

ECE276B: Planning & Learning in Robotics

Lecture 10: Infinite-Horizon Optimal Control

Nikolay Atanasov
natanasov@ucsd.edu

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering

Outline

Infinite-Horizon Optimal Control

Bellman Equations

Policy Evaluation

Value Iteration

Policy Iteration

Linear Programming

Finite-Horizon Stochastic Optimal Control

- Recall the finite-horizon stochastic optimal control problem:

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

| | |
|---|--|
| $\mathbf{x} \in \mathcal{X}$ | 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 |
| $q(\mathbf{x})$ | 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 Deterministic Optimal Control

- ▶ Finite-horizon deterministic optimal control (DOC) problem:

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

- ▶ An open-loop control sequence $\mathbf{u}_{\tau:T-1}^*$ is optimal for the DOC problem
- ▶ The DOC problem is equivalent to the deterministic shortest path (DSP) problem, which led to the forward Dynamic Programming and Label Correcting algorithms

Infinite-Horizon Stochastic Optimal Control

- ▶ In this lecture, we consider what happens with the stochastic optimal control problem as the planning horizon T goes to infinity
- ▶ We will consider two formulations of the infinite-horizon stochastic optimal control problem
 - ▶ **Discounted Problem:** obtained by letting $T \rightarrow \infty$ in the finite-horizon stochastic optimal control problem with $\gamma < 1$
 - ▶ **First-Exit Problem:** obtained by considering stochastic transitions in the shortest path problem and terminating when the goal region is reached
- ▶ Just like the DOC and DSP problems, the Discounted Problem and the First-Exit Problem are equivalent, i.e., one can be converted into the other

Discounted Problem

- ▶ Let $T \rightarrow \infty$ in the finite-horizon stochastic optimal control problem
- ▶ The terminal cost q is no longer necessary since the problem never terminates
- ▶ Assume the motion model p_f and the stage cost ℓ are time-invariant
- ▶ The discount factor γ must be < 1 to ensure that the infinite sum of stage costs is finite
- ▶ As $T \rightarrow \infty$, the time-invariant motion model and stage costs lead to **time-invariant** optimal value function $V^*(\mathbf{x}) = \min_{\pi} V^{\pi}(\mathbf{x})$ and associated optimal policy $\pi^*(\mathbf{x}) \in \arg \min_{\pi} V^{\pi}(\mathbf{x})$
- ▶ **Discounted Problem:**

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

s.t. $\mathbf{x}_{t+1} \sim p_f(\cdot \mid \mathbf{x}_t, \pi(\mathbf{x}_t))$,
 $\mathbf{x}_t \in \mathcal{X}, \pi(\mathbf{x}_t) \in \mathcal{U}$

First-Exit Problem

- ▶ Consider a stochastic shortest path problem with state space \mathcal{X} and transitions defined by $p_f(\mathbf{x}'|\mathbf{x}, \mathbf{u})$ with control $\mathbf{u} \in \mathcal{U}$
- ▶ Let $\mathcal{T} \subseteq \mathcal{X}$ be a set of **terminal states** with terminal cost $q(\mathbf{x})$ for $\mathbf{x} \in \mathcal{T}$
- ▶ **First-Exit Time**: terminate at $T := \min \{t \geq 0 \mid \mathbf{x}_t \in \mathcal{T}\}$, the first passage time from an initial state \mathbf{x}_0 to a terminal state $\mathbf{x}_t \in \mathcal{T}$
- ▶ Note that T is a **random variable** unlike in the finite-horizon problem
- ▶ **First-Exit Problem**:

$$V^*(\mathbf{x}) = \min_{\pi} V^{\pi}(\mathbf{x}) := \mathbb{E} \left[q(\mathbf{x}_T) + \sum_{t=0}^{T-1} \ell(\mathbf{x}_t, \pi(\mathbf{x}_t)) \mid \mathbf{x}_0 = \mathbf{x} \right]$$

s.t. $\mathbf{x}_{t+1} \sim p_f(\cdot \mid \mathbf{x}_t, \pi(\mathbf{x}_t))$,
 $\mathbf{x}_t \in \mathcal{X}, \pi(\mathbf{x}_t) \in \mathcal{U}$

From Discounted Problem to First-Exit Problem

- ▶ Given a Discounted Problem, we can define an equivalent First-Exit problem

- ▶ **Discounted Problem:** $\mathcal{X}, \mathcal{U}, p_f(\mathbf{x}'|\mathbf{x}, \mathbf{u}), \ell(\mathbf{x}, \mathbf{u}), \gamma$

- ▶ **First-Exit Problem:** $\tilde{\mathcal{X}}, \tilde{\mathcal{U}}, \tilde{p}_f(\mathbf{x}'|\mathbf{x}, \mathbf{u}), \tilde{\ell}(\mathbf{x}, \mathbf{u}), \tilde{q}(\mathbf{x}), \tilde{\mathcal{T}}$

- ▶ State space: $\tilde{\mathcal{X}} = \mathcal{X} \cup \{\tau\}$ and $\tilde{\mathcal{T}} = \{\tau\}$ where τ is a virtual terminal state
- ▶ Control space: $\tilde{\mathcal{U}} = \mathcal{U}$
- ▶ Motion model:

$$\tilde{p}_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) = \gamma p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) \quad \text{for } \mathbf{x}' \neq \tau$$

$$\tilde{p}_f(\tau | \mathbf{x}, \mathbf{u}) = 1 - \gamma,$$

$$\tilde{p}_f(\mathbf{x}' | \tau, \mathbf{u}) = 0, \quad \text{for } \mathbf{x}' \neq \tau$$

$$\tilde{p}_f(\tau | \tau, \mathbf{u}) = 1,$$

- ▶ Stage cost: $\tilde{\ell}(\mathbf{x}, \mathbf{u}) = \begin{cases} \ell(\mathbf{x}, \mathbf{u}) & \mathbf{x} \neq \tau \\ 0 & \mathbf{x} = \tau \end{cases}$

- ▶ Terminal cost: $\tilde{q}(\mathbf{x}) = 0$

- ▶ There is a one-to-one mapping between a policy $\tilde{\pi}$ of this first-exit problem and a policy π of the discounted problem:

$$\tilde{\pi}(\mathbf{x}) = \begin{cases} \pi(\mathbf{x}) & \mathbf{x} \neq \tau \\ \text{some } \mathbf{u} \in \mathcal{U}, & \mathbf{x} = \tau \end{cases}$$

From Discounted Problem to First-Exit Problem

- ▶ Next, we show that for all $\mathbf{x} \in \mathcal{X}$:

$$\tilde{V}^{\tilde{\pi}}(\mathbf{x}) = \mathbb{E} \left[\sum_{t=0}^{T-1} \tilde{\ell}(\tilde{\mathbf{x}}_t, \tilde{\pi}_t(\tilde{\mathbf{x}}_t)) \mid \tilde{\mathbf{x}}_0 = \mathbf{x} \right] = V^{\pi}(\mathbf{x}) = \mathbb{E} \left[\sum_{t=0}^{T-1} \gamma^t \ell(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) \mid \mathbf{x}_0 = \mathbf{x} \right]$$

