SEEM 3470: Dynamic Optimization and Applications **2013–14 Second Term**

Handout 11: Infinite–Horizon Dynamic Programming Problems

Instructor: Shiqian Ma March 31, 2014

Suggested Reading: Chapters 7 of Bertsekas, *Dynamic Programming and Optimal Control: Volume I (3rd Edition)*, Athena Scientific, 2005; Chapter 3 of Powell, *Approximate Dynamic Programming: Solving the Curse of Dimensionalty (2nd Edition)*, Wiley, 2010.

1 Introduction

In the previous handouts, we focused on dynamic programming (DP) problems with a finite horizon and developed algorithms for solving them. In this handout, we will study infinite–horizon DP problems. Such problems arise naturally if we want to understand the steady state of the underlying system, or if we do not know in advance the number of periods we need to make a decision (such as the Asset Pricing example in Handout 10). To solve infinite–horizon DP problems, a tempting idea would be to apply the DP algorithm. Recall that in the DP algorithm, we start with the final–stage problem and apply backward recursion. However, if there is an infinite number of stages, it is not clear what the "final–stage" problem would be. Hence, we need some new techniques in order to develop algorithms for solving infinite–horizon DP problems.

2 Examples

2.1 Example 1: Asset selling

Consider an infinite horizon version of the asset selling example of Handout 10, assuming the set of possible offers is finite. Here, if accepted, the amount x_k offered in period k, will be invested at a rate of interest r. By depreciating the sale amount to period 0 dollars, we view $(1+r)^{-k}x_k$ as the reward for selling the asset in period k at a price x_k, where $r > 0$ is the rate of interest. Then we have a total discounted reward problem with discount factor $\alpha = 1/(1+r)$. The analysis (we will learn shortly) shows that the optimal value function J^* is the unique solution of the Bellman's equation (optimality condition)

$$J^*(x) = \max\left[x, \frac{E\{J^*(w)\}}{1+r}\right].$$

The optimal reward function is characterized by the critical number

$$\bar{\alpha} = \frac{E\{J^*(w)\}}{1+r},$$

which can be calculated as in Handout 10. An optimal policy is to sell if and only if the current offer x_k is greater than or equal to $\bar{\alpha}$.

2.2 Example 2: Minimum expected time

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 than 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.

We view as state the distance between spider and fly. Then the problem can be formulated as a stochastic shortest path problem with states $0, 1, \ldots, n$, where n is the initial distance. State 0 is the termination state where the spider captures the fly. Let us denote $p_{1j}(M)$ and $p_{1j}(\bar{M})$ the transition probabilities from state 1 to state j if the spider moves and does not move, respectively, and let us denote by p_{ij} the transition probabilities from a state $i \geq 2$. We have

$$p_{ii} = p, \quad p_{i(i-1)} = 1 - 2p, \quad p_{i(i-2)} = p, \quad i \geq 2,$$

$$p_{11}(M) = 2p, \quad p_{10}(M) = 1 - 2p,$$

$$p_{12}(\bar{M}) = p, \quad p_{11}(\bar{M}) = 1 - 2p, \quad p_{10}(\bar{M}) = p,$$

with all other transition probabilities being 0.

For state $i \geq 2$, Bellman's equation is written as

$$J^*(i) = 1 + pJ^*(i) + (1 - 2p)J^*(i - 1) + pJ^*(i - 2), \quad i \geq 2, \tag{1}$$

where $J^*(0) = 0$ by definition. The only state where the spider has a choice is when it is one unit away from the fly, and for that state Bellman's equation is given by

$$J^*(1) = 1 + \min\left[2pJ^*(1), pJ^*(2) + (1 - 2p)J^*(1)\right], \tag{2}$$

where the first and the second expression within the bracket above are associated with the spider moving and not moving, respectively. By writing Eq. (1) for $i = 2$, we obtain

$$J^*(2) = 1 + pJ^*(2) + (1 - 2p)J^*(1),$$

from which

$$J^*(2) = \frac{1}{1 - p} + \frac{(1 - 2p)J^*(1)}{1 - p}. \tag{3}$$

Substituting this expression in (2), we obtain

$$J^*(1) = 1 + \min\left[2pJ^*(1), \frac{p}{1 - p} + \frac{p(1 - 2p)J^*(1)}{1 - p} + (1 - 2p)J^*(1)\right],$$

or equivalently,

$$J^*(1) = 1 + \min\left[2pJ^*(1), \frac{p}{1 - p} + \frac{(1 - 2p)J^*(1)}{1 - p}\right].$$

