Dynamic Programming & Optimal Control

Lecture 6
Solving the Bellman Equation (cont'd)

Fall 2023

Prof. Raffaello D'Andrea

ETH Zurich

Learning Objectives

Topic: Solving the Bellman Equation (cont'd)

Objectives

- You know how to solve a stochastic shortest path problem using *Linear Programming*.
- You know how to combine *Value Iteration* and *Policy Iteration*.
- You know how to solve *discounted infinite horizon* problems.

Outline

Solving the Bellman Equation (cont'd)

- Linear Programming (LP)
- Discounted Problems
- Additional reading material

Linear Programming (1/5)

Recall VI:

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

By VI, $V_l(i)$ converges to $J^*(\cdot)$, which satisfies the Bellman Equation:

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

Linear Programming (2/5)

The equalities in the previous slide are equivalent to the inequalities:

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

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

Furthermore, we can write:

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

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

Linear Programming (3/5)

Suppose that we use value iteration to generate a sequence of vectors V_l starting with an initial vector V_0 that satisfies

$$V_0(i) \leq \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_0(j) \right) \quad \forall i \in \mathcal{S}^+,$$

i.e.

$$V_0(i) \leq q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_0(j) \quad \forall \mathrm{u} \in \mathcal{U}(i),\ \forall i \in \mathcal{S}^+.$$

Linear Programming (4/5)

By the VI $V_{l+1}(i) = \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_l(j) \right)$ for all $i \in \mathcal{S}^+$. Thus, we have:

$$V_0(i) \leq V_1(i) \,, \quad \forall i \in \mathcal{S}^+.$$

Therefore:

$$\begin{aligned} V_2(i) &= \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_1(j) \right) \\ &\geq \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_0(j) \right) \\ &= V_1(i) \,, \quad \forall i \in \mathcal{S}^+. \end{aligned}$$

This leads to:

$$V_{l+1}(i) \geq V_l(i) \,, \quad \forall i \in \mathcal{S}^+, \, \forall l.$$

Linear Programming (5/5)

By VI we know that $V_l(i)$ converges to $J^*(i)$ as l goes to infinity. We thus have:

$$J^*(i) \geq V_0(i)\,, \quad \forall i \in \mathcal{S}^+.$$

Thus J^* is the “largest” V_0 that satisfies the constraint

$$V_0(i) \leq \left(q(i,\mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V_0(j) \right), \quad \forall \mathrm{u} \in \mathcal{U}(i),\ \forall i \in \mathcal{S}^+.$$

We can write this as an optimization problem as in the following theorem.

Theorem 6.1: Linear Programming

The solution to the optimization problem

$$\begin{aligned} \underset{V}{\text{maximize}} \quad & \sum_{i \in \mathcal{S}^+} V(i) \\ \text{subject to} \quad & V(i) \leq \left(q(i, \mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) V(j) \right), \forall \mathrm{u} \in \mathcal{U}(i),\ \forall i \in \mathcal{S}^+ \end{aligned}$$

also solves the Bellman Equation and yields the optimal cost J^* for the SSP problem.

Note that in the optimization problem, both the objective function and the constraints are linear in V. This is known as a *Linear Program*.

In general, Linear Programs can be solved efficiently and can handle millions of variables, with many mature solvers available.

Proof of Theorem 6.1

Let V^* be the solution to the linear program and thus satisfies the inequality constraint

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

By contradiction, assume that $V^* \neq J^*$, thus there exists a state $\bar{i} \in \mathcal{S}^+$ such that:

$$V^*(\bar{i}) < J^*(\bar{i}).$$

Thus,

$$\sum_{i \in \mathcal{S}^+} V^*(i) < \sum_{i \in \mathcal{S}^+} J^*(i).$$

But since J^* also satisfies the inequality constraint as it solves the BE, V^* is not the solution to the linear program, which is a contradiction.

Outline

Solving the Bellman Equation (cont'd)

Linear Programming (LP)

Discounted Problems

Additional reading material

Discounted Problems (1/4)

We now consider a class of infinite horizon problems where future stage costs are discounted exponentially, and there is no assumption of a termination state.

We will show that this is equivalent to an associated stochastic shortest path problem.

Dynamics

$$\tilde{x}_{k+1} = \tilde{w}_k, \quad \tilde{x}_k \in \mathcal{S}^+,\ k = 0, \ldots, N-1,$$
$$p_{\tilde{w}|\tilde{x},\tilde{u}}(j|i, \mathrm{u}) = \tilde{P}_{ij}(\mathrm{u}), \quad \mathrm{u} \in \tilde{\mathcal{U}}(\tilde{x}_k),$$

where $\mathcal{S}^+ = \{1, \ldots, n\}$ is a finite set and $\tilde{\mathcal{U}}(x)$ is a finite set for all $x \in \mathcal{S}^+$, and the $\tilde{w}_k$ are independent of all previous variables when conditioned on $\tilde{x}_k, \tilde{u}_k$.

