^t P_\pi^t \overline{R}_\pi, \tag{9}$$

where

$$\overline{R}_\pi(s) = \mathbb{E}\big[r_{t+1} | s_t = s, a_t = \pi(s)\big]. \tag{10}$$

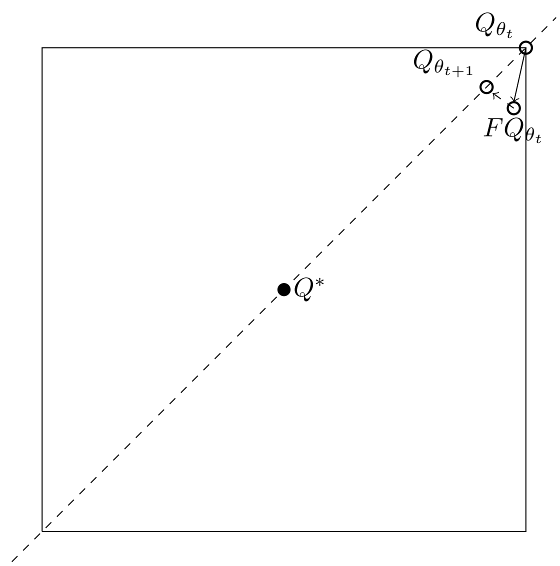

Figure 2: Convergence of approximate VI with Q-functions with specific setting

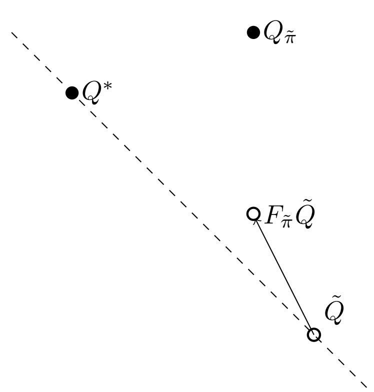

Figure 3: Performance bound

It can be shown to be mathematically equivalent to the previous case. Still in the context of the aggregation case, the real time Q-learning algorithm is outlined in Algorithm 1. In this algorithm, instead of sampling

Algorithm 1: Real time Q-learning

Initialize with some θ
Sample $s_0 \sim \rho$
for $t = 0, 1, 2 \cdots$ **do**
  Select a_t
  Observe r_{t+1}, s_{t+1}
  $\theta \leftarrow \theta + \gamma_t \big(\nabla_\theta Q_\theta(s_t, a_t)\big)\big(r_{t+1} + \alpha \max_{a' \in \mathcal{A}} Q(s', a') - Q_\theta(s_t, a_t)\big)$

each state-action pair according to μ, we simulate the real process and update θ as we actually visit the state-action pair. The algorithm does not guarantee convergence. However, if it converges, the performance of $\tilde{Q}$ is much better than that of the μ-sampled Q-learning.

2.2 Performance Bound for Real-Time Q-Learning

Firstly, we need to clarify a few notations. We let $\tilde{\mu}(s, a)$ be the relative frequency of (s, a) under policy $\tilde{\pi}$ (assuming it is consistent regardless of starting point). $\tilde{Q}$ is the fixed point with $\tilde{Q} = \Pi_{\tilde{\mu}} F \tilde{Q}$. Then the performance bound for the real-time Q-learning algorithm is given by

$$\tilde{\mu}^\top (Q^* - Q_{\tilde{\pi}}) \leqslant \frac{\alpha}{1-\alpha} \left\| Q^* - \Pi_{\tilde{\mu}} Q^* \right\|_\infty , \tag{11}$$

where the left hand side of the bound can be interpreted as the performance loss:

$$\tilde{\mu}^\top (Q^* - Q_{\tilde{\pi}}) = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \tilde{\mu}(s, a)((Q^*(s, a) - Q_{\tilde{\pi}}(s, a)). \tag{12}$$