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

- ▶ **Conclusion:** since $\tilde{V}^{\tilde{\pi}}(\mathbf{x}) = V^{\pi}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and $\tilde{\pi}$ maps to π , by solving the auxiliary First-Exit Problem, we can obtain an optimal policy and the optimal value for the Discounted Problem

From Discounted Problem to First-Exit Problem

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbf{x}}_{1:T}}[\tilde{\ell}(\tilde{\mathbf{x}}_t, \tilde{\pi}_t(\tilde{\mathbf{x}}_t)) \mid \mathbf{x}_0 = \mathbf{x}] &= \sum_{\bar{\mathbf{x}}_{1:T} \in \tilde{\mathcal{X}}^T} \tilde{\ell}(\bar{\mathbf{x}}_t, \tilde{\pi}_t(\bar{\mathbf{x}}_t)) \mathbb{P}(\tilde{\mathbf{x}}_{1:T} = \bar{\mathbf{x}}_{1:T} \mid \mathbf{x}_0 = \mathbf{x}) \\ &= \sum_{\bar{\mathbf{x}}_t \in \tilde{\mathcal{X}}} \tilde{\ell}(\bar{\mathbf{x}}_t, \tilde{\pi}_t(\bar{\mathbf{x}}_t)) \mathbb{P}(\tilde{\mathbf{x}}_t = \bar{\mathbf{x}}_t \mid \mathbf{x}_0 = \mathbf{x}) \\ \frac{\tilde{\ell}(\tau, \mathbf{u})=0}{\tilde{\mathcal{X}} = \mathcal{X} \cup \{\tau\}} &\sum_{\bar{\mathbf{x}}_t \in \mathcal{X}} \tilde{\ell}(\bar{\mathbf{x}}_t, \tilde{\pi}_t(\bar{\mathbf{x}}_t)) \mathbb{P}(\tilde{\mathbf{x}}_t = \bar{\mathbf{x}}_t, \tilde{\mathbf{x}}_t \neq \tau \mid \mathbf{x}_0 = \mathbf{x}) \\ &= \sum_{\bar{\mathbf{x}}_t \in \mathcal{X}} \tilde{\ell}(\bar{\mathbf{x}}_t, \tilde{\pi}_t(\bar{\mathbf{x}}_t)) \mathbb{P}(\tilde{\mathbf{x}}_t = \bar{\mathbf{x}}_t \mid \mathbf{x}_0 = \mathbf{x}, \tilde{\mathbf{x}}_t \neq \tau) \mathbb{P}(\tilde{\mathbf{x}}_t \neq \tau \mid \mathbf{x}_0 = \mathbf{x}) \\ \stackrel{(?)}{=} &\sum_{\bar{\mathbf{x}}_t \in \mathcal{X}} \tilde{\ell}(\bar{\mathbf{x}}_t, \tilde{\pi}_t(\bar{\mathbf{x}}_t)) \mathbb{P}(\mathbf{x}_t = \bar{\mathbf{x}}_t \mid \mathbf{x}_0 = \mathbf{x}) \gamma^t \\ &= \sum_{\bar{\mathbf{x}}_t \in \mathcal{X}} \ell(\bar{\mathbf{x}}_t, \pi_t(\bar{\mathbf{x}}_t)) \mathbb{P}(\mathbf{x}_t = \bar{\mathbf{x}}_t \mid \mathbf{x}_0 = \mathbf{x}) \gamma^t \\ &= \mathbb{E}_{\mathbf{x}_{1:T}} [\gamma^t \ell(\mathbf{x}_t, \pi_t(\mathbf{x}_t)) \mid \mathbf{x}_0 = \mathbf{x}]\end{aligned}$$

From Discounted Problem to First-Exit Problem

(?) Show that for transitions $\tilde{p}_f(\mathbf{x}' | \mathbf{x}, \mathbf{u})$ under $\tilde{\pi}$, $\mathbb{P}(\tilde{\mathbf{x}}_t \neq \tau | \mathbf{x}_0 = \mathbf{x}) = \gamma^t$

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

$$\mathbb{P}(\tilde{\mathbf{x}}_{t+1} \neq \tau | \tilde{\mathbf{x}}_t = \mathbf{x}) = 1 - \tilde{p}_f(\tau | \mathbf{x}, \mathbf{u}) = \gamma$$

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

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

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

From Discounted Problem to First-Exit Problem

(?) Show that $\mathbb{P}(\tilde{\mathbf{x}}_t = \bar{\mathbf{x}}_t \mid \mathbf{x}_0 = \mathbf{x}, \tilde{\mathbf{x}}_t \neq \tau) = \mathbb{P}(\mathbf{x}_t = \bar{\mathbf{x}}_t \mid \mathbf{x}_0 = \mathbf{x})$

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

$$\begin{aligned}\mathbb{P}(\tilde{\mathbf{x}}_{t+1} = \mathbf{x}' \mid \tilde{\mathbf{x}}_{t+1} \neq \tau, \tilde{\mathbf{x}}_t = \mathbf{x}, \tilde{\mathbf{u}}_t = \mathbf{u}) &= \frac{\mathbb{P}(\tilde{\mathbf{x}}_{t+1} = \mathbf{x}', \tilde{\mathbf{x}}_{t+1} \neq \tau \mid \tilde{\mathbf{x}}_t = \mathbf{x}, \tilde{\mathbf{u}}_t = \mathbf{u})}{\mathbb{P}(\tilde{\mathbf{x}}_{t+1} \neq \tau \mid \tilde{\mathbf{x}}_t = \mathbf{x}, \tilde{\mathbf{u}}_t = \mathbf{u})} \\ &= \frac{\tilde{p}_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u})}{\gamma} = p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}) = \mathbb{P}(\mathbf{x}_{t+1} = \mathbf{x}' \mid \mathbf{x}_t = \mathbf{x}, \mathbf{u}_t = \mathbf{u})\end{aligned}$$

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

$$\mathbb{P}(\tilde{\mathbf{x}}_t = \bar{\mathbf{x}}_t \mid \mathbf{x}_0 = \mathbf{x}, \tilde{\mathbf{x}}_t \neq \tau) = \mathbb{P}(\mathbf{x}_t = \bar{\mathbf{x}}_t \mid \mathbf{x}_0 = \mathbf{x})$$

Outline

Infinite-Horizon Optimal Control

Bellman Equations

Policy Evaluation

Value Iteration

Policy Iteration

Linear Programming

Bellman Equation

- ▶ Recall the Dynamic Programming algorithm for finite horizon T :

$$V_T(\mathbf{x}) = q(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}$$

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

- ▶ **Bellman Equation:** as $T \rightarrow \infty$, the sequence $\dots, V_{t+1}(\mathbf{x}), V_t(\mathbf{x}), \dots$ converges to a fixed point $V(\mathbf{x})$ of the dynamic programming recursion:

