Reinforcement Learning and Optimal Control

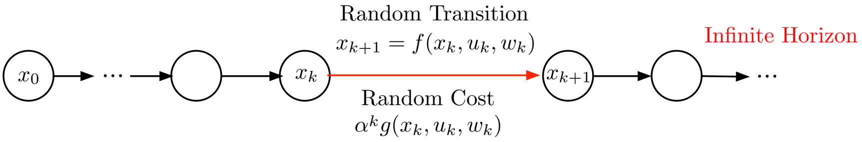
ASU, CSE 691, Winter 2020

Dimitri P. Bertsekas
dimitrib@mit.edu

Lecture 8

Outline

Stochastic DP Problems - Infinite Horizon



Infinite number of stages, and stationary system and cost

- System $x_{k+1} = f(x_k, u_k, w_k)$ with state, control, and random disturbance.
- Policies $\pi = \{\mu_0, \mu_1, \ldots\}$ with $\mu_k(x) \in U(x)$ for all x and k.
- Special scalar α with $0 < \alpha \leq 1$. If $\alpha < 1$ the problem is called discounted.
- Cost of stage k: $\alpha^k g\big(x_k, \mu_k(x_k), w_k\big)$.
- Cost of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$

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

- Optimal cost function $J^*(x_0) = \min_\pi J_\pi(x_0)$.
- If $\alpha = 1$ we assume a special cost-free termination state t. The objective is to reach t at minimum expected cost. The problem is called stochastic shortest path (SSP) problem.

Going from Finite to Infinite Horizon (Just Intuition - Proofs Needed)

Value iteration (VI) convergence: Fix horizon N, let terminal cost be 0

- Let $V_{N-k}(x)$ be the optimal cost starting at x with k stages to go, so
$$V_{N-k}(x) = \min_{u \in U(x)} E_w\Big\{\alpha^{N-k} g(x,u,w) + V_{N-k+1}\big(f(x,u,w)\big)\Big\} \quad \text{(Finite Horizon DP)}$$
- Reverse the time index: Define $J_k(x) = V_{N-k}(x)/\alpha^{N-k}$ and divide with α^{N-k}:
$$J_k(x) = \min_{u \in U(x)} E_w\Big\{g(x,u,w) + \alpha J_{k-1}\big(f(x,u,w)\big)\Big\} \qquad \text{(VI)}$$
- $J_N(x)$ is equal to $V_0(x)$, which is the N-stages optimal cost starting from x
- Hence, intuitively, J_N converges to J^*:
$$J^*(x) = \lim_{N\to\infty} J_N(x), \qquad \text{for all states } x \quad \text{(proof needed)}$$

The following Bellman equation holds: Take the limit in Eq. (VI)

$$J^*(x) = \min_{u \in U(x)} E_w\Big\{g(x,u,w) + \alpha J^*\big(f(x,u,w)\big)\Big\}, \qquad \text{for all states } x \quad \text{(proof needed)}$$

Optimality condition: Let $\mu(x)$ attain the min in the Bellman equation for all x

The policy $\{\mu, \mu, \ldots\}$ is optimal (??). (This type of policy is called stationary.)

Transition Probability Notation for Finite-Spaces Problems

- States: $i = 1, \ldots, n$. Successor states: j. (For SSP there is also the extra termination state t.)
- Probability of $i \to j$ transition under control u: $p_{ij}(u)$ (plays the role of the system equation)
- Cost of $i \to j$ transition under control u: $g(i, u, j)$

VI (translated to the new notation - note that $J_k(t) = 0$ for SSP)

$$J_{k+1}(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + \alpha J_k(j)\big) \qquad \text{(for discounted)}$$

$$J_{k+1}(i) = \min_{u \in U(i)} \left[p_{it}(u) g(i,u,t) + \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + J_k(j)\big) \right] \qquad \text{(for SSP)}$$

Bellman equation (translated to the new notation - note that $J^*(t) = 0$ for SSP)

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

$$J^*(i) = \min_{u \in U(i)} \left[p_{it}(u) g(i,u,t) + \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + J^*(j)\big) \right] \qquad \text{(for SSP)}$$

Finite-Spaces Discounted Problems - Statement of Main Results

Convergence of VI

Given any initial conditions $J_0(1), \ldots, J_0(n)$, the sequence $\{J_k(i)\}$ generated by VI

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

converges to $J^*(i)$ for each i.

Bellman's equation

The optimal cost function $J^* = \big(J^*(1), \ldots, J^*(n)\big)$ satisfies the equation

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

and is the unique solution of this equation.

Optimality condition

A stationary policy μ is optimal if and only if for every state i, $\mu(i)$ attains the minimum in the Bellman equation.

