6.231 Dynamic Programming and Stochastic Control
Fall 2008

6.231 DYNAMIC PROGRAMMING

LECTURE 19

LECTURE OUTLINE

- Undiscounted problems
- Stochastic shortest path problems (SSP)
- Proper and improper policies
- Analysis and computational methods for SSP
- Pathologies of SSP

UNDISCOUNTED PROBLEMS

- System: $x_{k+1} = f(x_k, u_k, w_k)$
- Cost of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$

$$J_\pi(x_0) = \lim_{N\to\infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k), w_k\big) \right\}$$

- Shorthand notation for DP mappings

$$(TJ)(x) = \min_{u\in U(x)} \mathop{E}_{w} \Big\{ g(x,u,w) + J\big(f(x,u,w)\big) \Big\}, \ \forall\, x$$

- For any stationary policy μ

$$(T_\mu J)(x) = \mathop{E}_{w} \Big\{ g\big(x,\mu(x),w\big) + J\big(f(x,\mu(x),w)\big) \Big\}, \ \forall\, x$$

- Neither T nor T_μ are contractions in general, but their monotonicity is helpful.
- SSP problems provide a "soft boundary" between the easy finite-state discounted problems and the hard undiscounted problems.
  - They share features of both.
  - Some of the nice theory is recovered because of the termination state.

SSP THEORY SUMMARY I

- As earlier, we have a cost-free term. state t, a finite number of states $1, \ldots, n$, and finite number of controls, but we will make weaker assumptions.

- Mappings T and T_μ (modified to account for termination state t):

$$(TJ)(i) = \min_{u \in U(i)} \left[g(i,u) + \sum_{j=1}^{n} p_{ij}(u) J(j) \right], \quad i = 1, \ldots, n,$$

$$(T_\mu J)(i) = g\big(i, \mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big) J(j), \quad i = 1, \ldots, n.$$

- **Definition:** A stationary policy μ is called **proper**, if under μ, from every state i, there is a positive probability path that leads to t.

- **Important fact:** If μ is proper, T_μ is contraction with respect to some weighted max norm

$$\max_i \frac{1}{v_i} |(T_\mu J)(i) - (T_\mu J')(i)| \leq \rho_\mu \max_i \frac{1}{v_i} |J(i) - J'(i)|$$

- T is similarly a contraction if all μ are proper (the case discussed in the text, Ch. 7, Vol. I).

SSP THEORY SUMMARY II

- The theory can be pushed one step further. Assume that:

  (a) There exists at least one proper policy

  (b) For each improper μ, $J_\mu(i) = \infty$ for some i

- Then T is not necessarily a contraction, but:
  - J^* is the unique solution of Bellman's Equ.
  - μ^* is optimal if and only if $T_{\mu^*} J^* = T J^*$
  - $\lim_{k\to\infty}(T^k J)(i) = J^*(i)$ for all i
  - Policy iteration terminates with an optimal policy, if started with a proper policy

- **Example:** Deterministic shortest path problem with a single destination t.
  - States <=> nodes; Controls <=> arcs
  - Termination state <=> the destination
  - Assumption (a) <=> every node is connected to the destination
  - Assumption (b) <=> all cycle costs > 0

SSP ANALYSIS I

- For a proper policy μ, J_μ is the unique fixed point of T_μ, and $T_\mu^k J \to J_\mu$ for all J (holds by the theory of Vol. I, Section 7.2)
- A stationary μ satisfying $J \geq T_\mu J$ for some J must be proper - true because

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

and some component of the term on the right blows up if μ is improper (by our assumptions).

- Consequence: T can have at most one fixed point.

Proof: If J and J' are two solutions, select μ and μ' such that $J = TJ = T_\mu J$ and $J' = TJ' = T_{\mu'} J'$. By preceding assertion, μ and μ' must be proper, and $J = J_\mu$ and $J' = J_{\mu'}$. Also

$$J = T^k J \leq T_{\mu'}^k J \to J_{\mu'} = J'$$

Similarly, $J' \leq J$, so $J = J'$.

SSP ANALYSIS II

- We now show that T has a fixed point, and also that policy iteration converges.
- Generate a sequence $\{\mu_k\}$ by policy iteration starting from a proper policy μ_0.
- μ_1 is proper and $J_{\mu_0} \geq J_{\mu_1}$ since

$$J_{\mu_0} = T_{\mu_0} J_{\mu_0} \geq T J_{\mu_0} = T_{\mu_1} J_{\mu_0} \geq T_{\mu_1}^k J_{\mu_0} \geq J_{\mu_1}$$

- Thus $\{J_{\mu_k}\}$ is nonincreasing, some policy μ will be repeated, with $J_\mu = T J_\mu$. So J_μ is a fixed point of T.
- Next show $T^k J \to J_\mu$ for all J, i.e., value iteration converges to the same limit as policy iteration. (Sketch: True if $J = J_\mu$, argue using the properness of μ to show that the terminal cost difference $J - J_\mu$ does not matter.)
- To show $J_\mu = J^*$, for any $\pi = \{\mu_0, \mu_1, \ldots\}$

$$T_{\mu_0} \cdots T_{\mu_{k-1}} J_0 \geq T^k J_0,$$

where $J_0 \equiv 0$. Take $\limsup$ as $k \to \infty$, to obtain $J_\pi \geq J_\mu$, so μ is optimal and $J_\mu = J^*$.