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

- ▶ Assuming convergence, $V(\mathbf{x})$ is equal to the optimal value $V^*(\mathbf{x})$
- ▶ Both $V^*(\mathbf{x})$ and the associated optimal 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)

Bellman Equation

- ▶ The optimal value function $V^*(\mathbf{x})$ satisfies:

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

- ▶ The value function $V^\pi(\mathbf{x})$ of policy π satisfies:

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

- ▶ The latter can be obtained from:

$$\begin{aligned} V^\pi(\mathbf{x}) &:= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \ell(\mathbf{x}_t, \pi(\mathbf{x}_t)) \mid \mathbf{x}_0 = \mathbf{x} \right] \\ &= \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E} \left[\sum_{t=1}^{\infty} \gamma^{t-1} \ell(\mathbf{x}_t, \pi(\mathbf{x}_t)) \mid \mathbf{x}_0 = \mathbf{x} \right] \\ &= \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi(\mathbf{x}))} [V^\pi(\mathbf{x}')] \end{aligned}$$

Action-Value (Q) Function

- ▶ **Value Function** $V^\pi(\mathbf{x})$: the expected long-term cost of following policy π starting from state \mathbf{x}
- ▶ **Q Function** $Q^\pi(\mathbf{x}, \mathbf{u})$: the expected long-term cost of taking action \mathbf{u} in state \mathbf{x} and following policy π afterwards:

$$\begin{aligned} Q^\pi(\mathbf{x}, \mathbf{u}) &:= \ell(\mathbf{x}, \mathbf{u}) + \mathbb{E} \left[\sum_{t=1}^{\infty} \gamma^t \ell(\mathbf{x}_t, \pi(\mathbf{x}_t)) \mid \mathbf{x}_0 = \mathbf{x} \right] \\ &= \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V^\pi(\mathbf{x}')] \\ &= \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} \underbrace{[Q^\pi(\mathbf{x}', \pi(\mathbf{x}'))]}_{V^\pi(\mathbf{x}')} \end{aligned}$$

- ▶ **Optimal Q Function**: $Q^*(\mathbf{x}, \mathbf{u}) := \min_{\pi} Q^\pi(\mathbf{x}, \mathbf{u})$

$$\begin{aligned} Q^*(\mathbf{x}, \mathbf{u}) &= \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V^*(\mathbf{x}')] \\ &= \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} \left[\min_{\mathbf{u}' \in \mathcal{U}} Q^*(\mathbf{x}', \mathbf{u}') \right] \end{aligned}$$

$$\pi^*(\mathbf{x}) \in \arg \min_{\mathbf{u} \in \mathcal{U}} Q^*(\mathbf{x}, \mathbf{u})$$

Bellman Equations Summary

► **Value Function:**

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

► **Optimal Value Function:**

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

► **Q Function:**

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

► **Optimal Q Function:**

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

Bellman Operators

▶ **Hamiltonian:**

$$H[\mathbf{x}, \mathbf{u}, V] = \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V(\mathbf{x}')]]$$

▶ **Policy Evaluation Operator:**

$$\mathcal{B}_\pi[V](\mathbf{x}) := \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi(\mathbf{x}))} [V(\mathbf{x}')]] = H[\mathbf{x}, \pi(\mathbf{x}), V(\cdot)]$$

▶ **Optimal Value Operator:**

$$\mathcal{B}_*[V](\mathbf{x}) := \min_{\mathbf{u} \in \mathcal{U}} \{ \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V(\mathbf{x}')]] \} = \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V(\cdot)]$$

▶ **Policy Q-Evaluation Operator:**

$$\mathcal{B}_\pi[Q](\mathbf{x}, \mathbf{u}) := \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [Q(\mathbf{x}', \pi(\mathbf{x}'))] = H[\mathbf{x}, \mathbf{u}, Q(\cdot, \pi(\cdot))]$$

▶ **Optimal Q-Value Operator:**

$$\mathcal{B}_*[Q](\mathbf{x}, \mathbf{u}) := \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} \left[\min_{\mathbf{u}' \in \mathcal{U}} Q(\mathbf{x}', \mathbf{u}') \right] = H[\mathbf{x}, \mathbf{u}, \min_{\mathbf{u}' \in \mathcal{U}} Q(\cdot, \mathbf{u}')]]$$

Finite-Horizon Problem

- ▶ Trajectories terminate at fixed $T < \infty$

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

- ▶ The optimal value $V_t^*(\mathbf{x})$ can be found with a single backward pass through time, initialized from $V_T^*(\mathbf{x}) = q(\mathbf{x})$ and following the recursion:

Bellman Equations (Finite-Horizon Problem)

Hamiltonian: $H[\mathbf{x}, \mathbf{u}, V(\cdot)] = \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V(\mathbf{x}')]$

Policy Evaluation: $V_t^{\pi}(\mathbf{x}) = Q_t^{\pi}(\mathbf{x}, \pi_t(\mathbf{x})) = H[\mathbf{x}, \pi_t(\mathbf{x}), V_{t+1}^{\pi}(\cdot)]$

Bellman Equation: $V_t^*(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}} Q_t^*(\mathbf{x}, \mathbf{u}) = \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V_{t+1}^*(\cdot)]$

Optimal Policy: $\pi_t^*(\mathbf{x}) = \arg \min_{\mathbf{u} \in \mathcal{U}} Q_t^*(\mathbf{x}, \mathbf{u}) = \arg \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V_{t+1}^*(\cdot)]$

Infinite-Horizon Discounted Problem

- ▶ Trajectories continue forever but costs are discounted via $\gamma \in [0, 1)$:

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

Bellman Equations (Discounted Problem)

Hamiltonian: $H[\mathbf{x}, \mathbf{u}, V(\cdot)] = \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V(\mathbf{x}')]]$

Policy Evaluation: $V^{\pi}(\mathbf{x}) = Q^{\pi}(\mathbf{x}, \pi(\mathbf{x})) = H[\mathbf{x}, \pi(\mathbf{x}), V^{\pi}(\cdot)]$

Bellman Equation: $V^*(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}} Q^*(\mathbf{x}, \mathbf{u}) = \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V^*(\cdot)]$

Optimal Policy: $\pi^*(\mathbf{x}) = \arg \min_{\mathbf{u} \in \mathcal{U}} Q^*(\mathbf{x}, \mathbf{u}) = \arg \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V^*(\cdot)]$

Infinite-Horizon First-Exit Problem

- ▶ Trajectories terminate at $T := \inf \{t \geq 1 | \mathbf{x}_t \in \mathcal{T}\}$, the first passage time from initial state \mathbf{x}_0 to a terminal state $\mathbf{x}_t \in \mathcal{T} \subseteq \mathcal{X}$:

$$\min_{\pi} V^{\pi}(\mathbf{x}) = \mathbb{E} \left[q(\mathbf{x}_T) + \sum_{t=0}^{T-1} \ell(\mathbf{x}_t, \pi(\mathbf{x}_t)) \mid \mathbf{x}_0 = \mathbf{x} \right]$$

