7.1

a) $V^\pi = (I - \gamma P^\pi)^{-1} R$

$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $\quad \pi = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$ $\quad \gamma = \frac{3}{4}$

$P^\pi = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix}$ $\quad R = \begin{bmatrix} 9 \\ -6 \\ 3 \end{bmatrix}$

$V^\pi = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{3}{4} \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix} \right)^{-1} \begin{pmatrix} 9 \\ -6 \\ 3 \end{pmatrix}$ $\quad$ * plugged into MATLAB

$\boxed{V^\pi = \begin{pmatrix} 12 \\ 0 \\ 4 \end{pmatrix}}$

b) $\pi'(s) = \underset{a}{\operatorname{argmax}} \left[Q^\pi(s,a) \right]$

$\pi'(1) = \underset{a}{\operatorname{argmax}} \left[Q^\pi(1,a) \right] = \underset{a}{\operatorname{argmax}} \left[R(1) + \gamma \sum_{s'} P(s'|1,1) V^\pi(s') \quad R(1) + \gamma \sum_{s'} P(s'|1,2) V^\pi(s') \right]$

$= \underset{a}{\operatorname{argmax}} \left[\sum_{s'} P(s'|1,1) V^\pi(s') \quad , \quad \sum_{s'} P(s'|1,2) V^\pi(s') \right]$

$= \underset{a}{\operatorname{argmax}} \left[\left(\frac{1}{3}, \frac{2}{3}, 0\right) \cdot (12, 0, 4) \quad , \quad \left(\frac{1}{3}, 0, \frac{2}{3}\right) \cdot (12, 0, 4) \right]$

$= \underset{a}{\operatorname{argmax}} \left[4 \quad , \quad \frac{20}{3} \right] = 2$

$\pi'(2) = \underset{a}{\operatorname{argmax}} \left[\sum_{s'} P(s'|2,1) V^\pi(s') \quad , \quad \sum_{s'} P(s'|2,2) V^\pi(s') \right]$

$= \underset{a}{\operatorname{argmax}} \left[\left(0, \frac{1}{3}, \frac{2}{3}\right) \cdot (12, 0, 4) \quad , \quad \left(\frac{2}{3}, \frac{1}{3}, 0\right) \cdot (12, 0, 4) \right]$

$= \underset{a}{\operatorname{argmax}} \left[\frac{8}{3} \quad , \quad \frac{24}{3} \right] = 2$

$\pi'(3) = \underset{a}{\operatorname{argmax}} \left[\sum_{s'} P(s'|3,1) V^\pi(s') \quad , \quad \sum_{s'} P(s'|3,2) V^\pi(s') \right]$

