MS&E338 Reinforcement Learning Lecture 4 - 04.11.2018

Q-Learning and Stochastic Approximation

Lecturer: Ben Van Roy *Scribe: Christopher Lazarus — Javier Sagastuy*

In this lecture we study the convergence of Q-Learning updates via stochastic approximation results.

1 Q-learning update

Assume we are given data samples of the following form: $(s_k, a_k, r_{k+1}, s_{k+1}) \quad k = 0, 1, 2, \ldots$ and we iteratively apply the following update to the state-action value function:

$$Q_{k+1}(s,a) = \begin{cases} (1-\gamma_k)Q_k(s,a) + \gamma_k \left(r_{k+1} + \max_{a' \in A} Q_k(s_{k+1}, a')\right) & \text{if } s = s_k, a = a_k \\ Q_k(s,a) & \text{otherwise} \end{cases} \tag{1}$$

Proposition 1. *If each $(s,a) \in \mathcal{S} \times \mathcal{A}$ is sampled infinitely often and we have $\{\gamma_k\}$ deterministic satisfying*

$$\forall (s,a) \sum_{k:(s_k,a_k)=(s,a)} \gamma_k = \infty, \qquad \sum_{k:(s_k,a_k)=(s,a)} \gamma_k^2 < \infty,$$

then $Q_k \to Q^*$

Originally, in class, there were a couple of issues with the formulation of the above proposition: we want each state-action pair to be sampled infinitely often. If an adversary is deciding that, it can try to game the system to just choose an update at particular times, for example:

- Adversary picks particular state action pairs when γ_k is zero.
- Adversary spaces updates, so that spaces get bigger and bigger and it can look as though the values for gamma are plummeting quickly.

We can rewrite the updating equation as

$$Q_{k+1}(s_k, a_k) = Q_k(s_k, a_k) + \gamma_k \overbrace{\left(r_{k+1} + \max_{a' \in A} Q_k(s'_{k+1}, a') - Q_k(s_k, a_k)\right)}^{\text{temporal difference}}$$

A temporal difference is a difference between predictions that you can make in consecutive time periods. It is composed of two parts: the current value of the state-action function and the proposed value by the greedy update to the Q function. The difference helps us realize if we over or underestimated the previous value of the Q function.

Note that the Q-learning update can be understood as an asynchronous version of stochastic approximation to value iteration. Thus, in the following section we study Stochastic Approximation as a way to understand the convergence properties of Q-Learning.

2 Stochastic Approximation

We are not going to learn a comprehensive foundation to prove stochastic approximation results. We will go over intuition for how formal analysis of these things goes.

The idea to stochastic approximation is:

$$x_{k+1} = x_k + \gamma_k s(x_k, w_k) \text{ where } w_k \text{ is a random disturbance.}$$

Let's assume that w_k is ergodic and in particular there is some stationary distribution for w_k so we can talk about expectation and distribution. In fact, if w_k is drawn form the steady state distribution, let

$$\bar{s}(x) = \mathbb{E}\left[s(x, w_k)\right]$$

Under various technical conditions when γ_k is "small" the sequence approximates an ODE of the form

$$\dot{x_t} = \bar{s}(x_t)$$

The idea is that we want a relationship from a continuous to a discrete sequence.

$$\begin{array}{ccc} \overbrace{x_t}^{\text{continuous}} & \longleftrightarrow & \overbrace{x_k}^{\text{discrete}} \\ t & \longleftrightarrow & \sum_{i=0}^{k-1} \gamma_i \end{array}$$

The γ is can be thought of like the dt in an ODE.

$$\begin{aligned} \frac{x_{t+dt} - x_t}{dt} &= \bar{s}(x_t) \\ x_{t+dt} &= x_t + dt\bar{s}(x_t) \end{aligned}$$

If $\sum \gamma_k = \infty, \quad \sum \gamma_k^2 < \infty$ then we would converge with this update rule. As k increases, this follows more and more closely the tracks of the ODE.

Let's look at a simple case: assume the sample mean of w_k i.i.d. with $\mathbb{E}\left[w_k^2\right] < \infty$

$$\begin{aligned} x_k &= \frac{1}{k}\sum_{i=0}^{k-1} w_i && \text{approximately, using the LLN} \\ x_{k+1} &= \frac{1}{k+1}\sum_{i=0}^{k} w_i = \frac{1}{k+1} w_k + \frac{k}{k+1} x_k \\ x_{k+1} &= \left(1 - \frac{1}{k+1}\right) x_k + \frac{1}{k+1} w_k = (1-\gamma_k)\, x_k + \gamma_k w_k && \text{if we let } \gamma_k = \frac{1}{k+1} \end{aligned}$$

By the law of large numbers, $x_k \to \mathbb{E}[w_0]$. In fact, as long as $\sum \gamma_k = \infty$ and $\sum \gamma_k^2 < \infty$, we get $x_k \to \mathbb{E}[w_0]$.

