16.2

$$Q_n(k) = \frac{1}{n}((n-1) \times Q_{n-1}(k) + v_n)$$

[推导]：

$$Q_n(k) = \frac{1}{n}\sum_{i=1}^{n} v_i = \frac{1}{n}\left(\sum_{i=1}^{n-1} v_i + v_n\right) = \frac{1}{n}((n-1)Q_{n-1}(k) + v_n)$$

16.4

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

τ越小则平均奖赏高的摇臂被选取的概率越高

[解析]：

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

16.7

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

[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

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

其中

$$
r_1 = \pi(x,a)P_{x\to x'}^{a}R_{x\to x'}^{a}
$$

最后一个等式用到了递归形式。

16.8

$$
V_{\gamma}^{\pi}(x) = \sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x\to x'}^{a}(R_{x\to x'}^{a} + \gamma V_{\gamma}^{\pi}(x'))
$$

[推导]：

$$
\begin{aligned}
V_{\gamma}^{\pi}(x) &= \mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^{t}r_{t+1} \mid x_0 = x] \\
&= \mathbb{E}_{\pi}[r_1 + \sum_{t=1}^{\infty}\gamma^{t}r_{t+1} \mid x_0 = x] \\
&= \mathbb{E}_{\pi}[r_1 + \gamma\sum_{t=1}^{\infty}\gamma^{t-1}r_{t+1} \mid x_0 = x] \\
&= \sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x\to x'}^{a}(R_{x\to x'}^{a} + \gamma\mathbb{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^{t}r_{t+1} \mid x_0 = x']) \\
&= \sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x\to x'}^{a}(R_{x\to x'}^{a} + \gamma V_{\gamma}^{\pi}(x'))
\end{aligned}
$$

16.16

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

[推导]：

$$
\begin{aligned}
V^{\pi}(x) &\leq Q^{\pi}(x,\pi'(x)) \\
&= \sum_{x'\in X}P_{x\to x'}^{\pi'(x)}(R_{x\to x'}^{\pi'(x)} + \gamma V^{\pi}(x')) \\
&\leq \sum_{x'\in X}P_{x\to x'}^{\pi'(x)}(R_{x\to x'}^{\pi'(x)} + \gamma Q^{\pi}(x',\pi'(x'))) \\
&= \sum_{x'\in X}P_{x\to x'}^{\pi'(x)}(R_{x\to x'}^{\pi'(x)} + \gamma\sum_{x'\in X}P_{x'\to x'}^{\pi'(x')}(R_{x'\to x'}^{\pi'(x')} + \gamma V^{\pi}(x'))) \\
&= \sum_{x'\in X}P_{x\to x'}^{\pi'(x)}(R_{x\to x'}^{\pi'(x)} + \gamma V^{\pi'}(x')) \\
&= V^{\pi'}(x)
\end{aligned}
$$

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

$$Q^{\pi}(x',\pi'(x'))=\sum_{x'\in X}P_{x'\rightarrow x'}^{\pi'(x')}(R_{x'\rightarrow x'}^{\pi'(x')}+\gamma V^{\pi}(x'))$$

于是，当前状态的最优值函数为

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

16.31

$$Q_{t+1}^{\pi}(x,a)=Q_{t}^{\pi}(x,a)+\alpha(R_{x\rightarrow x'}^{a}+\gamma Q_{t}^{\pi}(x',a')-Q_{t}^{\pi}(x,a))$$

[推导]：对比公式16.29

$$Q_{t+1}^{\pi}(x,a)=Q_{t}^{\pi}(x,a)+\frac{1}{t+1}(r_{t+1}-Q_{t}^{\pi}(x,a))$$

以及由

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

可知

$$r_{t+1}=R_{x\rightarrow x'}^{a}+\gamma Q_{t}^{\pi}(x',a')$$

而由γ折扣累积奖赏可估计得到。