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

Lecture 6: Convergence of Q-Learning and DQN

Lecturer: Daniel Jiang *Scribes: Kamal Basulaiman*

References:

D. P. Bertsekas, J. N. Tsitsiklis *Neuro-dynamic programming*, Athena Scientific, Belmont MA, 1996. (§4)

V. Mnih et. al., *Human-level control through deep reinforcement learning*, Nature, 518(7540), pp. 529-533, 2018.

6.1 Convergence of Q-Learning

A special case of last week's theorem is the algorithm

$$r_{t+1} = (1-\alpha_t)r_t + \alpha_t w_{t+1} = r_t + \alpha_t(w_{t+1} - r_t) \tag{6.1}$$

where $r_t \in \mathbb{R}$, w_{t+1} is mean zero, and $\mathbf{E}\left[w_{t+1}^2(i) \,\middle|\, \mathcal{F}_t\right] \leqslant A$.

Corollary 6.1. $r_t \to 0$ *w.p.*1.

Proof. Apply last week's theorem with $f(r) = r^2$. □

If $\mathbf{E}\left[w_{t+1}(i) \,\middle|\, \mathcal{F}_t\right] = \mu$, then use $f(r) = (r-\mu)^2$ in the above corollary.

Now let us consider a more general algorithm based a pseudo-contraction.

- "States" or "components of a vector" $i = 1, 2, \ldots, n$.
- The stochastic algorithm we consider is as follows. Start with some arbitrary estimate $r_0 \in \mathbb{R}^n$. Then, for $t \geqslant 1$,

  $$r_{t+1}(i) = (1-\alpha_t(i))r_t + \alpha_t(i)\left[(Hr_t)(i) + w_{t+1}(i) + u_{t+1}(i)\right],$$

  where H is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ with $(Hr_t)(i)$ being the i^{th} component new observation of Hr_t, w_{t+1} is unbiased noise (e.g. sampling error due to not computing $\mathbf{E}$ exactly), and u_{t+1} is biased noise (some sort of approximation error).

Let $\mathcal{F}_t = \{r_0(i), r_1(i), \dots, r_t(i), w_0(i), \dots, w_t(i), \alpha_0(i), \dots, \alpha_t(i)\}$, for $i = \{1, 2, \dots, n\}$.

Assumption 6.2. *We make the following assumptions.*

1. *Unbiasedness:* $\forall\ i, t$, $\mathbf{E}\left[w_{t+1}(i) \,\middle|\, \mathcal{F}_t\right] = 0$,
2. *Bound on variance:* $\exists\ A, B$ *s.t.* $\mathbf{E}\left[w_{t+1}^2(i) \,\middle|\, \mathcal{F}_t\right] \leqslant A + B\|r_t\|^2$,
3. *Stepsize:* $\forall\ i$, $\sum_t \alpha_t(i) = \infty, \sum_t \alpha_t^2(i) < \infty$, *and* $\alpha_t(i) = 0$ *if* i *not visited,*
4. *Pseudo-contraction:* $\exists\ r^*$ *and a scalar* $\beta \in [0, 1)$ *s.t.* $\|Hr_t - r^*\| \leqslant \beta\|r_t - r^*\|$.
5. *Disappearing bias:* $\exists\, \theta_t \to 0$ *such that* $|u_t(i)| \leqslant \theta_t\|r_t\|$
6. *Each state* i *is visited infinitely often with probability 1.*

Theorem 6.3. *For each state* i*, the iterates* $r_t(i) \to r^*(i)\ \ w.p.1$.

Proof. Assume $r^* = 0$ since we can just shift the coordinate system. Then by Prop. 4.7 of Bertsekas and Tsitsiklis (Neuro-DP), we know that r_t is bounded $w.p.1$. Because r_t is bounded, $\exists\ D_0$ s.t. $\|r_t\| \leqslant D_0$ for all t. Define $D_0 = (\beta + 2\epsilon)\ D_k, k \geqslant 0$ for some $\epsilon \geqslant 0$ s.t. $\beta + 2\epsilon < 1$, $D_k \to 0\ w.p.1$.

Induction: Suppose $\exists$ a random time t_k s.t. $\|r_t\| \leqslant D_k$ for all $t \geqslant t_k$, meaning r_t enters D_k forever at t_k.

