SUPPLEMENT 1 TO CHAPTER 19
A POLICY IMPROVEMENT ALGORITHM FOR FINDING OPTIMAL POLICIES

19S1-1.

Number of states: 3

Number of decisions: 2

Cost Matrix, C_{ik}: $\begin{bmatrix} 0 & 0 \\ -27 & -31 \\ -27 & -31 \end{bmatrix}$

$p_{ij}(1) = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0.6 & 0.4 \end{bmatrix}$ $\quad p_{ij}(2) = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.4 & 0.5 & 0.1 \\ 0 & 0.8 & 0.2 \end{bmatrix}$

Initial Policy:

$d_0(R_1) = 1$
$d_1(R_1) = 1$
$d_2(R_1) = 2$ $\quad$ Discount Factor = 1

Iteration # 1

Value Determination:

$$g(R_1) = 0 + 0.5v_0(R_1) + 0.5v_1(R_1) + 0v_2(R_1) - v_0(R_1)$$
$$g(R_1) = -27 + 0.3v_0(R_1) + 0.5v_1(R_1) + 0.2v_2(R_1) - v_1(R_1)$$
$$g(R_1) = -31 + 0v_0(R_1) + 0.8v_1(R_1) + 0.2v_2(R_1) - v_2(R_1)$$

Solution of Value Determination Equations:

$g(R_1) = -18.8$
$v_0(R_1) = 52.84$
$v_1(R_1) = 15.27$
$v_2(R_1) = 0$

Policy Improvement:

State 0:
$0 + 0.5(52.84) + 0.5(15.27) + (0) - (52.84) = -18.8$
$0 + 0.5(52.84) + 0.5(15.27) + (0) - (52.84) = -18.8$

State 1:
$-27 + 0.3(52.84) + 0.5(15.27) + (0) - (15.27) = -18.8$
$-31 + 0.4(52.84) + 0.5(15.27) + (0) - (15.27) = -17.5$

State 2:
$-27 + 0(52.84) + 0.6(15.27) + (0) - (0) = -17.8$
$-31 + 0(52.84) + 0.8(15.27) + (0) - (0) = -18.8$

Optimal Policy:
$d_0(R_2) = 1$
$d_1(R_2) = 1$
$d_2(R_2) = 2$

$g(R_2) = -18.8$
$v_0(R_2) = 52.84$
$v_1(R_2) = 15.27$
$v_2(R_2) = 0$

19S1-2.

Number of states = 2

Number of decisions = 5

Cost Matrix, C(ik):

$$\begin{bmatrix} 0 & 4.5 & 5 & ---- & ---- \\ ---- & ---- & ---- & 50 & 9 \end{bmatrix}$$

Transition Matrix, p(ij)[1]:

$$\begin{bmatrix} 0.9 & 0.1 \\ 0 & 0 \end{bmatrix}$$

Transition Matrix, p(ij)[2]:

$$\begin{bmatrix} 0.98 & 0.02 \\ 0 & 0 \end{bmatrix}$$

Transition Matrix, p(ij)[3]:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Transition Matrix, p(ij)[4]:

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

Transition Matrix, p(ij)[5]:

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Initial Policy:

d0(R1) = 1
d1(R1) = 4

Discount Factor = 1

ITERATION # 1

Value Determination:

```
g(R1) = 0    +  0.9v0(R1) +   0.1v1(R1) - v0(R1)
g(R1) = 50   +    1v0(R1) +     0v1(R1) - v1(R1)
```

Solution of Value Determination Equations:

```
g(R1)  = 4.545
v0(R1) = -45.5
v1(R1) = 0
```

Policy Improvement:

State 0:

```
0    + 0.9  (-45.5) + 0.1  (0) - (-45.5) = 4.545
4.5  + 0.98 (-45.5) + 0.02 (0) - (-45.5) = 5.409
5    + 1    (-45.5) + 0    (0) - (-45.5) = 5
---  + 0    (-45.5) + 0    (0) - (-45.5) = ---
---  + 0    (-45.5) + 0    (0) - (-45.5) = ---
```

