Lec 7: Policy gradient methods for tabular MDP (Based on Chi Jin's lecture)

MDP: $\{P_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}$

- objective function:

$$J(\pi) = V_1^{\pi}(s_1) = \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} r_h(s_h, a_h) \,\Big|\, s_1\right]$$

where $(s_1, a_1, s_2, a_2, \dots, s_H, a_H) \sim (P, \pi)$

$$a_1 \sim \pi_1(\cdot \mid s_1)$$
$$s_2 \sim P_1(\cdot \mid s_1, a_1)$$
$$\vdots$$
$$a_h \sim \pi_h(\cdot \mid s_h)$$
$$s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$$

- Parametrization

$$\pi = (\pi_1, \dots, \pi_H) \quad \text{where } \pi_h(\cdot \mid s) \in \Delta(A)$$

Consider the natural policy gradient method (mirror descent)

- initialize $\pi_h^1(\cdot|s)$ to be uniform over $\mathcal{A}$ for all s, h
- for $t = 1, 2, \dots, T$:

$$\pi^{t+1} = \mathrm{Proj}_{\Pi}\left(\pi^t + \frac{\eta}{2} \Delta J(\pi^t)\right)$$

$$= \arg\max_{\pi \in \Pi} \langle \Delta J(\pi^t), \pi - \pi^t \rangle - \frac{1}{\eta} \underbrace{D_{\Phi}(\pi, \pi^t)}_{\mathbb{E}_{\pi^t}\left[\sum_{h=1}^{H} KL\left[\pi_h(\cdot|s_h) \| \pi_h^t(\cdot|s_h)\right]\right]}$$

- return: $\pi^1, \dots, \pi^T$

- Compute the gradient $\dfrac{\partial J(\pi)}{\partial \pi_h(a|s)}$

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum_{h'=1}^{H} r_{h'}(s_{h'}, a_{h'})\right] = \underbrace{\mathbb{E}_{\pi}\left[\sum_{h'=1}^{h-1} r_{h'}(s_{h'}, a_{h'})\right]}_{\text{independent of } \pi_h} + \underbrace{\mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'})\right]}_{(\mathrm{I})}$$

$$(\mathrm{I}) = \mathbb{E}_{\substack{s \sim \Pr(s_h = s|\pi) \\ a \sim \pi_h(\cdot|s)}}\left[Q_h^{\pi}(s,a) \right]$$

$$\Rightarrow \frac{\partial J(\pi)}{\partial \pi_h(a|s)} = \underbrace{\Pr(s_h = s|\pi)}_{d_h^{\pi}(s)} Q_h^{\pi}(s,a)$$

Solve:

$$\arg\max_{\pi \in \Pi} \langle \nabla J(\pi^t), \pi - \pi^t \rangle - \frac{1}{\eta} D_{\Phi}(\pi, \pi^t)$$

Lagrangian multiplier:

$$L(\pi, \lambda) = \sum_{h,s,a} d_h^{\pi^t}(s,a)\, Q_h^{\pi^t}(s,a) \left(\pi_h(a|s) - \pi_h^t(a|s) \right) - \frac{1}{\eta} \sum_{h,s,a} d_h^{\pi}(s,a) \cdot \pi_h(a|s) \log \frac{\pi_h(a|s)}{\pi_h^t(a|s)}$$

$$- \sum_{h,s} \lambda_{h,s} \left(\sum_a \pi_h(a|s) - 1 \right)$$

$$\frac{\partial L(\pi, \lambda)}{\pi_h(a|s)} = d_h^{\pi^t}(s,a)\, Q_h^{\pi^t}(s,a) - \frac{1}{\eta} d_h^{\pi^t}(s,a) \cdot \left(\log \frac{\pi_h(a|s)}{\pi_h^t(a|s)} + 1 \right) - \lambda_{h,s} = 0$$

$$\Rightarrow \quad \pi_h(a|s) = \pi_h^t(a|s)\, e^{\eta Q_h^{\pi^t}(s,a)} / Z_t$$