2.1 Understanding the intuition behind the ODE approximation

We now reason about the same example in a continuous context. Let h be a small increment in the continuous time domain.

$$x_{t+h} = x_t + hs(x_t, w_t) \tag{2}$$

Suppose N large but $<< \frac{1}{h}$ so that $Nh << 1$. Then,

$$\begin{aligned}
x_{t+Nh} &= x_t + h\sum_{n=0}^{N-1} s(x_{t+nh}, w_{t+nh}) && \text{from equation (2) by induction} \\
&\simeq x_t + h\sum_{n=0}^{N-1} s(x_t, w_{t+nh}) && \text{we assume that } x_t \text{ doesn't change much since } nh \text{ is small} \\
&= x_t + \underbrace{hN}_{dt}\frac{1}{N}\sum_{n=0}^{N-1} s(x_t, w_{t+nh}) && \text{multiplying and dividing by } N \\
&= x_t + dt\left(\bar{s}(x_t) + O\left(\frac{1}{\sqrt{N}}\right)\right)
\end{aligned}$$

where in the last step, $O\left(\frac{1}{\sqrt{N}}\right)$ comes from the fact that the variance of the mean of N iid samples is on the order of $\frac{1}{\sqrt{N}}$.

We will leverage basic martingale convergence theorems and use that to prove how basic stochastic approximation theorems work.

2.2 A prototypical example of a stochastic approximation result

Proposition 2. *If w_k i.i.d, and there exist x^*, c_1, c_2 such that for all x*

1. $(x^* - x)^\intercal \bar{s}(x) \geq c_1 \left\|x^* - x\right\|_2^2$
2. $\mathbb{E}\left[\|s(x, w_k)\|_2^2\right] \leq c_2\left(1 + \|x^* - x\|_2^2\right)$

and if γ_k is deterministic with $\sum \gamma_k = \infty, \sum \gamma_k^2 < \infty$, then we have $x_k \to x^*$ with probability 1.

Intuitively, the two necessary conditions in proposition 2 can be thought as follows:

1. States that you are going in the right direction fast enough.
2. Gives a bound on the noise provided by w_k. It may be very noisy, but not too much. If you are far away you are allowed large noise cause that is compensated by the fact that you can take big steps.

Theorem 1. ***Supermartingale Convergence Theorem***
Let X_k, Y_k, Z_k $(k = 0, 1, 2, \dots)$ be nonnegative scalar random variables, with $\sum Y_k < \infty$. If

$$\mathbb{E}_k\left[X_{k+1}\right] \leq X_k + Y_k - Z_k \quad \forall\ k$$

then with probability 1, $\lim_{k\to\infty} X_k$ exists and is finite, and $\sum Z_k < \infty$.

In the above statement, the sub k on the expectation means conditioning on $X_0, \dots, X_k$. Note that the result presented in Theorem 1 is a stochastic generalization of convergence in real analysis.

We are not going to prove Theorem 1, it is pretty hard, but we will use it to prove proposition 2:

Proof. Let $\Delta_k = \|x_k - x^*\|_2^2$ then

$$\begin{aligned}
\Delta_{k+1} &= \|x_k + \gamma_k s(x_k, w_k) - x^*\|_2^2 \\
&= \Delta_k + \gamma_k^2 \|s(x_k, w_k)\|_2^2 - 2\gamma_k (x^* - x_k)^\intercal s(x_k, w_k) \\
\mathbb{E}_k\left[\Delta_{k+1}\right] &\leq \Delta_k + \gamma_k^2 c_2 \left(1 + \|x_k - x^*\|_2^2\right) - 2\gamma_k c_1 \|x_k - x^*\|_2^2 \quad \text{from the necessary conditions on prop. 2} \\
&= \Delta_k + \gamma_k^2 c_2 \left(1 + \Delta_k\right) - 2\gamma_k c_1 \Delta_k \\
&= \underbrace{\Delta_k}_{X_k} + \underbrace{\Delta_k (c_2 \gamma_k^2 - 2c_1 \gamma_k)}_{-Z_k} + \underbrace{c_2 \gamma_k^2}_{Y_k} \\
\mathbb{E}_k\left[X_{k+1}\right] &\leq X_k + Y_k - Z_k
\end{aligned}$$

We now want to see if the variables we just defined satisfy the conditions on Theorem 1. First, note that $\sum Y_k = c_2 \sum \gamma_k^2 < \infty$ since $\sum \gamma_k^2 < \infty$. Also, X_k and Y_k as defined are non-negative. However, note that $Z_k = \Delta_k(2C_1\gamma_k - C_2\gamma_k^2)$ could take on negative values. But since γ_k converges to 0, there will be some K such that $\forall k \geq K$, $Z_k \geq 0$. Thus, we can look at Z_k starting at index $k = K$.

