Lecture #25: FoSML

- The following result is similar to our final theorem from last class, and provides a related guarantee regarding stochastic optimization.

Theorem: Let $H: \mathbb{R}^N \to \mathbb{R}^N$, $\{w_t\}_{t \in \mathbb{N}}$ a sequence of $\mathbb{R}^N$-valued r.v., and $\{\alpha_t\}_{t \in \mathbb{N}}$ a sequence of reals. Let $\{x_t\}_{t \in \mathbb{N}}$ be a sequence defined inductively by

$$\forall s \in \{1, \dots, N\}, \quad x_{t+1}(s) = x_t(s) + \alpha_t(s)\left[H_t(x_t)(s) - x_t(s) + w_t(s)\right],$$

with $x_0 \in \mathbb{R}^N$ chosen arbitrarily. Let $\mathcal{F}_t = \{\{x_{t'}\}_{t' \le t}, \{w_{t'}\}_{t' \le t}, \{\alpha_{t'}\}_{t' \le t}\}$ be the history up to time t. Suppose the following conditions hold:

(a.) $\exists K_1, K_2 \in \mathbb{R}$ s.t. $\mathbb{E}\left(\|w_t\|^2(s) \mid \mathcal{F}_t\right) \le K_1 + K_2 \|x_t\|^2$, for some norm

(b) $\forall s \in \{1, \dots, N\}$, $\sum_{t=0}^{\infty} \alpha_t(s) = \infty$, $\sum_{t=0}^{\infty} \alpha_t^2(s) < \infty$

(c.) $\mathbb{E}\left(w_t \mid \mathcal{F}_t\right) = 0$

(d.) H is a $\|\cdot\|_\infty$-contraction with fixed point x^*.

Then $x_t \xrightarrow{a.s.} x^*$.

- Using these general results on stochastic approximation, we consider two methods for RL in unknown environments: TD(0) algorithm and

Q-learning.

- Recall the linear Bellman equations:

$$V_\pi(s) = \mathbb{E}\left(r(s,\pi(s))\right) + \gamma \sum_{s'} \mathbb{P}(s' \mid s, \pi(s)) V_\pi(s')$$

$$= \underbrace{\mathbb{E}}_{s'}\left(r(s,\pi(s)) + \gamma V_\pi(s') \mid s\right).$$

we don't know $\mathbb{P}(s' \mid s)$, since we don't know the underlying environment.

- To get around not knowing $\mathbb{P}(s' \mid s)$, we sample. More precisely, the TD(0) algorithm consists of iteratively performing the following two steps:

(1.) Sampling a state s'

(2.) Updating the value function as

$$V(s) \leftarrow \underbrace{(1-\alpha)V(s)}_{\text{old value}} + \alpha\left[r(s,\pi(s)) + \gamma V(s')\right]$$

$$= V(s) + \alpha\underbrace{\left[r(s,\pi(s)) + \gamma V(s') - V(s)\right]}_{\text{temporal difference in } V \text{ ("TD")}}$$

- By our theorems above, TD(0) is guaranteed to converge.
- Note that TD(0) may be understood as a special case of Q-learning!

Which we now introduce.

- Let us, WLOG, assume the optimal policy π^* is deterministic.
- Recall that the optimal policy π^* satisfies

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a)$$

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a)$$

- The Q-learning algorithm is based on the following familiar equations:

$$Q^*(s,a) = \mathbb{E}(r(s,a)) + \gamma \sum_{s' \in S} \mathbb{P}(s'|s,a) \cdot V^*(s')$$

$$= \mathop{\mathbb{E}}_{s'} \left(r(s,a) + \gamma \max_{a \in \mathcal{A}} Q^*(s',a) \right).$$

- Unfortunately, we don't know $\mathbb{P}(s'|s,a)$ nor $V^*(s')$. So, as in the TD(0) algorithm, we iteratively update in a 2-step ~~proceed~~ procedure:

(a.) Sampling a ~~new~~ new state s'

(b.) Updating $Q(s,a) \leftarrow$ [redacted] $(1-\alpha) Q(s,a)$

$$+ \alpha \left[r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} Q(s',a') \right],$$

where α may depend on the number of visits to state s.

- This is kind of like a stochastic value iteration, in which we need to sample.

Theorem: Consider a finite MDP. Suppose the property that for all $s \in S$ and $a \in A$, $\sum_{t=0}^{\infty} \alpha_t(s,a) = +\infty$, $\sum_{t=0}^{\infty} \alpha_t(s,a)^2 = \infty$ for $\alpha_t(s,a) \in [0,1]$ holds. Then Q-learning converges to Q^* a.s.