Finite-Spaces SSP Problems - Statement of Main Results

Assumption (Termination Inevitable Under all Policies)

There exists $m > 0$ such that regardless of the policy used and the initial state, there is positive probability that t will be reached within m stages; i.e., for all π

$$\max_{i=1,\ldots,n} P\{x_m \neq t \mid x_0 = i, \pi\} < 1.$$

VI Convergence: $J_k \to J^*$ for all initial conditions J_0, where

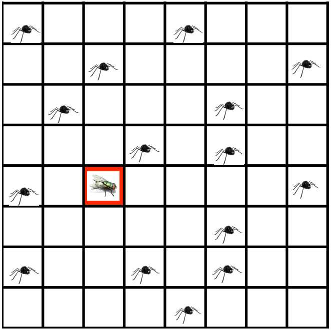
$$J_{k+1}(i) = \min_{u \in U(i)} \left[p_{it}(u) g(i,u,t) + \sum_{j=1}^{n} p_{ij}(u) \big(g(i,u,j) + J_k(j) \big) \right], \qquad i = 1, \ldots, n$$

Bellman's equation: J^* satisfies

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

and is the unique solution of this equation.
Optimality condition: μ is optimal if and only if for every i, $\mu(i)$ attains the minimum in the Bellman equation.

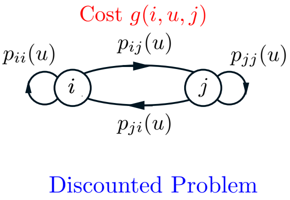
A Spiders-and-Fly SSP Example (or Search-and-Rescue)

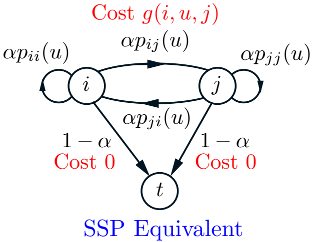


15 spiders move along 4 directions ($\leq$ 1 unit) w. perfect observation; fly moves randomly

- Objective is to catch the fly in minimum time.
- Is the “termination inevitable" assumption satisfied?
- There is a way to fix that (see next slide).
- One-step lookahead and rollout are impossible: $\approx 5^{15}$ Q-factors.
- Note for the future: We can reformulate one-step lookahead so that spiders move one-at-a-time. This will trade off state space and control space complexity.

SSP Analysis and Extensions (An Overview)





- A discounted problem can be converted to an SSP problem, since the stage k expected cost is identical in both problems, under the same policy.
- Proof line: Start with SSP analysis, get discounted analysis as special case.
- Key proof argument: The tail portion (k to ∞) of the infinite horizon cost diminishes to 0, as $k \to \infty$, at a geometric progression rate (so the finite horizon costs converge to the infinite horizon cost).

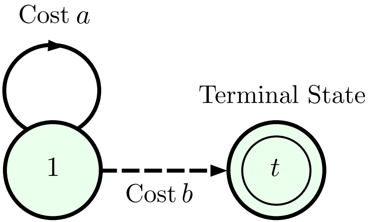
A more general assumption for SSP results: Nonterminating policies are "bad"

- Every stationary policy under which termination is not inevitable from some initial states is "bad," in the sense that it has ∞ cost for some initial states.
- There exists at least one stationary policy under which termination is inevitable.

A Word of Caution: SSP Problems can be Tricky

Without the assumption "nonterminating policies are bad"

- Bellman equation may have any number of solutions: one, infinitely many, or none.
- Bellman equation may have one or more solutions, but J^* may not be a solution.
- VI may converge to J^* from some initial conditions but not from others.



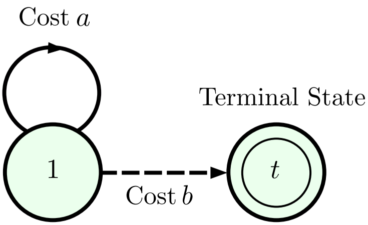
Deterministic one-state SSP

Two possible controls at state 1 (costs a and b)

Challenge questions: Consider the cases $a > 0$, $a = 0$, and $a < 0$

- What is $J^*(1)$?
- What is the solution set of Bellman's equation $J(1) = \min\big[b,\ a + J(1)\big]$?
- What is the limit of the VI algorithm $J_{k+1}(1) = \min\big[b,\ a + J_k(1)\big]$?

Answers to the Challenge Questions



Deterministic one-state SSP

Two possible controls at state 1 (costs a and b)

Bellman Eq: $J(1) = \min\big[b,\ a + J(1)\big]$; VI: $J_{k+1}(1) = \min\big[b,\ a + J_k(1)\big]$

- If $a > 0$ (positive cycle): $J^*(1) = b$ is the unique solution, and VI converges to $J^*(1)$. Here the “nonterminating policies are bad" assumption is satisfied.
- If $a = 0$ (zero cycle):
  - $J^*(1) = \min[0, b]$.
  - The solution set of the Bellman equation is $= (-\infty, b]$.
  - The VI algorithm, $J_{k+1}(1) = \min\big[b,\ J_k(1)\big]$, converges (in one step) to b starting from $J_0(1) \geq b$, and does not move from a starting value $J_0(1) \leq b$.
- If $a < 0$ (negative cycle): B-Eq has no solution, and VI diverges to $J^*(1) = -\infty$.

