MS&E338 Reinforcement Learning Lecture 8 - April 25, 2018

Episodic TD/Q-learning, Aggregation, Sampled Q-learning

Lecturer: Ben Van Roy *Scribe: Jiayue Wan, Yue Hui*

1 Episodic Temporal Difference Learning with Multiple Actions per State

In the previous lecture, we discussed the simplified case where there is only one possible action corresponding to each state. In that case, we used $V(s)$ in place of $Q(s,a)$ and gave a temporal difference learning algorithm. For such autonomous case (one action per state) we showed that

$$(\theta^* - \theta_t)^T \dot{\theta}_t \geq (1-\alpha)||\Phi\theta^* - \Phi\theta_t||_{2,\rho}^2.$$

Let $z_t = ||\theta^* - \theta_t||_2^2$. Then,

$$\dot{z_t} = -2(\theta^* - \theta_t)^T \dot{\theta}_t \leq -2(1-\alpha)||\Phi\theta^* - \Phi\theta_t||_{2,\rho}^2. \tag{1}$$

Hence, from equation 1, for z_t to converge, it is necessary to have that $||\Phi\theta^* - \Phi\theta_t||_{2,\rho}^2 \to 0$.

This time, we want to generalize to cases where there might be several actions per state. Below is an episodic temporal difference learning algorithm with multiple actions per state.

Algorithm 1 Episodic TD / Q-learning

initialize with some $\theta \in \mathbb{R}^K$
for $l = 0, 1, 2, \cdots$ **do**
  observe s_0
  for $t = 0, 1, 2, \cdots, \tau - 1$ **do**
    select a_t
    observe r_{t+1}, s_{t+1}
    $\theta := \theta + \gamma_l \nabla_\theta Q_\theta(s_t, a_t)\left(r_{t+1} + \max_{a \in A} Q_\theta(s_{t+1}, a) - Q_\theta(s_t, a_t)\right)$
  end for
end for

Today, we will prove similar results for the case with multiple actions per state in the aggregation case (where Algorithm 1 still have no known convergence result) and propose a sample-based $Q-$learning algorithm (Algorithm 2) based on the results. We will also compare Algorithm 1 and Algorithm 2 in this lecture.

2 Aggregation

Remark In all the followings, we are considering cases where, from each state, the termination probability is $1-\alpha$ for some $\alpha \in (0,1)$.

In this section, we describe "aggregation", a special case of the linear representation, where we partition the space into subsets $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K$, and the feature extractors are the indicator functions of the partition:

$$\phi_k(s,a) = \begin{cases} 1 & \text{if } (s,a) \in \mathcal{X}_k \\ 0 & \text{otherwise.} \end{cases}$$

One possible matrix representation of the "aggregation" could be

$$\Phi = \begin{bmatrix} 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ 0 & 1 & \cdots \\ \vdots & \vdots & \cdots \\ 0 & 0 & \cdots \end{bmatrix}.$$

2.1 Convergence Result

Now, we present a convergence result for this aggregation in the case where we sample (s, a) from a fixed distribution. We start the analysis with a proposition.

Proposition 1. *$\forall \mathcal{X}_1, \cdots, \mathcal{X}_K$, $\forall \mu$ s.t. $\sum_{(s,a)\in\mathcal{X}_k} \mu(s,a) > 0$ for all k, the operator Π where*

$$\Pi Q = \underset{Q_\theta}{argmin} \sum_{s,a} \mu(s,a) \left(Q(s,a) - Q_\theta(s,a)\right)^2$$

satisfies $||\Pi||_\infty \leq 1$.

