IE 3186: Approximate Dynamic Programming **Fall 2018**

Lecture 5: Q-Learning and Stochastic Approximation

Lecturer: Daniel Jiang *Scribes: Shaoning Han*

References:

D.P. Bertsekas, J.N. Tsitsklis. *Neuro-dynamic programming*, Athena Scientific, Belmont MA, 1996. (§4.2)

5.1 Introduction to Q-Learning

Last time we showed how to interchange $\mathbf{E}(\cdot)$ and $\min(\cdot)$, which results in:

$$Q^*(i,u) = \mathbf{E}\big[g(i,u,w) + \gamma \min_{u'} Q^*(f(i,u,w),u')\big],$$

a reformulation of the Bellman equation written for Q-factors. Recall that Q-iteration is simply V.I. (value iteration) for Q-factors:

$$Q_{k+1} = FQ_k$$

where F is the Q-Bellman operator. Next step: can we adjust this algorithm so that the expectation does not need to be computed at every iteration. Instead, what if we only had access to samples of the transition model?

Example 5.1 (Gridworld)**.** *This is the standard example from reinforcement learning, where Q-learning originated.*

- *States* = {locations on the grid}
- *Actions* = {move **North**, **South**, **East**, **West**}
- *Transitions: there is eastward wind, so you either go the intended direction or with small probability, you move* **East***.*
- *Reward* = 1 *if good (or goal),* −1 *bad state,* 0 *otherwise.*

Assumptions:

- *Imagine you are a robot and can only operate in the real environment;*

- *You do not know the transition model (DP is not possible even if the state space is small), but can try things and learn from the environment;*
- *You also may not know the cost/reward structure (but can see it through experiencing).*
- *Alternatively, you have a simulation model of the environment, e.g., of the wind, but still no exact knowledge of the wind distribution.*

Let $\alpha \in [0,1]$ be a stepsize parameter. The Q-iteration algorithm $Q_{k+1} = FQ_k$ can be generalized by doing "partial" updates:

$$Q_{k+1}(i,u) = (1-\alpha)Q_k(i,u) + \alpha\,\mathbf{E}[g(i,u,w) + \gamma \min_{u'} Q_k(f(i,u,w),u')].$$

Q-learning is an approximate version of this, where $\mathbf{E}$ is not computed, but sampled:

$$Q_{t+1}(i,u) = (1-\alpha_t(i,u))\,Q_t(i,u)) + \alpha_t(i,u)\left[g(i,u,\tilde{w}) + \min_{u'} Q_t(f(i,u,\tilde{w}),u')\right]$$

where $\tilde{w}$ sampled (or observed) from same distribution as w and at each iteration t, one state-action pair (i,u) is updated. Let $T^{i,u}$ be the iterations at which (i,u) is updated. Then, $\alpha_t(i,u) = 0$ if $t \notin T^{i,u}$. Typically it is assumed that $|T^{i,u}| = \infty$.

Convergence of Q-learning: if certain assumptions are satisfied, $Q_t(i,u) \to Q^*(i,u)$ w.p. 1. There are two parts to the analysis:

1. Understand some basic properties of stochastic approximation (SA) theory/stochastic gradient descent (SGD).
2. In conjunction with SA/SGD, use the contraction property of Bellman operator F to show convergence.

Let's start with stochastic approximation. Next time we discuss the second part.

5.2 Stochastic Approximation and Convergence

Example 5.2 (SGD)**.** *The stochastic gradient descent algorithm minimizes a cost function f by moving in (noisy) directions of its gradient.*

$$r_{t+1} = r_t - \alpha_t(\nabla f(r_t) + w_{t+1})$$

where $\mathbf{E}_t[w_{t+1}] = 0$*. Under some technical conditions,* $\lim_{t\to\infty} \nabla f(r_t) = 0$*. If f is convex, then it follows that a minimizer is found.*

Example 5.3 (Estimation of Unknown Mean)**.** *Let v_t be i.i.d random variables with unknown mean μ and finite variance. The "Robbins-Monro" SA (stochastic approximation) algorithm*

$$r_{t+1} = (1 - \alpha_t)\, r_t + \alpha_t v_{t+1}$$

can be thought of as averaging (i.e., set $\alpha_t = 1/t$). Under some conditions, $r_t \to r$.

