Chapter 8

Stochastic Shortest Path Problems

[1]In this chapter, we study a stochastic version of the shortest path problem of chapter 2, where only probabilities of transitions along different arcs can be controlled, and the objective is to minimize the expected length of the path. We discuss Bellman's equation, value and policy iteration, for the case of a finite state space. Technical difficulties arise because the DP operator is not necessarily a contraction as in the discounted cost problems of chapter 6.

8.1 Problem Formulation

The stochastic shortest path problem of this chapter is an infinite horizon problem where

1. There is no discounting, i.e., $\alpha = 1$ in the notation of chapter 6.

2. There is a special absorbing state, denoted t, representing the "destination": $p_{tt}(u) = 1, \forall u$.

3. The state space $\mathsf{X} = \{1, \dots, n, t\}$ and the control constraint sets $\mathsf{U}(i)$ are finite.

4. The destination state is cost-free: $c(t, u) = 0, \forall u$.

Clearly the cost of any policy starting from t is 0, hence we can omit t from the notation and define the vectors $J, c_\mu \in \mathbb{R}^n$ and $P_\mu \in \mathbb{R}^{n \times n}$ as in section 7.1. The difference however is that P_μ is substochastic instead of being stochastic. That is

$$\sum_{j=1}^{n} p_{ij}(u) = 1 - p_{it}(u) \leq 1,$$

[1] This version: November 22 2009.

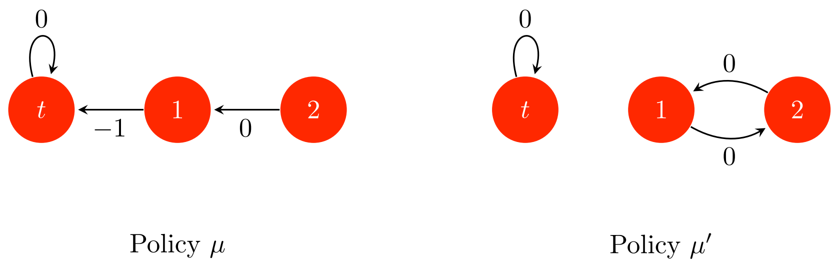

Figure 8.1: Transition diagrams and costs for two policies in a stochastic shortest path problem. The costs are written next to the arrows.

with strict inequality if $p_{it}(u) > 0$. The operators T, T_μ are defined as in chapter 6 with $\alpha = 1$, i.e.,

$$\begin{aligned} T_\mu J &= c_\mu + P_\mu J \\ TJ(i) &= \min_{u \in \mathsf{U}(i)} \left\{ c(i,u) + \sum_{j=1}^{n} p_{ij}(u) J(j) \right\}. \end{aligned}$$

Bellman's Equation Need Not Hold for SSP

In general, even under the assumptions above, the DP operator for SSP is not a contraction. Without additional assumptions, the theory developed in chapter 6 for discounted cost problems runs into difficulties. Consider a problem with state space $\{1, 2, t\}$ and two policies with transition probabilities and costs as given on Fig. 8.1. The equation $J = TJ$ can be written

$$\begin{aligned} J(1) &= \min\{-1, J(2)\} \\ J(2) &= J(1), \end{aligned}$$

and is satisfied for any J of the form $J(1) = J(2) = \delta$ with $\delta \leq -1$. The policy μ is clearly optimal with cost vector $J(1) = J(2) = -1$. The difficulty arises because the policy μ' starting from 1 or 2 never reaches the final state yet has a finite (zero) cost.

8.2 DP Theory

In view of the example of section 8.1, consider the following definition

Definition 8.2.1. A stationary policy μ is said to be *proper* if, when using this policy, there is a positive probability that the destination will be reached after at most n stages, regardless of the initial state. That is

$$\rho_\mu = \max_{i=1,\ldots,n} \mathbb{P}_\mu(x_n \neq t | x_0 = i) < 1. \tag{8.1}$$

A stationary policy that is not proper is said to be *improper*.

It is not hard to see that μ is a proper policy if and only if in the transition graph of the corresponding Markov chain, each state is connected to the destination by some path. We make the following assumptions.