State 1:

```
---  + 0    (-45.5) + 0    (0) - (0) = ---
---  + 0    (-45.5) + 0    (0) - (0) = ---
---  + 0    (-45.5) + 0    (0) - (0) = ---
50   + 1    (-45.5) + 0    (0) - (0) = 4.545
9    + 0    (-45.5) + 1    (0) - (0) = 9
```

```
Optimal Policy: g(R1)  = 4.545
  d0(R2) = 1    v0(R1) = -45.5
  d1(R2) = 4    v1(R1) = 0
```

19S1-3.

```
Number of states = 2            Cost Matrix, C(ik):
                                 |‾ 0      14   ‾|
Number of decisions = 2          |_ 75     14   _|

Transition Matrix, p(ij)[1]:    Transition Matrix, p(ij)[2]:
|‾ 0.125  0.875 ‾|              |‾ 0.875  0.125 ‾|
|_ 0.125  0.875 _|              |_ 0.875  0.125 _|


Initial Policy:

d0(R1) = 1
d1(R1) = 2

Discount Factor = 1

ITERATION # 1

Value Determination:

g(R1) = 0    +0.125v0(R1) + 0.875v1(R1) - v0(R1)
g(R1) = 14   +0.875v0(R1) + 0.125v1(R1) - v1(R1)

Solution of Value Determination Equations:
g(R1)  = 7
v0(R1) = -8
v1(R1) = 0

Policy Improvement:

State 0:
0    + 0.125(   -8) + 0.875(0) - (-8   ) = 7
14   + 0.875(   -8) + 0.125(0) - (-8   ) = 15

State 1:
75   + 0.125(   -8) + 0.875(0) - (0) = 74
14   + 0.875(   -8) + 0.125(0) - (0) = 7

Optimal Policy: g(R1)  = 7
  d0(R2) = 1    v0(R1) = -8
  d1(R2) = 2    v1(R1) = 0
```

19S1-4.

Number of states: 2

Number of decisions: 2

Cost Matrix, C_{ik}: $\begin{bmatrix} -0.12 & 0.292 \\ 0.5 & 0.417 \end{bmatrix}$

$p_{ij}(1) = \begin{bmatrix} 0.375 & 0.625 \\ 1 & 0 \end{bmatrix}$ $p_{ij}(2) = \begin{bmatrix} 0.875 & 0.125 \\ 1 & 0 \end{bmatrix}$

Initial Policy:

$d_0(R_1) = 1$
$d_1(R_1) = 1$

Iteration # 1

Value Determination:

$g(R_1) = -0.12+0.375v_0(R_1) + 0.625v_1(R_1) - v_0(R_1)$
$g(R_1) = 0.5 + 1v_0(R_1) + 0v_1(R_1) - v_1(R_1)$

Solution of Value Determination Equations:

$g(R_1) = 0.115$
$v_0(R_1) = -0.38$
$v_1(R_1) = 0$

Policy Improvement:

State 0:
$-0.12+ 0.375(-0.38) +(0) - (-0.38) = 0.115$
$0.292+ 0.875(-0.38) +(0) - (-0.38) = 0.34$

State 1:
$0.5 + 1(-0.38) +(0) - (0) = 0.115$
$0.417+ 1(-0.38) +(0) - (0) = 0.032$

New Policy:

$d_0(R_2) = 1$
$d_1(R_2) = 2$

Iteration # 2

Value Determination:

$g(R_2) = -0.12+0.375v_0(R_2) + 0.625v_1(R_2) - v_0(R_2)$
$g(R_2) = 0.417+ 1v_0(R_2) + 0v_1(R_2) - v_1(R_2)$

Solution of Value Determination Equations:

$g(R_2) = 0.083$
$v_0(R_2) = -0.33$
$v_1(R_2) = 0$

Policy Improvement:

State 0:
$-0.12+ 0.375(-0.33) +(0) - (-0.33) = 0.083$
$0.292+ 0.875(-0.33) +(0) - (-0.33) = 0.333$

