Class Tutorial 14

1. Review: Policy Gradient Algorithm

- Given initial parameters θ

Repeat:

- Simulate/implement a single episode $\boldsymbol{\tau} = (x_0, u_0, \ldots, x_T)$ of the controlled system with policy π_θ, with $x_0 \sim P(x_0)$.
- Compute $R(\boldsymbol{\tau}) = \sum_{t=0}^{T} r(x_t, u_t)$
- Compute $\hat{\nabla} J(\theta) = R(\boldsymbol{\tau}) \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(u_t \mid x_t)$
- Update parameters: (ϵ is a step size)

  $\theta := \theta + \epsilon \hat{\nabla} J(\theta)$

2. Policy Gradient algorithm – Softmax Policy

Consider the following policy representation:

$$\pi_\theta(u \mid x) = \frac{e^{\alpha\theta^\top \phi(x,u)}}{\sum_{u'} e^{\alpha\theta^\top \phi(x,u')}}$$

Where $\phi(x,u)$ are state-action features.

a. Write down the policy gradient (likelihood ratio method) algorithm with the softmax policy.

Solution:

a. All that we need to modify in the algorithm from the previous section is the gradient estimator:

$$\nabla_\theta \log \pi_\theta(u_t \mid x_t) = \nabla_\theta \log\left(\frac{e^{\alpha\theta^\top \phi(x,u)}}{\sum_{u'} e^{\alpha\theta^\top \phi(x,u')}}\right)$$
$$= \nabla_\theta\left(\alpha\theta^\top \phi(x,u) - \log \sum_{u'} e^{\alpha\theta^\top \phi(x,u')}\right)$$
$$= \alpha\phi(x,u) - \frac{\sum_{u'} \alpha\phi(x,u') e^{\alpha\theta^\top \phi(x,u')}}{\sum_{u'} e^{\alpha\theta^\top \phi(x,u')}}$$

3. The Policy Gradient Theorem

Consider an episodic and stationary MDP setting with a fixed initial state x_0, and a stationary parameterized policy π_θ, and let $J(\theta) = E^{\pi_\theta}(\sum_{t=0}^{T} R_t)$, where T is the time that a terminal state is reached.

a. Show that the following relation holds:

$$\nabla_\theta J(\theta) = \sum_{t=0}^{\infty}\sum_{x} P\left(x_t = x \mid \pi_\theta, x_0\right) \sum_{u} \nabla_\theta \pi_\theta\left(u \mid x\right) Q^\pi(x,u) \qquad (*)$$

b. Show that (*) is equivalent to

$$\nabla_\theta J(\theta) = E^\pi \sum_{t=0}^{T} \frac{\nabla_\theta \pi_\theta\left(u_t \mid x_t\right)}{\pi_\theta\left(u_t \mid x_t\right)} Q^\pi(x_t, u_t)$$

c. Consider tabular representation, i.e., we represent the true policy in its complete form, $\theta = \{\pi(a \mid s)\}_{s,a}$. Write (*) when assuming this representation.

d. What is the policy π' which maximizes $\langle \nabla_\pi J(\pi), \pi' \rangle$? i.e., the aligned in the direction in which the gradient is maximal.

Solution:

a. Recall that

$$Q^\pi(x,u) = E^\pi(\sum_{t=0}^{T} R_t \mid x_0 = x, u_0 = u)$$
$$V^\pi(x) = E^\pi(\sum_{t=0}^{T} R_t \mid x_0 = x)$$

And

$$V^{\pi}(x)=\sum_{u}\pi(u\mid x)Q^{\pi}(x,u).$$

Taking a gradient we have

$$\nabla_{\theta}V^{\pi}(x)=\sum_{u}\nabla_{\theta}\pi(u\mid x)Q^{\pi}(x,u)+\sum_{u}\pi(u\mid x)\nabla_{\theta}Q^{\pi}(x,u) \qquad (1)$$

From the Bellman equation, recall that

$$Q^{\pi}(x,u)=r(x,u)+\sum_{x'}P(x'\mid x,u)V^{\pi}(x'),$$

Therefore

$$\nabla_{\theta}Q^{\pi}(x,u)=\sum_{x'}P(x'\mid x,u)\nabla_{\theta}V^{\pi}(x'),$$

And plugging in (1) we obtain an equation for $\nabla_{\theta}V^{\pi}(x)$

$$\nabla_{\theta}V^{\pi}(x)=\sum_{u}\left(\nabla_{\theta}\pi(u\mid x)Q^{\pi}(x,u)+\pi(u\mid x)\sum_{x'}P(x'\mid x,u)\nabla_{\theta}V^{\pi}(x')\right)$$

Note that $J(\theta)=V^{\pi}(x_0)$, and we have

$$\begin{aligned}\nabla_{\theta}J(\theta)=\nabla_{\theta}V^{\pi}(x_0)&=\sum_{u}\nabla_{\theta}\pi(u\mid x_0)Q^{\pi}(x_0,u)\\&\quad+\sum_{u}\pi(u\mid x_0)\sum_{x'}P(x'\mid x_0,u)\nabla_{\theta}V^{\pi}(x')\\&=\sum_{u}\nabla_{\theta}\pi(u\mid x_0)Q^{\pi}(x_0,u)\\&\quad+\sum_{x'}P(x'\mid x_0,\pi)\nabla_{\theta}V^{\pi}(x')\end{aligned}$$

