1. a) 同步：根据 $v_{k+1}(s)=\max\limits_{a\in A}\left(r_s^a+\gamma\sum\limits_{s'\in S}P_{ss'}^a v_k(s')\right)$，计算得：

$V_2(a) = -8 + 0.3\times V_1(b) = -7.4$

2.6

$V_2(b) = \max\{2+0.3V_1(a),\ -2+0.3V_1(c)\} = \cancel{2.3}$

$V_2(c) = \max\{0.25(4+0.3V_1(a)) + 0.75(0+0.3V_1(c)),\ 8+0.3V_1(b)\} = 8.6$

由确定性贪心策略 $\pi_*(s)=\arg\max\limits_{a\in A}\left(r_s^a+\gamma\sum\limits_{s'\in S}P_{ss'}^a v_*(s')\right)$，得到策略 $\pi_2(a|s)$ 为：

$\pi_2(a=ab\mid s=A)=1$

$\pi_2(a=ba\mid s=B)=1$，$\pi_2(a=bc\mid s=B)=0$

$\pi_2(a=ca\mid s=C)=0$，$\pi_2(a=cb\mid s=C)=1$

b) 异步：根据 $v(s)=\max\limits_{a\in A}\left(r_s^a+\gamma\sum\limits_{s'\in S}P_{ss'}^a v(s')\right)$，计算得：

$V(a) = -8 + 0.3V(b) = -7.4$

$V(b) = \max\{2+0.3V(a),\ -2+0.3V(c)\} = -0.22$

$V(c) = \max\{0.25(4+0.3V(a)) + 0.75(0+0.3V(c)),\ 8+0.3V(b)\} = 7.934$

由确定性贪心策略，得到策略 $\pi_2'(a|s)$ 为：

$\pi_2'(a=ab\mid s=A)=1$

$\pi_2'(a=ba\mid s=B)=1$，$\pi_2'(a=bc\mid s=B)=0$

$\pi_2'(a=ca\mid s=C)=0$，$\pi_2'(a=cb\mid s=C)=1$

2. a) 根据状态价值贝尔曼期望方程 $v_\pi(s)=\sum\limits_{a\in A}\pi(a|s)\left(r_s^a+\gamma\sum\limits_{s'\in S}P_{ss'}^a v_\pi(s')\right)$

列出 $\begin{cases} v_\pi(A)=P_{AA}(3+v_\pi(A))+P_{AB}(-3+v_\pi(B)) \\ v_\pi(B)=P_{BA}(3+v_\pi(A))+P_{Bt}\cdot 0 \end{cases}$ 解得 $\begin{cases} v_\pi(A)=-1 \\ v_\pi(B)=1 \end{cases}$

∴ 状态价值函数 $v(A)=-1$，$v(B)=1$

b) 首次访问：

∵ $G_1(A)=+2+3-5+5-2=3$，$G_1(B)=-5+5-2=-2$

$G_2(A)=+3-3=0$，$G_2(B)=-2+3-3=-2$

∴ $V(A)=\frac{1}{2}(G_1(A)+G_2(A))=1.5$，$V(B)=\frac{1}{2}(G_1(B)+G_2(B))=-2$

每次访问：

∵ $G_{11}(A)=+2+3-5+5-2=3$，$G_{12}(A)=+3-5+5-2=+1$，$G_{13}(A)=+5-2=3$

$G_{11}(B)=-5+5-2=-2$，$G_{12}(B)=-2$

$G_{21}(A)=+3-3=0$，$G_{21}(B)=-2+3-3=-2$，$G_{22}(B)=-3$

∴ $V(A)=\frac{1}{4}(G_{11}(A)+G_{12}(A)+G_{13}(A)+G_{21}(A))=1.75$，$V(B)=\frac{1}{4}(G_{11}(B)+G_{12}(B)+G_{21}(B)+G_{22}(B))=-2.25$

综上，使用首次访问，估计状态价值函数 $v(A)=1.5$，$v(B)=-2$；

使用每次访问，估计 $v(A)=1.75$，$v(B)=-2.25$