Assumption 8.2.1. There exists at least one proper policy.

Assumption 8.2.2. For every improper policy μ, the corresponding cost $J_\mu(i)$ is infinite for at least one state i. That is, some component of $\sum_{k=0}^{N-1} P_\mu^k c_\mu$ diverges as $N \to \infty$.

In the case of a deterministic shortest path problem, assumption 8.2.1 is satisfied if and only if every node is connected to the destination by some path, and assumption 8.2.2 is satisfied if and only if each cycle that does not contain the destination has positive length. Assumption 8.2.2 is satisfied for example if $c(i,u)$ is positive for all $i \neq t$ and $u \in \mathsf{U}(i)$. Another important case where the assumptions are verified is when all policies are proper.

Note that if a policy is proper, then almost surely the destination must be reached after a finite number of steps. This is because a consequence of 8.1 is that

$$\begin{aligned}\mathbb{P}_\mu(x_{2n} \neq t | x_0 = i) &= \mathbb{P}_\mu(x_{2n} \neq t | x_n \neq t, x_0 = i) P_\mu(x_n \neq t | x_0 = i) \\ &\leq \rho_\mu^2,\end{aligned}$$

and more generally

$$\mathbb{P}_\mu(x_k \neq t | x_0 = i) \leq \rho_\mu^{\lfloor k/n \rfloor}, \;\; i = 1, \ldots, n. \tag{8.2}$$

We can then invoke the Borel-Cantelli lemma to conclude. Moreover under a proper policy the cost is finite, since

$$|J_\mu(i)| \leq \lim_{N \to \infty} \sum_{k=0}^{N-1} \rho_\mu^{\lfloor k/n \rfloor} \max_{j=1,\ldots,n} |c(j, \mu(j))| < \infty. \tag{8.3}$$

Under assumptions 8.2.1 and 8.2.2, the dynamic programming theory is similar to the one for the discounted cost problems of chapter 6. Note that here the constant ρ_μ in a sense plays the role of the discount factor, see (8.3). Note in particular that (8.2) implies that for all J, we have

$$\lim_{k \to \infty} P_\mu^k J = 0.$$

Proposition 8.2.1. *1. For a proper policy μ, the associated cost vector satisfies $\lim_{k\to\infty} T_\mu^k J = J_\mu$, for every vector J, and J_μ is the unique solution of the equation $J_\mu = T_\mu J_\mu$. Moreover, a policy μ such that $J \geq T_\mu J$ is satisfied for some vector J is proper.*

2. *The optimal cost vector is the unique solution of Bellman's equation $J^* = TJ^*$, and we have $\lim_{k\to\infty} T^k J = J^*$ for every vector J. A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$.*

Proof. We have

$$T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m c_\mu,$$

and so

$$\lim_{k\to\infty} T_\mu^k J = \lim_{k\to\infty} \sum_{m=0}^{k-1} P_\mu^m c_\mu = J_\mu,$$

by (8.2) (the fact that the limit exists is also a consequence of (8.2)). Then taking the limit as $k \to \infty$ in

$$T_\mu^{k+1} J = c_\mu + P_\mu T_\mu^k J$$

we obtain $J_\mu = T_\mu J_\mu$. Finally for uniqueness, suppose $J = T_\mu J$, then $J = T_\mu^k J$ for all k and letting $k \to \infty$ we get $J = J_\mu$. Now consider a policy μ for which $J \geq T_\mu J$ for some vector J. Then for all k

$$J \geq T_\mu^k J = P_\mu^k J + \sum_{m=0}^{k-1} P_\mu^m c_\mu,$$

so no component of the sum can diverge and so μ is not improper (hence proper) by our assumption 8.2.2.

Next consider the second statement. Note first that because the control sets are finite, for any vector J we can find a policy μ such that $TJ = T_\mu J$. Suppose Bellman's equation has two solutions J and J', with corresponding policies μ and μ' such that $J = T_\mu J$ and $J' = T_{\mu'} J'$. Then by the first statement μ and μ' must be proper, and so $J = J_\mu$ and $J' = J_{\mu'}$. Then $J = T^k J \leq T_{\mu'}^k J$ for all k so $J \leq J_{\mu'} = J'$. Similarly $J' \leq J$ so T has at most one fixed point.

