6.231 DYNAMIC PROGRAMMING

LECTURE 17

LECTURE OUTLINE

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

UNDISCOUNTED PROBLEMS

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

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

Note that $J_\pi(x_0)$ and $J^*(x_0)$ can be $+\infty$ or $-\infty$

- Shorthand notation for DP mappings

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

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

- T and T_μ need not be contractions in general, but their monotonicity is helpful (see Ch. 4, Vol. II of text for an analysis).
- SSP problems provide a "soft boundary" between the easy finite-state discounted problems and the hard undiscounted problems.
  - They share features of both.
  - Some nice theory is recovered thanks to the termination state, and special conditions.

SSP THEORY SUMMARY I

- As before, we have a cost-free term. state t, a finite number of states $1, \ldots, n$, and finite number of controls.

- Mappings T and T_μ (modified to account for termination state t). For all $i = 1, \ldots, n$:

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

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

or $T_\mu J = g_\mu + P_\mu J$ and $TJ = \min_\mu [g_\mu + P_\mu J]$.

- 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: (To be shown) If μ is proper, T_μ is contraction w. r. t. some weighted sup-norm

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

- 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. Instead of all policies being proper, assume that:

  (a) There exists at least one proper policy

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

- 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

- Note that T is not necessarily a contraction.

- The theory in summary is as follows:
  - J^* is the unique solution of Bellman's Eq.
  - μ^* is optimal if and only if $T_{\mu^*} J^* = T J^*$
  - VI converges: $T^k J \to J^*$ for all $J \in \Re^n$
  - PI terminates with an optimal policy, if started with a proper policy

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)

- Key Fact: A μ satisfying $J \geq T_\mu J$ for some $J \in \Re^n$ must be proper - true because

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

since $J_\mu = \sum_{m=0}^{\infty} P_\mu^m g_\mu$ and some component of the term on the right blows up as $k \to \infty$ if μ is improper (by our assumptions).

- Consequence: T can have at most one fixed point within $\Re^n$.

Proof: If J and J' are two fixed points, 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 first show that T has a fixed point, and also that PI converges to it.

- Use PI. Generate a sequence of proper policies $\{\mu^k\}$ starting from a proper policy μ^0.

- μ^1 is proper and $J_{\mu^0} \geq J_{\mu^1}$ since

$$J_{\mu^0} = T_{\mu^0} J_{\mu^0} \geq T J_{\mu^0} = T_{\mu^1} J_{\mu^0} \geq T^k_{\mu^1} J_{\mu^0} \geq J_{\mu^1}$$

- Thus $\{J_{\mu^k}\}$ is nonincreasing, some policy $\bar{\mu}$ is repeated and $J_{\bar{\mu}} = T J_{\bar{\mu}}$. So $J_{\bar{\mu}}$ is fixed point of T.

- Next show that $T^k J \to J_{\bar{\mu}}$ for all J, i.e., VI converges to the same limit as PI. (Sketch: True if $J = J_{\bar{\mu}}$, argue using the properness of $\bar{\mu}$ to show that the terminal cost difference $J - J_{\bar{\mu}}$ does not matter.)

- To show $J_{\bar{\mu}} = J^*$, 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 $J_\pi \geq J_{\bar{\mu}}$, so $\bar{\mu}$ is optimal and $J_{\bar{\mu}} = J^*$.

SSP ANALYSIS III

- Contraction Property: If all policies are proper (cf. 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

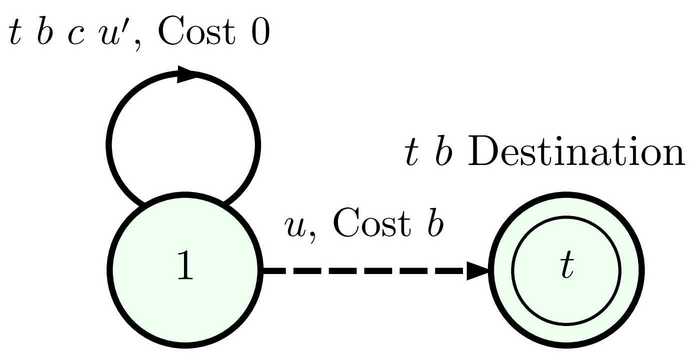
$$\rho = \max_{i=1,\ldots,n} \frac{v_i - 1}{v_i} < 1.$$

This implies T_μ and T are contractions of modulus ρ for norm $\|J\| = \max_{i=1,\ldots,n} |J(i)|/v_i$ (by the results of earlier lectures).

SSP ALGORITHMS

- All the basic algorithms have counterparts under our assumptions; see the text (Ch. 3, Vol. II)
- “Easy” case: All policies proper, in which case the mappings T and T_μ are contractions
- Even with improper (infinite cost) policies all basic algorithms have satisfactory counterparts
  - VI and PI
  - Optimistic PI
  - Asynchronous VI
  - Asynchronous PI
  - Q-learning analogs
- ** THE BOUNDARY OF NICE THEORY **
- Serious complications arise under any one of the following:
  - There is no proper policy
  - There is improper policy with finite cost $\forall\, i$
  - The state space is infinite and/or the control space is infinite [infinite but compact $U(i)$ can be dealt with]

PATHOLOGIES I: DETERM. SHORTEST PATHS



- Two policies, one proper (apply u), one improper (apply u')
- Bellman's equation is

$$J(1) = \min\bigl[J(1), b\bigr]$$

Set of solutions is $(-\infty, b]$.