Unrolling $\nabla_{\theta}V^{\pi}(x')$ once on the right hand side gives

$$\nabla_{\theta}J(\theta)=\sum_{u}\nabla_{\theta}\pi(u\mid x_0)Q^{\pi}(x_0,u)+\sum_{x}P(x_1=x\mid x_0,\pi)\sum_{u}\nabla_{\theta}\pi(u\mid x)Q^{\pi}(x,u)+\sum_{x'}P(x'\mid x_1,\pi)\nabla_{\theta}V^{\pi}(x')$$

After unrolling $\nabla_{\theta}V^{\pi}(x')$ again and again, we obtain

$$\nabla_{\theta}J(\theta)=\sum_{t=0}^{\infty}\sum_{x}P\left(x_t=x\mid x_0,\pi\right)\sum_{u}\nabla_{\theta}\pi_{\theta}\left(u\mid x\right)Q^{\pi}(x,u)$$

b. By dividing and multiplying by $\pi_\theta(u \mid x)$ we have

$$\begin{aligned}
&\sum_{t=0}^{\infty}\sum_{x} P(x_t = x \mid x_0, \pi) \sum_{u} \frac{\nabla_\theta \pi_\theta(u \mid x)}{\pi_\theta(u \mid x)} \pi_\theta(u \mid x) Q^\pi(x,u) = \\
&= \sum_{t=0}^{\infty}\sum_{x,u} P(x_t = x, u_t = u \mid x_0, \pi) \frac{\nabla_\theta \pi_\theta(u \mid x)}{\pi_\theta(u \mid x)} Q^\pi(x,u) = \\
&= E^\pi \sum_{t=0}^{\infty} \frac{\nabla_\theta \pi_\theta(u_t \mid x_t)}{\pi_\theta(u_t \mid x_t)} Q^\pi(x_t, u_t) \\
&= E^\pi \sum_{t=0}^{T} \frac{\nabla_\theta \pi_\theta(u_t \mid x_t)}{\pi_\theta(u_t \mid x_t)} Q^\pi(x_t, u_t)
\end{aligned}$$

Where the last equation holds since the value of the terminal state is zero.

c. When using this representation, we have that

$$\nabla_\theta \pi(u \mid x) = \nabla_{\pi(u' \mid x')} \pi(u \mid x) = \delta_{u,u'} \delta_{x,x'}.$$

Plugging this into (*), we get,

$$\begin{aligned}
&\nabla_{\pi(u' \mid x')} J(\pi) = \sum_{t=0}^{\infty}\sum_{x} P(x_t = x \mid \pi, x_0) \sum_{u} \nabla_{\pi(u' \mid x')} \pi(u \mid x) Q^\pi(x,u) \\
&= \sum_{t=0}^{\infty}\sum_{x}\sum_{u} P(x_t = x \mid \pi, x_0) \delta_{u,u'} \delta_{x,x'} Q^\pi(x,u) \\
&= \sum_{t=0}^{\infty} P(x_t = x' \mid \pi, x_0) Q^\pi(x', u')
\end{aligned}$$

See that $J(\pi)$ is a scalar, and that $\nabla_\pi J(\pi)$ is a vector in R^{SA}.

d. We calculate the projection explicitly, by using basic properties of inner product. Remember that:

$$\left\langle \sum_i a_i, b \right\rangle = \sum_i \langle a_i, b \rangle$$
$$\langle \alpha a, b \rangle = \alpha \langle a, b \rangle$$

For α a constant, and vectors a, b.

Using these, we get that for any x

$$\left\langle \nabla_{\pi(\bullet|x)} J(\pi), \pi'\left(\bullet| x\right) \right\rangle = \sum_{t=0}^{\infty} P\left(x_t = x \mid \pi, x_0\right) \left\langle Q^{\pi}(x,\bullet), \pi'\left(\bullet| x\right) \right\rangle.$$

Thus, to maximize the inner product, we need to find the policy which maximizes $\left\langle Q^{\pi}(x,\bullet), \pi'\left(\bullet| x\right) \right\rangle$ for any x. By definition, this policy is **the greedy policy**,

$$\begin{aligned}
\pi_G\left(\bullet| x\right) &= \arg\max_{\overline{\pi}} \sum_{u} \overline{\pi}(u \mid x)(r(x,u) + \sum_{x'} P(x' \mid x,u) V^{\pi}(x')) \\
&= \arg\max_{\overline{\pi}} \sum_{u} \overline{\pi}(u \mid x) Q^{\pi}(x,u) \\
&= \arg\max_{\overline{\pi}} \left\langle Q^{\pi}(x,\bullet), \overline{\pi}\left(\bullet| x\right) \right\rangle
\end{aligned}$$