强化学习与有监督学习的差别在于强化学习的loss函数不可导

16.7

假设$T=3$，则有随机变量a_1,a_2,a_3表示每一步采取的动作，x_1,x_2,x_3表示采取对应动作进入的状态。$a'_1,a'_2,a'_3,x'_1,x'_2,x'_3$表示对应的具体取值。为了统一标记，把原公式里的$x$记为$x'_0$

$$
\begin{aligned}
V_3^{\pi}(x'_0) &= \mathbb{E}_{\pi}[\frac{1}{3}\sum_{t=1}^{3} r_t | x_0 = x'_0] \\
&= \sum_{\substack{a'_1,a'_2,a'_3 \in A, \\ x'_1,x'_2,x'_3 \in X}} [p(a'_1,a'_2,a'_3,x'_1,x'_2,x'_3)\frac{1}{3}\sum_{t=1}^{3} R^{a'_t}_{x'_{t-1}\to x'_t}] \\
&= \sum_{\substack{a'_1,a'_2,a'_3 \in A, \\ x'_1,x'_2,x'_3 \in X}} [\pi(x'_0,a'_1)P^{a'_1}_{x'_0\to x'_1}\pi(x'_1,a'_2)P^{a'_2}_{x'_1\to x'_2}\pi(x'_2,a'_3)P^{a'_3}_{x'_2\to x'_3}\frac{1}{3}(R^{a'_1}_{x'_0\to x'_1}+R^{a'_2}_{x'_1\to x'_2}+R^{a'_3}_{x'_2\to x'_3})] \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{\substack{a'_2,a'_3 \in A, \\ x'_1,x'_2,x'_3 \in X}} [P^{a'_1}_{x'_0\to x'_1}\pi(x'_1,a'_2)P^{a'_2}_{x'_1\to x'_2}\pi(x'_2,a'_3)P^{a'_3}_{x'_2\to x'_3}\frac{1}{3}(R^{a'_1}_{x'_0\to x'_1}+R^{a'_2}_{x'_1\to x'_2}+R^{a'_3}_{x'_2\to x'_3})] \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{x'_1\in X}P^{a'_1}_{x'_0\to x'_1}\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}(R^{a'_1}_{x'_0\to x'_1}+R^{a'_2}_{x'_1\to x'_2}+R^{a'_3}_{x'_2\to x'_3})] \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{x'_1\in X}P^{a'_1}_{x'_0\to x'_1}\{\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}R^{a'_1}_{x'_0\to x'_1}]+\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}R^{a'_2}_{x'_1\to x'_2}]+\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}R^{a'_3}_{x'_2\to x'_3}]\} \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{x'_1\in X}P^{a'_1}_{x'_0\to x'_1}\{\frac{1}{3}R^{a'_1}_{x'_0\to x'_1}\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)]+\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}R^{a'_2}_{x'_1\to x'_2}]+\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}R^{a'_2}_{x'_2\to x'_3}]\} \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{x'_1\in X}P^{a'_1}_{x'_0\to x'_1}\{\frac{1}{3}R^{a'_1}_{x'_0\to x'_1}+\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}R^{a'_2}_{x'_1\to x'_2}]+\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}R^{a'_2}_{x'_2\to x'_3}]\} \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{x'_1\in X}P^{a'_1}_{x'_0\to x'_1}\{\frac{1}{3}R^{a'_1}_{x'_0\to x'_1}+\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{3}(R^{a'_2}_{x'_1\to x'_2}+R^{a'_3}_{x'_2\to x'_3})]\} \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{x'_1\in X}P^{a'_1}_{x'_0\to x'_1}\{\frac{1}{3}R^{a'_1}_{x'_0\to x'_1}+\frac{2}{3}\sum_{\substack{a'_2,a'_3 \in A, \\ x'_2,x'_3 \in X}} [p(a'_2,a'_3,x'_2,x'_3)\frac{1}{2}(R^{a'_2}_{x'_1\to x'_2}+R^{a'_3}_{x'_2\to x'_3})]\} \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{x'_1\in X}P^{a'_1}_{x'_0\to x'_1}(\frac{1}{3}R^{a'_1}_{x'_0\to x'_1}+\frac{2}{3}\mathbb{E}_{\pi}[\frac{1}{2}\sum_{t=1}^{2} r_t | x_0 = x'_1]) \\
&= \sum_{a'_1\in A}\pi(x'_0,a'_1)\sum_{x'_1\in X}P^{a'_1}_{x'_0\to x'_1}(\frac{1}{3}R^{a'_1}_{x'_0\to x'_1}+\frac{2}{3}V_2^{\pi}(x'_1))
\end{aligned}
$$