To solve the above equation, we consider the two cases where the first expression with in the bracket is larger and is smaller than the second expression. Thus we solve for $J^*(1)$ in the two cases where

$$J^*(1) = 1 + 2pJ^*(1), \tag{4}$$

$$2pJ^*(1) \leq \frac{p}{1-p} + \frac{(1-2p)J^*(1)}{1-p}, \tag{5}$$

and

$$J^*(1) = 1 + \frac{p}{1-p} + \frac{(1-2p)J^*(1)}{1-p}, \tag{6}$$

$$2pJ^*(1) \geq \frac{p}{1-p} + \frac{(1-2p)J^*(1)}{1-p}. \tag{7}$$

The solution of Eq. (4) is seen to be $J^*(1) = 1/(1-2p)$, and by substitution in Eq. (5), we find that this solution is valid when

$$\frac{2p}{1-2p} \leq \frac{p}{1-p} + \frac{1}{1-p},$$

or equivalently (after some calculation), $p \leq 1/3$. Thus for $p \leq 1/3$, it is optimal for the spider to move when it is one unit away from the fly.

Similarly, the solution of Eq. (6) is seen to be $J^*(1) = 1/p$, and by substitution in Eq. (7), we find that this solution is valid when

$$2 \geq \frac{p}{1-p} + \frac{1-2p}{p(1-p)},$$

or equivalently (after some calculation), $p \geq 1/3$. thus, for $p \geq 1/3$ it is optimal for the spider not to move when it is one unit away form the fly.

The minimal expected number of steps for capture when the spider is one unit away from the fly was calculated earlier to be

$$J^*(1) = \begin{cases} 1/(1-2p) & \text{if} \quad p \leq 1/3, \\ 1/p & \text{if} \quad p \geq 1/3. \end{cases}$$

Given the value of $J^*(1)$, we can calculate from Eq. (3) the minimal expected number of steps for capture when two units away, $J^*(2)$, and we can then obtain the remaining values $J^*(i), i = 3, \ldots, n$, from Eq. (1).

3 Infinite–Horizon DP Problems: Preliminaries

We begin with the setup of the problem. Let $\mathcal{S}$ be a discrete state space. Let S_k and W_k be the state and random parameter in period k, respectively. Given that the state and control in period k is $S_k = i$ and $x_k = x$, respectively, the next state S_{k+1} is specified by a probability distribution

$$p_{ij}(x) = \Pr(S_{k+1} = j \,|\, S_k = i, x_k = x), \tag{8}$$

and the cost incurred in period k is given by

$$\Lambda(S_k, x_k, W_k).$$

It should be noted that $p_{ij}(x)$ does not depend on the period k, and the cost function $\Lambda(\cdot, \cdot, \cdot)$ is the same in every period. In other words, both the transition probabilities and the cost function are *time homogeneous*. Moreover, the transition probabilities are *Markov*, i.e., $p_{ij}(x)$ depends only

on the current state S_k and not on any of the previous states $S_0, S_1, \ldots, S_{k-1}$. In the sequel, we shall assume that the cost function $\Lambda(\cdot, \cdot, \cdot)$ is non–negative.

Now, given a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ and an initial state $S_0 = s$, the total expected cost of the infinite–horizon problem associated with the above system is given by

$$J_0(s, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \Lambda(S_k, \mu_k(S_k), W_k) \,\middle|\, S_0 = s, \mu_0\right], \tag{9}$$

where $\gamma \in (0, 1]$ is a *discount factor*. The meaning of $\gamma < 1$ is that future costs matter to us less than the same costs incurred at the present time. As an example, think of k-th period dollars depreciated to initial period dollars by a factor of $(1 + r)^{-k}$, where r is a rate of interest; here $\gamma = 1/(1 + r)$.

Naturally, we are interested in choosing the policy that minimizes the above total expected cost, i.e., we would like to solve the problem

$$J_0(s) = \min_{\pi} J_0(s, \pi) \quad \text{for every state } s \in \mathcal{S}. \tag{10}$$

Towards that end, we shall make two simplifying assumptions:

1. The cost function does not depend on the random parameter W_k. In other words, we may write $\Lambda(S_k, \mu_k(S_k))$ instead of $\Lambda(S_k, \mu_k(S_k), W_k)$.

2. Since both the transition probabilities and cost function are time–homogeneous, it seems reasonable to expect that the optimal policy π^* that solves (10) is also time–homogeneous or *stationary*, i.e., π^* takes the form $\pi^* = \{\mu^*, \mu^*, \ldots\}$. Hence, we shall restrict our attention to stationary policies. In fact, in many cases, this assumption can be made without loss of generality. Since a stationary policy π is completely specified by the control function $\mu(\cdot)$, we shall abuse notation and use μ to denote a stationary policy.