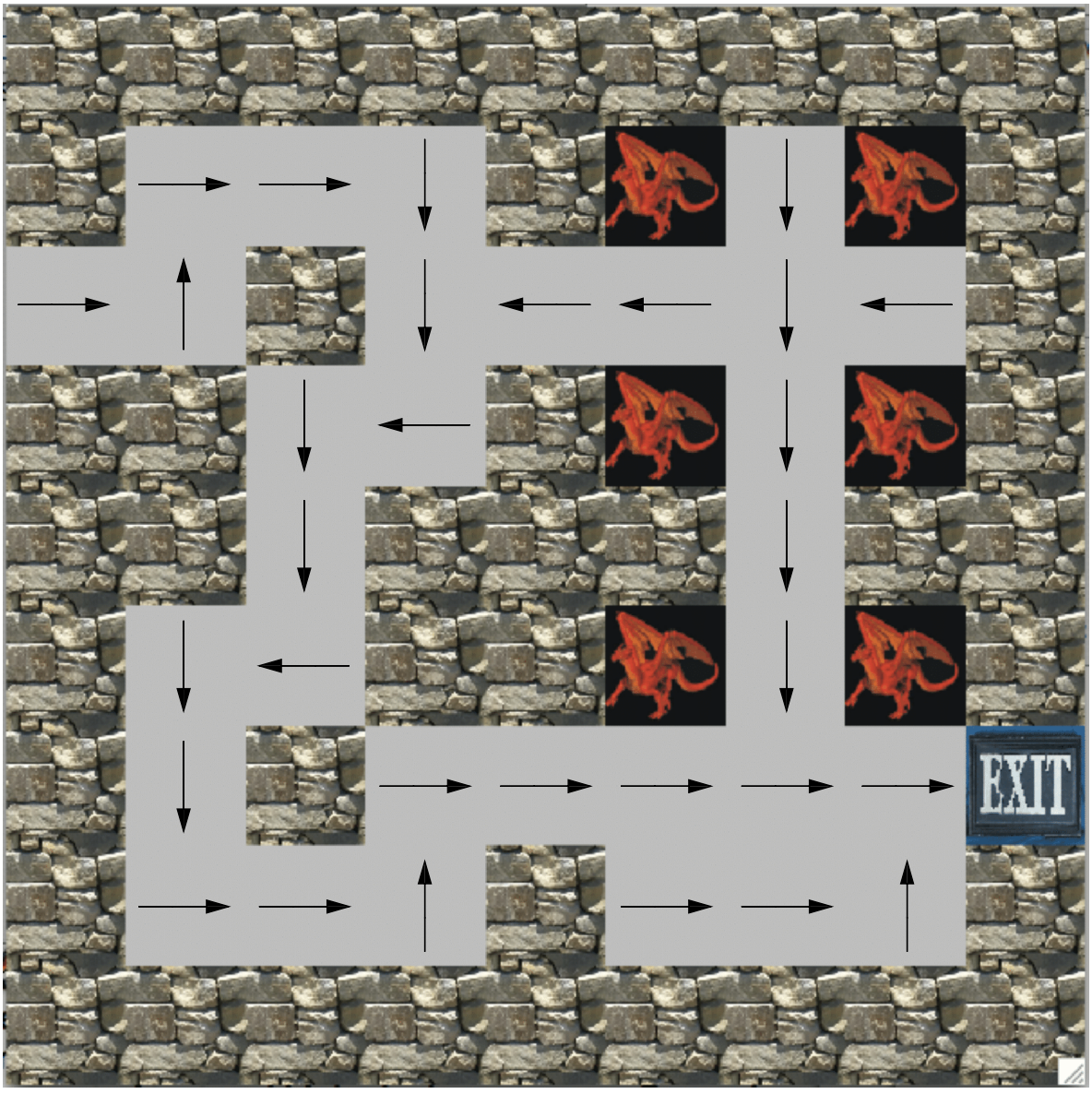
$= \underset{a}{\operatorname{argmax}} \left[\left(\frac{2}{3}, 0, \frac{1}{3}\right) \cdot (12, 0, 4) \quad , \quad \left(0, \frac{2}{3}, \frac{1}{3}\right) \cdot (12, 0, 4) \right]$

$= \underset{a}{\operatorname{argmax}} \left[\frac{28}{3} \quad , \quad \frac{4}{3} \right] = 1$

So $\pi'(1) = 2$
$\pi'(2) = 2$
$\pi'(3) = 1$

```
V(3) = 53.8585
V(11) = 55.3661
V(12) = 54.5973
V(15) = 61.9523
V(16) = 62.8250
V(17) = 63.7101
V(20) = 56.1463
V(22) = 59.4059
V(23) = 60.2428
V(24) = 61.0916
V(26) = 64.6076
V(29) = 56.9374
V(30) = 57.7397
V(31) = 58.5798
V(34) = 66.4406
V(35) = 65.5177
V(39) = 56.7579
V(43) = 67.3765
V(47) = -79.8588
V(48) = 49.2342
V(49) = -79.8588
V(51) = -79.8588
V(52) = 68.3255
V(53) = 71.4890
V(56) = 46.5581
V(57) = 54.0299
V(58) = 55.3042
V(59) = 63.4030
V(60) = 64.4911
V(61) = 73.2074
V(62) = 72.5547
V(65) = -79.8588
V(66) = 46.5581
V(67) = -79.8588
V(69) = -79.8588
V(70) = 74.5962
V(71) = 73.5605
V(79) = 79.8588

The optimal policy computed with both methods are the same!
With an initial policy of move east, the policy iteration converges in 5 iterations
With an initial policy of move west, the policy iteration also converges in 5 iterations
Note, with an initial policy of move south, the policy iteration converges in 14
iterations
>>
```


