1.

(1): $v_\pi(s) \le q_\pi(s, \pi'(s))$

$= E[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s)]$

$\le E_{\pi'}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s]$

$= E_{\pi'}[R_{t+1} + \gamma E_{\pi'}[R_{t+2} + \gamma v_\pi(S_{t+2}) \mid S_{t+1}, A_t = \pi'(s+1)] \mid S_t = s]$

$= E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) \mid S_t = s]$

$\le E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+3}) \mid S_t = s]$

$\vdots$

$\le E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s]$

$= E_{\pi'}[G_t \mid S_t = s]$

$= v_{\pi'}(s)$

(2) 由 $\pi'(a|s)$ 知, $\pi'(s) = \underset{a \in A}{\arg\max}\, q_\pi(s, a)$

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

则: $v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s, a) \le \sum_{a \in A} \pi(a|s) q_\pi(s, \pi'(s)) = q_\pi(s, \pi'(s)) \sum_{a \in A} \pi(a|s)$

故, $v_\pi(s) \le q_\pi(s, \pi'(s))$.

故由(1), $v_\pi(s) \le v_{\pi'}(s)$.

(3):

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

$$= \frac{\varepsilon}{|A|}\sum_{a\in A} q_\pi(s,a) + (1-\varepsilon)\max_{a\in A} q_\pi(s,a)$$

$$\geq \frac{\varepsilon}{|A|}\sum_{a\in A} q_\pi(s,a) + (1-\varepsilon)\sum_{a\in A}\frac{\pi(a|s)-\frac{\varepsilon}{|A|}}{1-\varepsilon}\cdot q_\pi(s,a)$$

$$= \frac{\varepsilon}{|A|}\sum_{a\in A} q_\pi(s,a) - \frac{\varepsilon}{|A|}\sum_{a\in A} q_\pi(s,a) + \sum_{a\in A}\pi(a|s) q_\pi(s,a)$$

$$= v_\pi(s).$$

故由(1)，$v_{\pi'}(s) \geq v_\pi(s)$.

4.

(1): $4\to 1$: 则: $V(4) = V(4) + \alpha\cdot(-1 + \gamma\cdot V(1) - V(4))$

$$\therefore V(4) = -\frac{1}{2}$$

$1\to 4$: $V(1) = V(1) + \alpha\cdot(-1 + \gamma\cdot V(4) - V(1))$

$$\therefore V(1) = -\frac{3}{4}$$

$4\to 7$. $V(4) = V(4) + \alpha\cdot(-1 + \gamma\cdot V(7) - V(4))$

$$\therefore V(4) = -\frac{3}{4}$$

$7\to$ terminal:

$$V(7) = V(7) + \alpha\cdot(-1 + \gamma\cdot V(\text{terminal}) - V(7))$$

$$\therefore V(7) = -\frac{1}{2}$$

$$\therefore V = \begin{bmatrix} 0 & -\frac{3}{4} & 0 \\ 0 & -\frac{3}{4} & 0 \\ 0 & -\frac{1}{2} & 0 \end{bmatrix}$$

(2):

初始为4，则由Q表知，下一步行动为"下"：

$R=-1$. $S'=7$. 则：$A'=$"左"

$Q(4,\downarrow)=Q(4,\downarrow)+\alpha[-1+\gamma Q(7,\leftarrow)-Q(4,\downarrow)]$

$\therefore Q(4,\downarrow)=-3$

令 $S\leftarrow 7$. $A\leftarrow$"左"

则：$R=-1$. $S'=6$. 则 $A'=$"上"

$Q(7,\leftarrow)=Q(7,\leftarrow)+\alpha[-1+\gamma Q(6,\uparrow)-Q(7,\leftarrow)]$

$\therefore Q(7,\leftarrow)=-3$

令 $S\leftarrow 6$. $A\leftarrow$"上"；

则. $R=-1$. $S'=3$. 则 $A'=$"上"

$Q(6,\uparrow)=Q(6,\uparrow)+\alpha[-1+\gamma Q(3,\uparrow)-Q(6,\uparrow)]$

$\therefore Q(6,\uparrow)=-2$

令 $S\leftarrow 3$, $A\leftarrow$"上"；

则 $R=-1$. $S'=$terminal.

