Dynamic Programming & Optimal Control

Lecture 4
Infinite Horizon Problems

Fall 2023

Prof. Raffaello D'Andrea

ETH Zurich

Learning Objectives

Topic: Infinite Horizon Problems

Objectives

- You know the *Infinite Horizon Problem* setup.
- You understand the *Bellman Equation* for infinite horizon problems.
- You know the definition of a *Stochastic Shortest Path* Problem.
- You know the main results of the DPA applied to a stochastic shortest path problem:
  - The DPA recursion converges to the optimal cost;
  - The optimal cost satisfies the Bellman equation;
  - The Bellman equation admits a unique solution.

Outline

Infinite Horizon Problems (1/5)

We will now look at how to solve the standard problem that we have been considering, but as the time horizon N goes to infinity.

We consider the time-invariant setup as follows:

Dynamics

- The state evolution is governed by the time-invariant system:

$$x_{k+1} = f\left(x_k, u_k, w_k\right),$$

where

$$x_k \in \mathcal{S},\ u_k \in \mathcal{U}(x_k),\ w_k \sim p_{w|x,u},\ k = 0, \ldots, N-1.$$

It is assumed that w_k is conditionally independent with all prior variables x_l, u_l, w_l, $l < k$, given x_k and u_k. Note that the PDF of w_k given x_k, u_k is time-invariant.

Infinite Horizon Problems (2/5)

- As usual, the control inputs u_k are generated by an admissible policy $\pi \in \Pi$:

$$\pi = (\mu_0(\cdot)\,, \mu_1(\cdot)\,, \ldots, \mu_{N-1}(\cdot)),$$

such that

$$u_k = \mu_k(x_k)\,, \;\; u_k \in \mathcal{U}(x_k),\; \forall x_k \in \mathcal{S},\; \forall k.$$

Cost

Let the cost function be a sum of time-invariant stage costs, namely:

$$\sum_{k=0}^{N-1} g(x_k, u_k, w_k)\,.$$

Infinite Horizon Problems (3/5)

Given $\mathrm{x} \in \mathcal{S}$, the expected closed loop cost of starting at x associated with policy $\pi \in \Pi$ is then

$$J_\pi(\mathrm{x}) = \underset{(X_1, W_0 | x_0 = \mathrm{x})}{\mathrm{E}} \left[\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k), w_k) \right],$$

where $X_1 := (x_1, \ldots, x_N)$, $W_0 := (w_0, \ldots, w_{N-1})$ and subject to

$$x_{k+1} = f(x_k, \mu_k(x_k), w_k).$$

Objective

Construct an optimal policy π^* with associated optimal cost $J^*(\mathrm{x}) = J_{\pi^*}(\mathrm{x})$ such that for all $\mathrm{x} \in \mathcal{S}$,

$$\pi^* = \arg\min_{\pi \in \Pi} J_\pi(\mathrm{x}).$$

Infinite Horizon Problems (4/5)

We can notice that as the time horizon N goes to infinity, the complexity of the solution collapses.

The intuition is that since the dynamics and stage costs are time-invariant, and the time horizon is infinite, the cost-to-go is itself time-invariant.

Let's write down the DPA and explore what happens as N goes to infinity:

Initialization

$$J_N(\mathrm{x}) = 0, \quad \forall \mathrm{x} \in \mathcal{S}.$$

Recursion

$$J_k(\mathrm{x}) = \min_{\mathrm{u} \in \mathcal{U}(\mathrm{x})} \mathop{\mathrm{E}}_{(w|x=\mathrm{x}, u=\mathrm{u})} \left[g(x,u,w) + J_{k+1}(f(x,u,w)) \right], \quad \forall \mathrm{x} \in \mathcal{S}, \quad \forall k = N-1, ..., 0.$$

Infinite Horizon Problems (5/5)

Let $l := N - k$ and $V_l(\cdot) := J_{N-l}(\cdot)$. Then:

$$V_0(\mathrm{x}) = 0, \quad \forall \mathrm{x} \in \mathcal{S}.$$

$$V_l(\mathrm{x}) = \min_{\mathrm{u} \in \mathcal{U}(\mathrm{x})} \underset{(w|x=\mathrm{x}, u=\mathrm{u})}{\mathrm{E}} \left[g(x,u,w) + V_{l-1}(f(x,u,w)) \right], \quad \forall \mathrm{x} \in \mathcal{S},\ \forall l = 1, ..., N. \quad (1)$$

Now assume that for each $\mathrm{x} \in \mathcal{S}$, the sequence $V_l(\mathrm{x})$ converges to some value $J(\mathrm{x})$ as N approaches infinity. Then, (1) becomes

$$J(\mathrm{x}) = \min_{\mathrm{u} \in \mathcal{U}(\mathrm{x})} \underset{(w|x=\mathrm{x}, u=\mathrm{u})}{\mathrm{E}} \left[g(x,u,w) + J(f(x,u,w)) \right], \quad \forall \mathrm{x} \in \mathcal{S}.$$

This equation is known as the **Bellman Equation** (BE). Assuming this convergence, $J(\mathrm{x})$ is the optimal cost-to-go $J^*(\mathrm{x})$.