State 1:
$0.5 + 1(-0.33) +(0) - (0) = 0.167$
$0.417+ 1(-0.33) +(0) - (0) = 0.083$

New Policy:

$d_0(R_3) = 2$
$d_1(R_3) = 1$

Iteration # 3

Value Determination:

$$g(R_3) = -0.12 + 0.375 v_0(R_3) + 0.625 v_1(R_3) - v_0(R_3)$$
$$g(R_3) = 0.5 + 1 v_0(R_3) + 1 v_1(R_3) - v_1(R_3)$$

Solution of Value Determination Equations:

$g(R_3) = 0.115$
$v_0(R_3) = -0.38$
$v_1(R_3) = 0$

Policy Improvement:

State 0:
$-0.12 + 0.375(-0.38) + (0) - (-0.38) = 0.115$
$0.292 + 0.875(-0.38) + (0) - (-0.38) = 0.34$

State 1:
$0.5 + 1(-0.38) + (0) - (0) = 0.115$
$0.417 + 1(-0.38) + (0) - (0) = 0.032$

Optimal Policy: $g(R_4) = 0.115$
$d_0(R_4) = 2$ $v_0(R_4) = -0.38$
$d_1(R_4) = 1$ $v_1(R_4) = 0$

19S1-5.

Initial Policy:

d0(R1) = 1
d1(R1) = 1
d2(R1) = 1

ITERATION # 1

Value Determination:

$$g(R1) = -22 + 0.4v0(R1) + 0.4v1(R1) + 0.2v2(R1) - v0(R1)$$
$$g(R1) = -10.5 + 0.3v0(R1) + 0.4v1(R1) + 0.3v2(R1) - v1(R1)$$
$$g(R1) = 16 + 0.1v0(R1) + 0.4v1(R1) + 0.5v2(R1) - v2(R1)$$

Solution of Value Determination Equations:

g(R1) = -4.37
v0(R1) = -54.3
v1(R1) = -37.4
v2(R1) = 0

```
Policy Improvement:

State 0:
-22   + 0.4  (-54.3) + 0.4  (-37.4) + 0.2  (0) - (-54.3) = -4.37
-9    + 0.4  (-54.3) + 0.4  (-37.4) + 0.2  (0) - (-54.3) = 8.629

State 1:
-10.5+ 0.3  (-54.3) + 0.4  (-37.4) + 0.3  (0) - (-37.4) = -4.37
-4.5 + 0.3  (-54.3) + 0.4  (-37.4) + 0.3  (0) - (-37.4) = 1.629

State 2:
16   + 0.1  (-54.3) + 0.4  (-37.4) + 0.5  (0) - (0) = -4.37
6.5  + 0.1  (-54.3) + 0.4  (-37.4) + 0.5  (0) - (0) = -13.9

New Policy:
  d0(R2) = 1
  d1(R2) = 1
  d2(R2) = 2

ITERATION # 2

Value Determination:

g(R2) = -22   +  0.4v0(R2) +   0.4v1(R2) +   0.2v2(R2) - v0(R2)
g(R2) = -10.5+  0.3v0(R2) +   0.4v1(R2) +   0.3v2(R2) - v1(R2)
g(R2) = 6.5   +  0.1v0(R2) +   0.4v1(R2) +   0.5v2(R2) - v2(R2)

Solution of Value Determination Equations:
g(R2)  = -7.63
v0(R2) = -40.7
v1(R2) = -25.1
v2(R2) = 0

Policy Improvement:

State 0:
-22   + 0.4  (-40.7) + 0.4  (-25.1) + 0.2  (0) - (-40.7) = -7.63
-9    + 0.4  (-40.7) + 0.4  (-25.1) + 0.2  (0) - (-40.7) = 5.371

State 1:
-10.5+ 0.3  (-40.7) + 0.4  (-25.1) + 0.3  (0) - (-25.1) = -7.63
-4.5 + 0.3  (-40.7) + 0.4  (-25.1) + 0.3  (0) - (-25.1) = -1.63

State 2:
16   + 0.1  (-40.7) + 0.4  (-25.1) + 0.5  (0) - (0) = 1.871
6.5  + 0.1  (-40.7) + 0.4  (-25.1) + 0.5  (0) - (0) = -7.63

Optimal Policy:
  d0(R3) = 1
  d1(R3) = 1
  d2(R3) = 2
```