EXIT

```
% Problem 7.2
clear;clc;close all

load('data.mat')
load('image.mat')
R = rewards;
n = 81;
gamma = 0.9875;
max_num_iter = 1000;
max_num_iter2 = 20;
V = zeros(n,max_num_iter+1);
I = eye(n);

% Construct transition matrices
A = zeros(n,n,4);
for i = 1:size(prob_a1,1)
    A(prob_a1(i,1),prob_a1(i,2),1) = prob_a1(i,3);
end
for i = 1:size(prob_a2,1)
    A(prob_a2(i,1),prob_a2(i,2),2) = prob_a2(i,3);
end
for i = 1:size(prob_a3,1)
    A(prob_a3(i,1),prob_a3(i,2),3) = prob_a3(i,3);
end
for i = 1:size(prob_a4,1)
    A(prob_a4(i,1),prob_a4(i,2),4) = prob_a4(i,3);
end

% Value iteration
best_action = ones(n,1);
for i = 1:max_num_iter+1
    for j = 1:n
        best_term = -Inf;
        best_action(j) = 0;
        for k = 1:4
            term = A(j,:,k)*V(:,i);
            if term > best_term
                best_term = term;
                best_action(j) = k;
            end
        end

        V(j,i+1) = R(j) + gamma*best_term;
    end

    % convergence criterion
    if var(V(:,i+1) - V(:,i)) < 0.000001
        last_iteration = i;
        break;
    end
end
```

```
for j = 1:n
    if V(j,last_iteration+1) ~= 0
        fprintf('V(%d) = %.4f\n',int64(j),V(j,last_iteration+1))
    end
end

% Draw arrows in image
% drawArrow = @(x,y) quiver(x(1), y(1), x(2)-x(1), y(2)-y(1), 0)
drawArrow = @(x,y) line([x(1), y(1)], [x(2), y(2)]) ;
hold on
imshow(maze, colormap, 'InitialMagnification', 50)
box_x_length = size(maze,2)/9;
box_y_length = size(maze,1)/9;
arrow_length = 0.75*box_x_length/2;
for i = 1:n
    if V(i,last_iteration+1) ~= 0 && ~ismember(i,[47 49 51 65 67 69 79])
        center = [box_x_length*ceil(i/9)-box_x_length/2 box_y_length*mod(i-1,9)
+box_y_length/2];
        switch best_action(i)
            case 1 % west
                arrow(center + [arrow_length 0], center - [arrow_length 0])
            case 2 % north
                arrow(center + [0 arrow_length], center - [0 arrow_length])
            case 3 % east
                arrow(center -[arrow_length 0], center + [arrow_length 0])
            case 4 % south
                arrow(center - [0 arrow_length], center + [0 arrow_length])
            otherwise
                disp('mistake')
        end
    end
end

% Policy iteration
V2 = zeros(n,max_num_iter2+1);
policy = zeros(n,max_num_iter2+1);
policy(:,1) = 4*ones(n,1); % east
for i = 1:max_num_iter2+1
    P = zeros(n);
    for j = 1:n
        P(j,:) = A(j,:,policy(j,i));
    end
    V2(:,i+1) = (I - gamma*P)\R;

    for j = 1:n
        best_term = -Inf;
        for k = 1:4
            term = A(j,:,k)*V2(:,i+1);
            if term > best_term
                best_term = term;
```

```
                policy(j,i+1) = k;
            end
        end
    end

    if policy(:,i+1) == policy(:,i)
        last_iteration2 = i;
        break;
    end
end

if all(policy(:,last_iteration2) == best_action)
    disp(' ')
    disp('The optimal policy computed with both methods are the same!')
end

disp('With an initial policy of move east, the policy iteration converges in 5
iterations')
disp('With an initial policy of move west, the policy iteration also converges in 5
iterations')
disp('Note, with an initial policy of move south, the policy iteration converges in 11
iterations')
```

