ECE276B: Planning & Learning in Robotics
Lecture 9: Infinite Horizon Problems and Stochastic Shortest Path

Lecturer:

Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistants:

Tianyu Wang: tiw161@eng.ucsd.edu

Yongxi Lu: yol070@eng.ucsd.edu

UC San Diego

JACOBS SCHOOL OF ENGINEERING

Electrical and Computer Engineering

Finite Horizon Optimal Control (Recap)

- Recall the (finite-state) **finite-horizon** optimal control problem:

$$\begin{aligned} \min_{\pi_{0:T-1}} \; & J_0^{\pi}(x_0) := \mathbb{E}_{x_{1:T}}\left[g_T(x_T) + \sum_{t=0}^{T-1} g_t(x_t, \pi_t(x_t)) \;\middle|\; x_0 \right] \\ \text{s.t.} \;\; & x_{t+1} \sim p_f(\cdot \mid x_t, \pi_t(x_t)), \qquad t = 0, \ldots, T-1 \\ & x_t \in \mathcal{X}, \;\; \pi_t(x_t) \in \mathcal{U}(x_t), \quad \forall x_t \in \mathcal{X} \end{aligned}$$

- The optimal cost $J_0^*(x_0) := \min_{\pi_{0:T-1}} J_0^{\pi}(x_0)$ and an optimal policy $\pi_{0:T-1}^*$ can be computed via the Dynamic Programming (DP) algorithm
- An open-loop policy is optimal for the deterministic finite-state (DFS) problem:

$$\begin{aligned} \min_{u_{0:T-1}} \; & g_T(x_T) + \sum_{t=0}^{T-1} g_t(x_t, u_t) \\ \text{s.t.} \;\; & x_{t+1} = f(x_t, u_t), \qquad t = 0, \ldots, T-1 \\ & x_t \in \mathcal{X}, \;\; u_t \in \mathcal{U}(x_t), \quad \forall x_t \in \mathcal{X} \end{aligned}$$

- The DFS problem is equivalent to the Shortest Path (SP) problem, which led to a **forward DP** algorithm and **label correcting** (LC) algorithms

Infinite Horizon Optimal Control

- In this lecture, we consider what happens with the standard optimal control problem as the planning horizon T goes to infinity
- To get a meaningful problem, we consider time-invariant stage-costs and no terminal cost:

$$
\begin{aligned}
\min_{\pi_{0:T-1}} \quad & J_0^{\pi}(x_0) := \mathbb{E}_{x_{1:T}}\left[\sum_{t=0}^{T-1} g(x_t, \pi_t(x_t)) \,\middle|\, x_0\right] \\
\text{s.t.} \quad & x_{t+1} \sim p_f(\cdot \mid x_t, \pi_t(x_t)), \quad t = 0, \ldots, T-1 \\
& x_t \in \mathcal{X}, \quad \pi_t(x_t) \in \mathcal{U}(x_t), \quad \forall x_t \in \mathcal{X}
\end{aligned}
$$

- As $T \to \infty$, the complexity collapses since the time-invariant dynamics and state costs lead to a **time-invariant** cost-to-go and associated optimal policy.

Infinite Horizon Dynamic Programming

- For fixed T, the DP algorithm is:

$$V_T(x) = 0, \quad \forall x \in \mathcal{X}$$
$$V_t(x) = \min_{u \in \mathcal{U}(x)} g(x,u) + \mathbb{E}_{x' \sim p_f(\cdot|x,u)}\left[V_{t+1}(x')\right], \quad \forall x \in \mathcal{X}, t = T-1, \ldots, 0$$

- **Bellman Equation**: as $T \to \infty$, the sequence $\ldots, V_{t+1}(x), V_t(x), \ldots$ converges to a fixed point $V(x)$ and the DP algorithm reduces to:

$$V(x) = \min_{u \in \mathcal{U}(x)} g(x,u) + \mathbb{E}_{x' \sim p_f(\cdot|x,u)}\left[V(x')\right], \quad \forall x \in \mathcal{X}$$

- Assuming this convergence, $V(x)$ is equal to the optimal cost-to-go $J^*(x)$, which suggests that both the value function and the opitmal policy are time-invariant, or **stationary**.

