1 Reinforcement Learning

1.1 Foundation

⋆ 动态规划算法的特性：

- 需要环境模型，即状态转移概率P_{sa}.
- 状态值函数的估计是自举的(bootstrapping)，即当前状态值函数的更新依赖于已知的其他状态值函数.

⋆ 蒙特卡罗方法的特点：

- 可以从经验中学习不需要环境模型.
- 状态值函数的估计是相互独立的.
- 只能用于episode tasks.

⋆ Monte Carlo

Monte Carlo的状态值函数更新公式如下：

$$V(s_t) \leftarrow V(s_t) + \alpha[R_t - V(s_t)] \tag{1}$$

其中R_t是每个episode结束后获得的实际累积回报，α 是学习率，这个式子的直观的理解就是用实际累积回报R_t 作为状态值函数$V(s_t)$的估计值。具体做法是对每个episode，考察实验中s_t的实际累积回报R_t和当前估计$V(s_t)$的偏差值，并用该偏差值乘以学习率来更新得到$V(S_t)$的新估值。

⋆ TD(0)

把等式1中R_t换成$r_{t+1} + \gamma V(s_{t+1})$，就得到了TD(0)的状态值函数更新公式：

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \tag{2}$$

为什么修改成这种形式呢，回忆一下状态值函数的定义：

$$V^{\pi}(s) = E_{\pi}[r(s^{'}|s,a) + \gamma V^{\pi}(s^{'})] \tag{3}$$

容易发现这其实是根据等式3的形式，利用真实的立即回报r_{t+1}和下个状态的值函数$V(s_{t+1})$来更新$V(s_t)$，这种方式就称为时间差分(temporal difference)。由于没有状态转移概率，所以要利用多次实验来得到期望状态值函数估值。类似MC方法，在足够多的实验后，状态值函数的估计是能够收敛于真实值的。

- aaa.
- BBB.

1. AAA.
2. BBB.

- B1.
- B2

BBB.

3. CCC.

section name goes here

★ term definition

DEF: term - and it's definition

★ an example

EX: example heading

★ a system of equations

$$\begin{cases} 2x + 4y = 2 \\ 2x + 6y = 3 \end{cases}$$

★ working a multistep problem

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \xrightarrow{R_1+R_2} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \Rightarrow \begin{cases} x = 1 \\ y = 2 \\ z = 3 \end{cases}$$

★ a vector in R^3

$$v = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

★ a multi-step process

$$A \xrightarrow{\text{do stuff}} B \xrightarrow{\text{more stuff}} C$$

★ an enumerated list

1. this is the first item in an enumerated list
2. this is the second item in an enumerated list

★ manually broken lines

the first line
the second line
the third line

* some math

$$\int_a^b f(x)\,dx \int f(x)\,dx \frac{\pi}{2} \sqrt{\theta}\; n = 1, 2, 3 \ldots 4$$