- ▶ At terminal states, $V^*(\mathbf{x}) = V^{\pi}(\mathbf{x}) = q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{T}$
- ▶ At other states, the following are satisfied:

Bellman Equations (First-Exit Problem)

Hamiltonian: $H[\mathbf{x}, \mathbf{u}, V(\cdot)] = \ell(\mathbf{x}, \mathbf{u}) + \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V(\mathbf{x}')]$

Policy Evaluation: $V^{\pi}(\mathbf{x}) = Q^{\pi}(\mathbf{x}, \pi(\mathbf{x})) = H[\mathbf{x}, \pi(\mathbf{x}), V^{\pi}(\cdot)]$

Bellman Equation: $V^*(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}} Q^*(\mathbf{x}, \mathbf{u}) = \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V^*(\cdot)]$

Optimal Policy: $\pi^*(\mathbf{x}) = \arg \min_{\mathbf{u} \in \mathcal{U}} Q^*(\mathbf{x}, \mathbf{u}) = \arg \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V^*(\cdot)]$

Bellman Equation Algorithms

- ▶ To determine the value function of policy π in either the Discounted or First-Exit Problem, we need to solve a **Policy Evaluation equation**:
 - ▶ Policy Evaluation: $V^\pi(\mathbf{x}) = H[\mathbf{x}, \pi(\mathbf{x}), V^\pi(\cdot)]$
 - ▶ Policy Q-Evaluation: $Q^\pi(\mathbf{x}, \mathbf{u}) = H[\mathbf{x}, \mathbf{u}, Q^\pi(\cdot, \pi(\cdot))]$
- ▶ A Policy Evaluation equation can be solved by:
 - ▶ Iterative Policy Evaluation
 - ▶ Linear System Solution (only for finite state space \mathcal{X})
- ▶ To determine the optimal value function in either the Discounted or First-Exit Problem, we need to solve a **Bellman equation**:
 - ▶ Bellman Equation: $V^*(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V^*(\cdot)]$
 - ▶ Q-Bellman Equation: $Q^*(\mathbf{x}, \mathbf{u}) = H[\mathbf{x}, \mathbf{u}, \min_{\mathbf{u}' \in \mathcal{U}} Q^*(\cdot, \mathbf{u}')]$
- ▶ A Bellman equation can be solved by:
 - ▶ Value Iteration
 - ▶ Policy Iteration
 - ▶ Linear Programming (only for finite state space \mathcal{X})

Outline

Infinite-Horizon Optimal Control

Bellman Equations

Policy Evaluation

Value Iteration

Policy Iteration

Linear Programming

Policy Evaluation

Policy Evaluation Theorem (Discounted Problem)

The value function $V^\pi(\mathbf{x})$ of policy π is the unique solution of:

$$V^\pi(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi(\mathbf{x}))} [V^\pi(\mathbf{x}')], \quad \forall \mathbf{x} \in \mathcal{X}.$$

If $\gamma \in [0, 1)$, for any initial condition $V_0(\mathbf{x})$, the sequence $V_k(\mathbf{x})$ generated by the recursion below converges to $V^\pi(\mathbf{x})$:

$$V_{k+1}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi(\mathbf{x}))} [V_k(\mathbf{x}')], \quad \forall \mathbf{x} \in \mathcal{X}.$$

- ▶ The PE algorithm requires infinite iterations for $V_k(\mathbf{x})$ to converge to $V^\pi(\mathbf{x})$
- ▶ In practice, the PE algorithm is terminated when $|V_{k+1}(\mathbf{x}) - V_k(\mathbf{x})| < \epsilon$ for all $\mathbf{x} \in \mathcal{X}$ and some threshold ϵ

Policy Evaluation

- ▶ **Proper policy for first-exit problem:** a policy π for which there exists an integer m such that $\mathbb{P}(\mathbf{x}_m \in \mathcal{T} \mid \mathbf{x}_0 = \mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$

Policy Evaluation Theorem (First-Exit Problem)

The value function $V^\pi(\mathbf{x})$ of policy π is the unique solution of:

$$\begin{aligned} V^\pi(\mathbf{x}) &= q(\mathbf{x}), & \forall \mathbf{x} \in \mathcal{T}, \\ V^\pi(\mathbf{x}) &= \ell(\mathbf{x}, \pi(\mathbf{x})) + \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi(\mathbf{x}))} [V^\pi(\mathbf{x}')], & \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}. \end{aligned}$$

If π is a proper policy, for any initial condition $V_0(\mathbf{x})$, the sequence $V_k(\mathbf{x})$ generated by the recursion below converges to $V^\pi(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$:

$$V_{k+1}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi(\mathbf{x}))} [V_k(\mathbf{x}')], \quad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}.$$

Policy Evaluation (Discounted Finite-State Problem)

- ▶ Let $\mathcal{X} = \{1, \dots, n\}$
- ▶ Let $\mathbf{v}_i := V^\pi(i)$, $\ell_i := \ell(i, \pi(i))$, $P_{ij} := p_f(j | i, \pi(i))$ for $i, j = 1, \dots, n$
- ▶ Policy evaluation:

$$\mathbf{v} = \ell + \gamma P \mathbf{v} \quad \Rightarrow \quad (I - \gamma P) \mathbf{v} = \ell$$

- ▶ Existence of solution: The matrix P has eigenvalues with modulus ≤ 1 . All eigenvalues of γP have modulus < 1 , so $(\gamma P)^T \rightarrow 0$ as $T \rightarrow \infty$ and $(I - \gamma P)^{-1}$ exists.
- ▶ The Policy Evaluation Theorem is an iterative solution to the linear system:

$$\mathbf{v}_1 = \ell + \gamma P \mathbf{v}_0$$

$$\mathbf{v}_2 = \ell + \gamma P \mathbf{v}_1 = \ell + \gamma P \ell + (\gamma P)^2 \mathbf{v}_0$$

\vdots

$$\mathbf{v}_k = (I + \gamma P + (\gamma P)^2 + \dots + (\gamma P)^{k-1}) \ell + (\gamma P)^k \mathbf{v}_0$$

\vdots

$$\mathbf{v}_\infty \rightarrow (I - \gamma P)^{-1} \ell$$

Policy Evaluation (First-Exit Finite-State Problem)

- ▶ Let $\mathcal{X} = \mathcal{N} \cup \mathcal{T}$ and $P_{ij} := p_f(j \mid i, \pi(i))$ for $i, j \in \mathcal{N} \cup \mathcal{T}$
- ▶ Let $\mathbf{q}_i := q(i)$ for $i \in \mathcal{T}$ and $\mathbf{v}_i := V^\pi(i)$, $\ell_i := \ell(i, \pi(i))$ for $i \in \mathcal{N}$
- ▶ Policy evaluation:

$$\mathbf{v} = \ell + P_{\mathcal{N}\mathcal{N}}\mathbf{v} + P_{\mathcal{N}\mathcal{T}}\mathbf{q} \quad \Rightarrow \quad (I - P_{\mathcal{N}\mathcal{N}})\mathbf{v} = \ell + P_{\mathcal{N}\mathcal{T}}\mathbf{q}$$

