Dynamic Programming Lecture #13

Outline:

- Infinite Horizon Preview
- Stochastic Shortest Path
- Bellman equation

Infinite Horizon

- System:

$$x_{k+1} = f_k(x_k, u_k, w_k)$$

- New cost:

$$\min E\left\{\sum_{k=0}^{\infty} g_k(x_k, u_k, w_k)\right\}$$

- Why deal with ∞ horizon?
  - Know game is finite, but don't know # stages.
  - ∞ horizon set-up may be simplifying!
- Important assumption: Make all terms stage-independent:

$$x^+ = f(x, u, w)$$
$$\min E\left\{\sum_{k=0}^{\infty} g(x_k, u_k, w_k)\right\}$$

Problems with Infinity

- Example:

$$x^+ = Ax + Bu + Lw$$
$$\min \sum_{k=0}^{\infty} x^T Q x + u^T u$$

Issue: Cost is infinite!

- Ways around ∞:
  - Stochastic shortest path

$$\text{cost} = E\left\{\sum_{k=0}^{\infty} g(x_k, u_k, w_k)\right\}$$

  but we WILL terminate.

  - Discounted cost:

$$\begin{aligned}\text{cost} &= E\left\{\sum_{k=0}^{\infty} \alpha^k g(x_k, u_k, w_k)\right\} \\ &= E\left\{g(x_0, u_0, w_0) + \alpha g(x_1, u_1, w_1) + \alpha^2 g(x_2, u_2, w_2) + \ldots\right\}\end{aligned}$$

  with $0 < \alpha < 1 \Rightarrow$ far future doesn't matter.

  - Average cost:

$$\text{cost} = \lim_{N\to\infty} \frac{1}{N} E\left\{\sum_{k=0}^{\infty} g(x_k, u_k, w_k)\right\}$$

  now near future doesn't matter.

  - Specialized analysis and monotonicity.

Expectations

- Policy is stage-independent

$$\{\mu_0, \mu_1, \mu_2, \ldots\} \text{ vs } \{\mu, \mu, \mu, \ldots\}$$

  Why whould 1,000,000 stages to go differ from 1,000,001 stages to go?

- DP recursions (new notation):

$$J_{k+1} = \min_u E\left\{g(x,u,w) + J_k(f(x,u,w))\right\}$$

  Before:

$$J_0 \leftarrow J_1 \ldots J_{N-1} \leftarrow J_N$$

  Now:

$$J_\infty(?) \leftarrow \ldots J_2 \leftarrow J_1 \leftarrow J_0$$

  Expect $J_k \rightarrow J^*$ for ANY J_0.

- New Bellman equation:

$$J^*(x) = \min_u E\left\{g(x,u,w) + J^*(f(x,u,w))\right\}$$

  i.e., a "fixed point" of DP iterations.

Set-up: Controlled Finite State Markov Chains

- State-space: $X = \{1, 2, \ldots, n\}$.
- Controls: For $i \in X$, must use $u \in U(i)$, where $U(i)$ is finite set.
- Transistion probabilities: $p_{ij}(u)$ or $P(u)$.
- Notation: $\pi = \{\mu_0, \mu_1, \mu_2, \ldots\}$.
- Infinite horizon cost:

$$\min_{\pi} \lim_{N \to \infty} E \left\{ \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k)) | x_0 = i \right\}$$

- Note: We do NOT write $\sum_0^\infty$
  1. Commit to π.
  2. Take limit.

Stochastic Shortest Path

- Assume a cost free termination state: t
  - $p_{tt}(u) = 1$
  - $p_{ti}(u) = 0$
  - $g(t, u) = 0$
- ASSUME: There exists an m s.t. for any π:

$$\rho_\pi = \max_{i=1,\ldots,n} Pr\,(\, x_m \neq t | x_o = i, \pi\,) < 1$$

  i.e., after m-steps, there is a non-zero probability that we will terminate.

- Fact: Since $U(i)$ is always a finite set, there is a finite number of policies over $[0, m-1]$, so

$$\max_\pi \rho_\pi = \rho < 1$$

- Assumption assures that probability of continuation decay exponentially.
  - What is $Pr\,(\, x_m \neq t \;\&\; x_{2m} \neq t\,)$?

$$Pr(x_{2m} \neq t | x_m \neq t) P(x_m \neq t) \leq \rho^2$$

  - Similarly $Pr\,(\, x_{km} \neq t\,) \leq \rho^k$.

