强化学习

值迭代

价值函数，给定当前状态 s，(当前行动 a)，和策略 π

$$V^{\pi}(s) = \mathop{E}_{\tau\sim\pi}[R(\tau)|s]$$

$$Q^{\pi}(s,a) = E_{\tau\sim\pi}[R(\tau)\mid s,a]$$

Bellman 方程

$$Q^*(s,a) = \max_{\pi} E\Big[\sum_{t=0}\gamma^t r_t \mid s_0=s,\ a_0=a,\ \pi\Big]$$

$$= \mathop{E}_{s'\sim P(\cdot|s,a)}\Big[r_0 + \max_{\pi} E\Big[\sum_{t=1}\gamma^t r_t \mid s_0=s,\ a_0=a,\ \pi\Big]$$

给定 s'，选 a'，使 $\gamma(\sum\gamma^{t-1}r_t)$ 最大

$$\Rightarrow a' = \arg\max_{a} Q^*(s',a)$$

$$= \mathop{E}_{s'\sim P(\cdot|s,a)}\Big[r_0(s',s,a) + \gamma \max_{a'} Q^*(s',a')\Big]$$

深度值迭代

$$\nabla_\theta L(\theta) = \nabla_\theta E_{s,a\sim P(\cdot)}\big[(y - Q(s,a;\theta))^2\big]$$

$$= E_{s,a\sim P(\cdot)}\big[-2(y - Q(s,a;\theta))\,\nabla_\theta Q(s,a;\theta)\big]$$