- ▶ Existence of solution: A unique solution for \mathbf{v} exists as long as π is a proper policy. By the Chapman-Kolmogorov equation, $[P^k]_{ij} = \mathbb{P}(\mathbf{x}_k = j \mid \mathbf{x}_0 = i)$ and since π is proper, $[P^k]_{ij} \rightarrow 0$ as $k \rightarrow \infty$ for all $i, j \in \mathcal{X} \setminus \mathcal{T}$. Since $P_{\mathcal{N}\mathcal{N}}^k$ vanishes as $k \rightarrow \infty$, all eigenvalues of $P_{\mathcal{N}\mathcal{N}}$ must have modulus less than 1 and $(I - P_{\mathcal{N}\mathcal{N}})^{-1}$ exists.
- ▶ The Policy Evaluation Theorem is an iterative solution to the linear system:

$$\mathbf{v}_1 = \ell + P_{\mathcal{N}\mathcal{T}}\mathbf{q} + P_{\mathcal{N}\mathcal{N}}\mathbf{v}_0$$

$$\mathbf{v}_2 = \ell + P_{\mathcal{N}\mathcal{T}}\mathbf{q} + P_{\mathcal{N}\mathcal{N}}\mathbf{v}_1 = \ell + P_{\mathcal{N}\mathcal{T}}\mathbf{q} + P_{\mathcal{N}\mathcal{N}}(\ell + P_{\mathcal{N}\mathcal{T}}\mathbf{q}) + P_{\mathcal{N}\mathcal{N}}^2\mathbf{v}_0$$

$$\mathbf{v}_\infty \rightarrow (I - P_{\mathcal{N}\mathcal{N}})^{-1}(\ell + P_{\mathcal{N}\mathcal{T}}\mathbf{q})$$

Outline

Infinite-Horizon Optimal Control

Bellman Equations

Policy Evaluation

Value Iteration

Policy Iteration

Linear Programming

Value Iteration

Value Iteration Theorem (Discounted Problem)

The optimal value function $V^*(\mathbf{x})$ is the unique solution of:

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

If $\gamma \in [0, 1)$, for any initial condition $V_0(\mathbf{x})$, the sequence $V_k(\mathbf{x})$ generated by the recursion below converges to $V^*(\mathbf{x})$:

$$V_{k+1}(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}} \left\{ \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V_k(\mathbf{x}')] \right\}, \quad \forall \mathbf{x} \in \mathcal{X}.$$

- ▶ The VI algorithm is an infinite-horizon equivalent of the DP algorithm ($V_0(\mathbf{x})$ in VI corresponds to $V_{T \rightarrow \infty}(\mathbf{x})$ in DP)
- ▶ VI requires infinite iterations for $V_k(\mathbf{x})$ to converge to $V^*(\mathbf{x})$
- ▶ In practice, the VI algorithm is terminated when $|V_{k+1}(\mathbf{x}) - V_k(\mathbf{x})| < \epsilon$ for all $\mathbf{x} \in \mathcal{X}$ and some threshold ϵ

Gauss-Seidel Value Iteration

- ▶ A regular VI implementation stores the values from a previous iteration and updates them for all states simultaneously:

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

- ▶ **Gauss-Seidel Value Iteration** updates the values in place:

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

- ▶ Gauss-Seidel VI converges and often leads to faster convergence and requires less memory than VI

Value Iteration

Value Iteration Theorem (First-Exit Problem)

The optimal value function $V^*(\mathbf{x})$ is the unique solution of:

$$\begin{aligned} V^*(\mathbf{x}) &= q(\mathbf{x}), & \forall \mathbf{x} \in \mathcal{T}, \\ V^*(\mathbf{x}) &= \min_{\mathbf{u} \in \mathcal{U}} \left\{ \ell(\mathbf{x}, \mathbf{u}) + \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V^*(\mathbf{x}')] \right\}, & \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}. \end{aligned}$$

If a proper policy exists, for any initial condition $V_0(\mathbf{x})$, the sequence $V_k(\mathbf{x})$ generated by the recursion below converges to $V^*(\mathbf{x})$:

$$\begin{aligned} V_k(\mathbf{x}) &= q(\mathbf{x}), & \forall \mathbf{x} \in \mathcal{T}, \forall k, \\ V_{k+1}(\mathbf{x}) &= \min_{\mathbf{u} \in \mathcal{U}} \left\{ \ell(\mathbf{x}, \mathbf{u}) + \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V_k(\mathbf{x}')] \right\}, & \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}. \end{aligned}$$

Contraction in Discounted Problems

Contraction Mapping

Let $\mathcal{F}(\mathcal{X})$ denote the linear space of bounded functions $V : \mathcal{X} \rightarrow \mathbb{R}$ with norm $\|V\|_\infty := \sup_{\mathbf{x} \in \mathcal{X}} |V(\mathbf{x})|$. A function $\mathcal{B} : \mathcal{F}(\mathcal{X}) \rightarrow \mathcal{F}(\mathcal{X})$ is called a *contraction mapping* if there exists a scalar $\alpha < 1$ such that:

$$\|\mathcal{B}[V] - \mathcal{B}[V']\|_\infty \leq \alpha \|V - V'\|_\infty \quad \forall V, V' \in \mathcal{F}(\mathcal{X})$$

Contraction Mapping Theorem

If $\mathcal{B} : \mathcal{F}(\mathcal{X}) \rightarrow \mathcal{F}(\mathcal{X})$ is a contraction mapping, then there exists a unique function $V^* \in \mathcal{F}(\mathcal{X})$ such that $\mathcal{B}[V^*] = V^*$.

Contraction in Discounted Problems

Properties of $\mathcal{B}_*[V]$

The operator $\mathcal{B}_*[V](\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}} \{ \ell(\mathbf{x}, \mathbf{u}) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V(\mathbf{x}')]] \}$ satisfies:

1. Monotonicity: $V(\mathbf{x}) \leq V'(\mathbf{x}) \Rightarrow \mathcal{B}_*[V](\mathbf{x}) \leq \mathcal{B}_*[V'](\mathbf{x})$
2. γ -Additivity: $\mathcal{B}_*[V + d](\mathbf{x}) = \mathcal{B}_*[V](\mathbf{x}) + \gamma d$ for $d \in \mathbb{R}$
3. Contraction: $\|\mathcal{B}_*[V] - \mathcal{B}_*[V']\|_\infty \leq \gamma \|V - V'\|_\infty$

► **Proof of Contraction:** Let $d = \sup_{\mathbf{x}} |V(\mathbf{x}) - V'(\mathbf{x})|$. Then:

$$V(\mathbf{x}) - d \leq V'(\mathbf{x}) \leq V(\mathbf{x}) + d, \quad \forall \mathbf{x} \in \mathcal{X}$$