- The Bellman Equation may seem simple but it needs to be solved for all $x \in \mathcal{X}$ simultaneously, which can be done analytically only for very few problems (e.g., the Linear Quadratic Regulator (LQR) problem).

The Stochastic Shortest Path (SSP) Problem

- The convergence on the previous slide does not hold for all problems
- The SSP problem is one instance in which the convergence holds and solving the Bellman Equation yields the optimal cost-to-go and an associated optimal stationary policy
- Consider a finite state problem with $\mathcal{X} := \{0, 1, \ldots, n\}$ and a finite control set $\mathcal{U}(i)$ for all $i \in \mathcal{X}$
- **Dynamics**: specified by matrices (correcting our previous notation): $P_{ij}^{u} = \mathbb{P}(x_{t+1} = j \mid x_t = i, u_t = u) = p_f(j \mid x_t = i, u_t = u)$
- **Terminal State Assumption**: Suppose that state 0 is a cost-free termination state (the goal), i.e., $P_{0,0}^{u} = 1$ and $g(0, u) = 0$, $\forall u \in \mathcal{U}(0)$

Existence of Solution to the SSP Problem

- **Proper Stationary Policy**: a policy π for which there exists an integer m such that $\mathbb{P}(x_m = 0 \mid x_0 = i) > 0$ for all $i \in \mathcal{X}$ subject to transitions governed by P_{ij}^u with $u = \pi(i)$
- **Proper Policy Assumption**: there exists at least one proper policy π. Furthermore, for every improper policy π', the corresponding cost function $J^{\pi'}(i)$ is infinite for at least one state $i \in \mathcal{X}$.
- The above assumption is required to ensure that:
  - there exists a unique solution to the Bellman Equation for SSP
  - a policy exists for which the probability of reaching the termination state goes to 1 as $T \to \infty$
  - policies that do not reach the termination state incur infinite cost (i.e., there are no non-positive cycles as in the SP problem)

Theorem: Bellman Equation for the SSP Problem

Under the termination state and proper policy assumptions, the following are true for the SSP problem:

1. Given any initial conditions $\bar{V}_0(1), \ldots, \bar{V}_0(n)$ (corresp. to $T = \infty$), the sequence $\bar{V}_k(i)$ generated by the iteration:

$$\bar{V}_{k+1}(i) = \min_{u \in \mathcal{U}(i)} \left[g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} \bar{V}_k(j) \right], \quad \forall i \in \mathcal{X} \setminus \{0\}$$

   converges to the optimal cost $J^*(i)$ for all $i \in \mathcal{X} \setminus \{0\}$

2. The optimal costs satisfy the **Bellman Equation**:

$$J^*(i) = \min_{u \in \mathcal{U}(i)} \left[g(i,u) + \sum_{j=1}^{n} P_{ij}^{u} J^*(j) \right], \quad \forall i \in \mathcal{X} \setminus \{0\}$$

3. The solution to the Bellman Equation is **unique**
4. The minimizing u of the Bellman Equation for each $i \in \mathcal{X} \setminus \{0\}$ gives an optimal policy, which is **stationary**

Theorem Intuition

- We give intuition under a stronger assumption: $\exists m \in \mathbb{N}$ such that for **any** admissible policy $\mathbb{P}(x_m = 0 \mid x_0 = i) > 0$, subject to transitions governed by P_{ij}^u with $u = \pi(i)$, i.e., there is a positive probability that the termination state will be reached regardless of the initial state.