The Bellman Equation

$$J(\mathrm{x}) = \min_{\mathrm{u} \in \mathcal{U}(\mathrm{x})} \underset{(w|x=\mathrm{x}, u=\mathrm{u})}{\mathrm{E}} \left[g(x,u,w) + J(f(x,u,w)) \right], \quad \forall \mathrm{x} \in \mathcal{S}$$

This suggests that the optimal policy is time-invariant, or *stationary* (that is, at each time period the *same feedback map* $\mu(\cdot)$ is used) and solving the minimization in u for every $\mathrm{x} \in \mathcal{S}$ yields that policy.

The BE may seem simple but it has to be solved for all $\mathrm{x} \in \mathcal{S}$ simultaneously. We can do this analytically only for a very small set of problems (e.g. the Linear Quadratic Regulator (LQR) problem).

Outline

The SSP Problem (1/3)

We now consider a subclass of problems, known as the stochastic shortest path (SSP) problem, for which solving the BE yields the optimal cost-to-go and an optimal stationary policy.

Dynamics

Consider the finite state, time-invariant system

$$\begin{aligned} &x_{k+1} = w_k, \qquad && x_k \in \mathcal{S}, \\ &\Pr\left(w_k = j | x_k = i, u_k = \mathrm{u}\right) = P_{ij}(\mathrm{u}), \qquad && \mathrm{u} \in \mathcal{U}(i), \end{aligned}$$

where $\mathcal{S}$ is a finite set and $\mathcal{U}(\mathrm{x})$ is a finite set for all $\mathrm{x} \in \mathcal{S}$.

Note that the transition probabilities are time-invariant as opposed to the general time-varying case of the previous lectures.

The SSP Problem (2/3)

Cost

Given an initial state $i \in \mathcal{S}$, the expected closed loop cost of starting at i associated with policy π becomes:

$$J_\pi(i) = \underset{(X_1, W_0 | x_0 = i)}{\mathrm{E}} \left[\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k), w_k) \right],$$

subject to

$$x_{k+1} = w_k,$$
$$\Pr(w_k = j | x_k = i, u_k = \mu_k(i)) = P_{ij}(\mu_k(i)).$$

The SSP Problem (3/3)

Assumption 4.1: Cost-free termination state

There exists a cost-free termination state, which we designate as state 0. In particular, there are $n+1$ states with $\mathcal{S} = \{0, 1, \ldots, n\}$, where

$$P_{00}(\mathrm{u}) = 1 \text{ and } g(0, \mathrm{u}, 0) = 0, \quad \forall \mathrm{u} \in \mathcal{U}(0).$$

This assumption is used to make the expected cost meaningful. We can think of the termination state as a destination state where we can land and stop accumulating cost.

Objective

As before, we want to construct an optimal policy π^* such that for all $i \in \mathcal{S}$,

$$\pi^* = \arg\min_{\pi \in \Pi} J_\pi(i).$$

We want to explore what happens as the time horizon N goes to infinity.

Outline

SSP and BE (1/2)

For simplicity, and with a slight abuse of notation, we let a stationary policy $\pi = (\mu(\cdot), \mu(\cdot), \dots)$ be represented by μ.

A stationary policy μ is said to be *proper* if, when using this policy, there exists an integer m such that:

$$\Pr\left(x_m = 0 | x_0 = i\right) > 0, \quad \forall i \in \mathcal{S},$$

subject to

$$\begin{aligned} & x_{k+1} = w_k, \\ & \Pr\left(w_k = j | x_k = i, u_k = \mu(i)\right) = P_{ij}(\mu(i)). \end{aligned}$$

A stationary policy that is not proper is said to be *improper*.

SSP and BE (2/2)

Assumption 4.2: Proper policy

There exists at least one proper policy $\mu \in \Pi$. Furthermore, for every improper policy μ' and at least one state $i \in \mathcal{S}$, the corresponding cost function is $J_{\mu'}(i) = +\infty$.

This assumption is required in order to guarantee that a unique solution to the BE exists for the SSP problem, which will then be the optimal cost.

