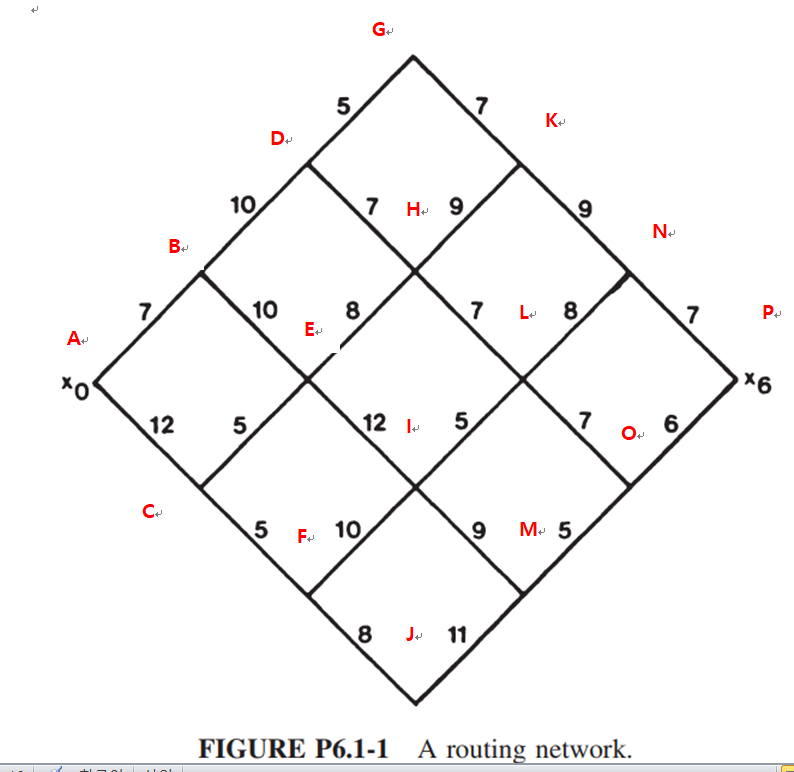
P.6.1-1 page 284

1. Dynamic Programming



-Solution

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stage | State | Bellman | Value | Optimal |
| 5 | N | V(N) | 7 | **V(N)** |
|  | O | V(O) | 6 | **V(O)** |
|  |  |  |  |  |
| 4 | K | 9+V(N) | 16 | **V(K)** |
|  | L | 8+V(N) | 15 |  |
|  |  | 7+V(O) | 13 | **V(L)** |
|  | M | 5+V(O) | 11 | **V(M)** |
|  |  |  |  |  |
| 3 | G | 7+V(K) | 23 | **V(G)** |
|  | H | 9+V(K) | 24 |  |
|  |  | 7+V(L) | 20 | **V(H)** |
|  | I | 5+V(L) | 18 | **V(I)** |
|  |  | 9+V(M) | 20 |  |
|  | J | 11+V(M) | 22 | **V(J)** |
|  |  |  |  |  |
| 2 | D | 5+V(G) | 27 | **V(D)** |
|  |  | 7+V(H) | 27 | **V(D)** |
|  | E | 8+V(H) | 28 | **V( E)** |
|  |  | 12+V(I) | 30 |  |
|  | F | 10+V(I) | 28 | **V(F)** |
|  |  | 8+V(J) | 30 |  |
|  |  |  |  |  |
| 1 | B | 10+V(D) | 37 | **V(B)** |
|  |  | 10+V( E) | 38 |  |
|  | C | 5+V( E) | 33 | **V( C)** |
|  |  | 5+V(F) | 33 | **V( C)** |
|  |  |  |  |  |
| 0 | A | 7+V(B) | 44 | **V(A)** |
|  |  | 12+V( C) | 45 |  |
|  |  |  |  | **Optimal path** |
|  |  |  |  | ABDGKNP |
|  |  |  |  | **Optimal Value** |
|  |  |  |  | 44 |

2. Reinforcement Learning

If the transition matrix is ergodic,

1. If LHP , which is a present stage
   1. Policy evaluation

Find for

Hence

Hence Given : the transition probability , : reward value

It can be determined

Now for following the same step except

* 1. Policy improvement

Find to minimize

For each state, find the optimal policy

1. What are the differences between 1 and 2 method?