ECE276B: Planning & Learning in Robotics
Lecture 10: Stochastic Shortest Path

Instructor:
Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistant:
Hanwen Cao: h1cao@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:

$$\min_{\pi_{\tau:T-1}} V_\tau^\pi(\mathbf{x}_\tau) := \mathbb{E}_{\mathbf{x}_{\tau+1:T}}\left[\gamma^{T-\tau}\mathfrak{q}(\mathbf{x}_T) + \sum_{t=\tau}^{T-1}\gamma^{t-\tau}\ell(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) \,\middle|\, \mathbf{x}_\tau\right]$$

$$\text{s.t. } \mathbf{x}_{t+1} \sim p_f(\cdot \mid \mathbf{x}_t, \pi_t(\mathbf{x}_t)), \qquad t = \tau, \ldots, T-1$$

$$\mathbf{x}_t \in \mathcal{X}, \pi_t(\mathbf{x}_t) \in \mathcal{U}(\mathbf{x}_t)$$

$\mathbf{x} \in \mathcal{X}$	state
$\mathbf{u} \in \mathcal{U}$	control
$p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u})$	motion model
$\mathbf{x}' = f(\mathbf{x}, \mathbf{u}, \mathbf{w})$	motion model
$\ell(\mathbf{x}, \mathbf{u})$	stage cost
$\mathfrak{q}(\mathbf{x})$	terminal cost
$T \in \mathbb{N}$	planning horizon
$\gamma \in [0, 1]$	discount factor
$\pi_t(\mathbf{x})$	policy function at time t
$V_t^\pi(\mathbf{x})$	value function at state $\mathbf{x}$, time t, under policy $\pi_{t:T-1}$

Finite-Horizon Stochastic Optimal Control (Recap)

- **Episode**: a sequence ρ_τ of random states and controls from the start state $\mathbf{x}_\tau$, following the motion model to termination under policy π:

$$\rho_\tau := \mathbf{x}_\tau, \mathbf{u}_\tau, \mathbf{x}_{\tau+1}, \mathbf{u}_{\tau+1}, \ldots, \mathbf{x}_{T-1}, \mathbf{u}_{T-1}, \mathbf{x}_T \sim \pi$$

- **Long-term cost**: a random variable defined as the sum of the discounted stage costs along an episode ρ_τ:

$$L_\tau(\rho_\tau) := \gamma^{T-\tau}\mathfrak{q}(\mathbf{x}_T) + \sum_{t=\tau}^{T-1} \gamma^{t-\tau}\ell_t(\mathbf{x}_t, \mathbf{u}_t)$$

- **Value function**: $V_t^\pi(\mathbf{x}) := \mathbb{E}_{\rho_t \sim \pi}\left[L_t(\rho_t) \mid \mathbf{x}_t = \mathbf{x}\right]$

- **Optimal value function**: $V_t^*(\mathbf{x}) := \min_\pi V_t^\pi(\mathbf{x})$

- **Optimal policy**: $\pi_{t:T-1}^* := \arg\min\limits_{\pi} V_t^\pi(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$

- The optimal value function and policy can be computed via the Dynamic Programming algorithm

Finite-Horizon Deterministic Optimal Control (Recap)

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

$$\begin{aligned}\min_{\mathbf{u}_{\tau:T-1}} \quad & V_\tau^{\mathbf{u}_{\tau:T-1}}(\mathbf{x}_\tau) := \gamma^{T-\tau}\mathfrak{q}(\mathbf{x}_T) + \sum_{t=\tau}^{T-1} \gamma^{t-\tau}\ell_t(\mathbf{x}_t, \mathbf{u}_t) \\ \text{s.t.} \quad & \mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t), \qquad t = \tau, \ldots, T-1 \\ & \mathbf{x}_t \in \mathcal{X}, \\ & \mathbf{u}_t \in \mathcal{U}(\mathbf{x}_t)\end{aligned}$$

- An open-loop control sequence $\mathbf{u}^*_{\tau:T-1}$ 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 Discounted Stochastic Optimal Control

- In this lecture, we will consider what happens with the stochastic optimal control problem as the planning horizon T goes to infinity

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

- The terminal cost $\mathfrak{q}(\mathbf{x}_T)$ is no longer necessary since we never reach a terminal time-step

- As $T \to \infty$, the time-invariant motion model and stage costs lead to a **time-invariant** optimal value function $V^*(\mathbf{x}) = \min_\pi V^\pi(\mathbf{x})$ and associated optimal policy $\pi^*(\mathbf{x}) = \arg\min_\pi V^\pi(\mathbf{x})$.