By the definition of the optimal state-action value function $Q^*(s, a)$, multiplying $(1 - \alpha)$ on the both sides, we can find that

$$(1-\alpha) Q^*(s, a) = (1-\alpha) \left(r(s, a) + \mathbb{E} \left[\sum_{t=1}^{\infty} \alpha^t r_{t+1} \mid s_0 = s, a_0 = a, a_t = \pi^*(s_t) \text{ for } t \geq 1 \right] \right). \tag{13}$$

As α approaches 1, we have the following observations

$$\lim_{\alpha \uparrow 1} (1-\alpha) Q^*(s, a) = \lambda^*(s, a) \overset{(a)}{=} \lambda^*; \tag{14}$$

$$\lim_{\alpha \uparrow 1} (1-\alpha) Q_{\tilde{\pi}}(s, a) = \lambda_{\tilde{\pi}}(s, a) \overset{(b)}{=} \lambda_{\tilde{\pi}}, \tag{15}$$

where λ is the expected average reward, and (a) and (b) follow from our assumption that the average reward does not depend on the starting state and action.

Applying (11) with (14) and (15), we may find

$$\lambda^* - \lambda_{\tilde{\pi}} = \lim_{\alpha \uparrow 1} (1-\alpha) \tilde{\mu}^\top (Q^* - Q_{\tilde{\pi}}) \leqslant \left\| Q^* - \Pi_{\tilde{\mu}} Q^* \right\|_\infty . \tag{16}$$

Recalling (6), the performance bound of the approximate Q-value iteration (i.e. μ-sampled Q-learning), we can rewrite (6) similarly to (16) as

$$\lambda^* - \lambda_{\tilde{\pi}} \leqslant \lim_{\alpha \uparrow 1} (1-\alpha) \left\| Q^* - Q_{\tilde{\pi}} \right\|_\infty \leqslant \lim_{\alpha \uparrow 1} \frac{2\alpha}{1-\alpha} \left\| Q^* - \Pi Q^* \right\|_\infty . \tag{17}$$

Note that in μ-sampled Q-learning, as α approaches 1 the upper bound of the performance bound becomes vacuous since $\lim_{\alpha\to 1}\frac{2\alpha}{1-\alpha}=\infty$, and the bound is actually tight. Real time Q-learning does not have this problem.

Now we proceed to prove the performance bound for the real-time Q-learning algorithm (11).

Proof: First, here are two key observations, which we will prove later:

1. $\tilde{\mu}^\top \tilde{Q} = \tilde{\mu}^\top Q_{\tilde{\pi}}$.
2. $\tilde{\mu}^\top \tilde{Q} = \tilde{\mu}^\top F\tilde{Q}$.

With these key observations we may continue the proof:

$$\begin{aligned}
\tilde{\mu}^\top (Q^* - Q_{\tilde{\pi}}) &\overset{(c)}{=} \tilde{\mu}^\top (Q^* - \tilde{Q}) \\
&\overset{(d)}{=} \tilde{\mu}^\top (Q^* - F\tilde{Q}) \\
&\overset{(e)}{\leqslant} \|Q^* - F\tilde{Q}\|_\infty \\
&\overset{(f)}{\leqslant} \alpha\|Q^* - \tilde{Q}\|_\infty \\
&\overset{(g)}{\leqslant} \frac{\alpha}{1-\alpha}\|Q^* - \Pi_{\tilde{\mu}}Q^*\|_\infty,
\end{aligned} \tag{18}$$

where (c) and (d) correspond with the key observations. (e) results from the fact that the weighted average cannot exceed the maximum value. (f) is true since operator F is a contraction mapping. (g) follows from Proposition 2 in lecture 8.

We proceed to prove key observation 1, $\tilde{\mu}^\top \tilde{Q} = \tilde{\mu}^\top Q_{\tilde{\pi}}$.