Proof: Let $\{Q_t(s,a)\}_{t \geq 0}$ be the sequence of state-action value functions generated by the Q-learning algorithm. By definition,

$$\text{(★)} \quad Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha \left[r(s_t,a_t) + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s_t,a_t) \right]$$

[illegible]

Let s,a be arbitrary, we get

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t(s,a) \left[r(s,a) + \gamma \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left(\max_{a'} Q_t(u,a') \right) - Q_t(s,a) \right]$$

$$+ \gamma \alpha_t(s,a) \cdot \left[\max_{a'} Q_t(s',a') - \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left(\max_{a'} Q_t(u,a') \right) \right].$$

where s' = next state after (s,a) and $\alpha_t(s,a) = \begin{cases} 0 & \text{if } (s,a) \neq (s_t,a_t) \\ \alpha_t(s_t,a_t) & \text{otherwise.} \end{cases}$

Let $w_t^{(s,a)} := \max_{a'} Q_t(s',a') - \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left(\max_{a'} Q_t(u,a') \right)$ define our noise vector, and let $\bar{Q}_t$ be the function Q_t, thought of as a vector. Similarly, let $H(\bar{Q}_t)$ be the vector with $H(\bar{Q}_t)(s,a) = r(s,a) + \gamma \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left(\max_{a' \in A} Q_t(u,a') \right)$.

We claim the theorem stated at the beginning of this lecture applies, implying that $Q^* \xrightarrow{a.s.} Q$.

Condition (b) holds by hypothesis, while condition (c) holds by construction of w_t.

To see (a.), note that $\forall s \in S$,

$$|w_t(s)| = \left| \max_{a'} Q_t(s',a') - \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left(\max_{a'} Q_t(u,a') \right) \right|$$

$$\leq \max_{a'} |Q_t(s',a')| + \left| \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left(\max_{a'} Q_T(u,a') \right) \right|$$

$$\leq 2 \cdot \max_{s'} \left| \max_{a'} Q_t(s',a') \right|$$

$$= 2 \|\bar{Q}_t\|_\infty$$

$$\Rightarrow \mathbb{E}\left(|w_t^2(s)| \mid \mathcal{F}_t \right) \leq 4 \cdot \|\bar{Q}_t\|_\infty^2, \text{ which gives (a.).}$$

To see (d), note that for any $\bar{Q}_{t_1}, \bar{Q}_{t_2} \in \mathbb{R}^{|S| \times |A|}$, and any $(s,a) \in S \times A$,

$$\left| H(Q_{t_2})(s,a) - H(Q_{t_1})(s,a) \right| = \left| \gamma \cdot \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left(\max_{a'} Q_{t_2}(u,a') - \max_{a'} Q_{t_1}(u,a') \right) \right|$$

$$\leq \gamma \cdot \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left(\left| \max_{a'} Q_{t_2}(u,a') - \max_{a'} Q_{t_1}(u,a') \right| \right)$$

$$\leq \gamma \cdot \mathbb{E}_{u \sim \mathbb{P}(\cdot|s,a)} \left| \max_{a'} Q_{t_2}(u,a') - \max_{a'} Q_{t_1}(u,a') \right|$$

$$\leq \gamma \cdot \|\bar{Q}_{t_2} - \bar{Q}_{t_1}\|_\infty$$

Since H is then a contraction, it admits a fixed point, which Q-learning converges to. ∎

Remark: Since $\sum_{t=0}^{\infty} \alpha_t(s,a) = \infty$ for all (s,a), the sampling procedure must sample each (s,a) pairs infinitely often.

. How the sampling procedure impacts the algorithm isn't part of the above theorem (as long as the conditions on $\alpha_t(s,a)$ hold).

Remark: One popular choice of sampling function is to use a Boltzmann distribution. That is, letting Q be the current state-action function, set

$$p_t(a|s,Q) = \frac{\exp\left(Q(s,a)/\tau_t\right)}{\sum_{a' \in A} \exp\left(Q(s,a')/\tau_t\right)},$$

for a sequence of temperatures $\{\tau_t\}_{t=0}^{\infty}$ with $\tau_t \to 0$ as $t \to \infty$; this ensures that for long time, $p_t(a|s,Q)$ localizes on $\operatorname{argmax}_{a'} Q(s,a')$, which makes sense since for large time, we have explored the environment well.