19S1-6.

```
Number of states = 2                              Cost Matrix, C(ik):

Number of decisions = 2                           |‾ 0        0      ‾|
                                                  |_ -1200   -1200   _|

Transition Matrix, p(ij)[1]:                      Transition Matrix, p(ij)[2]:

|‾ 0.4     0.6    ‾|                              |‾ 0.5     0.5    ‾|
|_ 0.6     0.4    _|                              |_ 0.4     0.6    _|

 Initial Policy:

 d0(R1) = 1                   Discount Factor = 1
 d1(R1) = 1
ITERATION # 1

Value Determination:

g(R1) = 0     +  0.4v0(R1) +   0.6v1(R1) - v0(R1)
g(R1) = -1200+  0.6v0(R1) +   0.4v1(R1) - v1(R1)

Solution of Value Determination Equations:
g(R1)  = -600
v0(R1) = 1000
v1(R1) = 0

Policy Improvement:

State 0:
0     + 0.4  ( 1000) + 0.6  (0) - (1000 ) = -600
0     + 0.5  ( 1000) + 0.5  (0) - (1000 ) = -500

State 1:
-1200+ 0.6  ( 1000) + 0.4  (0) - (0) = -600
-1200+ 0.4  ( 1000) + 0.6  (0) - (0) = -800

New Policy:
  d0(R2) = 1
  d1(R2) = 2

ITERATION # 2

Value Determination:

g(R2) = 0     +  0.4v0(R2) +   0.6v1(R2) - v0(R2)
g(R2) = -1200+  0.4v0(R2) +   0.6v1(R2) - v1(R2)

Solution of Value Determination Equations:
g(R2)  = -720
v0(R2) = 1200
v1(R2) = 0

Policy Improvement:

State 0:
0     + 0.4  ( 1200) + 0.6  (0) - (1200 ) = -720
0     + 0.5  ( 1200) + 0.5  (0) - (1200 ) = -600

-1200+ 0.6  ( 1200) + 0.4  (0) - (0) = -480
-1200+ 0.4  ( 1200) + 0.6  (0) - (0) = -720

Optimal Policy: g(R2)  = -720
  d0(R3) = 1    v0(R2) = 1200
  d1(R3) = 2    v1(R2) = 0
```

19S1-7.

```
Markovian Decision Processes Model:

Number of states = 3              Cost Matrix, C(ik):

Number of decisions = 3           |‾ 13.33  18.67  24    ‾|
                                  |  4      19     ----   |
                                  |_ 4      ----   ----  _|

Transition Matrix, p(ij)[1]:   Transition Matrix, p(ij)[2]:   Transition Matrix, p(ij)[3]:

|‾ 1      0      0     ‾|      |‾ 0.667  0.333  0     ‾|      |‾ 0.333  0.333  0.333 ‾|
|  0.667  0.333  0      |      |  0.333  0.333  0.333  |      |  0      0      0      |
|_ 0.333  0.333  0.333 _|      |_ 0      0      0     _|      |_ 0      0      0     _|

 Initial Policy:

 d0(R1) = 3
 d1(R1) = 1                    Discount Factor = 1
 d2(R1) = 1

Average Cost Policy Improvement Algorithm:

ITERATION # 1

Value Determination:

g(R1) = 24    +0.333v0(R1) + 0.333v1(R1) + 0.333v2(R1) - v0(R1)
g(R1) = 4     +0.667v0(R1) + 0.333v1(R1) +     0v2(R1) - v1(R1)
g(R1) = 4     +0.333v0(R1) + 0.333v1(R1) + 0.333v2(R1) - v2(R1)

Solution of Value Determination Equations:
g(R1)  = 12.89
v0(R1) = 20
v1(R1) = 6.667
v2(R1) = 0

Policy Improvement:

State 0:
13.33+ 1     (   20) + 0     (6.667) + 0     (0) - (20    ) = 13.33
18.67+ 0.667(    20) + 0.333(6.667) + 0     (0) - (20    ) = 14.22
24    + 0.333(   20) + 0.333(6.667) + 0.333(0) - (20    ) = 12.89

State 1:
4     + 0.667(   20) + 0.333(6.667) + 0     (0) - (6.667) = 12.89
19    + 0.333(   20) + 0.333(6.667) + 0.333(0) - (6.667) = 21.22
---   + 0     (   20) + 0     (6.667) + 0     (0) - (6.667) = ---

State 2:
4     + 0.333(   20) + 0.333(6.667) + 0.333(0) - (0) = 12.89
---   + 0     (   20) + 0     (6.667) + 0     (0) - (0) = ---
---   + 0     (   20) + 0     (6.667) + 0     (0) - (0) = ---

New Policy:
  d0(R2) = 3
  d1(R2) = 1
  d2(R2) = 1
```

