Chapter 12
Markov Decision Processes

12.1. (a) $\mathbf{g_1} = (800, 275, 300, 250)^T$

(b)

$$v^\alpha(a) = \max\left\{800 + 0.95(0.1, 0.3, 0.6, 0)\begin{pmatrix}8651.88\\8199.73\\8233.37\\8402.65\end{pmatrix};\right.$$
$$\left.600 + 0.95(0.6, 0.3, 0.1, 0)\begin{pmatrix}8651.88\\8199.73\\8233.37\\8402.65\end{pmatrix}\right\}$$
$$= \max\{8651.88, 8650.66\} = 8651.88$$

$$v^\alpha(b) = \max\left\{275 + 0.95(0, 0.2, 0.5, 0.3)\begin{pmatrix}8651.88\\8199.73\\8233.37\\8402.65\end{pmatrix};\right.$$
$$\left.75 + 0.95(0.75, 0.1, 0.1, 0.05)\begin{pmatrix}8651.88\\8199.73\\8233.37\\8402.65\end{pmatrix}\right\}$$
$$= \max\{8138.56, 8199.73\} = 8199.73$$

$$v^\alpha(c) = \max\{300 + 0.95(0,0.1,0.2,0.7)\begin{pmatrix}8651.88\\8199.73\\8233.37\\8402.65\end{pmatrix};$$
$$100 + 0.95(0.8,0.2,0,0)\begin{pmatrix}8651.88\\8199.73\\8233.37\\8402.65\end{pmatrix}\}$$
$$= \max\{8231.08, 8233.37\} = 8233.37$$

$$v^\alpha(d) = \max\{250 + 0.95(0.8,0.1,0,0.1)\begin{pmatrix}8651.88\\8199.73\\8233.37\\8402.65\end{pmatrix};$$
$$150 + 0.95(0.9,0.1,0,0)\begin{pmatrix}8651.88\\8199.73\\8233.37\\8402.65\end{pmatrix}\}$$
$$= \max\{8402.65, 8326.33\} = 8402.65$$

Since, for each $i \in E$, the maximum of the two values yields the given vector v^α, it is optimum.

(c) Using the value iteration algorithm, the optimal value function is
$\mathbf{v}_0 = (0,0,0,0)$
$\mathbf{v}_1 = (800,275,300,250)$
$\vdots$
$\mathbf{v}_{30} = (1676.76, 1170.67, 1222.76, 1366.59)$

(d) $\alpha = 1.0/1.12$
$\mathbf{a}_0 = (1,1,1,1) \implies \mathbf{v} = (4092,3581,3672,3835)$
$\mathbf{a}_1 = (1,2,1,1) \implies \mathbf{v} = (4170,3707,3742,3908)$
$\mathbf{a}_2 = (1,2,1,1)$; therefore $\mathbf{a}_2$ is optimal.

(e) $\min u_a + u_b + u_c + u_d$
subject to:

$$\begin{array}{lllll}
u_a \ge 800 & +0.089u_a & +0.268u_b & +0.535u_c & \\
u_a \ge 600 & +0.535u_a & +0.268u_b & +0.089u_c & \\
u_b \ge 275 & & +0.178u_b & +0.446u_c & +0.268u_d \\
u_b \ge 75 & +0.669u_a & +0.089u_b & +0.089u_c & +0.044u_d \\
u_c \ge 300 & & +0.089u_b & +0.178u_c & +0.625u_d \\
u_c \ge 100 & +0.714u_a & +0.178u_b & & \\
u_d \ge 250 & +0.714u_a & +0.089u_b & & +0.089u_d \\
u_d \ge 150 & +0.803u_a & +0.089u_b & &
\end{array}$$

12.3. (a) Let the action space $A = \{1,2,3\}$ where each denotes to vote the Labor Party, Worker's Choice Party and the independent candidates, respectively. The optimal policy is $\mathbf{a} = (1,2,3)$ with

$$\mathbf{P} = \begin{bmatrix} 0.75 & 0.2 & 0.05 \\ 0.2 & 0.6 & 0.2 \\ 0.05 & 0.4 & 0.55 \end{bmatrix}$$

$\mathbf{f} = (3.2, 2.3, 1.5)$ (in millions)

(b) The optimal policy is $\mathbf{a} = (2,1,1)$ with the value function $\mathbf{v}^\alpha = (36.8, 35.46, 33.71)$ (in trillions), and

$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.15 & 0.05 \\ 0.3 & 0.5 & 0.2 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}$$

$\mathbf{f} = (4, 3.5, 2.5)$ (in trillions)

(c) The optimal policy is $\mathbf{a} = (2,1,3)$ with the value function $\mathbf{v}^\alpha = (15.57, 16.57, 17.56)$ (in thousands), and

$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.15 & 0.05 \\ 0.3 & 0.5 & 0.2 \\ 0.05 & 0.4 & 0.55 \end{bmatrix}$$

$\mathbf{f} = (1.33, 1.75, 2.2)$ (in thousands)

12.5. State space E with j states, Markov matrix $\mathbf{P}$ and profit function f. Expanding the state space E with a new state Δ which has a profit of zero, the Markov decision process can be formulated as:
State space $E' = \{i \mid i \in E \text{ or } \Delta\}$ where Δ stands for the absorbing state of stopping.
Action space $A = \{1,2\}$ where 1 denotes the action of continuing and 2 denotes the action of stopping.
The profit vectors are $\mathbf{f}_1 = (0, \cdots, 0)$ and $\mathbf{f}_2 = (f(1), f(2), \cdots, f(j), 0)$
We will construct the new transition matrices the same way as we did in Example 3.4.

$$\mathbf{P}_1 = \begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}, \quad \mathbf{P}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{1} \\ \mathbf{0} & 1 \end{bmatrix}$$

where $\mathbf{0}$ and $\mathbf{1}$ are matrices or vectors of the proper dimenstion as the context requires.The linear programming formulation is (from Algorithm 3.11):

$\min \sum_{i \in E} u(i)$
subject to:
$u(i) \geq f_1(i) + \alpha \sum_{j \in E} P_1(i,j) u(j)$ for each $i \in E$
$u(i) \geq f(i) + \alpha \sum_{j \in E} P_2(i,j) u(j)$ for each $i \in E$.