SSP ANALYSIS III

- If all policies are proper (the assumption of Section 7.1, Vol. I), T_μ and T are contractions with respect to a weighted sup norm.

Proof: Consider a new SSP problem where the transition probabilities are the same as in the original, but the transition costs are all equal to -1. Let $\hat{J}$ be the corresponding optimal cost vector. For all μ,

$$\hat{J}(i) = -1 + \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\hat{J}(j) \leq -1 + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big)\hat{J}(j)$$

For $v_i = -\hat{J}(i)$, we have $v_i \geq 1$, and for all μ,

$$\sum_{j=1}^{n} p_{ij}\big(\mu(i)\big)v_j \leq v_i - 1 \leq \rho\, v_i, \qquad i = 1, \ldots, n,$$

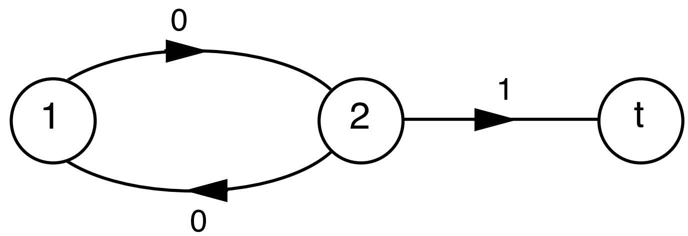
where

$$\rho = \max_{i=1,\ldots,n} \frac{v_i - 1}{v_i} < 1.$$

This implies contraction of T_μ and T by the results of the preceding lecture.

PATHOLOGIES I: DETERM. SHORTEST PATHS

- If there is a cycle with cost $= 0$, Bellman's equation has an infinite number of solutions. Example:



- We have $J^*(1) = J^*(2) = 1$.
- Bellman's equation is

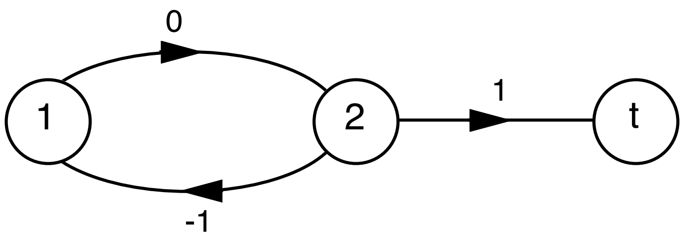
$$J(1) = J(2), \qquad J(2) = \min\big[J(1), 1\big].$$

- It has J^* as solution.
- Set of solutions of Bellman's equation:

$$\big\{J \mid J(1) = J(2) \leq 1\big\}.$$

PATHOLOGIES II: DETERM. SHORTEST PATHS

- If there is a cycle with cost < 0, Bellman's equation has no solution [among functions J with $-\infty < J(i) < \infty$ for all i]. Example:



- We have $J^*(1) = J^*(2) = -\infty$.
- Bellman's equation is

$$J(1) = J(2), \qquad J(2) = \min\big[-1 + J(1), 1\big].$$

- There is no solution [among functions J with $-\infty < J(i) < \infty$ for all i].
- Bellman's equation has as solution $J^*(1) = J^*(2) = -\infty$ [within the larger class of functions $J(\cdot)$ that can take the value $-\infty$ for some (or all) states]. This situation can be generalized (see Chapter 3 of Vol. II of the text).

PATHOLOGIES III: THE BLACKMAILER

- Two states, state 1 and the termination state t.
- At state 1, choose a control $u \in (0,1]$ (the blackmail amount demanded) at a cost $-u$, and move to t with probability u^2, or stay in 1 with probability $1-u^2$.
- Every stationary policy is proper, but the control set in not finite.
- For any stationary μ with $\mu(1) = u$, we have

$$J_\mu(1) = -u + (1-u^2)J_\mu(1)$$

from which $J_\mu(1) = -\frac{1}{u}$

- Thus $J^*(1) = -\infty$, and there is no optimal stationary policy.
- It turns out that a nonstationary policy is optimal: demand $\mu_k(1) = \gamma/(k+1)$ at time k, with $\gamma \in (0,1/2)$. (Blackmailer requests diminishing amounts over time, which add to ∞; the probability of the victim's refusal diminishes at a much faster rate.)