Example 5.4 (Minimizing a Sum)**.** *Minimize the cost*

$$f(r) = \frac{1}{K} \sum_{k=1}^{K} f_k(r).$$

Rather than compute $\nabla f(r)$, one might consider

$$r_{t+1} = r_t - \alpha_t \nabla f_{k(t)}(r_t)$$

where $k(t)$ is a random variable uniformly distributed over $\{1, 2, \ldots, K\}$.

Now, let us consider the more general algorithm:

$$r_{t+1} = r_t + \alpha_t\, s_{t+1}, \tag{5.1}$$

which cover the above examples. We need the following assumptions to show the convergence of Algorithm 5.1.

Assumption 5.5 (Stepsize)**.** *Suppose*

$$\alpha_t \geq 0, \quad \sum_{t=0}^{\infty} \alpha_t = \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2 \leq B < \infty,$$

where B is a deterministic constant (can be relaxed).

Remark 5.6. *If $\sum_{t=0}^{\infty} \alpha_t \leq A < \infty$, then it may be the case that we never find the desired solution because*

$$|r_t - r_0| \leq \sum_{\tau=0}^{t-1} \alpha_\tau |s_{\tau+1}| \leq c \cdot A,$$

where we suppose $s_{\tau+1}$ is bounded by c. If r_0 is chosen badly, r^* may never be found.

Remark 5.7. *Suppose $\alpha_t = \alpha$ for all t, where α is some constant, and the variance of s_{t+1} is constant, we can only hope to converge to neighborhood of r^*.*

Let $\mathcal{F}_t = \{r_0, \ldots, r_t, s_0, \ldots, s_t, \alpha_0, \ldots, \alpha_t\}$ be the history of the algorithm. Let $|| \cdot ||$ be the Euclidean norm.

Assumption 5.8 (Potential Function)**.** *There exists a function $f : \mathbb{R}^n \to \mathbb{R}$ s.t.*

1. *(Nonnegative) $f(r) \geq 0$ for all r.*
2. *(Lipschitz ∇f) The function f is continuously differeatiable and $\exists\, L$ s.t.*

$$||\nabla f(r) - \nabla f(r')|| \leq L||r - r'||$$

   for all $r, r' \in \mathbb{R}^n$.

3. *(Pseudo-gradient) $\exists$ a positive constant c s.t.*

$$c||\nabla f(r_t)||^2 \leq -\nabla f(r_t)^T \mathbf{E}[s_{t+1}|\mathcal{F}_t].$$

   "Expected update direction is roughly opposite of gradient and is a direction of cost decrease."

4. *$\exists$ positive constants $K_1,\ K_2$ s.t.*

$$\mathbf{E}\big[||s_{t+1}||^2\,|\,\mathcal{F}_t\big] \leq K_1 + K_2\,\|\nabla f\|^2.$$

Theorem 5.9. *Under these assumptions and (5.1), the following hold w.p. 1:*

1. *$f(r_t)$ converges;*
2. *$\lim_{t\to\infty} \nabla f(r_t) = 0$;*
3. *Every limit point of r_t is a stationary point of f.*

5.3 Proof of the Convergence Theorem

First, let's recall some helpful results from probability.

Lemma 5.10 (Supermartingale Convergence Theorem)**.** *Let X_t, Y_t, Z_t, $t = 0, 1, \ldots,$ be sequences of nonnegative random variables in $\mathcal{F}_t$. Suppose that:*

1. *For each t, we have*

$$\mathbf{E}[Y_{t+1}|\mathcal{F}_t] \leq Y_t - X_t + Z_t;$$

2. *It holds that $\sum_{t=0}^{\infty} Z_t$.*

Then $\sum_{t=0}^{\infty} X_t < \infty$ and $\{Y_t\}$ converges to some random variable Y w.p. 1.