<u>Consider the simplest case:</u> $\forall \pi \xrightarrow{\boxed{\text{oracle}}} Q_h^{\pi} \quad \forall h$

Theorem $\quad \text{Regret}(T) = \sum_{t=1}^{T} \left[V_1^{\pi^*}(s_1) - V_1^{\pi^t}(s_1) \right] \le \frac{H \log A}{\eta} + H^2$

(when $\eta \to \infty$, $\text{regret}(T) \to H^2 \Rightarrow$ policy iteration)

Lemma: (Performance difference Lemma) $\quad \forall$ policies π, π':

$$V_1^{\pi}(s_1) - V_1^{\pi'}(s_1) = \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} \langle Q_h^{\pi'}(s_h, \cdot), \pi_h(\cdot|s_h) - \pi_h'(\cdot|s_h) \rangle \,\Big|\, s_1 \right]$$

Proof: Let $\pi^{(0)} = \{\pi_1', \ldots, \pi_H'\}$, $\pi^{(1)} = \{\pi_1, \pi_2', \ldots, \pi_H'\}, \ldots, \pi^{(H)} = \{\pi_1, \ldots, \pi_H\}$

($\pi^{(t)}$: follows π for the first t steps and π' for the remaining $H-t$ steps)

Then: $\quad V_1^{\pi}(s_1) - V_1^{\pi'}(s_1) = V_1^{\pi^{(H)}}(s_1) - V_1^{\pi^{(0)}}(s_1) = \sum_{h=1}^{H} V_1^{\pi^{(h)}}(s_1) - V_1^{\pi^{(h-1)}}(s_1)$

But: $\quad V_1^{\pi^{(h)}}(s_1) - V_1^{\pi^{(h-1)}}(s_1) = \sum_{s_h} d_h^{\pi}(s_h) \left(\sum_a Q_h^{\pi'}(s_h, a)\, \pi_h(a|s_h) - V_h^{\pi'}(s_h) \right)$

$$= \mathbb{E}_{\pi}\left[\langle Q_h^{\pi'}(s_h, \cdot), \pi_h(\cdot|s_h) - \pi_h'(\cdot|s_h) \rangle \right]$$

- $$\pi_h^{t+1}(a|s) = \pi_h^t(a|s)\, e^{\eta Q_h^{\pi^t}(s,a)} / Z_{t,h,s}$$

$$\Rightarrow \quad Q_h^{\pi^t}(s,a) = \frac{1}{\eta} \log \frac{\pi_h^{t+1}(a|s)}{\pi_h^t(a|s)} + \frac{1}{\eta} \log Z_{t,h,s}$$

$$\Rightarrow \left\langle Q_h^{\pi^t}(s,\cdot), \pi_h^{t+1}(\cdot|s) \right\rangle = \frac{1}{\eta} KL\left[\pi_h^{t+1}(\cdot|s) \,\|\, \pi_h^t(\cdot|s)\right] + \frac{1}{\eta} \log Z_{t,h,s}$$

$$- \quad \left\langle Q_h^{\pi^t}(s,\cdot), \pi_h^{*}(\cdot|s) \right\rangle = \frac{1}{\eta}\left[KL\left[\pi^{*}(\cdot|s) \,\|\, \pi_h^t(\cdot|s)\right] - KL\left[\pi^{*}(\cdot|s) \,\|\, \pi_h^{t+1}(\cdot|s)\right]\right]$$

$$+ \frac{1}{\eta} \log Z_{t,h,s}$$

- $$\left\langle Q_h^{\pi^t}(s,\cdot), \pi_h^{t+1}(\cdot|s) - \pi_h^{*}(\cdot|s) \right\rangle = \frac{1}{\eta} KL_s(\pi_h^{t+1}, \pi_h^t)$$

$$- \frac{1}{\eta} KL_s(\pi_h^{*}, \pi_h^t) + \frac{1}{\eta} KL_s(\pi_h^{*}, \pi_h^{t+1})$$

$$\geq \frac{1}{\eta} KL_s(\pi_h^{*}, \pi_h^{t+1}) - \frac{1}{\eta} KL_s(\pi_h^{*}, \pi_h^t)$$