Based on the facts that $f(\Delta) = 0$ and Δ is an absorbing state, it follows that $u(\Delta) = 0$ and the previous definitions of $\mathbf{P}_1$ and $\mathbf{P}_2$, the above formulation reduces to the formulation given in Algorithm 3.18.

12.7. (a) The state space is $E = \{0, 1, 2, 3, 4, 5\}$ for the inventory at the end of Friday and the action space for the order up-to quantity is $A = \{0, 1, 2, 3, 4, 5\}$. According to each order up-to quantity, the expected profit function is the expected sales revenue minus costs. The corresponding Markov transition matrices for each order upto quantity can be obtained in a similar manner as in Ch. 2, Exercise 2.7. For example, when $k = 3$ the transition matrix is:

$$\mathbf{P}_3 = \begin{bmatrix} 0.58 & 0.22 & 0.15 & 0.05 & 0 & 0 \\ 0.58 & 0.22 & 0.15 & 0.05 & 0 & 0 \\ 0.58 & 0.22 & 0.15 & 0.05 & 0 & 0 \\ 0.58 & 0.22 & 0.15 & 0.05 & 0 & 0 \\ 0.36 & 0.22 & 0.22 & 0.15 & 0.05 & 0 \\ 0.18 & 0.18 & 0.22 & 0.22 & 0.15 & 0.05 \end{bmatrix}$$

and $\mathbf{f}_3 = (120.5, 620.5, 1120.5, 1720.5, 2008.5, 2152.5)$.

(b)

Friday's inventory	Oder up-to quan.
0	5
1	5
2	5
3	5
4	4
5	5

(c)

Friday's inventory	Order up-to quan.
0	5
1	5
2	5
3	3
4	4
5	5

(d) Note that the negative initial inventory denotes the number of items on back-order.

Friday's inventory	Order up-to quan.
-5	3
-4	3
-3	3
-2	3
-1	3
0	0
1	1
2	2
3	3
4	4
5	5

(e) The answers for part (b) and part (c) are the same as following under the average cost criterion:

Friday's inventory	Order up-to quan.
0	5
1	5
2	5
3	5
4	4
5	5

The answer for part (d) changes to:

Friday's inventory	Order up-to quan.
-5	5
-4	5
-3	5
-2	5
-1	5
0	5
1	1
2	2
3	3
4	4
5	5

12.9. (a) $0.069 + 0.931p$.
(b) $0.931(1-p)$.
(c) If $I_{n+1} = 0$, then $Z_{n+1} = (0.02 + 0.98p)/(0.069 + 0.931p)$; if $I_{n+1} = 1$, then $Z_{n+1} = 0$;
(d) For $p = 0$, there is only one possible decision, which yields

$$v(0) = 500 + 0.9 \times \left(0.069\, v(\tfrac{0.02}{0.069}) + 0.931 v(0) \right)$$

and for $p > 0$, we have

$$v(p) = \max\{ -475 + 0.9 \times ((0.069 + 0.931p)\, v(\tfrac{0.02+0.98p}{0.069+0.931p}) + 0.931(1-p)v(0)) ;$$
$$-2975 + 0.9 \times (0.069\, v(\tfrac{0.02}{0.069}) + 0.931 v(0)) \}$$

(e) Notice that the possible values of p are discrete being contained within the following set (depending a p^*): $\{0.0, 0.290, 0.897, 0.994, \cdots\}$. Thus, one way to solve the problem is to first let p^* be a number between 0.0 and 0.290 which will yield the following equations:

$$\begin{aligned} v(0) &= 500 + 0.062v(0.290) + 0.838v(0) \\ v(0.290) &= -2975 + 0.062v(0.290) + 0.838v(0) \end{aligned}$$

which yields $v(0) = 2842$. Next if p^* is a number between 0.290 and 0.897 the system of equations becomes:

$$\begin{aligned} v(0) &= 500 + 0.062v(0.290) + 0.838v(0) \\ v(0.290) &= -475 + 0.305v(0.897) + 0.595v(0) \\ v(0.897) &= -2975 + 0.062v(0.290) + 0.838v(0) \end{aligned}$$

which yields $v(0) = 3810$. Next if p^* is a number between 0.897 and 0.994 the system of equations becomes:

$$\begin{aligned} v(0) &= 500 + 0.062v(0.290) + 0.838v(0) \\ v(0.290) &= -475 + 0.305v(0.897) + 0.595v(0) \\ v(0.897) &= -475 + 0.814v(0.994) + 0.086v(0) \\ v(0.994) &= -2975 + 0.062v(0.290) + 0.838v(0) \end{aligned}$$

which yields $v(0) = 3774$. Thus, we would assert that p^* should be any value between 0.290 and 0.897. (In other words, replace whenever two bad products are produced in sequence.)