3.

$$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \ldots = (1 + \gamma + \gamma^2 + \ldots) \times 6 = \frac{6}{1-\gamma} = 60$$

$$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \ldots = 2 + (\gamma + \gamma^2 + \ldots) \times 6 = \frac{2+4\gamma}{1-\gamma} = 56$$

4.

	1	2
3	4	5
6	7	

补充条件：设左上右下分别为终止状态0，8，网格中9个格子对应状态{0,1,2,...,8}；同时只要智能体进行了动作，那么即使由于会移出网格而使得状态保持不变，该步回报也为-1，即$R_{t+1} = -1$（例如$r_3^{left} = -1$）（即定义得和Lecture12,13中一样）

由Bellman期望方程：

$$v^\pi = r^\pi + \gamma P^\pi v^\pi$$

需要求r^π, P^π. 由等概率随机策略，以及不能超出网格的约束，可得单步回报r^π为：

$$r^\pi = \begin{pmatrix} 0 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & 0 \end{pmatrix},$$
$$\text{或} = (0, -1, -1, -1, -1, -1, -1, -1, 0)^T$$

等概率随机策略下的转移概率矩阵P^π为：

$$P^\pi = (p^\pi_{ss'}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0.5 & 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0.25 & 0 & 0 & 0.25 & 0.25 & 0 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0 & 0.25 & 0 & 0.25 & 0 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0 & 0.25 & 0.25 & 0 & 0 & 0.25 \\ 0 & 0 & 0 & 0.25 & 0 & 0 & 0.5 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

注意到$\gamma=1$时$I-\gamma P^{\pi}$是奇异阵，上述Bellman期望方程是一个**不定方程**。但可以直接由定义得到$v_{\pi}(0)=E[0+0+0+\ldots+|S_t=0]=0=v_{\pi}(8)$. 代入原期望方程，解得

$$v^{\pi}=\begin{pmatrix}0&-7&-9\\-7&-8&-7\\-9&-7&0\end{pmatrix}$$

因此由$q_{\pi}(s,a)=r_s^a+\gamma\sum_{s'\subset S}p_{ss'}^a v_{\pi}(s')$，可得：

$$q_{\pi}(4,left)=-1+v^{\pi}(3)=-1-7=-8$$
$$q_{\pi}(7,right)=-1+v^{\pi}(8)=-1+0=-1$$

注：由于单步回报可以有不同理解，如果只要智能体的状态不变回报就为0，（例如$r_3^{left}=0$)则有：

$$r^{\pi}=\begin{pmatrix}0&-0.75&-0.5\\-0.75&-1&-0.75\\-0.5&-0.75&0\end{pmatrix},\ v^{\pi}=\begin{pmatrix}0&-5&-6\\-5&-6&-5\\-6&-5&0\end{pmatrix}$$

5.

（1） 由题意有$r_A=r_B=r_C=-1,r_D=0,\vec{r}=(-1,-1,-1,0)^T$

由Bellman期望方程：

$$v(s)=r_s+\sum_{s'\in S}p_{ss'}v(s'),\vec{v}=\vec{r}+\gamma P\vec{v}$$

转移概率矩阵为

$$P=\begin{pmatrix}0&0.5&0.5&0\\0&0.5&0&0.5\\0.5&0&0&0.5\\0&0&0&1\end{pmatrix}$$

解得：

$$\vec{v}=(I-\gamma P)^{-1}\vec{r}=(-1.6889,-1.3333,-1.4222,0)^T$$

即$v(A)=-1.6889,v(B)=-1.3333,v(C)=-1.4222,v(D)=0$

（2）策略评价问题，即求解稀疏线性方程组$\vec{v}=\vec{r}+\gamma P\vec{v}$. 可使用动态规划，即以Bellman方程从任意初值开始迭代，直至收敛到状态价值。具体如下：

设状态空间$S=s_1,s_2,\ldots,s_N$大小为N，给定收敛阈值θ，θ越小解越精确

1）对$\vec{v} = (v(s_1), v(s_2), ..., v(s_N))$设置任意初值：如$\vec{v}^{(0)} = (0, 0, ..., 0)$

2）迭代：$\vec{v}^{(k+1)} = \vec{r} + \gamma P \vec{v}^{(k)}$（可以使用同步迭代或异步迭代）

3）收敛判据：$\Delta = \max_{s \in S}\{|v^{(k+1)}(s) - v^k(s)|\} < \theta$

当然本质目标是解线性方程组$\vec{v} = \vec{r} + \gamma P \vec{v}$，因此也可以使用其他的迭代公式，如Jacobi迭代法、Gauss-Siedel迭代法或SOR迭代法。