Proof. We consider

$$\begin{aligned} &\min_\theta \sum_{(s,a)\in\mathcal{X}} \mu(s,a) \left(Q(s,a) - Q_\theta(s,a)\right)^2 \\ = \ &\min_\theta \left(\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{X}_k} \mu(s,a)(Q(s,a) - \theta_k)^2 \right). \end{aligned}$$

It is straightforward that the minimizer $\hat{\theta}_k$ is given by

$$\hat{\theta}_k = \frac{\sum_{(s,a)\in\mathcal{X}_k} \mu(s,a) Q(s,a)}{\sum_{(s,a)\in\mathcal{X}_k} \mu(s,a)}. \tag{2}$$

According to equation 2, we have that $|\hat{\theta}_k| \leq ||Q||_\infty$, which leads to $||\hat{\theta}||_\infty \leq ||Q||_\infty$. On the other hand, by our definition of the mapping Π, we also have that $||\Pi Q||_\infty = ||\hat{\theta}||_\infty$. Hence, we find that $||\Pi Q||_\infty \leq ||Q||_\infty$, which indicates that $||\Pi||_\infty \leq 1$ as desired. □

Remark Based on Proposition 1, it is straightforward that ΠF is a contraction mapping: for any $Q, \bar{Q}$, we find that

$$||\Pi F Q - \Pi F \bar{Q}||_\infty \leq ||FQ - F\bar{Q}||_\infty \leq \alpha ||Q - \bar{Q}||_\infty.$$

Hence, ΠF has a unique fixed point with respect to $||\cdot||_\infty$.

2.1.1 Error Bound for Fixed Point of ΠF

Now that we have shown that ΠF is a contraction mapping, we prove an error bound for fixed point of ΠF, analogous to our analysis in the last lecture.

Proposition 2. *Suppose $\tilde{Q} = \Pi F \tilde{Q}$ is the fixed point of ΠF. Then,*

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

Proof. We derive the error bound for $\tilde{Q}$,

$$\begin{aligned} ||Q^* - \tilde{Q}||_\infty &\leq ||Q^* - \Pi F Q^*||_\infty + ||\Pi F Q^* - \Pi F \tilde{Q}||_\infty \\ &\leq ||Q^* - \Pi Q^*||_\infty + \alpha ||Q^* - \tilde{Q}||_\infty, \end{aligned}$$

which leads to a upper bound in the error

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

□

Remark This error bound is tight.

2.1.2 Performance Bound

Now, we consider the performance of the greedy policy. Let $\tilde{\pi}$ be greedy policy w.r.t. $\tilde{Q}$. We have

$$\begin{aligned} ||Q_{\tilde{\pi}} - Q^*||_\infty &\leq ||Q_{\tilde{\pi}} - F_{\tilde{\pi}} \tilde{Q}||_\infty + ||F\tilde{Q} - Q^*||_\infty \\ &\leq \alpha ||Q_{\tilde{\pi}} - \tilde{Q}||_\infty + \alpha ||\tilde{Q} - Q^*||_\infty \\ &\leq \alpha ||Q_{\tilde{\pi}} - Q^*||_\infty + \alpha ||Q^* - \tilde{Q}||_\infty + \alpha ||\tilde{Q} - Q^*||_\infty, \end{aligned}$$

which leads to

$$||Q_{\tilde{\pi}} - Q^*||_\infty \leq \frac{2\alpha}{1-\alpha} ||Q^* - \tilde{Q}||_\infty. \tag{3}$$

Combining Proposition 2 and inequality 3, we find that

$$||Q_{\tilde{\pi}} - Q^*||_\infty \leq \frac{2\alpha}{(1-\alpha)^2} ||Q^* - \Pi Q^*||_\infty. \tag{4}$$

Remark If we look at this bound 4, the left hand side is the performance error for the value function that we obtained from the algorithm at the step where we find a fixed point of ΠF, and the right hand side represents some constant multiples of the best possible error. This constant multiple could be fairly large for large α, as there are two factors of $(1-\alpha)$ in the denominator.

We may want to normalize the value by a factor of $1-\alpha$ (since each episode lasts $1/(1-\alpha)$ time in expectation), so that we will have

$$(1-\alpha) ||Q_{\tilde{\pi}} - Q^*||_\infty \leq \frac{2\alpha}{1-\alpha} ||Q^* - \pi Q^*||_\infty.$$

However, we still have a factor of $1-\alpha$ left. There are several ways that we could eliminate one factor of $1-\alpha$, and hence improve the performance bound. One such method is described below.