With the above two assumptions, we can write $J_0(s, \pi)$ as

$$J_0(s, \mu) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \Lambda(S_k, \mu(S_k)) \,\middle|\, S_0 = s, \mu\right] \tag{11}$$

and

$$J_0(s) = \min_{\mu} J_0(s, \mu) \quad \text{for every state } s \in \mathcal{S} \tag{12}$$

(compare (11) with (9) and (12) with (10)).

As mentioned in the Introduction section, one cannot apply the DP algorithm for finite–horizon problems to (11) directly, as the definition of the "final–stage" problem in (11) is not clear. However, for any given integer $N \geq 0$, we can consider the *truncated version* of (11), i.e.,

$$Q_N(s, \mu) = \mathbb{E}\left[\sum_{k=0}^{N-1} \gamma^k \Lambda(S_k, \mu(S_k)) \,\middle|\, S_0 = s, \mu\right] \tag{13}$$

and set

$$Q_N(s) = \min_{\mu} Q_N(s, \mu) \quad \text{for every state } s \in \mathcal{S}. \tag{14}$$

The motivation for considering the above truncated problems is twofold. First, problem (14) is a *finite–horizon* DP by construction, and hence we can use the DP algorithms we developed earlier to solve it. Secondly, as we take $N \to \infty$, we have $Q_N(s,\mu) \to J_0(s,\mu)$. Hence, it seems intuitive that $Q_N(s) \to J_0(s)$. In other words, we can perhaps use the truncated, finite–horizon problem (14) to approximate the original, infinite–horizon problem (12), and at the limit, this approximation will be exact.

To implement the above idea, we need to establish a relationship between $Q_N(s)$ and $J_0(s)$. Let us begin by computing

$$\begin{aligned} Q_N(s,\mu) &= \Lambda(s,\mu(s)) + \sum_{k=1}^{N-1} \gamma^k \cdot \mathbb{E}\left[\Lambda(S_k,\mu(S_k)) \,|\, S_0 = s,\mu\right] && (15)\\ &= \Lambda(s,\mu(s)) + \sum_{k=1}^{N-1} \gamma^k \left[\sum_{t\in\mathcal{S}} \Big(\mathbb{E}\left[\Lambda(S_k,\mu(S_k)) \,|\, S_1 = t,\mu\right] \cdot \Pr(S_1 = t \,|\, S_0 = s,\mu)\Big)\right] && (16)\\ &= \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \left(p_{s,t}(\mu(s)) \cdot \mathbb{E}\left[\sum_{k=1}^{N-1} \gamma^{k-1} \Lambda(S_k,\mu(S_k)) \,\middle|\, S_1 = t,\mu\right]\right), && (17) \end{aligned}$$

where (15) and (17) follow from linearity of expectation, and (16) follows by conditioning on S_1 and using the Markov property of the transition probabilities, i.e.,

$$\begin{aligned} \mathbb{E}\left[\Lambda(S_k,\mu(S_k)) \,|\, S_0 = s,\mu\right] &= \sum_{t\in\mathcal{S}} \left[\mathbb{E}\left[\Lambda(S_k,\mu(S_k)) \,|\, S_1 = t, S_0 = s,\mu\right] \cdot \Pr(S_1 = t \,|\, S_0 = s,\mu)\right]\\ &= \sum_{t\in\mathcal{S}} \big(p_{s,t}(\mu(s)) \cdot \mathbb{E}\left[\Lambda(S_k,\mu(S_k)) \,|\, S_1 = t,\mu\right]\big). \end{aligned}$$

Now, observe that the problem

$$\mathbb{E}\left[\sum_{k=1}^{N-1} \gamma^{k-1} \Lambda(S_k,\mu(S_k)) \,\middle|\, S_1 = t,\mu\right]$$

in (17) has the same form as that in (13), except that the former has only $N-1$ stages and starts in the state t. Hence, we have

$$Q_{N-1}(t,\mu) = \mathbb{E}\left[\sum_{k=1}^{N-1} \gamma^{k-1} \Lambda(S_k,\mu(S_k)) \,\middle|\, S_1 = t,\mu\right].$$

Upon substituting this into (17), we conclude that

$$\boxed{Q_N(s,\mu) = \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \left[p_{s,t}(\mu(s)) \cdot Q_{N-1}(t,\mu)\right] \quad \text{for all } s \in \mathcal{S} \text{ and } \mu} \tag{18}$$