Lemma 5.11 (Martinale Convergence Theorem)**.** *Let X_t be a martingale, i.e., $\mathbf{E}[X_{t+1}\,|\,\mathcal{F}_t] = X_t]$ and $\mathbf{E}[|X_t|] \leq M$, for some positive M. Then, X_t converges to some X w.p.1.*

Proof of Theorem 5.9. (Part 1; lower bound on limit) First note that using the assumption on f,

$$f(\bar{r}) \leq f(r) + \nabla f(r)^T(\bar{r} - r) + \frac{L}{2}||\bar{r} - r||^2.$$

(See page 95 of Bertsekas & Tsitsiklis for derivation).

Applying this with $r = r_t, \bar{r} = r_{t+1} = r_t + \gamma_t\, s_{t+1}$:

$$f(r_{t+1}) \leq f(r_t) + \alpha_t \nabla f(r_t)^T s_{t+1} + \frac{L}{2}\alpha_t^2 ||s_{t+1}||^2$$

Take $\mathbf{E}[\cdot\,|\,\mathcal{F}_t]$ on both sides to obtain

$$\begin{aligned}
\mathbf{E}[f(r_{t+1})|\mathcal{F}_t] &\leq f(r_t) + \alpha_t \nabla f(r_t)^T \mathbf{E}[s_{t+1}\,|\,\mathcal{F}_t] + \frac{L}{2}\alpha_t^2\big(K_1 + K_2||\nabla f(r_t)||^2\big) \\
&\leq f(r_t) - \alpha_t\Big(c - \frac{LK_2\alpha_t}{2}\Big)||\nabla f(r_t)||^2 + \frac{LK_1 d_t^2}{2} \\
&= f(r_t) - X_t + Z_t
\end{aligned}$$

where

$$X_t := \max\Big(\alpha_t\Big(c - \frac{LK_2\alpha_t}{2}\Big)||\nabla f(r_t)||^2, 0\Big), \quad Y_t = f(r_t),$$

$$Z_t := \frac{LK_1\alpha_t^2}{2} - \min\Big(\alpha_t\Big(c - \frac{LK_2\alpha_t}{2}\Big)||\nabla f(r_t)||^2, 0\Big).$$

Clearly, $X_t, Y_t, Z_t \in \mathcal{F}$ are nonnegative and $\sum_{t=0}^{\infty} Z_t < \infty$ by assumption. By the supermartingale convergence theorem, $f(r_t)$ converges and $\sum_t X_t < \infty$. After some finite time when $LK_2\alpha_t \leq c$

$$X_t = \alpha_t\Big(c - \frac{LK_2\alpha_t}{2}\Big)||\nabla f(r_t)||^2 \geq \frac{c}{2}\alpha_t||\nabla f(r_t)||^2.$$

So we have $\sum_t \alpha_t||\nabla f(r_t)||^2 < \infty$ (otherwise, $\sum_t X_t = \infty$). Suppose $\exists\, t_0, \delta > 0$ s.t. $||\nabla f(r_t)||^2 \geq \delta$, $\forall\, t \geq t_0$. But then, we would have

$$\sum_t \alpha_t ||\nabla f(r_t)||^2 \geq \sum_t \alpha_t \delta = \infty$$

by the stepsize assumption. Contradiction! It follows that

$$\liminf_{t\to\infty} ||\nabla f(r_t)|| = 0.$$

(Part 2; upper bound on limit) Next, we want to show an upper bound on the limit is zero as well. Fix $\epsilon > 0$. A time interval $\{t, t+1, \ldots, \bar{t}\}$ is an <u>upcrossing interval</u> if

$$||\nabla f(r_t)|| < \epsilon/2, \;\; ||\nabla f(r_{\bar{t}})|| > \epsilon$$

and

$$||\nabla f(r_\tau)|| \in [\epsilon/2, \epsilon] \quad \forall\, t < \tau < \bar{t}.$$