Suppose we define the projection matrix Π_μ where $\mu = \mu_{\tilde{\pi}}$. If we solve the more complicated equation,

$$\tilde{Q} = \Pi_{\mu_{\tilde{\pi}}} F \tilde{Q},$$

then we will eliminate one factor of $1-\alpha$ in the performance bound. Next time we will talk about the episodic TD/Q-learning in detail, and we will discuss why this seems to solve the $1-\alpha$ problem.

3 μ-Sampled Q-learning Algorithm

Now, we are going to introduce a stochastic approximation version of this algorithm, where we do not have to have F explicitly.

For a fixed μ, we perform the following stochastic approximation algorithm.

Algorithm 2 μ-Sampled Q-learning

initialize with some θ
for $i = 1, 2, \cdots$ **do**
  sample $(s_i, a_i) \sim \mu$
  observe (r_i, s_i')
  $\theta := \theta + \gamma_i \nabla_\theta Q_\theta(s_i, a_i) \left(r_i + \max_{a \in \mathcal{A}} Q_\theta(s_i', a) - Q_\theta(s_i, a_i)\right)$
end for

To be more precise, the aggregation case is where

$$\nabla_{\theta_k} Q_\theta(s_i, a_i) = \begin{cases} 1 & \text{if } (s_i, a_i) \in \mathcal{X}_k \\ 0 & \text{otherwise.} \end{cases}$$

Algorithms 1 and 2 are very similar. However, they are different in that algorithm 1 is sampled real time, while algorithm 2 is sampled based on a a fixed distribution. Algorithm 1 could have much better performance if it converges, but we do not yet know how to prove that. The analysis of performance bound of Algorithm 2 in the aggregation case is done in Section 2.

3.1 General Stochastic Approximation Update in Aggregation

Now, what about a general stochastic approximation update?

$$\theta := \theta + \gamma_i f(\theta, (s_i, a_i, r_i, s_i')) \tag{5}$$

Then, we find that,

$$\mathbb{E}[f_k(\theta, (s_i, a_i, r_i, s_i')] = \mu(\mathcal{X}_k) \left((G\theta)_k - \theta_k\right) \tag{6}$$

where G is a contraction mapping applied to θ (for example, in the case we discussed before, G could be derived from ΠF s.t. $G\theta_0$ to be the θ from the equation $Q_\theta = \Pi F Q_{\theta_0}$ in our previous example).

Hence, this algorithm accords with the ODE we want as in Lecture 7 and we could have desired convergence.

4 Non-linear Representation

Throughout this lecture, we discussed linear representations of the $Q-$function generally. In nonlinear cases, things could explode and the convergence results we saw last time might no longer hold.

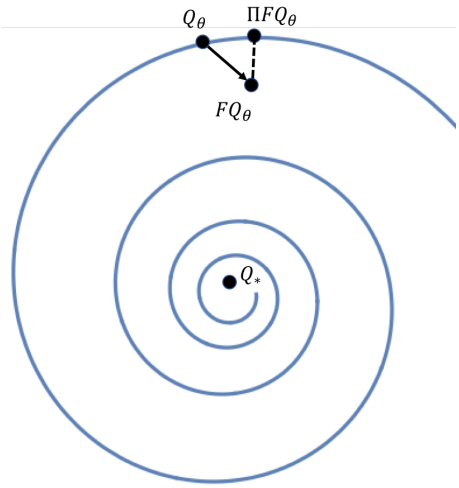

Figure 1: Example of divergence in non-linear representation.

The intuitive reason for this non-convergence example (as shown in Figure 1) is that if the restricted space behave badly (in this case, like a spiral), and then the projection of FQ could spiral out and hence becomes further and further away from Q_*.

Remark Nowadays, nonlinear representations raise increasing interests, for example, from neural networks (as shown in Figure 2).

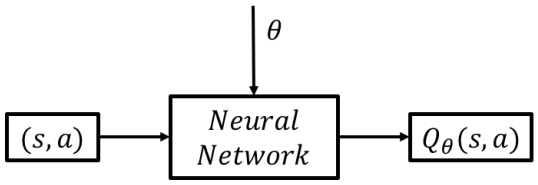

Figure 2: A nonlinear, neural network representation.

For projects, it would be interesting to explore some non-linear representations, where some sort of convergence holds.