Now, from Theorem 1 we know that $X_k = \Delta_k = \|x_k - x^*\|_2^2$ converges to a finite random variable with probability 1. We also know that $\sum Z_k < \infty$. Since $\sum Z_k = \sum \Delta_k c_2 \gamma_k^2 - \sum \Delta_k 2c_1 \gamma_k < \infty$ and we know that $\sum \Delta_k c_2 \gamma_k^2 < \infty$, then $\sum \Delta_k 2c_1 \gamma_k < \infty$. However, since c_1 is a constant and $\sum \gamma_k = \infty$, the only way the previous sum could take on a finite value is if Δ_k is converging to zero. If Δ_k converged to any value other than zero, the sum would not converge.

Now we finally know that $\Delta_k = \|x_k - x^*\|_2^2 \to 0$ with probability 1, which implies that $x_k \to x^*$ with probability 1, as desired. □

Example 1. Consider $F : \mathbb{R}^N \to \mathbb{R}^N$ s.t $\|F(x) - F(y)\|_2 \leq \alpha \|x - y\|_2, \quad \alpha \in (0, 1)$

$$x_{k+1} = x_k + \gamma_k (F(x_k) - x_k + w_k)$$

with w_k i.i.d. $\mathbb{E}\left[w_k\right] = 0$, $\mathbb{E}\left[\|w_k\|_2^2\right] < \infty$.

$$\begin{aligned}
s(x, w) &= F(x) - x + w \\
\bar{s}(x) &= F(x) - x \\
x^* &= F(x^*)
\end{aligned}$$

Then:

1.

$$\begin{aligned}
(x^* - x)^\intercal \bar{s}(x) &= (x^* - x)^\intercal (F(x) - x^*) + \underbrace{(x^* - x)^\intercal (x^* - x)}_{\|x^* - x\|^2} \\
&\geq -\|x^* - x\|_2 \|F(x) - x^*\|_2 + \|x^* - x\|_2^2 \\
&\geq -\alpha \|x^* - x\|_2^2 + \|x^* - x\|_2^2 \\
&= (1 - \alpha) \|x^* - x\|_2^2
\end{aligned}$$

2.

$$\begin{aligned}
\mathbb{E}\left[\|s(x,w_k)\|_2^2\right] &= \mathbb{E}\left[\|F(x)-x+w_k\|_2^2\right] \\
&\leq \mathbb{E}\left[\left(\|F(x)-x\|_2+\|w_k\|_2\right)^2\right] \\
&\leq \mathbb{E}\left[\left((1+\alpha)\|x^*-x\|_2+\|w_k\|_2\right)^2\right] \\
&\leq \mathbb{E}\left[2\left((1+\alpha)^2\|x-x^*\|_2^2+\|w_k\|_2^2\right)\right] \\
&\leq \underbrace{2(1+\alpha)^2\left(1+\mathbb{E}\left[\|w_k\|_2^2\right]\right)}_{C_2}\left(1+\|x-x^*\|_2^2\right)
\end{aligned}$$

Thus, by Proposition 2, $x_k \to x^*$.

One last problem to think about: in the last homework assignment we showed that asynchronous Value Iteration converges to the optimal value function. Now, what if instead of having a maximum norm contraction mapping, we had a contraction with respect to the euclidean norm? We can show that this does not guarantee convergence with an asynchronous update (although you do get it in the synchronous case). Come up with an example that there is an asynchronous process that does not get you to the fixed point with the euclidean contraction.

Homework # 2

Let $T : \mathbb{R}^N \to \mathbb{R}^N$ be such that $\exists \quad \alpha \in (0,1)$ s.t. $\forall \quad V, \bar{V}, \quad \left\|TV - T\bar{V}\right\|_2 \leq \alpha \left\|V - \bar{V}\right\|_2$. Consider the update

$$V_{k+1} = \begin{cases} (TV_k)(s) & \text{if } s = s_k \\ V_k(s) & \text{otherwise} \end{cases}$$

Show that V_k may not converge to V^* even if each $s \in S$ is selected infinitely often.

References

[1] John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202, 1994.