Apply \mathcal{B}_* to both sides and use monotonicity and γ -additivity:

$$\mathcal{B}_*[V](\mathbf{x}) - \gamma d \leq \mathcal{B}_*[V'](\mathbf{x}) \leq \mathcal{B}_*[V](\mathbf{x}) + \gamma d, \quad \forall \mathbf{x} \in \mathcal{X}$$

Proof of VI Convergence in Discounted Problems

- ▶ $\mathcal{B}_*[V]$ is monotone, γ -additive, and a contraction mapping
- ▶ By the contraction mapping theorem, there exists $V^*(\mathbf{x})$ such that $\mathcal{B}_*[V^*] = V^*$
- ▶ Value Iteration Algorithm:

$$\begin{aligned}V_0(\mathbf{x}) &\equiv 0 \\V_{k+1}(\mathbf{x}) &= \mathcal{B}_*[V_k](\mathbf{x})\end{aligned}$$

- ▶ Since $\mathcal{B}_*[V]$ is a contraction, the sequence V_k is Cauchy, i.e.,
 $\|V_{k+1} - V_k\|_\infty \leq \gamma^k \|V_1 - V_0\|_\infty$
- ▶ If $(\mathcal{F}(\mathcal{X}), \|\cdot\|_\infty)$ is a complete metric space, then V_k has a limit $V^* \in \mathcal{F}(\mathcal{X})$ and V^* is a fixed point of \mathcal{B}_*

Outline

Infinite-Horizon Optimal Control

Bellman Equations

Policy Evaluation

Value Iteration

Policy Iteration

Linear Programming

Discounted Problem Policy Iteration (PI)

- ▶ PI is an alternative algorithm to VI for computing $V^*(\mathbf{x})$
- ▶ PI iterates over policies instead of values
- ▶ **Policy Iteration:** repeat until $V^{\pi'}(\mathbf{x}) = V^\pi(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$:
 1. **Policy Evaluation:** given a policy π , compute V^π :

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

2. **Policy Improvement:** given V^π , obtain a new policy π' :

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

First-Exit Problem Policy Iteration (PI)

► **Policy Iteration:** repeat until $V^{\pi'}(\mathbf{x}) = V^{\pi}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$:

1. **Policy Evaluation:** given a policy π , compute V^{π} :

$$V^{\pi}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V^{\pi}(\mathbf{x}')], \quad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

2. **Policy Improvement:** given V^{π} , obtain a new policy π' :

$$\pi'(\mathbf{x}) = \arg \min_{\mathbf{u} \in \mathcal{U}} \{ \ell(\mathbf{x}, \mathbf{u}) + \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V^{\pi}(\mathbf{x}')], \quad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

Policy Improvement Theorem

Let π and π' be such that $V^\pi(\mathbf{x}) \geq Q^\pi(\mathbf{x}, \pi'(\mathbf{x}))$ for all $\mathbf{x} \in \mathcal{X}$. Then, π' is at least as good as π , i.e., $V^\pi(\mathbf{x}) \geq V^{\pi'}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$

► Proof:

$$\begin{aligned} V^\pi(\mathbf{x}) &\geq Q^\pi(\mathbf{x}, \pi'(\mathbf{x})) = \ell(\mathbf{x}, \pi'(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi'(\mathbf{x}))} [V^\pi(\mathbf{x}')] \\ &\geq \ell(\mathbf{x}, \pi'(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi'(\mathbf{x}))} [Q^\pi(\mathbf{x}', \pi'(\mathbf{x}'))] \\ &= \ell(\mathbf{x}, \pi'(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \pi'(\mathbf{x}))} \{ \ell(\mathbf{x}', \pi'(\mathbf{x}')) + \gamma \mathbb{E}_{\mathbf{x}'' \sim p_f(\cdot | \mathbf{x}', \pi'(\mathbf{x}'))} V^\pi(\mathbf{x}'') \} \\ &\geq \dots \geq \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t \ell(\mathbf{x}_t, \pi'(\mathbf{x}_t)) \middle| \mathbf{x}_0 = \mathbf{x} \right] = V^{\pi'}(\mathbf{x}) \end{aligned}$$

Theorem: Optimality of PI

Suppose that \mathcal{X} is finite and:

- $\gamma \in [0, 1)$ (Discounted Problem),
- there exists a proper policy (First-Exit Problem).

Then, the Policy Iteration algorithm converges to an optimal policy after a finite number of steps.

Proof of Optimality of PI (First-Exit Problem)

- ▶ Let π be a proper policy with value V^π obtained from Policy Evaluation
- ▶ Let π' be the policy obtained from Policy Improvement
- ▶ By definition of Policy Improvement: $V^\pi(\mathbf{x}) \geq Q^\pi(\mathbf{x}, \pi'(\mathbf{x}))$ for all $\mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$
- ▶ By the Policy Improvement Thm., $V^\pi(\mathbf{x}) \geq V^{\pi'}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$
- ▶ Since π is proper, $V^\pi(\mathbf{x}) < \infty$ for all $\mathbf{x} \in \mathcal{X}$, and hence π' is proper
- ▶ Since π' is proper, the Policy Evaluation step has a unique solution $V^{\pi'}$
- ▶ Since the number of stationary policies is finite, eventually $V^\pi = V^{\pi'}$ after a finite number of steps
- ▶ Once V^π has converged, it follows from the Policy Improvement step:

$$V^{\pi'}(\mathbf{x}) = V^\pi(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}} \left\{ \ell(\mathbf{x}, \mathbf{u}) + \sum_{\mathbf{x}' \in \mathcal{X}} \tilde{p}_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) V^\pi(\mathbf{x}') \right\}, \quad \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

- ▶ Since this is the Bellman equation for the first-exit problem, we have converged to an optimal policy $\pi^* = \pi$ with optimal value $V^* = V^\pi$

Generalized Policy Iteration

- ▶ PI and VI have a lot in common

- ▶ Rewrite VI as follows:

2. **Policy Improvement:** Given $V_k(\mathbf{x})$ obtain a policy:

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

1. **Value Update:** Given $\pi(\mathbf{x})$ and $V_k(\mathbf{x})$, compute

$$V_{k+1}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \mathbb{E}_{\mathbf{x}' \sim p_f(\cdot | \mathbf{x}, \mathbf{u})} [V_k(\mathbf{x}')], \quad \forall \mathbf{x} \in \mathcal{X}$$

- ▶ Value Update is a single step of the iterative Policy Evaluation algorithm
- ▶ PI solves the Policy Evaluation equation completely, which is equivalent to running the Value Update step of VI an infinite number of times
- ▶ **Generalized Policy Iteration:** assuming the Value Update and Policy Improvement steps are executed an infinite number of times for all states, all combinations of the following converge:
 - ▶ Any number of Value Update steps in between Policy Improvement steps
 - ▶ Any number of states updated at each Value Update step
 - ▶ Any number of states updated at each Policy Improvement step