Define inner product

$$\left\langle Q, \bar{Q}\right\rangle_{\tilde{\mu}} = \sum_{s,a} \tilde{\mu}(s,a) Q(s,a)\bar{Q}(s,a). \tag{19}$$

Then we can have the following results

1. Balance equation: $\left\langle 1, \tilde{P}_{\tilde{\pi}} Q\right\rangle_{\tilde{\mu}} = \langle 1, Q\rangle_{\tilde{\mu}}$, where $\tilde{P}_{\tilde{\pi}} = \tilde{P}_{\tilde{\pi}_{(s,a),(s',a')}}$ denotes the probability of transition between state-action pairs under greedy policy $\tilde{\pi}$.
2. Leveraging the self-adjointness of the projection operator, $\langle 1, \Pi_{\tilde{\mu}} Q\rangle_{\tilde{\mu}} = \langle \Pi_{\tilde{\mu}} 1, Q\rangle_{\tilde{\mu}} = \langle 1, Q\rangle_{\tilde{\mu}}$. The latter equality results from the fact that 1 is in the span of Φ.
3. $$\langle 1, Q_{\tilde{\pi}}\rangle_{\tilde{\mu}} = \left\langle 1, \sum_{t=0}^{\infty} \alpha^t \tilde{P}_{\tilde{\pi}}^t \bar{R}\right\rangle_{\tilde{\mu}} = \sum_{t=0}^{\infty}\alpha^t \left\langle 1, \tilde{P}_{\tilde{\pi}}^t \bar{R}\right\rangle_{\tilde{\mu}} = \frac{1}{1-\alpha}\left\langle 1, \bar{R}\right\rangle_{\tilde{\mu}}$$

   where the first equality uses the definition of Q value following policy $\tilde{\pi}$, the second equality uses the linearity of the inner product, the last equation uses result 1 listed above.

Then we can derive

$$\begin{aligned}
\left\langle 1, \tilde{Q}\right\rangle_{\tilde{\mu}} &\overset{(h)}{=} \left\langle 1, \Pi_{\tilde{\mu}} F\tilde{Q}\right\rangle_{\tilde{\mu}} \\
&\overset{(i)}{=} \left\langle 1, F\tilde{Q}\right\rangle_{\tilde{\mu}} \\
&\overset{(j)}{=} \left\langle 1, \bar{R} + \alpha\tilde{P}_{\tilde{\pi}}\tilde{Q}\right\rangle_{\tilde{\mu}} \\
&\overset{(k)}{=} \left\langle 1, \bar{R}\right\rangle_{\tilde{\mu}} + \alpha\left\langle 1, \tilde{P}_{\tilde{\pi}}\tilde{Q}\right\rangle_{\tilde{\mu}} \\
&\overset{(l)}{=} \left\langle 1, \bar{R}\right\rangle_{\tilde{\mu}} + \alpha\left\langle 1, \tilde{Q}\right\rangle_{\tilde{\mu}},
\end{aligned} \tag{20}$$

where (h) is based on $\tilde{Q}$ being the fixed point with $\tilde{Q} = \Pi_{\tilde{\mu}} F\tilde{Q}$; (i) results from the result 2 listed above; (j) uses the definition of the Bellman operator F: $F\tilde{Q}$ being the addition of the immediate reward $\bar{R}$ and the discounted future reward under greedy policy $\tilde{\pi}$; (k) is based on the linearity of inner product; (l) follows the balance equation listed as result 1. Rearranging gives

$$\left\langle 1, \tilde{Q} \right\rangle_{\tilde{\mu}} = \frac{1}{1-\alpha} \left\langle 1, \bar{R} \right\rangle_{\tilde{\mu}}. \tag{21}$$

Then combining result 3 with (21), we may derive

$$\langle 1, Q_{\tilde{\pi}} \rangle_{\tilde{\mu}} = \left\langle 1, \tilde{Q} \right\rangle_{\tilde{\mu}}. \tag{22}$$

Expanding the inner products in (22) yields

$$\tilde{\mu}^\top Q_{\tilde{\pi}} = \tilde{\mu}^\top \tilde{Q}. \tag{23}$$

Thus, we proved key observation 1. Key observation 2 follows from equation (i) in (20). The proof of the performance bound of the real-time Q-learning is complete.