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

Instructor:

Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistants:

Zhichao Li: zhl355@eng.ucsd.edu

Ehsan Zobeidi: ezobeidi@eng.ucsd.edu

Ibrahim Akbar: iakbar@eng.ucsd.edu

UC San Diego

JACOBS SCHOOL OF ENGINEERING

Electrical and Computer Engineering

Finite-Horizon Stochastic Optimal Control (Recap)

- Recall the **finite-horizon** stochastic optimal control problem:

$$
\begin{aligned}
\min_{\pi_{t:T-1}} \; & V_t^{\pi}(x_t) := \mathbb{E}_{x_{t+1:T}}\left[\gamma^{T-t}\mathfrak{q}(x_T) + \sum_{\tau=t}^{T-1}\gamma^{\tau-t}\ell_\tau(x_\tau, \pi_\tau(x_\tau)) \,\middle|\, x_t\right] \\
\text{s.t. } \; & x_{\tau+1} \sim p_f(\cdot \mid x_\tau, \pi_\tau(x_\tau)), \qquad \tau = t, \ldots, T-1 \\
& x_\tau \in \mathcal{X}, \\
& \pi_\tau(x_\tau) \in \mathcal{U}(x_\tau)
\end{aligned}
$$

$x \in \mathcal{X}$	state
$u \in \mathcal{U}(x)$	control
$p_f(x' \mid x, u)$	motion model
$x' = f(x, u, w)$	motion model
$\ell(x, u)$	stage cost
$\mathfrak{q}(x)$	terminal cost
T, γ	planning horizon and discount factor
$\pi_\tau(x)$	policy function at time τ
$V_\tau^{\pi}(x)$	value function at state x, time τ, under policy $\pi_{\tau:T-1}$

Finite-Horizon Stochastic Optimal Control (Recap)

- **Episode**: a random sequence ρ_t of states and controls from the start state x_t, following the system dynamics to termination under policy π:

$$\rho_t := x_t, u_t, x_1, u_1, \ldots, x_{T-1}, u_{T-1}, x_T \sim \pi$$

- **Long-term cost**: a random variable defined as the sum of the discounted stage costs along an episode $\rho_t := x_{t+1:T}, u_{t:T-1}$:

$$L_t(\rho_t) := L_t(x_{t+1:T}, u_{t:T-1}) := \gamma^{T-t}\mathfrak{q}(x_T) + \sum_{\tau=t}^{T-1} \gamma^{\tau-t}\ell_\tau(x_\tau, u_\tau)$$

- **Value function**: $V_t^\pi(x) := \mathbb{E}_{\rho_t\sim\pi}\left[L_t(\rho_t) \mid x_t = x\right]$
- **Optimal value function**: $V_t^*(x) := \min_\pi V_t^\pi(x)$
- **Optimal policy**: $\pi_{t:T-1}^* := \arg\min_\pi V_t^\pi(x)$ for all $x \in \mathcal{X}$
- The optimal value function and policy can be computed via the Dynamic Programming (DP) algorithm

Finite-Horizon Deterministic Optimal Control (Recap)

- Deterministic finite-state (DFS) optimal control problem:

$$\begin{aligned}
\min_{u_{t:T-1}} \quad & V_t^{u_{t:T-1}}(x_t) := \gamma^{T-t}\mathfrak{q}(x_T) + \sum_{\tau=t}^{T-1} \gamma^{\tau-t}\ell_\tau(x_\tau, u_\tau) \\
\text{s.t.} \quad & x_{\tau+1} = f(x_\tau, u_\tau), \qquad \tau = t, \dots, T-1 \\
& x_\tau \in \mathcal{X}, \\
& \pi_\tau(x_\tau) \in \mathcal{U}(x_\tau)
\end{aligned}$$

- An open-loop policy is optimal for the DFS problem
- The DFS problem is equivalent to the deterministic shortest path (DSP) problem, which led to the **forward DP** and **label correcting** algorithms

Infinite-Horizon Stochastic Optimal Control

- In this lecture, we will consider what happens with the 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:

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

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

Infinite-Horizon Dynamic Programming

- For fixed T, the DP algorithm is:

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

- **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:

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

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

Value Iteration Algorithm

- 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).
- Let $\bar{V}_0(x) := V_T(x)$. Below, $\bar{V}_0(x)$ corresponds to the terminal value function as $T \to \infty$
- **Value Iteration** (VI) algorithm: applies the DP recursion with an arbitrary initialization $\bar{V}_0(x)$ for all $x \in \mathcal{X}$:

$$\bar{V}_{t+1}(x) = \min_{u \in \mathcal{U}(x)} \left[\ell(x, u) + \gamma \sum_{x' \in \mathcal{X}} p_f(x' \mid x, u) \bar{V}_t(x') \right], \qquad \forall x \in \mathcal{X}$$