Results Involving Q-Factors - Discounted Problems

VI for Q-factors (finite horizon optimal Q-factors converge to infinite horizon Q-factors)

$$Q_{k+1}(i,u) = \sum_{j=1}^{n} p_{ij}(u) \left(g(i,u,j) + \alpha \min_{v \in U(j)} Q_k(j,v) \right)$$

converges to $Q^*(i,u)$ for each (i,u).

Bellman's equation for Q-factors

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

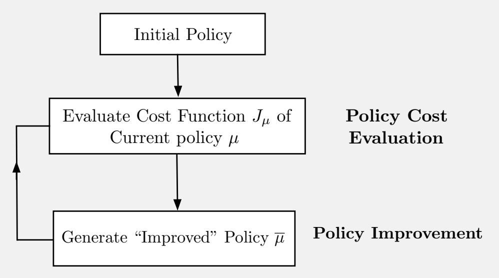
Q^* is the unique solution of this equation, and we have

$$J^*(i) = \min_{u \in U(i)} Q^*(i,u) \tag{1}$$

Optimality condition

A stationary policy μ is optimal if and only if $\mu(i)$ attains the minimum in Eq. (1) for every state i.

Policy Iteration (PI) Algorithm (Perpetual Rollout): From Base Policy to Rollout Policy and Repeat



Given the current (base) policy μ^k, a PI consists of two phases:

- Policy evaluation computes $J_{\mu^k}(i)$, $i = 1, \ldots, n$, as the solution of the (linear) Bellman equation system (or by some form of simulation)

$$J_{\mu^k}(i) = \sum_{j=1}^{n} p_{ij}\big(\mu^k(i)\big)\Big(g\big(i, \mu^k(i), j\big) + \alpha J_{\mu^k}(j)\Big), \quad i = 1, \ldots, n$$

- Policy improvement then computes a new (the rollout) policy μ^{k+1} as

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

Fundamental Policy Improvement Property - Intuition: Acting Optimally for One Step and then Using μ^k Should Improve on μ^k

PI finite-step convergence: PI generates an improving sequence of policies, i.e., $J_{\mu^{k+1}}(i) \leq J_{\mu^k}(i)$ for all i and k, and terminates with an optimal policy.

Proof: We will show that $J_{\tilde{\mu}} \leq J_{\mu}$, where $\tilde{\mu}$ is obtained from μ by PI

- Denote by J_N the cost function of a policy that applies $\tilde{\mu}$ for the first N stages and applies μ thereafter.
- We have the Bellman equation $J_\mu(i) = \sum_{j=1}^n p_{ij}(\mu(i))\Big(g(i,\mu(i),j) + \alpha J_\mu(j)\Big)$, so

$$J_1(i) = \sum_{j=1}^n p_{ij}(\tilde{\mu}(i))\Big(g(i,\tilde{\mu}(i),j) + \alpha J_\mu(j)\Big) \leq J_\mu(i) \qquad \text{(by policy improvement eq.)}$$

- From the definition of J_2 and J_1, monotonicity, and the preceding relation, we have

$$J_2(i) = \sum_{j=1}^n p_{ij}(\tilde{\mu}(i))\Big(g(i,\tilde{\mu}(i),j)+\alpha J_1(j)\Big) \leq \sum_{j=1}^n p_{ij}(\tilde{\mu}(i))\Big(g(i,\tilde{\mu}(i),j)+\alpha J_\mu(j)\Big) = J_1(i)$$

  so $J_2(i) \leq J_1(i) \leq J_\mu(i)$ for all i.
- Continuing similarly, we obtain $J_{N+1}(i) \leq J_N(i) \leq J_\mu(i)$ for all i and N. Since $J_N \to J_{\tilde{\mu}}$ (VI for $\tilde{\mu}$ converges), it follows that $J_{\tilde{\mu}} \leq J_\mu$.

Illustration Movies: A Single Step of Policy Iteration for a Four-Spiders and Two-Flies Problem

Base Policy Rollout Policy

We want to minimize the time to catch both flies

- Base policy (each spider follows the shortest path): Time is 85
- Rollout (all spiders move at once, 625 Q-factors/move choices): Time is 34
- We can reduce the number of Q-factors using multiagent/one-spider at-a-time rollout (will return to this later)

About the Next Lecture

We will cover:

- Infinite horizon policy iteration: extensions and approximations
- Rollout and parametric approximation methods
- We will likely need more that one lecture

PLEASE READ AS MUCH OF Chapter 4 AS YOU CAN

PLEASE DOWNLOAD THE LATEST VERSIONS FROM MY WEBSITE