Induction step: Assume this works for k, prove existence of t_{k+1} satisfying the condtition with $k \leftarrow k + 1$. Define an "accumulated noise" process started at τ by $W_{\tau,\tau}(i) = 0$, and

$$W_{t+1,\tau}(i) = (1 - \alpha_t(i))\, W_{t,\tau}(i) + \alpha_t(i) w_{t+1}(i), \quad \forall\ t \geqslant \tau,$$

which averages noise terms together. By Corollary 6.1, it follows that

$$\lim_{t\to\infty} W_{t,\tau}(i) = 0 \quad \forall\, \tau, i.$$

By the induction hypothesis, the biased noise satisfies $|u_t(i)| \leqslant \theta_t\|r_t\| \leqslant \theta_t D_t$, which implies $|u_t(i)| \to 0$, since our assumption said that $\theta \to 0$. Let $\tau_k \geqslant t_k$ be a future time at which $|u_t(i)| \leqslant \epsilon D_k$. This is a point where the noise is small enough that we can start analyzing the convergence. Define

$$Y_{\tau_k}(i) = D_k \quad \text{and} \quad Y_{t+1}(i) = (1 - \alpha_t(i))Y_t(i) + \alpha_t(i)(\beta + \epsilon)D_k$$

Note, by Corollary 6.1, $Y_t(i) \to (\beta + \epsilon)D_k$.

Claim 6.4. $\forall\, i$ *and* $t \geqslant \tau_k$, $-Y_t(i) + W_{t,\tau_k}(i) \leqslant r_t(i) \leqslant Y_t(i) + W_{t,\tau_k}(i)$.

Proof. We proceed by induction on t.

Base case ($t = \tau_k$): $Y_{\tau_k}(i) = D_k$ and $W_{\tau_k,\tau_k}(i) = 0$. So, it is clear that the statement is true. Assume it is true for t. We want to show it is true for $t+1$.

Induction step:

$$\begin{aligned} r_{t+1}(i) &= (1-\alpha_t(i))r_t(i) + \alpha_t(i)\left[(Hr_t)(i) + w_{t+1}(i) + u_{t+1}(i)\right] \\ &\leqslant (1-\alpha_t(i))(Y_t(i) + W_{t,\tau_k}(i)) + \alpha_t(i)(Hr_t)(i) + \alpha_t(i)w_{t+1}(i) + \alpha_t(i)u_{t+1}(i) \\ &\leqslant Y_{t+1}(i) + W_{t+1,\tau_k}(i), \end{aligned}$$

where we used $(Hr_t) \leqslant \beta\|r_t\| \leqslant \beta D_k$ and $u_{t+1}(i) \leqslant \epsilon D_k$. Symmetrically, it can be shown that,

$$-Y_{t+1}(i) + W_{t+1,\tau_k}(i) \leqslant r_{t+1}(i) \leqslant Y_{t+1}(i) + W_{t+1,\tau_k}(i),$$

which completes the proof. □

Since, $Y_t(i) \to (\beta+\epsilon)D_k$, and $W_{t,\tau_k}(i) \to 0$, then $\limsup_{t\to\infty}\|r_t\| \leqslant (\beta+\epsilon)D_k \leqslant D_{k+1}$. □

6.2 Connection to Q-Learning

1. $r_t(i) \iff Q_t(i,u)$.
2. $H \iff F$ (Bellman Operator).
3. $w_{t+1} \iff \gamma \min_{u'} Q_t(f(i,u,\tilde{w}) - \gamma\, \mathbf{E}\left[\min_{u'} Q_t(f(i,u,w))\right]$.
4. $u_{t+1}(i) \iff 0$.

Here we show that F is a γ-contraction in the maximum norm.

$$\|FQ - FQ'\|_\infty = \max_{(i,u)} \left| g(i,u) + \gamma\, \mathbf{E}\left[\min_{u'} Q(f(i,u,w'))\right] - g(i,u) - \gamma\, \mathbf{E}\left[\min_{u'} Q'(f(i,u,w'))\right] \right|$$

$$= \gamma \max_{(i,u)} \left| \mathbf{E}\left[\min_{u'} Q(f(i,u',w)) - \min_{u'} Q'(f(i,u',w))\right] \right|$$

$$\leqslant \gamma \max_{(i,u)} \mathbf{E} \left| \min_{u'} Q(f(i,u',w)) - \min_{u'} Q'(f(i,u',w)) \right|$$

$$
\begin{aligned}
&\leqslant \gamma \max_{(i,u)} \mathbf{E}\left[\max_{u'} |Q'(f(i,u',w)) - Q(f(i,u',w))|\right] \\
&\leqslant \gamma \|Q' - Q\|,
\end{aligned}
$$

where we used $|\min f - \min g| \leqslant \max |f - g|$. Therefore, we can apply the theorem to see that $Q_t \to Q^*$ $w.p.1$.

6.3 DQN Paper Discussion (Ziyue)

1. Each period between updates to the target parameter vector can be thought of as one Q-iteration (i.e., the value iteration algorithm applied using F).
2. During this time, DQN tries to approximate FQ by minimizing the loss function.
3. C can be thought of as number of SGD steps taken to fit $\hat{Q}$-network per Q-iteration.