- VI requires infinite iterations for $\bar{V}_t(x)$ to converge to $V^*(x)$
- In practice, define a threshold for $|\bar{V}_{t+1}(x) - \bar{V}_t(x)|$ for all $x \in \mathcal{X}$

The Stochastic Shortest Path (SSP) Problem

- The convergence on the previous slide does not hold for all problems when $\gamma = 1$
- 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 $\tilde{\mathcal{X}} := \{0, 1, \ldots, n\}$ and a finite control set $\tilde{\mathcal{U}}(x)$ for all $x \in \tilde{\mathcal{X}}$
- **Motion model**: specified by matrices:

$$\tilde{P}^u_{ij} = \mathbb{P}(x_{t+1} = j \mid x_t = i, u_t = u) = \tilde{p}_f(j \mid x_t = i, u_t = u)$$

- **Terminal State Assumption**: Suppose that state 0 is a cost-free termination state (the goal), i.e., $\tilde{P}^u_{0,0} = \tilde{p}_f(0 \mid 0, u) = 1$ and $\tilde{\ell}(0, u) = 0$, $\forall u \in \tilde{\mathcal{U}}(0)$

Existence of Solutions to the SSP Problem

- **Proper Stationary Policy**: a policy π for which there exists an integer m such that $\mathbb{P}(x_m = 0 \mid x_0 = x) > 0$ for all $x \in \tilde{\mathcal{X}}$ subject to transitions governed by the motion model and policy π.
- **Proper Policy Assumption**: there exists at least one proper policy π. Furthermore, for every improper policy π', the corresponding value function $V^{\pi'}(x)$ is infinite for at least one state $x \in \tilde{\mathcal{X}}$.
- The above assumption is required to ensure that:
  - there exists a unique solution to the Bellman Equation for the 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 DSP problem)

Theorem: Bellman Equation for the SSP

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

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

$$\bar{V}_{t+1}(x) = \min_{u \in \tilde{\mathcal{U}}(x)} \left[\tilde{\ell}(x,u) + \sum_{x' \in \tilde{\mathcal{X}} \setminus \{0\}} \tilde{p}_f(x' \mid, x, u) \bar{V}_t(x') \right], \quad \forall x \in \tilde{\mathcal{X}} \setminus \{0\}$$

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

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

$$V^*(x) = \min_{u \in \tilde{\mathcal{U}}(x)} \left[\tilde{\ell}(x,u) + \sum_{x' \in \tilde{\mathcal{X}} \setminus \{0\}} \tilde{p}_f(x' \mid, x, u) V^*(x') \right], \quad \forall x \in \tilde{\mathcal{X}} \setminus \{0\}$$

3. The solution to the Bellman Equation is **unique**
4. The minimizing u of the Bellman Equation for each $x \in \tilde{\mathcal{X}} \setminus \{0\}$ 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 $\mathbb{P}(x_m = 0 \mid x_0 = x) > 0$, subject to transitions governed by the motion model and π, 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:

$$V_0^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{T-1} \tilde{\ell}(x_t, \pi_t(x_t)) + \bar{V}_0(x_T) \;\middle|\; x_0 = x\right]$$

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:

$$\bar{V}_{k+1}(x) = \min_{u \in \tilde{\mathcal{U}}(x)} \left(\tilde{\ell}(x, u) + \sum_{x' \in \tilde{\mathcal{X}} \setminus \{0\}} \tilde{p}_f(x' \mid x, u) \bar{V}_k(x')\right), \;\; \forall x \in \tilde{\mathcal{X}} \setminus \{0\}, k = 0, \ldots, T \quad (*)$$

where state 0 can be excluded because $\tilde{\ell}(0, u) = 0$ by assumption and $\tilde{p}_f(x' \mid 0, u) = 0$ for all $x' \in \tilde{\mathcal{X}} \setminus \{0\}$.

Theorem Intuition

1. Thus, $\bar{V}_T(x) = V_0^*(x)$ 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 $\bar{J}_0(1), \ldots, \bar{J}_0(n)$ and $\bar{V}_0(1), \ldots, \bar{V}_0(n)$ 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.