1. Let $\bar{V}_0(0) = 0$ and consider the following finite-horizon problem:

$$J_0^\pi(i) = \mathbb{E}\left[\sum_{t=0}^{T-1} g(x_t, \pi_t(x_t)) + \bar{V}_0(x_T) \,\middle|\, x_0 = i\right]$$

where $\bar{V}_0(x_T)$ is the terminal cost. As $T \to \infty$, the probability that state 0 is reached approaches 1 for all policies and, since $\bar{V}_0(0) = 0$, the terminal cost does not influence the solution. The DP algorithm with re-labeled time index $k := T - t$ applied to this problem is:

$$\bar{V}_{k+1}(i) = \min_{u \in \mathcal{U}(i)} \left(g(i,u) + \sum_{j=1}^{n} P_{ij}^u \bar{V}_k(j) \right), \quad \forall i \in \mathcal{X} \setminus \{0\}, k = 0, \dots, T \quad (*)$$

where state 0 can be excluded because $g(0,u) = 0$ by assumption and $P_{0,j}^u = 0$ for all $j \in \mathcal{X} \setminus \{0\}$.

Theorem Intuition

1. Thus, $\bar{V}_T(i) = J_0^*(i)$ is the optimal cost for the finite horizon problem and as $T \to \infty$ it converges to the optimal cost of the infinite horizon problem due to the assumption that the terminal state is reached in finite time.

2. Follows from taking limits of both sides of (*) above.

3. Let $J_0(1), \ldots, J_0(n)$ and $\bar{J}_0(1), \ldots, \bar{J}_0(n)$ be two different solutions to the Bellman Equation. If both are used as initial conditions for (*) above, they both converge after 1 iteration. This leads to two different optimal costs which is a contradiction.

The Discounted Problem

- A class of infinite horizon problems in which there is no terminal state assumption but future stage costs are discounted (i.e., multiplied by γ^t for $\gamma \in [0, 1)$). This turns out to be equivalent to the SSP problem.
- Finite state space $\mathcal{X} := \{1, \ldots, n\}$ (no need for a terminal state) and finite control set $\mathcal{U}(i)$ for all $i \in \mathcal{X}$
- **Dynamics**: specified by matrices $P^u_{ij} := \mathbb{P}(x_{t+1} = j \mid x_t = i, u_t = u)$
- **Discounted Infinite Horizon Problem**: solve the following optimization as $T \to \infty$

$$\begin{aligned} \min_{\pi_{0:T-1}} \quad & J_0^{\pi}(i) := \mathbb{E}_{x_{1:T}} \left[\sum_{t=0}^{T-1} \gamma^t g(x_t, \pi_t(x_t)) \;\middle|\; x_0 = i \right] \\ \text{s.t.} \quad & x_{t+1} \sim p_f(\cdot \mid x_t, \pi_t(x_t)), \quad t = 0, \ldots, T-1 \\ & x_t \in \mathcal{X}, \quad \pi_t(x_t) \in \mathcal{U}(x_t), \quad \forall x_t \in \mathcal{X} \end{aligned}$$

- We define an auxiliary SSP problem and show that it is equivalent to the discounted problem

Equivalence between Discounted and SSP Problems

- **States**: $x_t \in \tilde{\mathcal{X}} := \mathcal{X} \cup \{0\}$, where 0 is a virtual terminal state
- **Control**: $u_t \in \tilde{\mathcal{U}}(x_t)$ where $\tilde{\mathcal{U}}(x_t) = \mathcal{U}(x_t)$ for $x_t \in \mathcal{X}$ and $\tilde{\mathcal{U}}(0) = \{stay\}$
- **Dynamics**:

$$\begin{aligned}
\tilde{P}^u_{ij} &= \gamma P^u_{ij}, && \text{for } u \in \tilde{\mathcal{U}}(i) \text{ and } i, j \in \mathcal{X} \\
\tilde{P}^u_{i,0} &= 1 - \gamma, && \text{for } u \in \tilde{\mathcal{U}}(i) \text{ and } i \in \mathcal{X} \\
\tilde{P}^u_{0,j} &= 0, && \text{for } u = stay \text{ and } j \in \mathcal{X} \\
\tilde{P}^u_{0,0} &= 1, && \text{for } u = stay
\end{aligned}$$

