Proposition 1: Under Assumption 1, the following hold for the stochastic shortest path problem:

a. Given any initial conditions $J_0(1), \dots, J_0(n)$ the sequence $J_k(i)$ generated by the DP iteration

$$J_{k+1}(i) = \min_{u \in U(i)} \left\{ \bar{g}(i,u) + \sum_{j=1}^{n} p(i,u,j)\, J_k(j) \right\}, \quad i \in [n] \qquad -(1)$$

converges to the optimal cost $J^*(i)$ for each i.

b. The optimal cost $J^*(1), \dots, J^*(n)$ satisfy Bellmans equation,

$$J^*(i) = \min_{u \in U(i)} \left\{ \bar{g}(i,u) + \sum_{i=1}^{n} p(i,u,j) \cdot J^*(j) \right\}, \quad i \in [n] \qquad -(2)$$

and in fact they are the unique solution of this equation.

c. For any stationary policy μ, the costs $J_\mu(1), \dots, J_\mu(n)$ are the unique solution of the equation

$$J_\mu(i) = \bar{g}(i, \mu(i)) + \sum_{i=1}^{n} p(i, \mu(i), j)\, J_\mu(j), \quad i = 1, \dots, n \qquad -(3)$$

Furthermore, given any initial conditions $J_0(1), \dots, J_0(n)$, the sequence $J_k(i)$ generated by the DP iteration

$$J_{k+1}(i) = \bar{g}(i, \mu(i)) + \sum_{i=1}^{n} p(i, \mu(i), j)\, J_k(j), \quad i = 1, \dots, n \qquad -(4)$$

converges to the cost $J_\mu(i)$ for each i.

d. A stationary policy μ is optimal if and only if for every state i, $\mu(i)$ attains the minimum in Bellmans equation (2).

Proof:

a. For every positive integer K, initial state x_0, and policy $\pi = \{\mu_0, \mu_1, \dots\}$, we break $J_\pi(x_0)$ in two parts as follows:

$$J_\pi(x_0) = \lim_{N \to \infty} \mathbb{E}\left[\sum_{k=0}^{N-1} \bar{g}(x_k, \mu_k(x_k)) \right]$$

$$= \mathbb{E}\left[\sum_{k=0}^{mK-1} \bar{g}(x_k, \mu_k(x_k)) \right] + \lim_{N \to \infty} \mathbb{E}\left[\sum_{k=mK}^{N-1} g(x_k, \mu_k(x_k)) \right] \qquad -(i)$$

Let $M := m \cdot \max_{\substack{i \in [n] \\ u \in U(i)}} |\bar{g}(i,u)|$. Then

Let $M := m \cdot \max_{\substack{i\in[n]\\ u\in U(i)}} |\bar{g}(i,u)|$. Then — (i)

$$\left| \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=mK}^{N-1} g(x_k, \mu_k(x_k))\right] \right| \leq M \sum_{k=K}^{\infty} \rho^k = \frac{\rho^K M}{1-\rho} \qquad -(ii)$$

Also, denoting $J_0(t) = 0$, let us view J_0 as a terminal cost function and bound its expected value under π after mK stages. We have

$$\begin{aligned} |\mathbb{E}[J_0(x_{mK})]| &= \left|\sum_{i=1}^{n} P(x_{mK} = i \mid x_0, \pi)\, J_0(i)\right| \\ &\leq \left(\sum_{i=1}^{n} P(x_{mK} = i \mid x_0, \pi)\right) \cdot \max_{i\in[n]} |J_0(i)| \\ &\leq \rho^K \cdot \max_{i\in[n]} |J_0(i)| \qquad — (iii) \end{aligned}$$

From (i) and (ii) we get that

$$J_\pi(x_0) - \frac{\rho^K M}{1-\rho} \leq \mathbb{E}\left[\sum_{k=0}^{mK-1} g(x_k, \mu_k(x_k))\right] \leq J_\pi(x_0) + \frac{\rho^K M}{1-\rho} \qquad — (iv)$$

Adding (iii) to (iv) we get

$$J_\pi(x_0) - \frac{\rho^K M}{1-\rho} - \rho^K \max_{i\in[n]} |J_0(i)| \leq J_{mK}(x_0) \leq J_\pi(x_0) + \frac{\rho^K M}{1-\rho} + \rho^K \max_{i\in[n]} |J_0(i)| \qquad — (v)$$

Taking the minimum over π we get

$$J^*(x_0) - \rho^K\left(\frac{M}{1-\rho} + \max_{i\in[n]} |J_0(i)|\right) \leq J_{mK}(x_0) \leq J^*(x_0) + \rho^K\left(\frac{M}{1-\rho} + \max_{i\in[n]} |J_0(i)|\right)$$

Now as taking limit $K \to \infty$, we get that

$$J^*(x_0) = \lim_{K\to\infty} J_{mK}(x_0)$$

Since $|J_{mK+q}(x_0) - J_{mK}(x_0)| \leq \rho^K M$, $q \in [m]$ we see that $\lim_{K\to\infty} J_{mK+q}(x_0)$ is the same for all $q = 1, \ldots, m$, so that we have $\lim_{k\to\infty} J_k(x_0) = J^*(x_0)$

b. By taking limit as $k \to \infty$ in the DP iteration

$$J_{k+1}(i) = \min_{u\in U(i)} \left\{ \bar{g}(i,u) + \sum_{j=1}^{n} P(i,u,j)\, J_k(j) \right\}, \quad i \in [n]$$

and using the result of a, we are done. (* Exchange of limit

and using the result of a, we are done. (* Exchange of limit & min)

c. Consider the admissible control set at state i to be $\bar{U}(i) = \{\mu(i)\}$ instead of $U(i)$. Result follows from a, b.

d We have that $\mu(i)$ attains the minimum in (2) if and only if we have

$$J^*(i) = \min_{u \in U(i)} \left\{ \bar{g}(i,u) + \sum_{j=1}^{n} p(i,u,j)\, J^*(j) \right\}$$

$$= \bar{g}(i,\mu(i)) + \sum_{j=1}^{n} p(i,\mu(i)\, j)\, J^*(j)$$

By c, $J^* = J_\mu$. Conversely if $J^* = J_\mu$ then part b & c imply the above equation. ▨

Exercise

A spider and a fly move along a straight line at times $k = 0, 1, \ldots$. The initial positions of the fly and the spider are integer. At each time period, the fly moves one unit to the left with probability p, one unit to the right with probability p, and stays where it is with probability $1 - 2p$. The spider, knows the position of the fly at the beginning of each period, and will always move one unit towards the fly if its distance from the fly is more that one unit. If the spider is one unit away from the fly, it will either move one unit towards the fly or stay where it is. If the spider and the fly land in the same position at the end of a period, then the spider captures the fly and the process terminates, The spider's objective is to capture the fly in minimum expected time.

Formulate this problem as stochastic shortest path problem and find the value function of the following two policies.

- Policy 1: If the distance is more than 1 unit then spider moves towards the fly and if the distance is 1 unit then the spyder doesn't move.
- Policy 2: If the distance is not 0 then spider moves towards the fly.

Can you tell which policy is better?