19S1-8.

When the number of pints of blood delivered can be specified at the time of delivery, the starting number of pints including the delivery will never exceed the largest possible demand in a period, so we can restrict our attention to states $i = 0, 1, 2, 3$. The admissible actions in state i are to order $0 \leq k \leq 3 - i$. Given a decision k, the transition probabilities and the immediate cost are computed as follows:

$$p_{ij}(k) = P\{D = i + k - j\} \text{ if } j \geq 1$$

$$p_{i0}(k) = P\{D \geq i + k\}$$

$$C_{ik} = 50k + E[100(i + k - D)^+].$$

Initialization: $d_i(R_1) = 1$ for $i = 0, 1, 2$ and $d_3(R_1) = 0$

$$\mathrm{P}(R_1) = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 \\ 0.3 & 0.3 & 0.4 & 0 \\ 0.1 & 0.2 & 0.3 & 0.4 \\ 0.1 & 0.2 & 0.3 & 0.4 \end{pmatrix} \quad C(R_1) = \begin{pmatrix} 90 \\ 60 \\ 50 \\ 0 \end{pmatrix}$$

<u>Iteration 1:</u>

Step 1: Value determination:

$$g(R_1) = 90 + 0.6v_0(R_1) + 0.4v_1(R_1) - v_0(R_1)$$

$$g(R_1) = 60 + 0.3v_0(R_1) + 0.3v_1(R_1) + 0.4v_2(R_1) - v_1(R_1)$$

$$g(R_1) = 50 + 0.1v_0(R_1) + 0.2v_1(R_1) + 0.3v_2(R_1) + 0.4v_3(R_1) - v_2(R_1)$$

$$g(R_1) = \ \ 0 + 0.1v_0(R_1) + 0.2v_1(R_1) + 0.3v_2(R_1) + 0.4v_3(R_1) - v_3(R_1)$$

$$v_3(R_1) = 0$$

$$\Rightarrow g(R_1) = 57.8, v_0(R_1) = 196.3, v_1(R_1) = 115.9, v_2(R_1) = 50, v_3(R_1) = 0$$

Step 2: Policy improvement:

$$\text{minimize} \begin{pmatrix} 100 + v_0(R_1) - v_0(R_1) = 100 \\ 90 + 0.6v_0(R_1) + 0.4v_1(R_1) - v_0(R_1) = 57.8 \\ 110 + 0.3v_0(R_1) + 0.3v_1(R_1) + 0.4v_2(R_1) - v_0(R_1) = 27.36 \\ 150 + 0.1v_0(R_1) + 0.2v_1(R_1) + 0.3v_2(R_1) + 0.4v_3(R_1) - v_0(R_1) = \mathbf{11.51} \end{pmatrix}$$