7.3

$$\sum_{n \geq t} \gamma^n r_n \leq \frac{1}{1-\gamma} e^{-t(1-\gamma)}$$ * suppose the worst case, where $r_n = 1 \ \forall n \geq t$

$$\sum_{n \geq t} \gamma^n \leq \frac{1}{1-\gamma} e^{-t(1-\gamma)}$$

$$\sum_{n \geq t} n \log \gamma \leq \log\left(\frac{1}{1-\gamma}\right) - t(1-\gamma)$$ * take log of both sides

* apply $\log \gamma \in \gamma - 1$

$$\sum_{n \geq t} n(\gamma - 1) \leq \log\left(\frac{1}{1-\gamma}\right) - t(1-\gamma)$$

$$\sum_{n \geq t} n(\gamma - 1) - t(\gamma - 1) \leq \log\left(\frac{1}{1-\gamma}\right)$$

$$\sum_{n \geq t+1} n(\gamma - 1) \leq -\log(1-\gamma)$$ * simplify

$$\sum_{n \geq t+1} n(\gamma - 1) \leq -(1 - \gamma - 1)$$ * apply $\log \gamma \leq \gamma - 1$ again

$$\sum_{n \geq t+1} n(\gamma - 1) \leq \gamma$$

* $\gamma \in [0, 1)$, so $\gamma - 1 < 0$, therefore LHS is negative and RHS is positive!

74

$V_{k+1}(s) \leftarrow R(s) + \gamma \sum_{s'} P(s'|s, \pi(s)) V_k(s')$

$\Delta_k = \max_s |V_k(s) - V^\pi(s)|$

We see that as $k \to \infty$,

$V_{k+1} = \left(\sum_{i=0}^{k+1} (\gamma P)^i + (\gamma P)^{i}\right) \vec{R} = V_k$ because $\gamma < 1$,

So $\lim_{k \to \infty} V_k = V^\pi$. Given that,

$V_k - V^\pi = V_k - \lim_{k \to \infty} V_k$

$= \left(\sum_{k+1}^{\infty} (\gamma P)^i\right) \vec{R}$

So, $\Delta_k = \max_s |V_k(s) - V^\pi(s)| \leq \left(\sum_{k+1}^{\infty} (\gamma P)^i\right) \vec{R}$

$V_0 = \vec{0}$

$V_1 = \vec{R}$

$V_2 = \vec{R} + \gamma P V_1 = \vec{R} + \gamma P \vec{R} = (I + \gamma P) \vec{R}$

$V_3 = \vec{R} + \gamma P V_2 = \vec{R} + \gamma P (I + \gamma P) \vec{R} = (I + \gamma P + \gamma$