Infinite-Horizon Discounted Stochastic Optimal Control

- As $T \to \infty$, it is sufficient to optimize over stationary value functions $V^\pi(\mathbf{x})$ and stationary policites $\pi(\mathbf{x})$
- **Infinite-Horizon Discounted Stochastic Optimal Control Problem**:

$$\begin{aligned} V^*(\mathbf{x}) = \min_\pi \; & V^\pi(\mathbf{x}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(\mathbf{x}_t, \pi(\mathbf{x}_t)) \;\middle|\; \mathbf{x}_0 = \mathbf{x}\right] \\ \text{s.t.} \;\; & \mathbf{x}_{t+1} \sim p_f(\cdot \mid \mathbf{x}_t, \pi(\mathbf{x}_t)), \\ & \mathbf{x}_t \in \mathcal{X}, \\ & \pi(\mathbf{x}_t) \in \mathcal{U}(\mathbf{x}_t) \end{aligned}$$

Infinite-Horizon Dynamic Programming

- Recall the Dynamic Programming algorithm for fixed T:

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

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

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

- Assuming convergence, $V(\mathbf{x})$ is equal to the optimal value $V^*(\mathbf{x})$
- Both $V^*(\mathbf{x})$ and the associated opitmal policy $\pi^*(\mathbf{x})$ are **stationary**
- The Bellman Equation needs to be solved for all $\mathbf{x} \in \mathcal{X}$ simultaneously, which can be done analytically only for very few problems (e.g., the Linear Quadratic Regulator (LQR) problem)

Value Iteration Algorithm

- Dynamic Programming for fixed T:

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

- Change the time index to $\bar{V}_t(\mathbf{x}) := V_{T-t}(\mathbf{x})$ so that $\bar{V}_0(\mathbf{x})$ corresponds to the terminal value function as $T \to \infty$

- **Value Iteration** (VI): applies dynamic programming with an arbitrary initialization $\bar{V}_0(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$:

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

- VI requires infinite iterations for $\bar{V}_t(\mathbf{x})$ to converge to $V^*(\mathbf{x})$

- In practice, the VI algorithm is terminated when $|\bar{V}_{t+1}(\mathbf{x}) - \bar{V}_t(\mathbf{x})| < \epsilon$ for all $\mathbf{x} \in \mathcal{X}$ and some threshold ϵ

Finite-State Stochastic Shortest Path Problem

- The VI algorithm does not always converge when $\gamma = 1$
- The **stochastic shortest path** (SSP) problem is one instance in which the VI algorithm converges to the optimal value function and corresponding optimal policy.
- **State space**: $\tilde{\mathcal{X}} := \{0, 1, \ldots, n\}$ (finite)
- **Control space**: $\tilde{\mathcal{U}}(x)$ (finite) for all $x \in \tilde{\mathcal{X}}$
- **Stage cost**: $\tilde{\ell}(x, u)$
- **Motion model**: specified by matrices $\tilde{P}^u$ for each u with elements:

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

- **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 Finite-State 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)

Finite-State Stochastic Shortest Path Problem

$$
\begin{aligned}
V^*(x) = \min_{\pi} V^{\pi}(x) &:= \mathbb{E}\left[\sum_{t=0}^{\infty} \tilde{\ell}(x_t, \pi_t(x_t)) \,\middle|\, x_0 = x\right] \\
\text{s.t. } & x_{t+1} \sim \tilde{p}_f(\cdot \mid x_t, \pi(x_t)), \\
& x_t \in \tilde{\mathcal{X}} := \{0, 1, \ldots, n\}, \\
& \pi(x_t) \in \tilde{\mathcal{U}}(x_t)
\end{aligned}
$$

- **Assumptions**:
  - **Terminal State**: $\tilde{p}_f(0 \mid 0, u) = 1$ and $\tilde{\ell}(0, u) = 0$, $\forall u \in \tilde{\mathcal{U}}(0)$
  - **Proper Policy**: there exists at least one proper policy π such that $\mathbb{P}(x_m = 0 \mid x_0 = x) > 0$ for some integer m under π. For every improper policy π', $V^{\pi'}(x) = \infty$ for some $x \in \tilde{\mathcal{X}}$.