It ensures that a policy exists for which the probability of reaching the termination state goes to one as the time horizon N goes to infinity.

It also ensures that the policies for which this does not occur incur infinite cost, which ensures that there are no non-positive cycles.

Theorem 4.1: SSP and BE

Under the cost-free termination state and proper policy assumptions, the following are true for the SSP:

1) Given *any* initial conditions $V_0(1), \ldots, V_0(n)$, the sequence $V_l(i)$ generated by the iteration

$$V_{l+1}(i) = \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_l(j) \right), \quad \forall\, i \in \mathcal{S}^+,$$

where $\mathcal{S}^+ := \mathcal{S} \backslash \{0\}$ and

$$q(i, \mathrm{u}) := \mathop{\mathrm{E}}_{(w | x=i, u=\mathrm{u})} \left[g(x, u, w) \right],$$

converges to the optimal cost $J^*(i)$ for all $i \in \mathcal{S}^+$;

Theorem 4.1: SSP and BE

Under the cost-free termination state and proper policy assumptions, the following are true for the SSP:

2) The optimal costs satisfy the Bellman Equation

$$J^*(i) = \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) J^*(j) \right) \quad \forall\, i \in \mathcal{S}^+;$$

3) The solution to the BE is unique;

4) The minimizing u for each $i \in \mathcal{S}^+$ of the BE gives an optimal policy, which is proper.

Proof: SSP and BE

We will not provide a rigorous proof. Instead, an intuition will be given under the stronger Assumption 4.3 (below) in place of Assumption 4.2.

Assumption 4.3: Any admissible policy is proper

There exists an integer m such that for **any** admissible policy $\pi \in \Pi$:

$$\Pr\left(x_m = 0 | x_0 = i\right) > 0 \quad \forall i \in \mathcal{S},$$

subject to

$$\begin{aligned} &x_{k+1} = w_k, \\ &\Pr\left(w_k = j | x_k = i, u_k = \mu(i)\right) = P_{ij}(\mu(i)). \end{aligned}$$

Thus, there is a positive probability that the termination state will be reached for any admissible policy regardless of the initial state.

Proof: SSP and BE (Part 1 of 4)

1) Given *any* initial conditions $V_0(1), \ldots, V_0(n)$, the sequence $V_l(i)$ generated by the iteration

$$V_{l+1}(i) = \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_l(j) \right), \quad \forall\, i \in \mathcal{S}^+$$

$$q(i, \mathrm{u}) := \mathop{\mathrm{E}}_{(w|x=i, u=\mathrm{u})} \left[g(x, u, w) \right],$$

converges to the optimal cost $J^*(i)$, for all $i \in \mathcal{S}^+$.

Let $V_0(0) = 0$. For fixed N, consider the following augmented cost:

$$J_\pi(i) = \mathop{\mathrm{E}}_{(X_1, W_0 | x_0 = i)} \left[\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k), w_k) + V_0(x_N) \right],$$

subject to

$$x_{k+1} = w_k,$$
$$\Pr\left(w_k = j | x_k = i, u_k = \mu(i)\right) = P_{ij}(\mu(i)).$$

Proof: SSP and BE (Part 1 of 4)

$$J_\pi(i) = \underset{(X_1, W_0 | x_0 = i)}{\mathrm{E}} \left[\sum_{k=0}^{N-1} g(x_k, \mu_k(x_k), w_k) + V_0(x_N) \right],$$

subject to

$$\begin{aligned} & x_{k+1} = w_k, \\ & \Pr(w_k = j | x_k = i, u_k = \mu(i)) = P_{ij}(\mu(i)). \end{aligned}$$

The $V_0(x_N)$ can be interpreted as a terminal cost.

By Assumption 4.3, as N goes to infinity, the probability that the termination state 0 is reached approaches one for all policies.

Thus $V_0(x_N)$ does not influence the augmented cost in the limit, since x_N will be 0 with probability 1, and $V_0(0) = 0$.

Proof: SSP and BE (Part 1 of 4)

The corresponding DPA is then:

$$J_N(i) = V_0(i)\,, \quad \forall i \in \mathcal{S}.$$

$$\begin{aligned} J_k(i) &= \min_{\mathrm{u}\in\mathcal{U}(i)} \mathop{\mathrm{E}}_{(w|x=i,u=\mathrm{u})} [g(x,u,w) + J_{k+1}(w)] \\ &= \min_{\mathrm{u}\in\mathcal{U}(i)} \left(q(i,\mathrm{u}) + \sum_{j=0}^{n} P_{ij}(\mathrm{u}) J_{k+1}(j) \right), \quad \forall i \in \mathcal{S}, \quad k = N-1, ..., 0. \end{aligned}$$