From (18), we have

$$\begin{aligned} Q_N(s,\mu) &\geq \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \left[p_{s,t}(\mu(s)) \cdot \left(\min_{\mu} Q_{N-1}(t,\mu)\right)\right]\\ &= \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \left[p_{s,t}(\mu(s)) \cdot Q_{N-1}(t)\right] && (19) \end{aligned}$$

Upon minimizing both sides of (19) with respect to $\mu(s)$, we have

$$Q_N(s) = \min_{\mu} Q_N(s,\mu) \geq \min_{\mu(s)} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot Q_{N-1}(t)\big] \right\}. \tag{20}$$

On the other hand, let

$$\bar{\mu} = \arg\min_{\mu} Q_{N-1}(t,\mu),$$

so that $Q_{N-1}(t) = Q_{N-1}(t,\bar{\mu})$. Then, from (18),

$$\begin{aligned} Q_N(s) &= \min_{\mu} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot Q_{N-1}(t,\mu)\big] \right\} \\ &\leq \Lambda(s,\bar{\mu}(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\bar{\mu}(s)) \cdot Q_{N-1}(t)\big]. \end{aligned}$$

Upon minimizing both sides with respect to $\bar{\mu}(s)$, we obtain

$$Q_N(s) \leq \min_{\mu(s)} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot Q_{N-1}(t)\big] \right\}.$$

This, together with (20), implies that

$$\boxed{Q_N(s) = \min_{\mu(s)} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot Q_{N-1}(t)\big] \right\} \quad \text{for all } s \in \mathcal{S}} \tag{21}$$

Now, we are ready to derive the optimality condition governing $J_0(s)$ using the finite–horizon approximation $Q_N(s)$. Recall that as $N \to \infty$, we have $Q_N(s,\mu) \to J_0(s,\mu)$ for every state $s \in \mathcal{S}$ and control $\mu(\cdot)$. Hence, upon taking $N \to \infty$ on both sides of (19), we obtain

$$J_0(s,\mu) \geq \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot J_0(t)\big].$$

Upon minimizing both sides with respect to μ and using the fact that $J_0(s) = \min_{\mu} J_0(s,\mu)$, we obtain

$$J_0(s) \geq \min_{\mu(s)} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot J_0(t)\big] \right\}. \tag{22}$$

On the other hand, let

$$\bar{\mu} = \arg\min_{\mu} J_0(t,\mu),$$

so that $J_0(t) = J_0(t,\bar{\mu})$. Upon taking $N \to \infty$ on both sides of (18), we have

$$J_0(s,\mu) = \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot J_0(t,\mu)\big].$$

Hence,

$$J_0(s) = \min_{\mu} J_0(s,\mu) \le \Lambda(s,\bar{\mu}(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\bar{\mu}(s)) \cdot J_0(t)\big].$$

Upon minimizing both sides with respect to $\bar{\mu}(s)$, we get

$$J_0(s) \le \min_{\mu(s)} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot J_0(t)\big] \right\}.$$

This, together with (22), shows that $J_0(\cdot)$ must satisfy the following *optimality equation*:

$$\boxed{J_0(s) = \min_{\mu(s)} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot J_0(t)\big] \right\} \quad \text{for all } s \in \mathcal{S}} \tag{23}$$

4 The Value Iteration Algorithm

Although the optimality equation (23) tells us the structure of the optimal total expected cost of the infinite–horizon problem (11), it does not yet yield an algorithm for computing the optimal value $J_0(s)$ or the optimal policy

$$\mu^*(s) = \arg\min_{\mu} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \big[p_{s,t}(\mu(s)) \cdot J_0(t)\big] \right\}.$$

The difficulty lies in the fact that J_0 is involved in both sides of the optimality equation (23). To circumvent this difficulty, we could try to decouple the two copies of J_0 in (23). In fact, equation (21) already gives a clue on how this could be done. Specifically, starting from an initial choice of Q_0, we could apply (21) iteratively to obtain a sequence $Q_0, Q_1, \ldots$. This gives rise to the so–called *value iteration algorithm*:

Algorithm 1 Value Iteration Algorithm for Finding J_0

1: **initialization**: set $Q_0(s) = 0$ for all $s \in \mathcal{S}$ and $N = 1$
2: **repeat**
3: find $Q_N(s)$ using (21) for every $s \in \mathcal{S}$
4: set $N \leftarrow N + 1$
5: **until convergence**

Let us illustrate the value iteration algorithm with an example.