- **Terminal state and proper policy assumptions**: since $\gamma < 1$, there is a non-zero probability to go to state 0 regardless of the control input and initial state and hence the SSP assumptions are satisfied.
- **Cost**:

$$\begin{aligned}
\tilde{g}(x_t, u_t) &= g(x_t, u_t), && \text{for } u \in \tilde{\mathcal{U}}(x_t), x_t \in \mathcal{X} \\
\tilde{g}(0, stay) &= 0
\end{aligned}$$

Equivalence between Discounted and SSP Problems

- There is a one-to-one mapping between a policy $\tilde{\pi}$ of the auxiliary SSP to a policy π of the discounted problem since $\tilde{\pi}$ just trivially assigns $\tilde{\pi}_t(0) = \textit{stay}$ while the rest remains the same
- Next, we show that for all $i \in \mathcal{X}$:

$$\tilde{J}^{\tilde{\pi}}(i) = \mathbb{E}\left[\sum_{t=0}^{T-1} \tilde{g}(\tilde{x}_t, \tilde{\pi}_t(\tilde{x}_t)) \;\middle|\; x_0 = i\right] = J^{\pi}(i) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t g(x_t, \pi_t(x_t)) \;\middle|\; x_0 = i\right]$$

  where the expectations are over $\tilde{x}_{1:T}$ and $x_{1:T}$ and subject to transitions induced by $\tilde{\pi}$ and π, respectively.

- **Conclusion**: since $\tilde{J}^{\tilde{\pi}}(i) = J^{\pi}(i)$ for all $i \in \mathcal{X}$ and the mapping of $\tilde{\pi}$ to π minimizes $J^{\pi}(i)$, by solving the Bellman Equation for the auxiliary SSP, we can obtain an optimal policy and the optimal cost-to-go for the infinite-horizon discounted problem.

Equivalence between Discounted and SSP Problems

$$\mathbb{E}_{\tilde{x}_{1:T}}[\tilde{g}(\tilde{x}_t,\tilde{\pi}_t(\tilde{x}_t)) \mid x_0 = i] = \sum_{\bar{x}_{1:T} \in \tilde{\mathcal{X}}^T} \tilde{g}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t))\mathbb{P}(\tilde{x}_{1:T} = \bar{x}_{1:T} \mid x_0 = i)$$

$$= \sum_{\bar{x}_t \in \tilde{\mathcal{X}}} \tilde{g}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t))\mathbb{P}(\tilde{x}_t = \bar{x}_t \mid x_0 = i)$$

$$\overset{\tilde{g}(0,stay,0)=0}{=\!=\!=} \sum_{\bar{x}_t \in \mathcal{X}} \tilde{g}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t))\mathbb{P}(\tilde{x}_t = \bar{x}_t, \tilde{x}_t \neq 0 \mid x_0 = i)$$

$$= \sum_{\bar{x}_t \in \mathcal{X}} \tilde{g}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t))\mathbb{P}(\tilde{x}_t = \bar{x}_t \mid x_0 = i, \tilde{x}_t \neq 0)\mathbb{P}(\tilde{x}_t \neq 0 \mid x_0 = i)$$

$$\overset{(?)}{=\!=} \sum_{\bar{x}_t \in \mathcal{X}} \tilde{g}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t))\mathbb{P}(x_t = \bar{x}_t \mid x_0 = i)\gamma^t$$

$$= \sum_{\bar{x}_t \in \mathcal{X}} g(\bar{x}_t, \pi_t(\bar{x}_t))\mathbb{P}(x_t = \bar{x}_t \mid x_0 = i)\gamma^t$$

$$= \mathbb{E}_{x_{1:T}}\left[\gamma^t g(x_t, \pi_t(x_t)) \mid x_0 = i\right]$$

Equivalence between Discounted and SSP Problems

(?) Show that for transitions $\tilde{P}^u_{ij}$ under $\tilde{\pi}$, $\mathbb{P}(\tilde{x}_t \neq 0 \mid x_0 = i) = \gamma^t$