Complexity of VI and PI

- ▶ Consider the complexity of VI and PI for a finite state space \mathcal{X}
- ▶ **Complexity of VI per Iteration:** $O(|\mathcal{X}|^2|\mathcal{U}|)$: evaluating the expectation (i.e., sum over \mathbf{x}') requires $|\mathcal{X}|$ operations and there are $|\mathcal{X}|$ minimizations over $|\mathcal{U}|$ possible control inputs
- ▶ **Complexity of PI per Iteration:** $O(|\mathcal{X}|^2(|\mathcal{X}| + |\mathcal{U}|))$: the Policy Evaluation step requires solving a system of $|\mathcal{X}|$ equations in $|\mathcal{X}|$ unknowns ($O(|\mathcal{X}|^3)$), while the Policy Improvement step has the same complexity as one iteration of VI
- ▶ PI is more computationally expensive than VI
- ▶ Theoretically it takes an infinite number of iterations for VI to converge
- ▶ PI converges in $|\mathcal{U}|^{|\mathcal{X}|}$ iterations (all possible policies) in the worst case

Value Iteration

- ▶ V^* is a fixed point of \mathcal{B}_* : $V_0, \mathcal{B}_*[V_0], \mathcal{B}_*^2[V_0], \mathcal{B}_*^3[V_0], \dots \rightarrow V^*$

Algorithm Value Iteration

- 1: Initialize V_0
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: $V_{k+1} = \mathcal{B}_*[V_k]$
-

- ▶ Q^* is a fixed point of \mathcal{B}_* : $Q_0, \mathcal{B}_*[Q_0], \mathcal{B}_*^2[Q_0], \mathcal{B}_*^3[Q_0], \dots \rightarrow Q^*$

Algorithm Q-Value Iteration

- 1: Initialize Q_0
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: $Q_{k+1} = \mathcal{B}_*[Q_k]$
-

Policy Iteration

► Policy Evaluation: $V_0, \mathcal{B}_\pi[V_0], \mathcal{B}_\pi^2[V_0], \mathcal{B}_\pi^3[V_0], \dots \rightarrow V^\pi$

Algorithm Policy Iteration

- 1: Initialize V_0
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: $\pi_{k+1}(\mathbf{x}) = \arg \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V_k(\cdot)]$ ▷ Policy Improvement
 - 4: $V_{k+1} = \mathcal{B}_{\pi_{k+1}}^\infty [V_k]$ ▷ Policy Evaluation
-

► Policy Q-Evaluation: $Q_0, \mathcal{B}_\pi[Q_0], \mathcal{B}_\pi^2[Q_0], \mathcal{B}_\pi^3[Q_0], \dots \rightarrow Q^\pi$

Algorithm Q-Policy Iteration

- 1: Initialize Q_0
 - 2: **for** $k = 0, 1, 2 \dots$ **do**
 - 3: $\pi_{k+1}(\mathbf{x}) = \arg \min_{\mathbf{u} \in \mathcal{U}} Q_k(\mathbf{x}, \mathbf{u})$ ▷ Policy Improvement
 - 4: $Q_{k+1} = \mathcal{B}_{\pi_{k+1}}^\infty [Q_k]$ ▷ Policy Evaluation
-

Generalized Policy Iteration

Algorithm Generalized Policy Iteration

- 1: Initialize V_0
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: $\pi_{k+1}(\mathbf{x}) = \arg \min_{\mathbf{u} \in \mathcal{U}} H[\mathbf{x}, \mathbf{u}, V_k(\cdot)]$ ▷ Policy Improvement
 - 4: $V_{k+1} = \mathcal{B}_{\pi_{k+1}}^n [V_k], \quad \text{for } n \geq 1$ ▷ Policy Evaluation
-

Algorithm Generalized Q-Policy Iteration

- 1: Initialize Q_0
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: $\pi_{k+1}(\mathbf{x}) = \arg \min_{\mathbf{u} \in \mathcal{U}} Q_k(\mathbf{x}, \mathbf{u})$ ▷ Policy Improvement
 - 4: $Q_{k+1} = \mathcal{B}_{\pi_{k+1}}^n [Q_k], \quad \text{for } n \geq 1$ ▷ Policy Evaluation
-

Example: Frozen Lake Problem

- ▶ Winter is here
- ▶ You and your friends were tossing around a frisbee at the park when you made a wild throw that left the frisbee out in the middle of the lake
- ▶ The water is mostly frozen but there are a few holes where the ice has melted
- ▶ If you step into one of those holes, you fall into the freezing water
- ▶ There is an international frisbee shortage so it is absolutely imperative that you navigate across the lake and retrieve the disc
- ▶ However, the ice is slippery so you cannot always move in the direction you intend

Example: Frozen Lake Problem

| | | | |
|---|---|---|---|
| S | F | F | F |
| F | H | F | H |
| F | F | F | H |
| H | F | F | G |

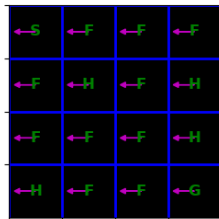
- ▶ S : starting point, safe
- ▶ F : frozen surface, safe
- ▶ H : hole, fall to your doom
- ▶ G : goal, where the frisbee is located
- ▶ $\mathcal{X} = \{0, 1, \dots, 15\}$
- ▶ $\mathcal{U} = \{\text{Left}(0), \text{Down}(1), \text{Right}(2), \text{Up}(3)\}$
- ▶ You receive a reward of 1 if you reach the goal, and zero otherwise

- ▶ An input $u \in \mathcal{U}$ succeeds 80% of the time. A neighboring control is executed in the other 50% of the time due to slip, e.g.,

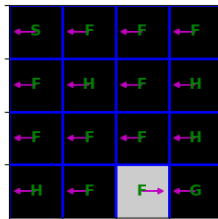
$$x' \mid x = 9, u = 1 = \begin{cases} 13, & \text{with prob. } 0.8 \\ 8, & \text{with prob. } 0.1 \\ 10, & \text{with prob. } 0.1 \end{cases}$$

- ▶ The state remains unchanged if a control leads outside of the map
- ▶ An episode ends when you reach the goal or fall in a hole