Example 1 *Consider a system with state space $\mathcal{S} = \{1, 2\}$ and control set $\mathcal{C} = \{x^1, x^2\}$. The transition probabilities associated with the controls are given by*

$$P(x^1) = \big[p_{i,j}(x^1)\big] = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix}, \quad P(x^2) = \big[p_{i,j}(x^2)\big] = \begin{bmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{bmatrix}.$$

Furthermore, the transition costs are given by

$$\begin{aligned} \Lambda(1,x^1) &= 2, \quad \Lambda(1,x^2) = 1/2, \\ \Lambda(2,x^1) &= 1, \quad \Lambda(2,x^2) = 3. \end{aligned}$$

Consider the infinite–horizon DP:

$$J_0(s) = \min_{\mu} \ \mathbb{E}\left[\sum_{k=0}^{\infty}(0.9)^k \cdot \Lambda(S_k, \mu(S_k)) \,\middle|\, S_0 = s, \mu\right], \ \textit{where } s \in \mathcal{S}. \tag{24}$$

In particular, we have $\gamma = 0.9$. Applying the value iteration algorithm to (24), the results for the first two iterations are given as follows:

Initialization. $Q_0(1) = Q_0(2) = 0$.
Iteration 1. *Since there are only two possible controls x^1, x^2, we have, from (21),*

$$\begin{aligned} Q_1(1) &= \min\left\{\Lambda(1,x^1) + 0.9\left(p_{1,1}(x^1)Q_0(1) + p_{1,2}(x^1)Q_0(2)\right),\right. \\ &\qquad \left.\Lambda(1,x^2) + 0.9\left(p_{1,1}(x^2)Q_0(1) + p_{1,2}(x^2)Q_0(2)\right)\right\} \\ &= \min\{2, 0.5\} \\ &= 0.5, \quad \textit{with the corresponding control } x^2. \end{aligned}$$

Similarly,

$$\begin{aligned} Q_1(2) &= \min\left\{\Lambda(2,x^1) + 0.9\left(p_{2,1}(x^1)Q_0(1) + p_{2,2}(x^1)Q_0(2)\right),\right. \\ &\qquad \left.\Lambda(2,x^2) + 0.9\left(p_{2,1}(x^2)Q_0(1) + p_{2,2}(x^2)Q_0(2)\right)\right\} \\ &= \min\{1, 3\} \\ &= 1, \quad \textit{with the corresponding control } x^1. \end{aligned}$$

Iteration 2. *We have*

$$\begin{aligned} Q_2(1) &= \min\left\{\Lambda(1,x^1) + 0.9\left(p_{1,1}(x^1)Q_1(1) + p_{1,2}(x^1)Q_1(2)\right),\right. \\ &\qquad \left.\Lambda(1,x^2) + 0.9\left(p_{1,1}(x^2)Q_1(1) + p_{1,2}(x^2)Q_1(2)\right)\right\} \\ &= \min\{2.5625, 1.2875\} \\ &= 1.2875, \quad \textit{with the corresponding control } x^2, \end{aligned}$$

and

$$\begin{aligned} Q_2(2) &= \min\left\{\Lambda(2,x^1) + 0.9\left(p_{2,1}(x^1)Q_1(1) + p_{2,2}(x^1)Q_1(2)\right),\right. \\ &\qquad \left.\Lambda(2,x^2) + 0.9\left(p_{2,1}(x^2)Q_1(1) + p_{2,2}(x^2)Q_1(2)\right)\right\} \\ &= \min\{1.5625, 3.7875\} \\ &= 1.5625, \quad \textit{with the corresponding control } x^1. \end{aligned}$$

Note that in the above example, we can continue running the algorithm for more iterations. A natural question then is when we can stop. One stopping criterion is when the difference between successive iterates Q_{N-1} and Q_N is small. More precisely, we can choose an accuracy threshold $\epsilon > 0$ and stop the algorithm when

$$\sum_{s\in\mathcal{S}} |Q_N(s) - Q_{N-1}(s)|^2 < \epsilon. \tag{25}$$

Now, in order for the value iteration algorithm to be well defined, we must show that the stopping criterion (25) will be satisfied eventually. Moreover, recall that we use the value iteration algorithm to compute approximations of J_0 in (23). Hence, it would be good to know how well Q_N approximates J_0 as $N \to \infty$. Both of these issues are addressed in the following theorem:

Theorem 1 *Suppose that* $0 \leq \Lambda(\cdot,\cdot) \leq B$ *for some constant* B *and* $0 < \gamma < 1$. *Then,* $J_0(s) = \lim_{N\to\infty} Q_N(s)$ *for all* $s \in \mathcal{S}$.