- For any $i \in \mathcal{X}$ and $u \in \tilde{\mathcal{U}}(i)$:

$$\mathbb{P}(\tilde{x}_{t+1} \neq 0 \mid \tilde{x}_t = i) = 1 - P^u_{i,0} = \gamma$$

- Similarly, for any $i \in \mathcal{X}$

$$\begin{aligned}
\mathbb{P}(\tilde{x}_{t+2} \neq 0 \mid \tilde{x}_t = i) &= \sum_{j \in \mathcal{X}} \mathbb{P}(\tilde{x}_{t+2} \neq 0 \mid \tilde{x}_{t+1} = j, \tilde{x}_t = i)\mathbb{P}(\tilde{x}_{t+1} = j \mid \tilde{x}_t = i) \\
&= \sum_{j \in \mathcal{X}} \mathbb{P}(\tilde{x}_{t+2} \neq 0 \mid \tilde{x}_{t+1} = j)\mathbb{P}(\tilde{x}_{t+1} = j \mid \tilde{x}_t = i) \\
&= \gamma \sum_{j \in \mathcal{X}} \tilde{P}^{\tilde{\pi}(i)}_{i,j} = \gamma^2
\end{aligned}$$

- Similarly, we can show that for any $m > 0$: $\mathbb{P}(\tilde{x}_{t+m} \neq 0 \mid x_t = i) = \gamma^m$

Equivalence between Discounted and SSP Problems

(?) Show that $\mathbb{P}(\tilde{x}_t = \bar{x}_t \mid x_0 = i, \tilde{x}_t \neq 0) = \mathbb{P}(x_t = \bar{x}_t \mid x_0 = i)$

- For any $i, j \in \mathcal{X}$ and $u = \tilde{\pi}_t(i) = \pi_t(i)$, we have

$$\mathbb{P}(\tilde{x}_{t+1} = j \mid \tilde{x}_{t+1} \neq 0, \tilde{x}_t = i, \tilde{u}_t = u) = \frac{\mathbb{P}(\tilde{x}_{t+1} = j, \tilde{x}_{t+1} \neq 0 \mid \tilde{x}_t = i, \tilde{u}_t = u)}{\mathbb{P}(\tilde{x}_{t+1} \neq 0 \mid \tilde{x}_t = i, \tilde{u}_t = u)}$$

$$= \frac{\tilde{P}_{ij}^u}{\gamma} = P_{ij}^u = \mathbb{P}(x_{t+1} = j \mid x_t = i, u_t = u)$$

- Similarly, it can be shown that for $\bar{x}_t \in \mathcal{X}$:

$$\mathbb{P}(\tilde{x}_t = \bar{x}_t \mid x_0 = i, \tilde{x}_t \neq 0) = \mathbb{P}(x_t = \bar{x}_t \mid x_0 = i)$$

Bellman Equation for the Discounted Problem

- **Discounted Infinite Horizon Problem**:

$$J^*(x) = \min_{\pi} \; J^\pi(x) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g(x_t, \pi(x_t)) \;\middle|\; x_0 = x\right]$$

$$\text{s.t.} \;\; x_{t+1} \sim p_f(\cdot \mid x_t, \pi(x_t)),$$

$$x_t \in \mathcal{X}, \quad \pi(x_t) \in \mathcal{U}(x_t), \quad \forall x_t \in \mathcal{X}$$

- The optimal cost of the Discounted problem satisfies the **Bellman Equation** (via the equivalence to the SSP problem):

$$J^*(i) = \min_{u \in \mathcal{U}(i)} \Big(g(i,u) + \gamma \sum_{j=1}^{n} P_{ij}^u J^*(j) \Big), \quad \forall i \in \mathcal{X}$$

- There exist several methods to solve the Bellman Equation for the Discounted and SSP problems:
  - Value Iteration
  - Policy Iteration
  - Linear Programming