Note that there is no explicit termination state required.

Discounted Problems (2/4)

As usual, the control inputs $\tilde{u}_k$ are generated by an admissible policy $\tilde{\pi} \in \tilde{\Pi}$:

$$\tilde{\pi} = (\tilde{\mu}_0(\cdot)\,, \tilde{\mu}_1(\cdot)\,, \ldots, \tilde{\mu}_{N-1}(\cdot)),$$

such that

$$\tilde{u}_k = \tilde{\mu}_k(\tilde{x}_k)\,, \;\; \tilde{u}_k \in \tilde{\mathcal{U}}(\tilde{x}_k), \; \forall \tilde{x}_k \in \mathcal{S}^+, \; k = 0, \ldots, N-1.$$

Discounted Problems (3/4)

Cost

Given an initial state $i \in \mathcal{S}^+$, the expected closed loop cost of starting at i associated with policy $\tilde{\pi} \in \tilde{\Pi}$ is:

$$\tilde{J}_{\tilde{\pi}}(i) = \underset{(\tilde{X}_1, \tilde{W}_0 | \tilde{x}_0 = i)}{\mathrm{E}} \left[\sum_{k=0}^{N-1} \alpha^k \tilde{g}(\tilde{x}_k, \tilde{\mu}_k(\tilde{x}_k), \tilde{w}_k) \right],$$

where $\tilde{X}_1 := (\tilde{x}_1, \ldots, \tilde{x}_N)$, $\tilde{W}_0 := (\tilde{w}_0, \ldots, \tilde{w}_{N-1})$, and $\alpha \in (0,1)$ is called the *discount factor*, and subject to

$$\tilde{x}_{k+1} = \tilde{w}_k, \quad \tilde{x}_k \in \mathcal{S}^+,\ k = 0, \ldots, N-1,$$
$$p_{\tilde{w}|\tilde{x},\tilde{u}}(j | i, \tilde{\mu}_k(i)) = \tilde{P}_{ij}(\tilde{\mu}_k(i)).$$

Discounted Problems (4/4)

Objective

Construct an optimal policy $\tilde{\pi}^*$ with associated optimal cost $\tilde{J}^*(i) = \tilde{J}_{\tilde{\pi}^*}(i)$ such that for all $i \in \mathcal{S}^+$,

$$\tilde{\pi}^* = \arg\min_{\tilde{\pi} \in \tilde{\Pi}} \tilde{J}_{\tilde{\pi}}(i),$$

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

We will define an auxiliary stochastic shortest path problem and show that it is equivalent to the discounted problem.

Auxiliary SSP problem (1/5)

- *State*:
$$x_k \in \mathcal{S} = \mathcal{S}^+ \cup \{0\} = \{0, 1, \ldots, n\},$$
where 0 is a virtual terminal state.
- *Control*:
$$u_k \in \mathcal{U}(x_k),\ \forall x_k \in \mathcal{S},$$
where
$$\begin{aligned} \mathcal{U}(x_k) &:= \tilde{\mathcal{U}}(x_k)\ \forall x_k \in \mathcal{S}^+, \\ \mathcal{U}(0) &:= \{\texttt{stay}\}. \end{aligned}$$
`stay` is a virtual control action that is applied when the state is the virtual termination state. 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
$$u_k = \mu_k(x_k)\,,\ \ u_k \in \mathcal{U}(x_k),\ \forall x_k \in \mathcal{S},\ \forall k.$$

Auxiliary SSP problem (2/5)

- *Dynamics*:

$$x_{k+1} = w_k, \quad x_k \in \mathcal{S},$$

where the transition probabilities are generated from

$$\begin{aligned}
p_{w|x,u}(j|i,\mathrm{u}) &= P_{ij}(\mathrm{u}) := \alpha \tilde{P}_{ij}(\mathrm{u}), \quad \mathrm{u} \in \mathcal{U}(i), \quad \forall i, j \in \mathcal{S}^+, \\
p_{w|x,u}(0|i,\mathrm{u}) &= P_{i0}(\mathrm{u}) := 1 - \alpha, \quad \mathrm{u} \in \mathcal{U}(i), \quad \forall i \in \mathcal{S}^+, \\
p_{w|x,u}(j|0,\mathrm{u}) &= P_{0j}(\mathrm{u}) := 0, \quad \mathrm{u} = \texttt{stay}, \quad \forall j \in \mathcal{S}^+, \\
p_{w|x,u}(0|0,\mathrm{u}) &= P_{00}(\mathrm{u}) := 1, \quad \mathrm{u} = \texttt{stay}.
\end{aligned}$$