Now we show the existence of a fixed point for T. By assumption 8.2.1 there is at least one proper policy μ_0, with cost J_{μ_0}. We use this policy to start policy iteration. Assume that at stage k we have a proper policies $\mu_0, \ldots, \mu_k$ such that

$$J_{\mu_0} \geq TJ_{\mu_0} \geq J_{\mu_1} \geq \ldots \geq J_{\mu_{k-1}} \geq TJ_{\mu_{k-1}} \geq J_{\mu_k}. \tag{8.4}$$

Then we choose μ_{k+1} such that

$$TJ_{\mu_k} = T_{\mu_{k+1}} J_{\mu_k}.$$

Hence $T_{\mu_{k+1}} J_{\mu_k} \leq T_{\mu_k} J_{\mu_k} = J_{\mu_k}$ so μ_{k+1} is proper by the first statement. Also by monotonicity of $T_{\mu_{k+1}}$ we have

$$J_{\mu_{k+1}} = \lim_{k\to\infty} T_{\mu_{k+1}}^k J_{\mu_k} \leq T_{\mu_{k+1}} J_{\mu_k} = TJ_{\mu_k} \leq J_{\mu_k},$$

which completes the induction. Since the set of proper policies is finite, some policy μ must be repeated within the sequence $\{\mu_k\}$. By the inequalities (8.4) we get $J_\mu = TJ_\mu$ for this policy.

We still have to show that the unique fixed point J_μ of T just constructed is the optimal cost vector J^*, and that $T^k J \to J_\mu = J^*$ for all J. We start with the second assertion. Let $\delta > 0, e = [1, \ldots, 1]^T$ and $\hat{J}$ be the vector satisfying

$$\hat{J} = T_\mu \hat{J} + \delta e.$$

There is a unique such vector $\hat{J}$, which is the cost of the proper policy μ with the cost c_μ replaced by $c_\mu + \delta e$. This interpretation also implies $J_\mu \le \hat{J}$ and so

$$J_\mu = TJ_\mu \le T\hat{J} \le T_\mu \hat{J} = \hat{J} - \delta e \le \hat{J}.$$

Using the monotonicity of T, we have then

$$J_\mu = T^k J_\mu \le T^k \hat{J} \le T^{k-1}\hat{J} \le \hat{J}, \ \ k \ge 1.$$

Hence the sequence $T^k \hat{J}$ converges to some vector $\tilde{J}$. We can see that the mapping T is continuous, so

$$T\tilde{J} = T(\lim_{k\to\infty} T^k \hat{J}) = \lim_{k\to\infty} T^{k+1}\hat{J} = \tilde{J}.$$

By the uniqueness of the fixed point of T, we conclude $\tilde{J} = J_\mu$. Now take any vector J. Recalling the interpretation of $\hat{J}$ above, we can always find $\delta > 0$ such that $J \le \hat{J}$ and in fact such that the following stronger condition holds

$$J_\mu - \delta e \le J \le \hat{J}. \tag{8.5}$$

We already know $\lim_{k\to\infty} T^k \hat{J} = J_\mu$. Moreover, since $P_{\mu'} e \le e$ for any policy μ', we have $T(J' - \delta e) \ge TJ' - \delta e$ for any vector J'. Thus

$$J_\mu - \delta e = TJ_\mu - \delta e \le T(J_\mu - \delta e) \le TJ_\mu = J_\mu.$$

Hence $T^k(J_\mu - \delta e)$ is monotonically increasing and bounded above, so as before we conclude that $\lim_{k\to\infty} T^k(J_\mu - \delta e) = J_\mu$. Finally by monotonicity of T and (8.5) we obtain $\lim_{k\to\infty} T^k J = J_\mu$. We can now show that $J_\mu = J^*$. Consider a policy $\pi = \{\mu_0, \mu_1, \ldots\}$, and let J_0 be the zero vector. Then

$$T_{\mu_0} T_{\mu_1} \ldots T_{\mu_{k-1}} J_0 \ge T^k J_0,$$

and taking the $\limsup$ on both sides we obtain $J_\pi \ge J_\mu$, hence $J_\mu = J^*$.