Bounding the Cost

- Let $J_\pi(i)$ = cost of policy π.
- Set

$$G = \max_{\substack{i=1,\ldots,n \\ u \in U(i)}} |g(i,u)|$$

  - $[0, m-1] : E\left\{\sum g\right\} \leq mG$
  - $[m, 2m-1] : E\left\{\sum g\right\} \leq mGPr(x_m \neq t) \leq \rho mG$
  - $[2m, 3m-1] : E\left\{\sum g\right\} \leq \rho^2 mG$

  Total cost bound:

$$|J_\pi(i)| \leq mG(1 + \rho + \rho^2 + \ldots) = mG\frac{1}{1-\rho}$$

Main Result: Value Iteration

- Note: Suppose $x = i$ and $x^+ = j$.

$$E_j\left\{F(j,u)|i\right\} = \sum_{j=1}^{n} p_{ij}(u) f(j,u)$$

- Theorem: For ANY starting $J_0(1), \ldots, J_0(n)$, the value iteration

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

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

- Furthermore, $J^*(i)$ satisfies the Bellman equation:

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

  i.e., $J^*(\cdot)$ is a fixed-point of value iteration.

Proof

- Divide time $[0, N-1]$ into intervals:

$$[0, m-1], [m, 2m-1], \ldots, [(K-1)m, Km-1], [Km, N-1]$$

- Cost of π:

$$\begin{aligned} J_\pi(x_0) &= \lim_{N\to\infty} E\left\{\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k))\right\} \\ &= E\left\{\sum_{k=0}^{Km-1} g(x_k, \mu_k(x_k))\right\} + \lim_{N\to\infty} E\left\{\sum_{k=Km}^{N-1} g(x_k, \mu_k(x_k))\right\} \end{aligned}$$

- Second term bound:

$$\left|E\left\{\sum_{k=Km}^{N-1} g(x_k, \mu_k(x_k))\right\}\right| \leq mG\frac{\rho^K}{1-\rho}$$

- Consider using $J_0(\cdot)$ as start of value iterations (terminal penalty). The diminishing role can be seen by:

$$|E\{J_0(x_{Km})\}| \leq \rho^K \max_i |J_0(i)|$$

Proof, cont (2)

- Approximate value of $J_\pi(x_0)$:

$$
\begin{aligned}
J_\pi(x_0) &= E\left\{\sum_{k=0}^{Km-1} g(x_k, \mu_k(x_k))\right\} \pm mG\frac{\rho^K}{1-\rho} \\
&\quad + E\left\{J_0(x_{Km})\right\} - E\left\{J_0(x_{Km})\right\} \\
&= E\left\{J_0(x_{Km}) + \sum_{k=0}^{Km-1} g(x_k, \mu_k(x_k))\right\} \\
&\quad \pm mG\frac{\rho^K}{1-\rho} \pm \rho^K \max_i |J_0(i)|
\end{aligned}
$$

- Minimize both sides over (different) π:

$$
J^*(x_0) = J_{Km}(x_0) \pm mG\frac{\rho^K}{1-\rho} \pm \rho^K \max_i |J_0(i)|
$$

so as $K \to \infty$:

$$
J_{Km}(x_0) \to J^*(x_0)
$$

- Could repeat analysis using any intervals of width $[0, m+q-1]$ to get desired result.

Proof, cont (3)

- Now to show $J^*(\cdot)$ satisfies Bellman equation:

$$J_{k+1}(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_j p_{ij}(u) J_k(j) \right\}$$

$$\lim_{k \to \infty} \Rightarrow$$

$$J^*(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_j p_{ij}(u) J_k(j) \right\}$$

- Issue: Is $J^*(\cdot)$ unique solution to Bellman equation?
- Given another solution, $\tilde{J}$, could start with $J_0 = \tilde{J}$ but still converge to $J^* \rightarrow \tilde{J} = J^*$.

Stationary Policy Derivation

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

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

  Then

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

  and $J_\mu(\cdot)$ can be computed as limit of

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

- How? Apply prior results with $U(i) = \mu(i)$.
- 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

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