推广一下就成为书里的公式：

$$
\begin{aligned}
V_T^{\pi}(x_0') &= \mathbb{E}_{\pi}[\frac{1}{T}\sum_{t=1}^{T} r_t | x_0 = x_0'] \\
&= \sum_{\substack{a_1',\dots,a_T' \in A, \\ x_1',\dots,x_T' \in X}} [p(a_1',\dots,a_T',x_1',\dots,x_T')\frac{1}{T}\sum_{t=1}^{T} R^{a_t'}_{x_{t-1}' \to x_t'}] \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} \sum_{\substack{a_2',\dots,a_T' \in A, \\ x_2',\dots,x_T' \in X}} [p(a_2',\dots,a_T',x_2',\dots,x_T')\frac{1}{T}\sum_{t=1}^{T} R^{a_t'}_{x_{t-1}' \to x_t'}] \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} \sum_{\substack{a_2',\dots,a_T' \in A, \\ x_2',\dots,x_T' \in X}} [p(a_2',\dots,a_T',x_2',\dots,x_T')\frac{1}{T} R^{a_1'}_{x_0' \to x_1'} + p(a_2',\dots,a_T',x_2',\dots,x_T')\frac{1}{T}\sum_{t=2}^{T} R^{a_t'}_{x_{t-1}' \to x_t'}] \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} \{\frac{1}{T} R^{a_1'}_{x_0' \to x_1'} + \sum_{\substack{a_2',\dots,a_T' \in A, \\ x_2',\dots,x_T' \in X}} [p(a_2',\dots,a_T',x_2',\dots,x_T')\frac{1}{T}\sum_{t=2}^{T} R^{a_t'}_{x_{t-1}' \to x_t'}]\} \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} \{\frac{1}{T} R^{a_1'}_{x_0' \to x_1'} + \frac{T-1}{T} \sum_{\substack{a_2',\dots,a_T' \in A, \\ x_2',\dots,x_T' \in X}} [p(a_2',\dots,a_T',x_2',\dots,x_T')\frac{1}{T-1}\sum_{t=2}^{T} R^{a_t'}_{x_{t-1}' \to x_t'}]\} \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} \{\frac{1}{T} R^{a_1'}_{x_0' \to x_1'} + \frac{T-1}{T} V_{T-1}^{\pi}(x_1')\}
\end{aligned}
$$

16.8

$$
\begin{aligned}
V_{\gamma}^{\pi}(x_0') &= \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid x_0 = x_0'] \\
&= \mathbb{E}_{\pi}[r_1 + \sum_{t=1}^{\infty} \gamma^t r_{t+1} \mid x_0 = x_0'] \\
&= \mathbb{E}_{\pi}[r_1 + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} r_{t+1} \mid x_0 = x_0'] \\
&= \sum_{\substack{a_1',\dots,a_T' \in A, \\ x_1',\dots,x_T' \in X}} [p(a_1',\dots,a_T',x_1',\dots,x_T')(R^{a_1'}_{x_0' \to x_1'} + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} R^{a_1'}_{x_t' \to x_{t+1}'})] \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} \sum_{\substack{a_2',\dots,a_T' \in A, \\ x_2',\dots,x_T' \in X}} [p(a_2',\dots,a_T',x_2',\dots,x_T')(R^{a_1'}_{x_0' \to x_1'} + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} R^{a_1'}_{x_t' \to x_{t+1}'})] \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} (R^{a_1'}_{x_0' \to x_1'} + \sum_{\substack{a_2',\dots,a_T' \in A, \\ x_2',\dots,x_T' \in X}} [p(a_2',\dots,a_T',x_2',\dots,x_T')(\gamma \sum_{t=1}^{\infty} \gamma^{t-1} R^{a_1'}_{x_t' \to x_{t+1}'})]) \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} (R^{a_1'}_{x_0' \to x_1'} + \gamma \mathbb{E}_{\pi}[\sum_{t=1}^{\infty} \gamma^t r_{t+1} \mid x_0 = x_1']) \\
&= \sum_{a_1' \in A} \pi(x_0', a_1') \sum_{x_1' \in X} P^{a_1'}_{x_0' \to x_1'} (R^{a_1'}_{x_0' \to x_1'} + \gamma V_{\gamma}^{\pi}(x_1'))
\end{aligned}
$$

16.29

把$\frac{1}{t+1}$替代为α

$$
\begin{aligned}
Q_{t+1}^{\pi}(x,a) &= Q_t^{\pi}(x,a) + \alpha(r_{t+1} - Q_t^{\pi}(x,a)) \\
&= (1-\alpha)Q_t^{\pi}(x,a) + \alpha r_{t+1} \\
&= (1-\alpha)((1-\alpha)Q_{t-1}^{\pi}(x,a) + \alpha r_t) + \alpha r_{t+1} \\
&= (1-\alpha)^2 Q_{t-1}^{\pi}(x,a) + \alpha((1-\alpha)r_t + r_{t+1}) \\
&= \alpha(r_{t+1} + (1-\alpha)r_t + (1-\alpha)^2 r_{t-1} + (1-\alpha)^3 r_{t-2} + \ldots + (1-\alpha)^t t_1)
\end{aligned}
$$

每一步奖赏的系数之和：

$$
S = \alpha\frac{1-(1-\alpha)^{t+1}}{1-(1-\alpha)} = \alpha\frac{1-(1-\alpha)^{t+1}}{\alpha} = 1-(1-\alpha)^{t+1} \to 1
$$