Dynamic Programming Lecture #14

Outline:

- Stochastic Shortest Path Example
- Policy Iteration

SSP Recap

- System: Controlled finite-state Markov chain with transition probabilities $p_{ij}(u)$.
- Key assumption: Termination state t & horizon window $[0, m-1]$ s.t. for any initial $x_0 = i$:

$$Pr(x_m \neq t | x_0 = i) \leq \rho < 1$$

- Cost:

$$E\left\{\sum_{k=0}^{\infty} g(x_k, \mu_k(x_k))\right\}$$

- Value iteration: Initialize with any $J_0(i)$. Iterate

$$J_{k+1}(i) = \min_{u \in U(i)} g(i,u) + \sum_{j=1}^{n} p_{ij}(u) J(j)$$

- Main results:
  - $J_k(i) \rightarrow J^*(i)$
  - J^* is unique solution to Bellman equation:

$$J^*(i) = \min_{u \in U(i)} g(i,u) + \sum_{j=1}^{n} p_{ij}(u) J^*(j)$$

Stationary Policy Derivation

- Let J_μ denote the cost of the STATIONARY (stage-independent) policy:

$$\pi = \{\mu, \mu, \mu, \ldots\}$$

- Main results:
  - J_μ can be computed by "μ-specific" value iteration:

$$J_{k+1}(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) J_k(j)$$

  - J_μ is unique solution to "μ-specific" Bellman equation:

$$J_\mu(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) J_\mu(j)$$

  - A stationary policy is optimal $\Leftrightarrow$

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

    * Proof ($\Leftarrow$): If μ achieves the minimum, then $J_\mu = J^*$ (apply above result).
    * Proof ($\Rightarrow$): If $J^* = J_\mu$, then

$$\begin{aligned} J_\mu(i) &= g(i, \mu(i)) + \sum_j p_{ij}(\mu(i)) J^*(j) \\ &= J^*(i) \\ &= \min_{u \in U(i)} \left\{ g(i,u) + \sum_j p_{ij}(u) J^*(j) \right\} \end{aligned}$$

Example: Random Pursuit

- Consider fly & spider hoping along string of nodes.
- Fly movement:
  - Left with probability p.
  - Right with probability p.
  - Stationary with probability $1 - 2p$.
- Spider movement:
  - Towards fly if distance ≥ 2.
  - Towards fly or still (decision) if distance $= 1$.
- State, i: Distance between spider and fly (finite!). $i = 0 \Rightarrow$ termination.
- Let initial distance $= n$.
- Optimization: Minimize time to termination $\Rightarrow g(i, u) = 1$.

Example, cont (2)

- For $i \geq 2$:
  - $p_{ii} = p$ (fly moves away)
  - $p_{i(i-2)} = p$ (fly moves towards)
  - $p_{i(i-1)} = 1 - 2p$ (fly stationary)
- For $i = 1$, probabilities depend on spider decision: $\begin{cases} M & \text{move} \\ S & \text{still} \end{cases}$.
  - $p_{11}(M) = 2p$ (fly moves towards or away)
  - $p_{10}(M) = 1 - 2p$ (fly stationary)
  - $p_{10}(S) = p$ (fly moves towards)
  - $p_{11}(S) = 1 - 2p$ (fly stationary)
  - $p_{12}(S) = p$ (fly moves away)
- For $n = 3$:

$$P(M) = \begin{pmatrix} 2p & 0 & 0 \\ 1-2p & p & 0 \\ p & 1-2p & p \end{pmatrix}, \quad P(S) = \begin{pmatrix} 1-2p & p & 0 \\ 1-2p & p & 0 \\ p & 1-2p & p \end{pmatrix}$$

Example, cont (3)

- Apply DP:

$$J_{k+1}(i) = \min_u \left(1 + \sum p_{ij}(u) J_k(j) \right)$$

  - For $i \geq 2$:

$$J_{k+1}(i) = 1 + pJ_k(i) + pJ_k(i-2) + (1-2p)J_k(i-1)$$

  - For $i = 1$:

$$J_{k+1}(1) = 1 + \min_{M/S} \{2pJ_k(1), (1-2p)J_k(1) + pJ_k(2)\}$$

- Bellman equation:

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

$$\Rightarrow$$

$$J^*(2) = 1 + pJ^*(2) + pJ^*(0) + (1-2p)J^*(1)$$

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

Solve for $J^*(2)$ above $\Rightarrow$

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

Example, cont (4)