$Q(3,\uparrow)=Q(3,\uparrow)+\alpha[-1+0-Q(3,\uparrow)]$

$\therefore Q(3,\uparrow)=-1$

$S\leftarrow$terminal. 达到终态，

此时Q表：

-4	-3	-1	-3	-4	-2	-4
-3	-3	-2	-4	-2	-3	-3
-4	-3	-4	-3	-2	-3	-4
-3	-2	-3	-3	-4	-3	-3

5,

(1) 状态: $S=\{$西双, 普洱, 墨江, 元江, 石屏, 峨山, 玉溪, 晋宁$\}$.
（依次标为 b c d e f g h i）

行动: $A=\{$向上, 向下$\}$

(2) 转移矩阵

	b	c	d	e	f	g	h	i
b	1							
c	0.5		0.5					
d		0.5		0.5				
e			0.5		0.5			
f				0.5		0.5		
g					0.5		0.5	
h						0.5		0.5
i								1

(3) $v_\pi(c)=\frac{1}{2}\cdot(0+v_\pi(b))+\frac{1}{2}(0+v_\pi(d))$

$v_\pi(d)=\frac{1}{2}(0+v_\pi(c))+\frac{1}{2}(0+v_\pi(e))$

$v_\pi(e)=\frac{1}{2}(0+v_\pi(f))+\frac{1}{2}(0+v_\pi(d))$

$v_\pi(f)=\frac{1}{2}(0+v_\pi(g))+\frac{1}{2}(0+v_\pi(e))$

$v_\pi(g)=\frac{1}{2}(0+v_\pi(h))+\frac{1}{2}(0+v_\pi(f))$

$v_\pi(h)=\frac{1}{2}(100+v_\pi(i))+\frac{1}{2}(0+v_\pi(g))$.

b、i为终止态, 则 $v_\pi(b)=v_\pi(i)=0$

故由上述方程组解得:

$V_\pi(c)=100$. $V_\pi(d)=200$.

$V_\pi(e)=300$. $V_\pi(f)=400$

$V_\pi(g)=500$、$V_\pi(h)=600$

(4): 由于b和i为终止点, 故在这2个状态不会有后续行动、

则: $q_\pi(c,\uparrow)=0+1\cdot1\cdot V_\pi(d)=200$.

$q_\pi(c,\downarrow)=0+1\cdot1\cdot V_\pi(b)=0$

$q_\pi(d,\uparrow)=0+V_\pi(e)=300$.

$q_\pi(d,\downarrow)=0+V_\pi(c)=100$.

$q_\pi(e,\uparrow)=0+V_\pi(f)=400$

$q_\pi(e,\downarrow)=0+V_\pi(d)=200$

$q_\pi(f,\uparrow)=0+V_\pi(g)=500$

$q_\pi(f,\downarrow)=0+V_\pi(e)=300$

$q_\pi(g,\uparrow)=0+V_\pi(h)=600$

$q_\pi(g,\downarrow)=0+V_\pi(f)=400$

$q_\pi(h,\uparrow)=700+V_\pi(i)=700$

$q_\pi(h,\downarrow)=0+V_\pi(g)=500$

5. 可以采用SARSA算法来解决这个问题, 令一个学习过程如下:

首先初始化家驿所在地(非终点). 初始化Q(S.A)表格

之后, 对于初始态, 根据Q表以及ε-贪心策略选择

行动A

之后进入循环迭代，

根据A获得单步行动回报R以及下一状态S'.

根据ε-贪心，选择下一动作A'.

之后更新Q(S,A)：

$Q(S,A) \mathrel{+}= \alpha[R+\gamma Q(S',A')-Q(S,A)]$，

更新$S \leftarrow S'$，$A \leftarrow A'$.

直到达到终点.

不断进行上述学习过程，则Q(S,A)表逐渐收敛.

利用最终得到的Q(S,A)表格，若知象棋初始位置，即可由贪心策略给出一条达到目的地的最优方案.

(6) 可以采用动态规划的思想求解.

同时可以利用异步迭代的方法.

对于大象当前的状态s及迭代次数k.

首先进行价值迭代：$v_{k+1}(s) \leftarrow \max_a \left[R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_k(s') \right]$

之后立刻对$v_{k+1}(s)$更新，并用确定性贪心来改进$\pi(s)$.

选择大象每走一步便更新之前的状态价值函数$V(s)$

以及策略$\pi(s)$.

可以较方便地选择何时停止更新，并以此时的$V(s)$近似状态价值，

当前策略$\pi(s)$近似为最优策略，

故可解决状态过多的问题.