$$\Rightarrow d_0(R_2) = 3$$

$$\text{minimize} \begin{pmatrix} 40 + 0.6v_0(R_1) + 0.4v_1(R_1) - v_1(R_1) = 88.24 \\ 60 + 0.3v_0(R_1) + 0.3v_1(R_1) + 0.4v_2(R_1) - v_1(R_1) = 57.8 \\ 100 + 0.1v_0(R_1) + 0.2v_1(R_1) + 0.3v_2(R_1) + 0.4v_3(R_1) - v_1(R_1) = \mathbf{41.91} \end{pmatrix}$$

$$\Rightarrow d_1(R_2) = 2$$

$$\text{minimize} \begin{pmatrix} 10 + 0.3v_0(R_1) + 0.3v_1(R_1) + 0.4v_2(R_1) - v_2(R_1) = 73.66 \\ 50 + 0.1v_0(R_1) + 0.2v_1(R_1) + 0.3v_2(R_1) + 0.4v_3(R_1) - v_2(R_1) = \mathbf{57.8} \end{pmatrix}$$

$$\Rightarrow d_2(R_2) = 1$$

R_2 is not identical to R_1, so optimality test fails.

Iteration 2:

Step 1: Value determination:

$$g(R_2) = 150 + 0.1v_0(R_2) + 0.2v_1(R_2) + 0.3v_2(R_2) + 0.4v_3(R_2) - v_0(R_2)$$
$$g(R_2) = 100 + 0.1v_0(R_2) + 0.2v_1(R_2) + 0.3v_2(R_2) + 0.4v_3(R_2) - v_1(R_2)$$
$$g(R_2) = 50 + 0.1v_0(R_2) + 0.2v_1(R_2) + 0.3v_2(R_2) + 0.4v_3(R_2) - v_2(R_2)$$
$$g(R_2) = \ \ 0 + 0.1v_0(R_2) + 0.2v_1(R_2) + 0.3v_2(R_2) + 0.4v_3(R_2) - v_3(R_2)$$
$$v_3(R_2) = 0$$

$\Rightarrow g(R_2) = 50, v_0(R_2) = 150, v_1(R_2) = 100, v_2(R_2) = 50, v_3(R_2) = 0$

Step 2: Policy improvement:

$$\text{minimize} \begin{pmatrix} 100 + v_0(R_2) - v_0(R_2) = 100 \\ 90 + 0.6v_0(R_2) + 0.4v_1(R_2) - v_0(R_2) = 70 \\ 110 + 0.3v_0(R_2) + 0.3v_1(R_2) + 0.4v_2(R_2) - v_0(R_2) = 55 \\ 150 + 0.1v_0(R_2) + 0.2v_1(R_2) + 0.3v_2(R_2) + 0.4v_3(R_2) - v_0(R_2) = \mathbf{50} \end{pmatrix}$$

$\Rightarrow d_0(R_3) = 3$

$$\text{minimize} \begin{pmatrix} 40 + 0.6v_0(R_2) + 0.4v_1(R_2) - v_1(R_2) = 70 \\ 60 + 0.3v_0(R_2) + 0.3v_1(R_2) + 0.4v_2(R_2) - v_1(R_2) = 55 \\ 100 + 0.1v_0(R_2) + 0.2v_1(R_2) + 0.3v_2(R_2) + 0.4v_3(R_2) - v_1(R_2) = \mathbf{50} \end{pmatrix}$$

$\Rightarrow d_1(R_3) = 2$

$$\text{minimize} \begin{pmatrix} 10 + 0.3v_0(R_1) + 0.3v_1(R_1) + 0.4v_2(R_1) - v_2(R_1) = 55 \\ 50 + 0.1v_0(R_1) + 0.2v_1(R_1) + 0.3v_2(R_1) + 0.4v_3(R_1) - v_2(R_1) = \mathbf{50} \end{pmatrix}$$

$\Rightarrow d_2(R_3) = 1$

R_3 is identical to R_2, so it is optimal to start every period with 3 pints of blood after delivery of the order.