Proof By definition (see (9) and (13)), we have

$$J_0(s,\mu) = Q_N(s,\mu) + \mathbb{E}\left[\sum_{k=N}^{\infty} \gamma^k \Lambda(S_k, \mu(S_k)) \,\middle|\, S_0 = s, \mu\right].$$

Thus, if $0 \leq \Lambda(\cdot,\cdot) \leq B$ and $0 < \gamma < 1$, then

$$Q_N(s,\mu) \leq J_0(s,\mu) \leq Q_N(s,\mu) + B\sum_{k=N}^{\infty} \gamma^k = Q_N(s,\mu) + \frac{B\gamma^N}{1-\gamma}.$$

Upon optimizing with respect to μ, we have

$$\min_{\mu} Q_N(s,\mu) \leq \min_{\mu} J_0(s,\mu) \leq \min_{\mu} Q_N(s,\mu) + \frac{B\gamma^N}{1-\gamma},$$

or equivalently (using the definitions of $J_0(s)$ and $Q_N(s)$ in (10) and (14)),

$$Q_N(s) \leq J_0(s) \leq Q_N(s) + \frac{B\gamma^N}{1-\gamma}. \tag{26}$$

The desired result now follows by taking $N \to \infty$. □

Theorem 1 not only shows that the stopping criterion (25) will be satisfied eventually, but also gives an error bound on how far the iterates $Q_0, Q_1, \ldots,$ produced by the value iteration algortihm are from the target J_0. Specifically, the inequalities in (26) show that

$$|Q_N(s) - J_0(s)| \leq \frac{B\gamma^N}{1-\gamma} \quad \text{for every state } s \in \mathcal{S}.$$

5 Policy Iteration Algorithm

In the last section, we study the value iteration algorithm, whose N–th iteration is essentially computing an N–stage approximation Q_N to the infinite–horizon cost J_0. In this section, we will develop another algorithm, called the *policy iteration algorithm*, for computing J_0. The main idea is again to decouple the two copies of J_0 in (23). In the value iteration algorithm, we achieve this decoupling by approximating the J_0 on the right–hand side of (23) by Q_N and optimizing the resulting expression. In the policy iteration algorithm, we first fix a policy μ (which may not be optimal). Then, we approximate the J_0 on the right–hand side of (23) by $J_0(\cdot,\mu)$ and optimize the resulting expression. Specifically, in the N–th iteration, we solve the following problem:

$$\mu^{N+1}(s) = \arg\min_{\mu(s)} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \left[p_{s,t}(\mu(s)) \cdot J_0(t,\mu^N)\right] \right\} \quad \text{for all } s \in \mathcal{S},$$

where we initialize the algorithm with an arbitrary policy μ^0. However, in order to implement this algorithm, we must know how to compute $J_0(\cdot, \mu)$ for any given policy μ. Towards that end, let us first recall the definition of $J_0(s, \mu)$ from (9):

$$J_0(S, \mu) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \Lambda(S_k, \mu(S_k)) \,\middle|\, S_0 = s, \mu\right]. \tag{27}$$

Consider the $k = 1$ term of (27). We have

$$\begin{aligned} \mathbb{E}\left[\Lambda(S_1, \mu(S_1)) \,|\, S_0 = s, \mu\right] &= \sum_{t \in \mathcal{S}} \left[\Lambda(t, \mu(t)) \cdot \Pr(S_1 = t \,|\, S_0 = s, \mu)\right] \\ &= \sum_{t \in \mathcal{S}} \left[p_{s,t}(\mu(s)) \cdot \Lambda(t, \mu(t))\right]. \end{aligned} \tag{28}$$

If we let $P(\mu) = [p_{i,j}(\mu(i))]$ be the $|\mathcal{S}| \times |\mathcal{S}|$ transition matrix associated with μ and $w(\mu)$ be the $|\mathcal{S}|$–dimensional column vector whose s–th entry is $\Lambda(s, \mu(s))$, where $s \in \mathcal{S}$, then we can write (28) more compactly as

$$\mathbb{E}\left[\Lambda(S_1, \mu(S_1)) \,|\, S_0 = s, \mu\right] = [P(\mu)w(\mu)]_s,$$

where $[P(\mu)w(\mu)]_s$ is the s–th row of the vector $P(\mu)w(\mu)$.