Note that $q(0,\mathrm{u}) = 0$ and $P_{0j}(\mathrm{u}) = 0$ for all $j \in \mathcal{S}^+$ by Assumption 4.1, and we have initialized $J_N(0) = V_0(0) = 0$. Therefore, $J_k(0) = 0$ for $k = 0, ..., N$. Thus:

$$J_k(i) = \min_{\mathrm{u}\in\mathcal{U}(i)} \left(q(i,\mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) J_{k+1}(j) \right), \quad \forall i \in \mathcal{S}^+, \quad k = N-1, ..., 0.$$

Proof: SSP and BE (Part 1 of 4)

Let $l := N - k$ and $V_l(\cdot) := J_{N-l}(\cdot)$. Then the DPA recursion

$$J_k(i) = \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) J_{k+1}(j) \right), \quad \forall i \in \mathcal{S}^+, \quad k = N-1, ..., 0,$$

becomes

$$V_l(i) = \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_{l-1}(j) \right), \quad \forall i \in \mathcal{S}^+, \quad l = 1, ..., N,$$

which is precisely the iteration in 1).

Furthermore, $V_N(i) = J_0(i)$ is the optimal cost for the augmented cost function and the original cost function in the limit.

Because of Assumption 4.3, it can be shown that $V_l(i)$ converges, and we thus have $\lim_{l \to \infty} V_l(i) = J^*(i)$.

Proof: SSP and BE (Part 2 of 4)

2) The optimal costs satisfy the Bellman Equation:

$$J^*(i) = \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) J^*(j) \right), \quad \forall\, i \in \mathcal{S}^+.$$

2) follows from taking limits of both sides of the iteration in part 1).

Proof: SSP and BE (Part 3 of 4)

3) The solution to the BE is unique.

Let $J_0(1), \ldots, J_0(n)$ and $\bar{J}_0(1), \ldots, \bar{J}_0(n)$ be two different solutions of Bellman Equation.

If we use both solutions as initial conditions for iteration in part 1), they both converge after 1 iteration of the the DP recursion.

This leads to two different optimal costs which is a contradiction.

Proof: SSP and BE (Part 4 of 4)

4) The minimizing u for each $i \in \mathcal{S}^+$ of the BE gives an optimal policy, which is proper.

Let μ be the stationary policy inferred from minimizing u in the BE. We then have:

$$J^*(i) = \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) J^*(j) \right)$$

$$= \left(q(i, \mu(i)) + \sum_{j=1}^{n} P_{ij}(\mu(i)) J^*(j) \right), \quad \forall\, i \in \mathcal{S}^+.$$

Consider the modified problem where $\tilde{\mathcal{U}}(i) = \{\mu(i)\}$ for all $i \in \mathcal{S}$.

By 1), with arbitrary initialization we converge to the optimal cost J_μ for the modified problem. The BE for the modified problem has the following form

$$J_\mu(i) = \left(q(i, \mu(i)) + \sum_{j=1}^{n} P_{ij}(\mu(i)) J_\mu(j) \right), \quad \forall\, i \in \mathcal{S}^+.$$

Proof: SSP and BE (Part 4 of 4)

The BE for the modified problem has the following form:

$$J_\mu(i) = \left(q(i, \mu(i)) + \sum_{j=1}^{n} P_{ij}(\mu(i)) J_\mu(j) \right), \quad \forall\, i \in \mathcal{S}^+.$$

Note that we do not need the minimization here since there is only one choice in the allowable control space.

By part 3), this BE for the modified problem has a unique solution.

Thus $J_\mu(i)$ is equal to $J^*(i)$ for all $i \in \mathcal{S}^+$. μ therefore incurs the optimal cost $J^*(i)$ and is an optimal policy.

Outline

Additional reading material

Next week we will see how to solve the BE exactly, but it is worth noticing that Reinforcement Learning has its foundations in the BE.

Given an *estimate* $\hat{J}$ of the optimal cost function, we can compute the optimal policy as

$$\mu^*(i) \in \arg\min_{\mathrm{u}\in\mathcal{U}(i)} \left(q(i,\mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u})\hat{J}(j) \right).$$

The Q-learning algorithm is based on the equation:

$$Q(i,u) = q(i,\mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) \min_{\mathrm{u}'\in\mathcal{U}(j)} Q(j,\mathrm{u}').$$

Can you spot the similarity?

- How to estimate $\hat{J}$ or Q? Some keywords: (Multi-Step) Temporal difference, Monte-Carlo Tree Search, SARSA.
- How to avoid local optima? Some keywords: exploration, exploitation, ε-greedy strategies.