Finally if μ is optimal, then $J_\mu = J^*$ and by our assumptions μ must be proper. Hence

$$T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*.$$

Conversely if $TJ^* = T_\mu J^*$, since we also have $J^* = TJ^*$ it follows from the first statement that μ is proper. By unicity of the fixed point of T_μ we get $J_\mu = J^*$ so μ is optimal. □

8.3 Value and Policy Iteration

Value Iteration

The convergence of the value iteration algorithm is shown in proposition 8.2.1. Moreover, there are error bounds similar to the ones discussed in section 7.1 for discounted cost problems, and the Gauss-Seidel value iteration method works and typically converges faster than the ordinary value iteration method.

Moreover, there are certain SSP problems for which we have better convergence results. Note for example that for deterministic shortest path problems, value iteration terminates in a finite number of step. More generally, let us assume that the transition graph corresponding to some optimal policy μ^* is acyclic. This requires in particular that there are no positive self-transitions $p_{ii}(\mu^*(i)) > 0$ for $i \neq t$. This last property can always be achieved however, as follows. Define a new SSP problem with transition probabilities

$$\tilde{p}_{ij}(u) = \begin{cases} 0 & \text{if } j = i, \\ \frac{p_{ij}(u)}{1-p_{ii}(u)} & \text{if } j \neq i, \end{cases}$$

for $i = 1, \ldots, n$ and costs $\tilde{g}(i,u) = g(i,u)/(1 - p_{ii}(u))$, $i = 1, \ldots, n$. This new SSP is equivalent to the original one in the sense that it has the same optimal costs and policies. Moreover it satisfies the assumption $\tilde{p}_{ii}(\mu^*(i)) = 0$.

Under the preceding acyclicity assumption, the value iteration method yields J^* after at most n iterations when started from the vector J such that $J(i) = \infty$ for all $i = 1, \ldots, n$. Hence such a vector could be a good choice to start VI in any case, even if the acyclicity property is not clear. Indeed consider the sets of states

$$\begin{aligned} S_0 &= \{t\}, \\ S_{k+1} &= \left\{ i \;\middle|\; p_{ij}(\mu^*(i)) = 0, \text{ for all } j \notin \cup_{m=0}^{k} S_m \right\}, k = 0, 1, \ldots \end{aligned}$$

Hence S_{k+1} is the set of states which from which we go to $\cup_{m=0}^{k} S_m$ at the next step with probability 1. A consequence of the acyclicity assumption is that the sets are non empty until we reach $\bar{k}$ such that $\cup_{m=0}^{\bar{k}} S_m = \{1, \ldots, n, t\}$. Indeed suppose $S_k = \emptyset$ for $k < \bar{k}$, and take $i \notin \mathcal{S} := \cup_{m=0}^{k-1} S_m$. Then there is an edge starting from i that does not enter $\mathcal{S}$. Follow this edge to reach a new state $i_1 \notin \mathcal{S}$, and repeat, until you come back to i, contradiction (note that because an optimal policy is proper, the path constructed cannot end at some state different from t). Then we show by induction that the value iteration method starting with the vector $J = \infty$ above yields $T^k J(i) = J^*(i)$ for all $i \in \cup_{m=0}^{k} S_m, i \neq t$. This is vacuously true for $k = 0$. Then we have by the monotonicity of T that $J^* \leq T^{k+1} J$, and for all $i \in \cup_{m=0}^{k+1} S_m$

$$\begin{aligned} T^{k+1} J(i) &\leq c(i, \mu^*(i)) + \sum_{j \in \cup_{m=0}^{k} S_m} p_{ij}(\mu^*(i)) J^*(j) \\ &= J^*(i). \end{aligned}$$

This completes the induction. Hence in this case value iteration gives the optimal costs for all states after $\bar{k}$ iterations.

Policy Iteration

Policy iteration and approximate policy iteration work similarly to the discounted cost case.