Now, consider the term $k = 2$. Upon conditioning on S_1 and using the fact that the transition probabilities are Markov, we have

$$\begin{aligned} \mathbb{E}\left[\Lambda(S_2, \mu(S_2)) \,|\, S_0 = s, \mu\right] &= \sum_{t \in \mathcal{S}} \left[\Lambda(t, \mu(t)) \cdot \Pr(S_2 = t \,|\, S_0 = s, \mu)\right] \\ &= \sum_{t \in \mathcal{S}} \left[\Lambda(t, \mu(t)) \sum_{t' \in \mathcal{S}} \Pr(S_2 = t \,|\, S_1 = t', \mu) \cdot \Pr(S_1 = t' \,|\, S_0 = s, \mu)\right] \\ &= \sum_{t \in \mathcal{S}} \left[\Lambda(t, \mu(t)) \sum_{t' \in \mathcal{S}} p_{t',t}(\mu(t')) \cdot p_{s,t'}(\mu(s))\right]. \end{aligned} \tag{29}$$

Observe that

$$\sum_{t' \in \mathcal{S}} p_{t',t}(\mu(t')) \cdot p_{s,t'}(\mu(s)) = [P^2(\mu)]_{s,t}.$$

Hence, we can write (29) as

$$\mathbb{E}\left[\Lambda(S_2, \mu(S_2)) \,|\, S_0 = s, \mu\right] = [P^2(\mu)w(\mu)]_s.$$

Continuing in this fashion, one can prove that

$$\mathbb{E}\left[\Lambda(S_k, \mu(S_k)) \,|\, S_0 = s, \mu\right] = [P^k(\mu)w(\mu)]_s \quad \text{for } k = 1, 2, \ldots.$$

Substituting this into (27), we obtain

$$J_0(s, \mu) = \Lambda(s, \mu(s)) + \sum_{k=1}^{\infty} \gamma^k [P^k(\mu)w(\mu)]_s \quad \text{for all } s \in \mathcal{S}. \tag{30}$$

Byletting J_0^μ to be the $|S|$–dimensional column vector whose s–th entry is $J_0(s,\mu)$, where $s \in \mathcal{S}$, we can rewrite (30) as

$$J_0^\mu = w(\mu) + \sum_{k=1}^{\infty} \gamma^k P^k(\mu) w(\mu) = w(\mu) + \gamma P(\mu) \underbrace{\sum_{k=1}^{\infty} \gamma^{k-1} P^{k-1}(\mu) w(\mu)}_{=J_0^\mu} = w(\mu) + \gamma P(\mu) J_0^\mu. \quad (31)$$

From (31), we can derive two useful identities. First, upon expanding (31) and recalling the definitions of J_0^μ and $w(\mu)$, we obtain

$$\boxed{J_0(s,\mu) = \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \left[p_{s,t}(\mu(s)) \cdot J_0(t,\mu)\right] \quad \text{for all } s \in \mathcal{S}.} \quad (32)$$

Secondly, we can compute J_0^μ via

$$\boxed{J_0^\mu = (I - \gamma P(\mu))^{-1} w(\mu).} \quad (33)$$

The above identities allow us to finish the description of the policy iteration algorithm.

Algorithm 2 Policy Iteration Algorithm for Finding J_0

1: **initialization**: select an arbitrary policy μ^0 and set $N = 1$
2: **repeat**
3: compute $P(\mu^{N-1})$ and $w(\mu^{N-1})$
4: compute $J_0^{\mu^{N-1}}$ via (33)
5: solve

$$\mu^{N+1}(s) = \arg\min_{\mu(s)} \left\{ \Lambda(s,\mu(s)) + \gamma \sum_{t\in\mathcal{S}} \left[p_{s,t}(\mu(s)) \cdot J_0(t,\mu^N)\right] \right\} \quad \text{for each } s \in \mathcal{S}$$

6: set $N \leftarrow N + 1$
7: **until** $\mu^N(s) = \mu^{N+1}(s)$ for all $s \in \mathcal{S}$

Again, let us illustrate the policy iteration algorithm by applying it to Example 1.

Example 2 *Using the data in Example 1, we run the policy iteration algorithm as follows:*

Initialization. *Set* $\mu^0(1) = \mu^0(2) = x^1$.