Theorem: Bellman Equation for the Finite-State SSP

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

1. Given any initial conditions $\bar{V}_0(1), \ldots, \bar{V}_0(n)$, 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 value function $V^*(x)$ for all $x \in \tilde{\mathcal{X}} \setminus \{0\}$

2. The optimal value function satisfies 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:

$$\bar{V}_{k+1}(x) = \min_{u \in \tilde{\mathcal{U}}(x)} \Big(\tilde{\ell}(x, u) + \sum_{x' \in \tilde{\mathcal{X}} \setminus \{0\}} \tilde{p}_f(x' \mid x, u) \bar{V}_k(x')\Big), \;\; \forall x \in \tilde{\mathcal{X}} \setminus \{0\}, \quad (*)$$

for $k = 0, \ldots, T$. 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

- The infinite-horizon discounted problem with $\gamma \in [0, 1)$ 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

$$
\begin{aligned}
\mathbb{E}_{\tilde{x}_{1:T}}[\tilde{\ell}(\tilde{x}_t, \tilde{\pi}_t(\tilde{x}_t)) \mid x_0 = x] &= \sum_{\bar{x}_{1:T} \in \tilde{\mathcal{X}}^T} \tilde{\ell}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t)) \mathbb{P}(\tilde{x}_{1:T} = \bar{x}_{1:T} \mid x_0 = x) \\
&= \sum_{\bar{x}_t \in \tilde{\mathcal{X}}} \tilde{\ell}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t)) \mathbb{P}(\tilde{x}_t = \bar{x}_t \mid x_0 = x) \\
&\overset{\tilde{\ell}(0, stay, 0) = 0}{=\!=\!=} \sum_{\bar{x}_t \in \mathcal{X}} \tilde{\ell}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t)) \mathbb{P}(\tilde{x}_t = \bar{x}_t, \tilde{x}_t \neq 0 \mid x_0 = x) \\
&= \sum_{\bar{x}_t \in \mathcal{X}} \tilde{\ell}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t)) \mathbb{P}(\tilde{x}_t = \bar{x}_t \mid x_0 = x, \tilde{x}_t \neq 0) \mathbb{P}(\tilde{x}_t \neq 0 \mid x_0 = x) \\
&\overset{(?)}{=\!=} \sum_{\bar{x}_t \in \mathcal{X}} \tilde{\ell}(\bar{x}_t, \tilde{\pi}_t(\bar{x}_t)) \mathbb{P}(x_t = \bar{x}_t \mid x_0 = x) \gamma^t \\
&= \sum_{\bar{x}_t \in \mathcal{X}} \ell(\bar{x}_t, \pi_t(\bar{x}_t)) \mathbb{P}(x_t = \bar{x}_t \mid x_0 = x) \gamma^t \\
&= \mathbb{E}_{x_{1:T}} \left[\gamma^t \ell(x_t, \pi_t(x_t)) \mid x_0 = x \right]
\end{aligned}
$$

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

- **Infinite-Horizon Discounted Problem**:

$$V^*(\mathbf{x}) = \min_{\pi} \quad V^{\pi}(\mathbf{x}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(\mathbf{x}_t, \pi(\mathbf{x}_t)) \;\middle|\; \mathbf{x}_0 = \mathbf{x}\right]$$

$$\text{s.t.} \quad \mathbf{x}_{t+1} \sim p_f(\cdot \mid \mathbf{x}_t, \pi(\mathbf{x}_t)),$$

$$\mathbf{x}_t \in \mathcal{X},$$

$$\pi(\mathbf{x}_t) \in \mathcal{U}(\mathbf{x}_t)$$

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

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

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