SS 2019: Human Reinforcement Learning **30/04/2019 (revised on 07/05/2019)**

Prof. Dr. Stefan Riezler Mayumi Ohta, ICL

Supplementary Material #1

1. **Bellman Equation**: Let's derive the Bellman Equation for the state-value function step by step.

$$\begin{aligned}
\mathbf{v}^{\pi}(s) &\stackrel{\text{def}}{=} \mathbb{E}\left[G_t | S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} | S_t = s\right] + \gamma \mathbb{E}\left[R_{t+2} | S_t = s\right] + \gamma^2 \mathbb{E}\left[R_{t+3} | S_t = s\right] + \cdots
\end{aligned}$$

Bellman Equation is expressed by a recursion, that is, we aim to produce $\mathbf{v}^{\pi}(s') = \mathbb{E}\left[G_{t+1} | S_{t+1} = s'\right]$ on the right hand side. To this end, we rewrite each term on the right hand side as follows:

$$\begin{aligned}
\mathbb{E}\left[R_{t+1} = r | S_t = s\right] &= \sum_{s'} \sum_{a} p(s', a | s) r \\
&= \sum_{s'} \sum_{a} \frac{p(s', a, s)}{p(s)} r \\
&= \sum_{s'} \sum_{a} \frac{p(s', a, s)}{p(s)} \frac{p(a, s)}{p(a, s)} r \\
&= \sum_{s'} \sum_{a} \frac{p(s', a, s)}{p(a, s)} \frac{p(a, s)}{p(s)} r \\
&= \sum_{s'} \sum_{a} p(s' | s, a) p(a | s) r \\
&= \sum_{a} \pi(a | s) \sum_{s'} p(s' | s, a) r \\
\mathbb{E}\left[R_{t+1} | S_t = s\right] &= \sum_{a} \pi(a | s) \sum_{s'} \sum_{r} p(s', r | s, a) r \\
\mathbb{E}\left[R_{t+2} = r' | S_t = s\right] &= \sum_{s''} \sum_{a'} \sum_{s'} \sum_{a} p(s'', a' | s') p(s', a | s) r' \\
&= \sum_{s'} \sum_{a} p(s', a | s) \sum_{s''} \sum_{a'} p(s'', a' | s') r' \\
&= \sum_{s'} \sum_{a} p(s | s', a) p(a | s) \sum_{s''} \sum_{a'} p(s'', a' | s') r' \\
&= \sum_{a} \pi(a | s) \sum_{s'} p(s' | s, a) \mathbb{E}\left[R_{t+2} = r' | S_{t+1} = s'\right] \\
\mathbb{E}\left[R_{t+2} | S_t = s\right] &= \sum_{a} \pi(a | s) \sum_{s'} \sum_{r} p(s', r | s, a) \mathbb{E}\left[R_{t+2} | S_{t+1} = s'\right]
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}\left[R_{t+3}=r''|S_t=s\right] &= \sum_{s'''}\sum_{a''}\sum_{s''}\sum_{a'}\sum_{s'}\sum_{a} p(s''',a''|s'')p(s'',a'|s')p(s',a|s)r'' \\
&= \sum_{s'}\sum_{a} p(s',a|s)\sum_{s'''}\sum_{a''}\sum_{s''}\sum_{a'} p(s''',a''|s')p(s'',a'|s')r'' \\
&= \sum_{s'}\sum_{a} p(s'|s,a)p(a|s)\sum_{s'''}\sum_{a''}\sum_{s''}\sum_{a'} p(s''',a''|s'')p(s'',a'|s')r'' \\
&= \sum_{a}\pi(a|s)\sum_{s'} p(s'|s,a)\mathbb{E}\left[R_{t+3}=r''|S_{t+1}=s'\right] \\
\mathbb{E}\left[R_{t+3}|S_t=s\right] &= \sum_{a}\pi(a|s)\sum_{s'}\sum_{r} p(s',r|s,a)\mathbb{E}\left[R_{t+3}|S_{t+1}=s'\right]
\end{aligned}
$$

Repeat the same procedure for $t+4, t+5, \cdots$, and plug them back in.

$$
\begin{aligned}
\mathbf{v}^{\pi}(s) &= \mathbb{E}\left[R_{t+1}|S_t=s\right]+\gamma\mathbb{E}\left[R_{t+2}|S_t=s\right]+\gamma^2\mathbb{E}\left[R_{t+3}|S_t=s\right]+\cdots \\
&= \sum_{a}\pi(a|s)\sum_{s'}\sum_{r} p(s',r|s,a)\left\{r+\gamma\mathbb{E}\left[R_{t+2}+\gamma R_{t+3}+\cdots|S_{t+1}=s'\right]\right\} \\
&= \sum_{a}\pi(a|s)\sum_{s'}\sum_{r} p(s',r|s,a)\left\{r+\gamma\mathbf{v}^{\pi}(s')\right\}
\end{aligned}
$$

Recall the notation we introduced in the lecture:

$$
\begin{aligned}
\mathcal{P}^{a}_{ss'} &\stackrel{\text{def}}{=} P\left[S_{t+1}=s'|S_t=s, A_t=a\right] = \sum_{r} p(s',r|s,a) \\
\mathcal{R}^{a}_{s} &\stackrel{\text{def}}{=} \mathbb{E}\left[R_{t+1}=r|S_t=s, A_t=a\right] = \sum_{s'}\sum_{r} p(s',r|s,a)r
\end{aligned}
$$

Thus, we get the desired equation:

$$
\mathbf{v}^{\pi}(s) = \sum_{a}\pi(a|s)\left(\mathcal{R}^{a}_{s}+\gamma\sum_{s'}\mathcal{P}^{a}_{ss'}\mathbf{v}^{\pi}(s')\right)
$$

2. (exercise) Derive the Bellman Equation for the action-value function, analogously.

$$
\begin{aligned}
\mathbf{q}^{\pi}(s,a) &= \mathbb{E}\left[G_t|S_t=s, A_t=a\right] \\
&= \ \vdots \\
&= \mathcal{R}^{a}_{s}+\gamma\sum_{s'}\mathcal{P}^{a}_{ss'}\sum_{a'}\pi(a'|s')\mathbf{q}^{\pi}(s',a')
\end{aligned}
$$

Special Thanks to Michael. Indeed, the Markov property need to be taken into account. That is, the probability to reach a state s'' in time step $t+2$ starting from a state s in t should be $p(s'',a'|s')p(s',a|s)$.