Approximate Dynamic Programming

1. Approximate Value Iteration

考虑对bellman最优方程：

$$\mathcal{T}V_k = \arg\max_a \mathbb{E}[r + \gamma V_k],$$

当我们使用函数近似时，更新后的价值$\hat{V}_{k+1}$未必完全等于$\mathcal{T}V_k$，即存在误差$\epsilon_{k+1} = \hat{V}_{k+1} - \mathcal{T}V_k$，误差会随着更新逐轮累计，导致学到的$\hat{V}$偏离$V^*$:

$$\begin{aligned}
\|\hat{V}_k - V^*\|_\infty &= \|\mathcal{T}V_{k-1} + \epsilon_k - \mathcal{T}V^*\|_\infty \\
&\leq \gamma\|\hat{V}_{k-1} - V^*\|_\infty + \|\epsilon_k\|_\infty \\
&\leq \gamma^k\|V_0 - V^*\|_\infty + \sum_{i=1}^{k}\gamma^{k-i}\|\epsilon_i\|_\infty.
\end{aligned}$$

若$\sup_k \|\epsilon_k\|_\infty \leq \epsilon$，那么有：

$$\lim_{k\to\infty}\|\hat{V}_k - V^*\|_\infty = \frac{\epsilon}{1-\gamma}. \tag{1}$$

接下来我们考虑的问题是，给定价值估计$\hat{V}$和最优价值V^*的误差$\|\hat{V} - V^*\|_\infty$，那么基于$\hat{V}$导出的贪心策略π的价值V^π和V^*的差距有多大？

$$\begin{aligned}
\|V^* - V^\pi\|_\infty &\leq \|V^* - \mathcal{T}^\pi\hat{V}\|_\infty + \|\mathcal{T}^\pi\hat{V} - \mathcal{T}^\pi V^\pi\|_\infty \\
&\leq \|\mathcal{T}V^* - \mathcal{T}\hat{V}\|_\infty + \gamma\|\hat{V} - V^\pi\|_\infty \\
&\leq \gamma\|V^* - \hat{V}\|_\infty + \gamma\left(\|\hat{V} - V^*\|_\infty + \|V^* - V^\pi\|_\infty\right) \\
&\leq \frac{2\gamma}{1-\gamma}\|V^* - \hat{V}\|_\infty.
\end{aligned} \tag{2}$$

将(1)和(2)式合并在一起可得，当使用approximate value iteration时，得到的策略π的价值和最优价值的差距为：

$$\lim_{k\to\infty} \|V^{\pi_k} - V^*\|_\infty \leq \frac{2\gamma}{1-\gamma}\epsilon^2. \tag{3}$$