Note that this is a valid transition probability distribution since for any $i \in \mathcal{S}^+$, and for any $\mathrm{u} \in \mathcal{U}(i)$,

$$\sum_{j \in \mathcal{S}} P_{ij}(\mathrm{u}) = \sum_{j \in \mathcal{S}^+} \alpha \tilde{P}_{ij}(\mathrm{u}) + P_{i0}(\mathrm{u}) = \alpha \cdot 1 + (1 - \alpha) = 1,$$

and for $i = 0$, $\mathrm{u} = \texttt{stay}$,

$$\sum_{j \in \mathcal{S}} P_{0j}(\mathrm{u}) = \sum_{j \in \mathcal{S}^+} P_{0j}(\mathrm{u}) + P_{00}(\mathrm{u}) = 0 + 1 = 1.$$

Auxiliary SSP problem (3/5)

- *Cost*:
  The stage costs are defined as:

$$g(x_k, u_k, w_k) = \alpha^{-1}\tilde{g}(x_k, u_k, w_k)\,, \quad \forall u_k \in \mathcal{U}(x_k),\ \forall x_k, w_k \in \mathcal{S}^+,$$
$$g(x_k, u_k, 0) = 0, \quad \forall u_k \in \mathcal{U}(x_k),\ \forall x_k \in \mathcal{S}.$$

  The total expected closed loop cost starting at $r \in \mathcal{S}$ associated with policy $\pi \in \Pi$ is:

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

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

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

Auxiliary SSP problem (4/5)

Note that there is a one-to-one mapping between a policy π of the auxiliary problem to a policy $\tilde{\pi}$ of the discounted problem.

Indeed, the feedback law of the auxiliary problem just trivially assigns $\mu_k(0) = \texttt{stay}$, and for the rest of the states $x_k \in \mathcal{S}$ they remain the same.

We proceed in three steps to prove $J_\pi(i) = \tilde{J}_{\tilde{\pi}}(i)\,, \forall i \in \mathcal{S}^+$ (see the Lecture Notes for the details):

1. $p_{x_k,w_k|x_0}(i,j|r) = \alpha^{k+1} p_{\tilde{x}_k,\tilde{w}_k|\tilde{x}_0}(i,j|r)$;
2. $\underset{(X_1,W_0|x_0=r)}{\mathrm{E}} \left[g(x_k, \mu_k(x_k)\,, w_k)\right] = \underset{(\tilde{X}_1,\tilde{W}_0|\tilde{x}_0=r)}{\mathrm{E}} \left[\alpha^k \tilde{g}(\tilde{x}_k, \tilde{\mu}_k(\tilde{x}_k)\,, \tilde{w}_k)\right]$; and
3. $J_\pi(i) = \tilde{J}_{\tilde{\pi}}(i)\,, \forall i \in \mathcal{S}^+$, using the above.

Auxiliary SSP problem (5/5)

Once we proved that $J_\pi(i) = \tilde{J}_{\tilde{\pi}}(i)\,, \forall i \in \mathcal{S}^+$, the mapping of the policy that minimizes $J_\pi(i)$ minimizes $\tilde{J}_{\tilde{\pi}}(i)$. Thus, by solving the Bellman Equation for the auxiliary problem, we also obtain an optimal policy and optimal cost-to-go for the infinite horizon discounted problem.

From the Bellman Equation for the auxiliary problem we can derive the Bellman Equation for the discounted problem:

$$\begin{aligned} J^*(i) &= \min_{\mathrm{u} \in \mathcal{U}(i)} \left(q(i,\mathrm{u}) + \sum_{j=1}^{n} P_{ij}(\mathrm{u}) J^*(j) \right) \\ &= \min_{\mathrm{u} \in \tilde{\mathcal{U}}(i)} \left(q(i,\mathrm{u}) + \alpha \sum_{j=1}^{n} \tilde{P}_{ij}(\mathrm{u}) J^*(j) \right), \quad \forall i \in \mathcal{S}^+, \end{aligned}$$

where

$$q(i,\mathrm{u}) = \sum_{j=1}^{n} P_{ij}(\mathrm{u}) g(i,\mathrm{u},j) = \sum_{j=1}^{n} (\alpha \tilde{P}_{ij}(\mathrm{u}))(\alpha^{-1} \tilde{g}(i,\mathrm{u},j)) = \sum_{j=1}^{n} \tilde{P}_{ij}(\mathrm{u}) \tilde{g}(i,\mathrm{u},j).$$

Outline

Solving the Bellman Equation (cont'd)

- Linear Programming (LP)
- Discounted Problems
- Additional reading material

Additional reading material

The story of linear programming is tightly coupled with many branches of science, and it is rich in Nobel prizes and Fields medals.

- The origins date back to Gaspard Monge, who formulated the *Optimal Transport problem* in 1781.
- However, his formulation was too "hard", and no progress on the problem was made for almost 200 years!
- It was only in 1937 with the Soviet mathematician and economist Leonid Kantorovich (Nobel prize) that the theory was unlocked: Linear Programming was born.
- Nowadays, Optimal Transport has various applications in optimization, machine learning, image processing, biology, and many more.

An entertaining snapshot: `https://www.imaginary.org/sites/default/files/snapshots/snapshots-2018-013.pdf`