1、(1) 老策略 π，新策略 π'

贪心策略：$\pi'(s) = \underset{a\in A}{\arg\max}\, q_\pi(s,a)$

$$q_\pi(s,\pi'(s)) = \max_{a\in A} q_\pi(s,a)$$

$$\begin{aligned} V_\pi(s) &= \sum_{a\in A} \pi(a|s)\, q_\pi(s,a) \\ &\le \sum_{a\in A} \pi(a|s)\, q_\pi(s,\pi'(s)) \\ &= q_\pi(s,\pi'(s)) \sum_{a\in A} \pi(a|s) \\ &= q_\pi(s,\pi'(s)) \end{aligned}$$

引理※

$$\begin{aligned} q_\pi(s,\pi'(s)) &= E[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = \pi'(s)] \\ &= E_{\pi'}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &\le E_{\pi'}[R_{t+1} + \gamma \underbrace{q_\pi(S_{t+1}, \pi'(S_{t+1}))}_{\text{继续套用本不等式}} \mid S_t = s] \\ &\le \cdots\cdots \\ &\le E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s] = V_{\pi'}(s) \end{aligned}$$

综上，$V_\pi(s) \le V_{\pi'}(s)$

(2) 老策略 π，新策略 π'

ε-贪心策略：
$$\pi'(a|s) = \begin{cases} 1-\varepsilon+\dfrac{\varepsilon}{|A|}, & a = \underset{a}{\arg\max}\, q_\pi(s,a) \\ \dfrac{\varepsilon}{|A|}, & \text{otherwise} \end{cases}$$

$$q_\pi(s,\pi'(s)) = \frac{\varepsilon}{|A|}\sum_a q_\pi(s,a) + (1-\varepsilon)\max_a q_\pi(s,a)$$

由于 $\sum_{a\in A}\pi(a|s)=1$ $\quad\therefore \sum_{a\in A}\frac{\pi(a|s)-\frac{\varepsilon}{|A|}}{1-\varepsilon}=1$

$$q_\pi(s,\pi'(s)) \geqslant \frac{\varepsilon}{|A|}\sum_a q_\pi(s,a) + (1-\varepsilon)\sum_a \frac{\pi(a|s)-\frac{\varepsilon}{|A|}}{1-\varepsilon}\cdot q_\pi(s,a)$$

$$= \sum_a \pi(a|s) q_\pi(s,a) = v_\pi(s)$$

$\therefore$ $q_\pi(s,\pi'(s)) \geqslant v_\pi(s)$，再套用(1)中引理※，可证 $v_\pi(s)\leq v_{\pi'}(s)$

(3) ε-greedy 避免策略陷入"死循环"，加入了一定的"探索"成份，虽然可能对最优性的利用有小幅损失，但换取了更多探索真正最优策略的机会。因此也更稳定.

2、 (1) 同步

$$\begin{bmatrix} V_2(a) \\ V_2(b) \\ V_2(c) \end{bmatrix} = \begin{bmatrix} -8+0.5\times V_1(b) \\ \max\{-2+0.5V_1(c),\ 2+0.5V_1(a)\} \\ \max\{8+0.5V_1(b),\ \frac{1}{4}\times 4+0.5(\frac{V_1(a)}{4}+\frac{3V_1(c)}{4})\} \end{bmatrix} = \begin{bmatrix} -7 \\ 3 \\ 9 \end{bmatrix}$$

$$\therefore \pi_2(a|s) = \begin{cases} ab, & s=a \\ ba, & s=b \\ cb, & s=c \end{cases}$$

(2) 异步：

$$V_2(a) = -8+0.5\times V_1(b) = -7$$

$$V_2(b) = \max\{-2+0.5V_1(c),\ 2+0.5V_2(a)\} = -1$$

$$V_2(c) = \max\{8+0.5V_2(b),\ \frac{1}{4}\times 4+0.5(\frac{V_2(a)}{4}+\frac{3V_1(c)}{4})\} = 7.5$$

$$\therefore \pi'_2(a|s) = \begin{cases} ab, & s=a \\ bc, & s=b \\ cb, & s=c \end{cases}$$