- Case $b > 0$, $J^* = 0$: VI does not converge to J^* except if started from J^*. PI may get stuck starting from the inferior proper policy
- Case $b < 0$, $J^* = b$: VI converges to J^* if started above J^*, but not if started below J^*. PI can oscillate (if started with u' it generates u, and if started with u it can generate u')

PATHOLOGIES II: BLACKMAILER'S DILEMMA

- Two states, state 1 and the termination state t.
- At state 1, choose $u \in (0,1]$ (the blackmail amount demanded) at a cost $-u$, and move to t with prob. u^2, or stay in 1 with prob. $1 - u^2$.
- Every stationary policy is proper, but the control set in not finite (also not compact).
- 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.
- 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, so the probability that the victim stays forever compliant is strictly positive.

SSP UNDER WEAK CONDITIONS I

- Assume there exists a proper policy, and J^* is real-valued. Let

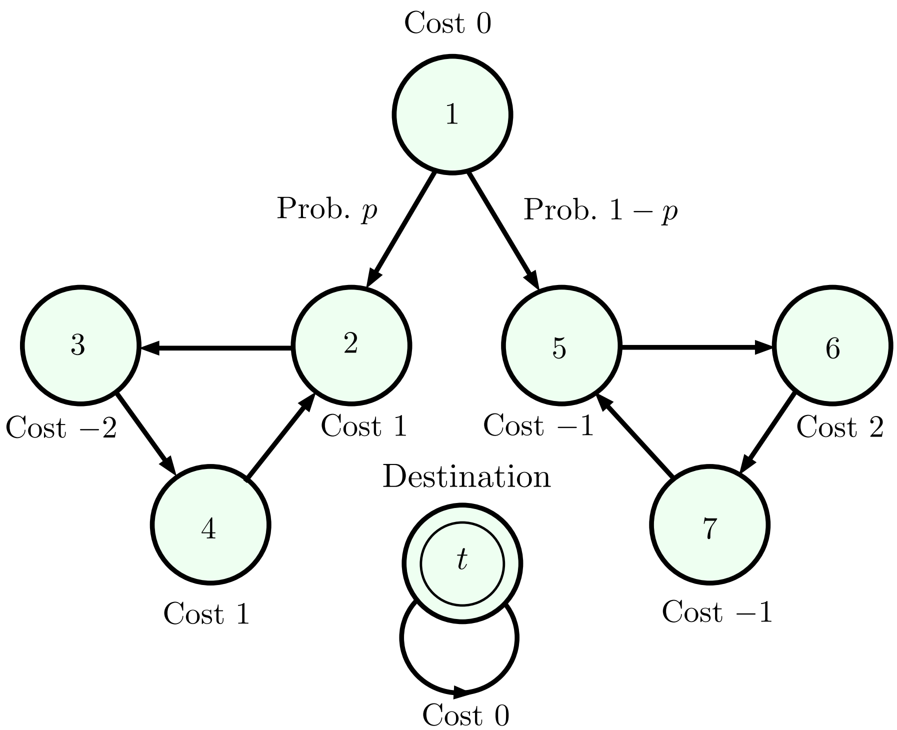
$$\hat{J}(i) = \min_{\mu:\,\text{proper}} J_\mu(i), \qquad i = 1, \ldots, n$$

Note that we may have $\hat{J} \neq J^*$ [i.e., $\hat{J}(i) \neq J^*(i)$ for some i].

- It can be shown that $\hat{J}$ is the unique solution of Bellman's equation within the set $\{J \mid J \geq \hat{J}\}$
- Also VI converges to $\hat{J}$ starting from any $J \geq \hat{J}$
- The analysis is based on the δ-perturbed problem: adding a small $\delta > 0$ to g. Then:
  - All improper policies have infinite cost for some states in the δ-perturbed problem
  - All proper policies have an additional $O(\delta)$ cost for all states
  - The optimal cost J_δ^* of the δ-perturbed problem converges to $\hat{J}$ as $\delta \downarrow 0$
- There is also a PI method that generates a sequence $\{\mu^k\}$ with $J_{\mu^k} \to \hat{J}$. Uses sequence $\delta_k \downarrow 0$, and policy evaluation based on the δ_k-perturbed problems with $\delta_k \downarrow 0$.

SSP UNDER WEAK CONDITIONS II

- J^* need not be a solution of Bellman's equation! Also J_μ for an improper policy μ.



- For $p = 1/2$, we have

$$J_\mu(1) = 0,\ J_\mu(2) = J_\mu(5) = 1,\ J_\mu(3) = J_\mu(7) = 0,\ J_\mu(4) = J_\mu(6) = 2,$$

Bellman Eq. at state 1, $J_\mu(1) = \frac{1}{2}\big(J_\mu(2)+J_\mu(5)\big)$, is violated.

- References: Bertsekas, D. P., and Yu, H., 2015. "Stochastic Shortest Path Problems Under Weak Conditions," Report LIDS-2909; Math. of OR, to appear. Also the on-line updated Ch. 4 of the text.

6.231 Dynamic Programming and Stochastic Control
Fall 2015