(replace π^* by π^t)

- $$\left\langle Q_h^{\pi^t}(s,\cdot), \pi^{t+1}(\cdot|s) - \pi_h^t(\cdot|s) \right\rangle \geq 0$$

$$\Longrightarrow \left\langle Q_h^{\pi^t}(s,\cdot), \pi_h^t(\cdot|s) - \pi_h^*(\cdot|s) \right\rangle \geq \left\langle Q_h^{\pi^t}(s,\cdot), \pi_h^t(\cdot|s) - \pi_h^{t+1}(\cdot|s) \right\rangle$$

$$+ \frac{1}{\eta} KL_s(\pi_h^*, \pi_h^{t+1}) - \frac{1}{\eta} KL_s(\pi_h^*, \pi_h^t)$$

<u>We have:</u>

$$V_1^{\pi^*}(s_1) - V_1^{\pi^t}(s_1) = \mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H} \left\langle Q_h^{\pi^t}(\cdot|s_h), \pi_h^*(\cdot|s_h) - \pi_h^t(\cdot|s_h) \right\rangle\right]$$

$$\leq \frac{1}{\eta}\sum_{h=1}^{H}\left[KL_{s_h}(\pi_h^*, \pi_h^t) - KL_{s_h}(\pi_h^*, \pi_h^{t+1})\right]$$

$$+ \mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H} \left\langle Q_h^{\pi^t}(s_h,\cdot), \pi_h^{t+1}(\cdot|s_h) - \pi_h^t(\cdot|s_h) \right\rangle\right]$$

<u>Note that:</u>

$$V_h^{\pi^{t+1}}(s_h) - V_h^{\pi^t}(s_h) = \mathbb{E}_{\pi^{t+1}}\left[\sum_{h'=h}^{H} \left\langle Q_{h'}^{\pi^t}(\cdot|s_{h'}), \pi_{h'}^{t+1}(\cdot|s_{h'}) - \pi_{h'}^t(\cdot|s_{h'}) \right\rangle\right]$$

$$\geq \left\langle Q_h^{\pi^t}(\cdot|s_h), \pi_h^{t+1}(\cdot|s_h) - \pi_h^t(\cdot|s_h) \right\rangle$$

$$V_1^{\pi^*}(s_1) - V_1^{\pi^t}(s_1) \le \frac{1}{\eta} \mathbb{E}_{\pi^*} \sum_{h=1}^{H} \left[KL_{s_h}(\pi_h^*, \pi_h^t) - KL_{s_h}(\pi_h^*, \pi_h^{t+1}) \right]$$

$$+ \mathbb{E}_{\pi^*} \left[\sum_{h=1}^{H} V_h^{\pi^{t+1}}(s_h) - V_h^{\pi^t}(s_h) \right]$$

$$\sum_{t=1}^{T} \text{RHS} \le \frac{1}{\eta} H \log A + H(H)$$

<u>Summary</u> (unknown r and p)

reference	Algorithm	Setting	regret
[Azar et al. '17]	VI_UCB	stochastic reward	$O(\sqrt{\text{poly}(H)\, SAT})$
[Efroni et al. '20]	NPG	stochastic reward	$O(\sqrt{\text{poly}(H)\, S^2AT})$
[Jin et al. '19]	Upper Occupancy Bound	adversarial reward	$O(\sqrt{\text{poly}(H)\, S^2AT})$
[Efroni et al. '20]	NPG	adversary reward	$O(T^{2/3})$