Equivalence between Discounted and SSP Problems

- It turns out that the infinite-horizon discounted problem (no terminal state assumption but future stage costs are discounted by γ^t for $\gamma \in [0, 1)$) is equivalent to the SSP problem.
- Given a Discounted problem, we can define an auxiliary SSP problem and show that it is equivalent
- **Discounted Problem**: $\mathcal{X} := \{1, \ldots, n\}, \mathcal{U}(x)$, $p_f(x' \mid x, u)$, $\ell(x, u)$
- **SSP**: $\tilde{\mathcal{X}} := \mathcal{X} \cup \{0\}$, where 0 is a virtual terminal state,

$$\tilde{\mathcal{U}}(x) := \begin{cases} \mathcal{U}(x), & x \in \mathcal{X} \\ \{stay\}, & x = 0 \end{cases}$$

Equivalence between Discounted and SSP Problems

- **SSP motion model**:

$$
\begin{aligned}
\tilde{p}_f(x' \mid x, u) &= \gamma p_f(x' \mid x, u) && \text{for } u \in \tilde{\mathcal{U}}(x) \text{ and } x, x' \in \mathcal{X} \\
\tilde{p}_f(0 \mid x, u) &= 1 - \gamma, && \text{for } u \in \tilde{\mathcal{U}}(x) \text{ and } x \in \mathcal{X} \\
\tilde{p}_f(x' \mid 0, u) &= 0, && \text{for } u = \mathit{stay} \text{ and } x' \in \mathcal{X} \\
\tilde{p}_f(0 \mid 0, u) &= 1, && \text{for } u = \mathit{stay}
\end{aligned}
$$

- **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.

- **SSP Cost**:

$$
\begin{aligned}
\tilde{\ell}(x, u) &= \ell(x, u), && \text{for } u \in \tilde{\mathcal{U}}(x), x \in \mathcal{X} \\
\tilde{\ell}(0, \mathit{stay}) &= 0
\end{aligned}
$$

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 $x \in \mathcal{X}$:

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

  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 $\tilde{V}^{\tilde{\pi}}(x) = V^{\pi}(x)$ for all $x \in \mathcal{X}$ and the mapping of $\tilde{\pi}$ to π minimizes $V^{\pi}(x)$, 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

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

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

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

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

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

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

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

Equivalence between Discounted and SSP Problems

(?) Show that for transitions $\tilde{p}_f(x' \mid x, u)$ under $\tilde{\pi}$, $\mathbb{P}(\tilde{x}_t \neq 0 \mid x_0 = x) = \gamma^t$

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

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

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

$$\begin{aligned}
\mathbb{P}(\tilde{x}_{t+2} \neq 0 \mid \tilde{x}_t = x) &= \sum_{x' \in \mathcal{X}} \mathbb{P}(\tilde{x}_{t+2} \neq 0 \mid \tilde{x}_{t+1} = x', \tilde{x}_t = x)\mathbb{P}(\tilde{x}_{t+1} = x' \mid \tilde{x}_t = x) \\
&= \sum_{x' \in \mathcal{X}} \mathbb{P}(\tilde{x}_{t+2} \neq 0 \mid \tilde{x}_{t+1} = x')\mathbb{P}(\tilde{x}_{t+1} = x' \mid \tilde{x}_t = x) \\
&= \gamma \sum_{x' \in \mathcal{X}} \tilde{p}_f(x' \mid x, \tilde{\pi}(x)) = \gamma^2
\end{aligned}$$

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

Equivalence between Discounted and SSP Problems

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

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

$$\mathbb{P}(\tilde{x}_{t+1} = x' \mid \tilde{x}_{t+1} \neq 0, \tilde{x}_t = x, \tilde{u}_t = u) = \frac{\mathbb{P}(\tilde{x}_{t+1} = x', \tilde{x}_{t+1} \neq 0 \mid \tilde{x}_t = x, \tilde{u}_t = u)}{\mathbb{P}(\tilde{x}_{t+1} \neq 0 \mid \tilde{x}_t = x, \tilde{u}_t = u)}$$

$$= \frac{\tilde{p}_f(x' \mid x, u)}{\gamma} = p_f(x' \mid x, u) = \mathbb{P}(x_{t+1} = x' \mid x_t = x, u_t = u)$$

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

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

Bellman Equation for the Discounted Problem

- **Discounted Infinite-Horizon Problem**:

$$V^*(x) = \min_{\pi} \; V^{\pi}(x) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(x_t, \pi(x_t)) \;\middle|\; x_0 = x\right]$$
$$\text{s.t.} \;\; x_{t+1} \sim p_f(\cdot \mid x_t, \pi(x_t)),$$
$$x_t \in \mathcal{X},$$
$$\pi(x_t) \in \mathcal{U}(x_t)$$

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

$$V^*(x) = \min_{u \in \mathcal{U}(x)} \Big(\ell(x, u) + \gamma \sum_{x' \in \mathcal{X}} p_f(x' \mid x, u) V^*(x') \Big), \quad \forall x \in \mathcal{X}$$

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