- Which term is smaller? Suppose first term (M):

$$J^*(1) = 1 + 2pJ^*(1) \Rightarrow J^*(1) = \frac{1}{1-2p}$$

Check whether

$$2pJ^*(1) \leq \frac{1-2p}{1-p} J^*(1) + \frac{p}{1-p}$$
$$\Leftrightarrow$$
$$p \leq 1/3$$

$$\Rightarrow$$

$$\mu^*(1) = \begin{cases} M & p \leq 1/3 \\ S & p \geq 1/3 \end{cases}$$

Policy Iteration

- Very important alternative to value iteration.
- Main steps:
  1. Pick any policy, μ.
  2. Compute J_μ.
  3. Policy improvement step:

$$\mu^+(i) = \arg \min_{u \in U(i)} g(i,u) + \sum_{j=1}^{n} p_{ij}(u) J_\mu(j)$$

  4. Repeat 2–4.
- How to compute J_μ?
  - Perform μ-specific value iteration.
  - Compute directly from μ-specific Bellman equation as a large linear system of equations:

$$\underbrace{J_\mu(i)}_{\text{unknown}} = g(i,u) + \sum_{j=1}^{n} p_{ij}(\mu(i)) \underbrace{J_\mu(i)}_{\text{unknown}}$$

Policy Iteration, cont

- Key issue: How does J_{μ^+} compare with J_μ?
- FACT: $J_{\mu^+}(i) \leq J_\mu(i)$.
- PROOF: Consider μ^+-specific value iteration with J_μ initial cost:

$$J_\mu^+ \hookleftarrow \ldots \hookleftarrow \tilde{J}_2 \hookleftarrow \tilde{J}_1 \hookleftarrow \tilde{J}_\mu$$

where

$$\begin{aligned} \tilde{J}_1 &= \min_u g(i,u) + \sum_j p_{ij}(u) J_\mu(j) \\ &\stackrel{\text{def}}{=} g(i,\mu^+(i)) + \sum_j p_{ij}(\mu^+(i)) J_\mu(j) \\ &\leq J_\mu(i) \end{aligned}$$

So $\tilde{J}_1 \leq J_\mu \Rightarrow J_{\mu^+} \leq J_\mu$ by monotonicity.

- FACT: Eventually $J_{\mu^+} = J_\mu \Rightarrow J_\mu = J^*$. (Proof: There are only a finite number of policies.)

Example: Spider/Fly

- Transition probabilities:

$$p_{ii} = p, \quad p_{i(i-1)} = 1 - 2p, \quad p_{i(i-2)} = p, \quad i \geq 2$$
$$p_{11}(M) = 2p, \quad p_{10}(M) = 1 - 2p$$
$$p_{12}(S) = p, \quad p_{11}(S) = 1 - 2p, \quad p_{10}(S) = p$$

- There are only 2 policies to compare: $\mu(1) = M$ or S.
- Compute J_M via M-specific Bellman equation:

$$J_M(1) = 1 + p_{11}(M)J_M(1) + p_{10}(M)J_M(0)$$
$$\Rightarrow$$
$$J_M(1) = \frac{1}{1-2p}$$

since $J_M(0) = 0$. Similarly,

$$J_M(2) = 1 + p_2 2 J_M(2) + p_{21}J_M(1) + p_{20}J_M(0)$$
$$\Rightarrow$$
$$J_M(2) = \frac{2}{1-p}$$

Example, cont (2)

- Policy improvement:

$$\mu^+(i) = \arg\min g(i,u) + \sum_j p_{ij}(u) J_\mu(j)$$

$$\Rightarrow$$

$$\begin{aligned}
\mu^+(1) &= \arg\min_{u=M/S} \begin{cases} 1 + pJ_\mu(2) + (1-2p)J_\mu(1) + pJ_\mu(0), & u = S \\ 1 + 2pJ_\mu(1) + (1-2p)J_\mu(0), & u = M \end{cases} \\
&= \arg\min_{u=M/S} \begin{cases} 1 + p\frac{2}{1-p} + (1-2p)\frac{1}{1-2p} + p \cdot 0, & u = S \\ 1 + 2p\frac{1}{1-2p} + (1-2p) \cdot 0, & u = M \end{cases} \\
&= \arg\min_{u=M/S} \begin{cases} 2 + \frac{2p}{1-p}, & u = S \\ 1 + \frac{2p}{1-2p}, & u = M \end{cases}
\end{aligned}$$

  Must compare

$$\underbrace{\frac{2}{1-p}}_{u=S} \leq \underbrace{\frac{1}{1-2p}}_{u=M}$$

  Change to $\mu + (1) = S$ if $p \geq 1/3$.

- Repeat for J_S, etc.