Iteration 1. *We compute*

$$P(\mu^0) = \begin{bmatrix} 3/4 & 1/4 \\ 3/4 & 1/4 \end{bmatrix}, \quad w(\mu^0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad J_0^{\mu^0} = \begin{bmatrix} 17.75 \\ 16.75 \end{bmatrix}.$$

Since

$$\min_{x\in\{x^1,x^2\}} \left\{\Lambda(1,x) + 0.9\left(p_{1,1}(x)J_0(1,\mu^0) + p_{1,2}(x)J_0(2,\mu^0)\right)\right\} = \min\{17.75, 15.8\} = 15.8,$$

we have $\mu^1(1) = x^2$. *Similarly, since*

$$\min_{x\in\{x^1,x^2\}} \left\{\Lambda(2,x) + 0.9\left(p_{2,1}(x)J_0(1,\mu^0) + p_{2,2}(x)J_0(2,\mu^0)\right)\right\} = \min\{16.75, 18.3\} = 16.75,$$

we have $\mu^1(2) = x^1$.

Iteration 2. *We compute*

$$P(\mu^1) = \begin{bmatrix} 1/4 & 3/4 \\ 3/4 & 1/4 \end{bmatrix}, \quad w(\mu^1) = \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}, \quad J_0^{\mu^1} = \begin{bmatrix} 7.32759 \\ 7.67241 \end{bmatrix}.$$

Since

$$\min_{x\in\{x^1,x^2\}} \left\{\Lambda(1,x) + 0.9\left(p_{1,1}(x)J_0(1,\mu^1) + p_{1,2}(x^1)J_0(2,\mu^1)\right)\right\} = \min\{17.75, 15.8\} = 15.8,$$

we have $\mu^2(1) = x^2$. *Also,*

$$\min_{x\in\{x^1,x^2\}} \left\{\Lambda(2,x) + 0.9\left(p_{2,1}(x)J_0(1,\mu^1) + p_{2,2}(x)J_0(2,\mu^1)\right)\right\} = \min\{16.75, 18.3\} = 16.75,$$

and hence $\mu^2(2) = x^1$. *Since* $\mu^2(1) = \mu^1(1)$ *and* $\mu^2(2) = \mu^1(2)$, *the algorithm terminates.*

As in our investigation of the value iteration algorithm, we need to establish the correctness of the above policy iteration algorithm. Towards that end, consider the sequence of policies $\mu^0, \mu^1, \ldots$ and the associated cost vectors $J_0^{\mu^0}, J_0^{\mu^1}, \ldots$ generated by the algorithm. It can be shown that $J_0^{\mu^N} \geq J_0^{\mu^{N+1}}$ for all $N \geq 0$, i.e.,

$$J_0(s,\mu^N) \geq J_0(s,\mu^{N+1}) \quad \text{for all } s \in \mathcal{S} \text{ and } N \geq 0; \tag{34}$$

see Problem 2 of Homework 4. Since $J_0(s,\mu^N) \geq J_0(s) = \min_\mu J_0(s,\mu)$, the sequence $\{J_0^{\mu^N}\}$ will converge. Moreover, if there is only a finite number of states and controls, then the number of different policies is finite. In this case, (34) implies that the policy iteration algorithm will eventually settle on a policy, and hence the algorithm will terminate.

6 The Linear Programming Approach

As it turns out, besides the two iterative approaches mentioned in the previous sections, one can also solve (23) via linear programming. Specifically, it can be shown that J_0 is the optimal solution to the following linear program:

$$\begin{aligned} \text{maximize} \quad & \sum_{s\in\mathcal{S}} J(s) \\ \text{subject to} \quad & J(s) \leq \Lambda(s,x) + \gamma \sum_{t\in\mathcal{S}} p_{s,t}(x)J(t) \qquad \text{for all } s \text{ and } x. \end{aligned}$$

As an illustration, consider Example 1 again. Since $\mathcal{S} = \{1,2\}$ and $\mathcal{C} = \{x^1, x^2\}$, the linear program corresponding to the problem of finding J_0 is given by

$$\begin{aligned} \text{maximize} \quad & J(1) + J(2) \\ \text{subject to} \quad & J(1) \leq \Lambda(1,x^1) + \gamma\left(p_{1,1}(x^1)J(1) + p_{1,2}(x^1)J(2)\right), \\ & J(1) \leq \Lambda(1,x^2) + \gamma\left(p_{1,1}(x^2)J(1) + p_{1,2}(x^2)J(2)\right), \\ & J(2) \leq \Lambda(2,x^1) + \gamma\left(p_{2,1}(x^1)J(1) + p_{2,2}(x^1)J(2)\right), \\ & J(2) \leq \Lambda(2,x^2) + \gamma\left(p_{2,1}(x^2)J(1) + p_{2,2}(x^2)J(2)\right). \end{aligned}$$