Goal: show there are only a finite number of upcrossings from $\epsilon/2$ to ϵ. Let

$$\bar{s}_t = \mathbf{E}[s_{t+1} \,|\, \mathcal{F}_t], \quad w_{t+1} = s_{t+1} - \bar{s}_t.$$

Note for later that

$$||s_{t+1}^2|| = ||w_{t+1}||^2 + ||\bar{s}_t||^2 + 2w_{t+1}^T \bar{s}_t,$$

so

$$||\bar{s}_t||^2 + \mathbf{E}[||w_{t+1}||^2 \,|\, \mathcal{F}_t] = \mathbf{E}[||s_{t+1}^2|| \,|\, \mathcal{F}_t] \leq K_1 + K_2 \,||\nabla f({}_t r)||^2. \tag{5.2}$$

Let $\chi_t = \mathbb{1}_{\{||\nabla f(r_t)|| \leq \epsilon\}}$. We need to pause the main proof here for a lemma.

Lemma 5.12 (Technical Lemma)**.** *The sequence u_t defined by*

$$u_t = \sum_{\tau=0}^{t-1} \chi_t \alpha_t w_{\tau+1}$$

converges w.p.1.

Proof. Note $\mathbf{E}[\chi_t \alpha_t w_{t+1} \,|\, \mathcal{F}_t] = \chi_t \alpha_t \mathbf{E}[w_{t+1} \,|\, \mathcal{F}_t] = 0$ by unbiasedness of w_{t+1}, so

$$\mathbf{E}[u_{t+1}|\mathcal{F}_t] = \mathbf{E}[u_t + \chi_t \alpha_t w_{t+1}|\mathcal{F}_t] = u_t.$$

Each component of u_t is a martingale if we can show $\exists\, M$ s.t. $\mathbf{E}|u_t(i)| \leq M$ for all t.

$$||u_{t+1}||^2 = ||u_t||^2 + 2u_t^T \chi_t \alpha_t w_{t+1} + \chi_t^2 \alpha_t^2 ||w_{t+1}||^2$$

$$\begin{aligned} \mathbf{E}[||u_{t+1}||^2|\mathcal{F}_t] &= ||u_t||^2 + 2u_t^T \chi_t \alpha_t \mathbf{E}[w_{t+1}|\mathcal{F}_t] + \chi_t^2 \alpha_t^2 \mathbf{E}[||w_{t+1}||^2|\mathcal{F}_t] \\ \text{by (5.2)} \qquad &\leq ||u_t||^2 + \chi_t^2 \alpha_t^2 (K_1 + K_2 ||\nabla f(r_t)||^2) \\ &\leq ||u_t||^2 + \alpha_t^2 (K_1 + K_2 \epsilon^2). \end{aligned}$$

Taking unconditional expectations and then iterating:

$$\mathbf{E}[||u_t||^2] \leq (K_1 + K_2 \epsilon^2) \mathbf{E}\Big[\sum_{\tau=0}^{\infty} \alpha_\tau^2\Big] \leq B$$

Since $||u_t|| \leq 1 + ||u_t||^2$, $\sup_t \mathbf{E}[||u_t||] \leq$ some M. So we can apply martingale convergence to conclude. □

Back to main proof: *(Part 2; upper bound on limit, continued)* Suppose we have a sample path with infinite number of upcrossings from $\epsilon/2$ to ϵ. Let $\{t_k, \dots, \bar{t}_k\}$ be the k^{th} upcrossing interval

$$\sum_{t=t_k}^{\bar{t}_k-1} \chi_t \alpha_t w_{t+1} = \sum_{t=t_k}^{\bar{t}_k-1} \alpha_t w_{t+1}.$$