$V_k = \left(\sum_{i=0}^{k} (\gamma P)^i\right) \vec{R}$

7.5

a) $V^{\pi}(s) = R(s) + \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi}(s')$

$$\boxed{V^{\pi}(s) = s + \gamma\left(\frac{3}{5}V^{\pi}(s) + \frac{2}{5}V^{\pi}(s+1)\right)}$$

b) $V^{\pi}(s) = as + b$ *substitute into result from (a)

$$\Downarrow$$

$$as + b = s + \gamma\left(\frac{3}{5}(as+b) + \frac{2}{5}(a(s+1)+b)\right)$$

$$as + b = s + \frac{3}{5}\gamma as + \frac{3}{5}\gamma b + \frac{2}{5}as + \frac{2}{5}a + \frac{2}{5}b$$

$$as + b = \left(1 + \frac{3}{5}\gamma a + \frac{2}{5}a\right)s + \left(\frac{3}{5}\gamma b + \frac{2}{5}a + \frac{2}{5}b\right)$$

$$\Downarrow$$

$$a = 1 + \frac{3}{5}\gamma a + \frac{2}{5}a$$

$$a - \frac{3}{5}\gamma a - \frac{2}{5}a = 1$$

$$a\left(1 - \frac{3}{5}\gamma - \frac{2}{5}\right) = 1$$

$$a = \frac{1}{\frac{3}{5} - \frac{3}{5}\gamma}$$

$$\boxed{a = \frac{5}{3 - 3\gamma}}$$

$$b = \frac{3}{5}\gamma b + \frac{2}{5}a + \frac{2}{5}b$$

$$b = \frac{3}{5}\gamma b + \frac{2}{5}\left(\frac{5}{3-3\gamma}\right) + \frac{2}{5}b$$

$$b - \frac{3}{5}\gamma b - \frac{2}{5}b = \frac{2}{3-3\gamma}$$

$$\left(\frac{3}{5} - \frac{3}{5}\gamma\right)b = \frac{2}{3-3\gamma}$$

$$b = \frac{2}{(3-3\gamma)\left(\frac{3}{5} - \frac{3}{5}\gamma\right)}$$

$$\boxed{b = \frac{10}{(3-3\gamma)^2}}$$

a) i) $\sum_{k=1}^{\infty} \alpha_k = \sum_{k=1}^{\infty} \frac{1}{k}$

* Using the integral test, we know $\int_1^{\infty} \frac{1}{k}\,dk = \left[\log k\right]_1^{\infty} = \log(\infty) - \log(1) = \infty$

$$\sum_{k=1}^{\infty} \frac{1}{k} = \infty$$

* this implies that the original sum diverges

ii) $\sum_{k=1}^{\infty} \left(\frac{1}{k}\right)^2 = \sum_{k=1}^{\infty} \frac{1}{k^2}$

* again using the integral test, we know $\int_1^{\infty} \frac{1}{k^2}\,dk = \left[-\frac{1}{2k}\right]_1^{\infty} = 0 + \frac{1}{2} = \frac{1}{2}$

$$\sum_{k=1}^{\infty} \frac{1}{k^2} < \infty$$

* this implies that the original sum converges

b) $\mu_k \leftarrow \mu_{k-1} + \alpha_k (x_k - \mu_{k-1})$

$\mu_k \leftarrow (1-\alpha_k)\mu_{k-1} + \alpha_k x_k$

$\mu_k \leftarrow \left(\frac{k-1}{k}\right)\mu_{k-1} + \left(\frac{1}{k}\right)x_k$

$\mu_{k-1} \leftarrow \left(\frac{k-2}{k-1}\right)\mu_{k-2} + \left(\frac{1}{k-1}\right)x_{k-1}$

$\Downarrow$

$\mu_k \leftarrow \left(\frac{k-1}{k}\right)\left(\left(\frac{k-2}{k-1}\right)\mu_{k-2} + \left(\frac{1}{k-1}\right)x_{k-1}\right) + \frac{1}{k}x_k$

$$\mu_k \leftarrow \left(\frac{k-2}{k}\right)\mu_{k-2} + \frac{x_{k-1}}{k} + \frac{x_k}{k}$$

* we can continue this process until $k=0$, we obtain

$$\mu_k \leftarrow \frac{1}{k}\left(x_1 + \ldots + x_k\right)$$

sample averages

$\mu_k = \frac{1}{k}\left(x_1 + \ldots + x_k\right)$

μ_k