16.2

$$Q_n(k)=rac{1}{n}((n-1) imes Q_{n-1}(k)+v_n)$$

[推导]:

$$Q_n(k) = rac{1}{n} \sum_{i=1}^n v_i = rac{1}{n} \Biggl(\sum_{i=1}^{n-1} v_i + v_n \Biggr) = rac{1}{n} ((n-1)Q_{n-1}(k) + v_n)$$

16.4

$$P(k) = rac{e^{rac{Q(k)}{ au}}}{\sum_{i=1}^K e^{rac{Q(i)}{ au}}}$$

au越小则平均奖赏高的摇臂被选取的概率越高

[解析]:

$$P(k) = rac{e^{rac{Q(k)}{ au}}}{\sum_{i=1}^K e^{rac{Q(i)}{ au}}} \propto e^{rac{Q(k)}{ au}} \propto rac{Q(k)}{ au} \propto rac{1}{ au}$$

16.7

$$egin{aligned} V_T^\pi(x) &= \mathbb{E}_\pi[rac{1}{T}\sum_{t=1}^T r_t \mid x_0 = x] \ &= \mathbb{E}_\pi[rac{1}{T}r_1 + rac{T-1}{T}rac{1}{T-1}\sum_{t=2}^T r_t \mid x_0 = x] \ &= \sum_{a \in A} \pi(x,a) \sum_{x' \in X} P_{x
ightarrow x'}^a(rac{1}{T}R_{x
ightarrow x'}^a + rac{T-1}{T}\mathbb{E}_\pi[rac{1}{T-1}\sum_{t=1}^{T-1} r_t \mid x_0 = x']) \ &= \sum_{a \in A} \pi(x,a) \sum_{x' \in X} P_{x
ightarrow x'}^a(rac{1}{T}R_{x
ightarrow x'}^a + rac{T-1}{T}V_{T-1}^\pi(x')]) \end{aligned}$$

[解析]:

因为

$$\pi(x, a) = P(action = a | state = x)$$

表示在状态x下选择动作a的概率,又因为动作事件之间两两互斥且和为动作空间,由全概率展开公式

$$P(A) = \sum_{i=1}^{\infty} P(B_i) P(A \mid B_i)$$

可得

$$egin{aligned} &= \mathbb{E}_{\pi}[rac{1}{T}r_{1} + rac{T-1}{T}rac{1}{T-1}\sum_{t=2}^{T}r_{t}\mid x_{0} = x] \ &= \sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x o x'}^{a}(rac{1}{T}R_{x o x'}^{a} + rac{T-1}{T}\mathbb{E}_{\pi}[rac{1}{T-1}\sum_{t=1}^{T-1}r_{t}\mid x_{0} = x']) \end{aligned}$$

其中

$$r_1=\pi(x,a)P^a_{x
ightarrow x'}R^a_{x
ightarrow x'}$$

最后一个等式用到了递归形式。

16.8

$$V^\pi_\gamma(x) = \sum_{a \in A} \pi(x,a) \sum_{x' \in X} P^a_{x
ightarrow x'}(R^a_{x
ightarrow x'} + \gamma V^\pi_\gamma(x'))$$

[推导]:

$$egin{aligned} V^\pi_\gamma(x) &= \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_{t+1} \mid x_0 = x] \ &= \mathbb{E}_\pi[r_1 + \sum_{t=1}^\infty \gamma^t r_{t+1} \mid x_0 = x] \ &= \mathbb{E}_\pi[r_1 + \gamma \sum_{t=1}^\infty \gamma^{t-1} r_{t+1} \mid x_0 = x] \ &= \sum_{a \in A} \pi(x,a) \sum_{x' \in X} P^a_{x o x'}(R^a_{x o x'} + \gamma \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_{t+1} \mid x_0 = x']) \ &= \sum_{x \in A} \pi(x,a) \sum_{x' \in X} P^a_{x o x'}(R^a_{x o x'} + \gamma V^\pi_\gamma(x')) \end{aligned}$$

16.16

$$V^{\pi}(x) \leq V^{\pi'}(x)$$

[推导]:

$$egin{align*} V^{\pi}(x) & \leq Q^{\pi}(x,\pi'(x)) \ & = \sum_{x' \in X} P^{\pi'(x)}_{x o x'}(R^{\pi'(x)}_{x o x'} + \gamma V^{\pi}(x')) \ & \leq \sum_{x' \in X} P^{\pi'(x)}_{x o x'}(R^{\pi'(x)}_{x o x'} + \gamma Q^{\pi}(x',\pi'(x'))) \ & = \sum_{x' \in X} P^{\pi'(x)}_{x o x'}(R^{\pi'(x)}_{x o x'} + \gamma \sum_{x' \in X} P^{\pi'(x')}_{x' o x'}(R^{\pi'(x')}_{x' o x'} + \gamma V^{\pi}(x'))) \ & = \sum_{x' \in X} P^{\pi'(x)}_{x o x'}(R^{\pi'(x)}_{x o x'} + \gamma V^{\pi'}(x')) \ & = V^{\pi'}(x) \end{split}$$

其中,使用了动作改变条件

$$Q^{\pi}(x,\pi'(x)) \geq V^{\pi}(x)$$

以及状态-动作值函数

$$Q^{\pi}(x',\pi'(x')) = \sum_{x' \in X} P^{\pi'(x')}_{x' o x'} (R^{\pi'(x')}_{x' o x'} + \gamma V^{\pi}(x'))$$

于是, 当前状态的最优值函数为

$$V^*(x) = V^{\pi'}(x) \geq V^\pi(x)$$

16.31

$$Q^{\pi}_{t+1}(x,a) = Q^{\pi}_{t}(x,a) + lpha(R^{a}_{x o x'} + \gamma Q^{\pi}_{t}(x',a') - Q^{\pi}_{t}(x,a))$$

[推导]: 对比公式16.29

$$Q^\pi_{t+1}(x,a) = Q^\pi_t(x,a) + rac{1}{t+1}(r_{t+1} - Q^\pi_t(x,a))$$

以及由

$$\frac{1}{t+1} = \alpha$$

可知

$$r_{t+1} = R^a_{x
ightarrow x'} + \gamma Q^\pi_t(x',a')$$

而由v折扣累积奖赏可估计得到。