We claim that

$$\lim_{k\to\infty} \sum_{t=t_k}^{\bar{t}_k-1} \alpha_t w_{t+1} = 0. \tag{5.3}$$

If not, u_t cannot converge since $u_t \approx \sum_k \sum_{t=t_k}^{\bar{t}_k-1} \alpha_t w_{t+1}$. Similarly

$$\lim_{k\to\infty} \alpha_{t_k} w_{t_k+1} = 0. \tag{5.4}$$

Now we have

$$\begin{aligned} ||\nabla f(r_{t_k+1})|| - ||\nabla f(r_{t_k})|| &\le ||\nabla f(r_{t_k+1}) - \nabla f(r_{t_k})|| \\ &\le L||r_{t_k+1} - r_{t_k}|| \\ &= \alpha_{t_k} L||\bar{s}_{t_k} + w_{t_k+1}|| \\ &\le \alpha_{t_k} L[||\bar{s}_{t_k}|| + ||w_{t_k+1}||] \\ &\le \alpha_{t_k} L(K_1 + K_2\epsilon^2) + L\alpha_{t_k}||w_{t_k+1}|| \\ \text{by (5.2) and (5.4)} \qquad &\to 0 \text{ as } k \to \infty. \end{aligned}$$

Note $||\nabla f(r_{t_k+1})|| \ge \epsilon/2$, we can suppose that for large enough k, $||\nabla f(r_{t_k})||$ will be close, say $\ge \epsilon/4$.

For any k, it is also true that

$$\begin{aligned} \frac{\epsilon}{2} &\le ||\nabla f(r_{\bar{t}_k})|| - ||\nabla f(r_{t_k})|| \\ &\le ||\nabla f(r_{\bar{t}_k}) - \nabla f(r_{t_k})|| \\ &\le L||r_{\bar{t}_k} - r_{t_k}|| \\ &\le \underbrace{L\sum_{t=t_k}^{\bar{t}_k-1} \alpha_t ||\bar{s}_t||}_{\text{denoted by } (*)} + \underbrace{L||\sum_{t=t_k}^{\bar{t}_k-1} \alpha_t w_{t+1}||}_{\to 0 \text{ by } (5.3)} \end{aligned}$$

Notice

$$(*) \le L\sum_{t=t_k}^{\bar{t}_k-1} \alpha_t(1 + ||\bar{s}_t||^2) \le L\sum_{t=t_k}^{\bar{t}_k-1} (1 + K_1 + K_2\epsilon^2)\alpha_t \le Ld\sum_{t=t_k}^{\bar{t}_k-1} \alpha_t$$

where K_1, K_2, d are some positive constants. It follows that

$$\liminf_{k\to\infty} \sum_{t=t_k}^{\bar{t}_k-1} \alpha_t \geq \frac{\epsilon}{2Ld}$$

Combining with the previous lower bound of $\epsilon/4$ on $||\nabla f(r_t)||$ for large k,

$$\liminf_{k\to\infty} \sum_{t=t_k}^{\bar{t}_k-1} \alpha_t ||\nabla f(r_t)||^2 \geq \frac{\epsilon}{2Ld}\left(\frac{\epsilon}{4}\right)^2$$

which implies

$$\sum_{t=0}^{\infty} \alpha_t ||\nabla f(r_t)||^2 = \infty,$$

a contradiction from the beginning of the proof where we used supermartingale convergence. Thus, there are a finite number of upcrossing intervals.

$$\limsup_{t\to\infty} ||\nabla f(r_t)|| \leq \epsilon$$

Because ϵ is arbitrary, $\lim_{t\to\infty} ||\nabla f(r_t)|| = 0$. Let r be a limit point of r_t. Then $\nabla f(r)$ is the limit of a subsequence of $\nabla f(r_t)$, so also $= 0$. □