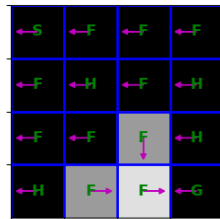
Value Iteration on Frozen Lake



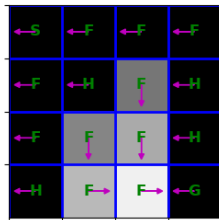
(a) $t = 0$



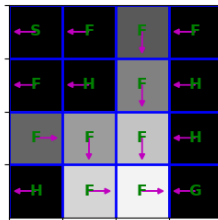
(b) $t = 1$



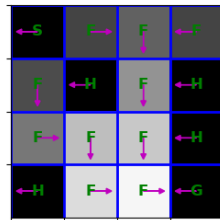
(c) $t = 2$



(d) $t = 3$



(e) $t = 4$

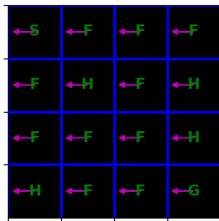


(f) $t = 5$

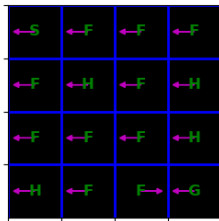
Value Iteration on Frozen Lake

| Iteration | $\max_x V_{t+1}(x) - V_t(x) $ | # changed actions | $V(0)$ |
|-----------|--------------------------------|-------------------|--------|
| 0 | 0.80000 | 0 | 0.000 |
| 1 | 0.60800 | 1 | 0.000 |
| 2 | 0.51984 | 2 | 0.000 |
| 3 | 0.39508 | 2 | 0.000 |
| 4 | 0.30026 | 2 | 0.000 |
| 5 | 0.25355 | 2 | 0.254 |
| 6 | 0.10478 | 1 | 0.345 |
| 7 | 0.09657 | 0 | 0.442 |
| 8 | 0.03656 | 0 | 0.478 |
| 9 | 0.02772 | 0 | 0.506 |
| 10 | 0.01111 | 0 | 0.517 |
| 11 | 0.00735 | 0 | 0.524 |
| 12 | 0.00310 | 0 | 0.527 |
| 13 | 0.00190 | 0 | 0.529 |
| 14 | 0.00083 | 0 | 0.530 |
| 15 | 0.00049 | 0 | 0.531 |
| 16 | 0.00022 | 0 | 0.531 |
| 17 | 0.00013 | 0 | 0.531 |
| 18 | 0.00006 | 0 | 0.531 |
| 19 | 0.00003 | 0 | 0.531 |

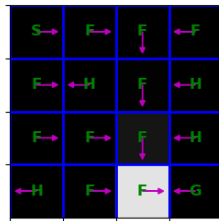
Policy Iteration on Frozen Lake



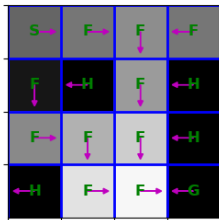
(a) $t = 0$



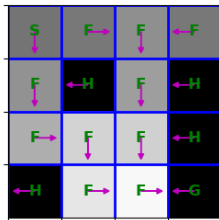
(b) $t = 1$



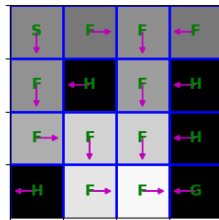
(c) $t = 2$



(d) $t = 3$



(e) $t = 4$

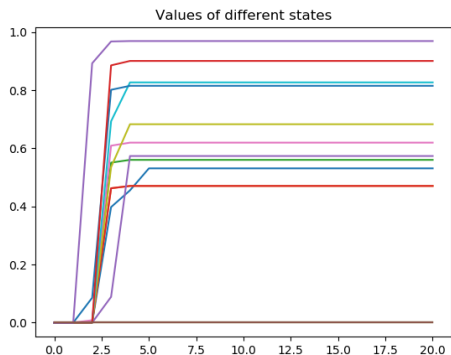
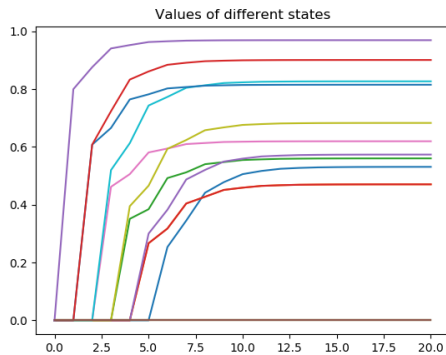


(f) $t = 5$

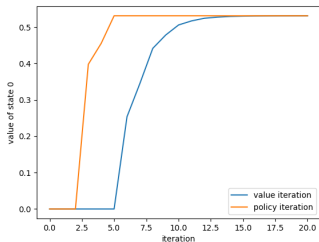
Policy Iteration on Frozen Lake

| Iteration | $\max_x V_{t+1}(x) - V_t(x) $ | # changed actions | $V(0)$ |
|-----------|--------------------------------|-------------------|--------|
| 0 | 0.00000 | 0 | 0.000 |
| 1 | 0.89296 | 1 | 0.000 |
| 2 | 0.88580 | 9 | 0.398 |
| 3 | 0.48504 | 2 | 0.455 |
| 4 | 0.07573 | 1 | 0.531 |
| 5 | 0.00000 | 0 | 0.531 |
| 6 | 0.00000 | 0 | 0.531 |
| 7 | 0.00000 | 0 | 0.531 |
| 8 | 0.00000 | 0 | 0.531 |
| 9 | 0.00000 | 0 | 0.531 |
| 10 | 0.00000 | 0 | 0.531 |
| 11 | 0.00000 | 0 | 0.531 |
| 12 | 0.00000 | 0 | 0.531 |
| 13 | 0.00000 | 0 | 0.531 |
| 14 | 0.00000 | 0 | 0.531 |
| 15 | 0.00000 | 0 | 0.531 |
| 16 | 0.00000 | 0 | 0.531 |
| 17 | 0.00000 | 0 | 0.531 |
| 18 | 0.00000 | 0 | 0.531 |
| 19 | 0.00000 | 0 | 0.531 |

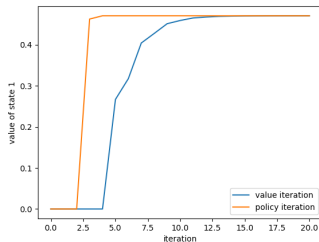
Value Iteration vs Policy Iteration



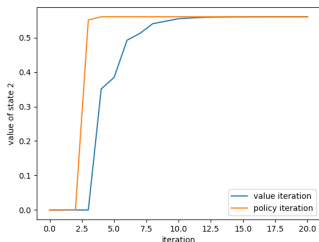
Value Iteration vs Policy Iteration



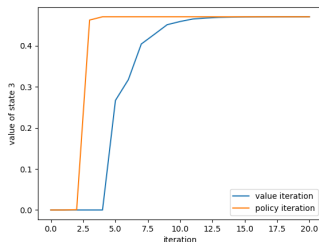
(a) State 0



(b) State 1



(c) State 2



(d) State 3

Outline

Infinite-Horizon Optimal Control

Bellman Equations

Policy Evaluation

Value Iteration

Policy Iteration

Linear Programming

Linear Programming Solution to the Bellman Equation

- ▶ Consider a Discounted Problem with finite state space \mathcal{X}
- ▶ Suppose we initialize VI with V_0 that satisfies a relaxed Bellman equation condition:

$$V_0(\mathbf{x}) \leq \min_{\mathbf{u} \in \mathcal{U}} \left(\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) V_0(\mathbf{x}') \right), \quad \forall \mathbf{x} \in \mathcal{X}$$

- ▶ Since \mathcal{B}_* is monotone, applying VI to V_0 leads to:

$$V_1(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}} \left(\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) V_0(\mathbf{x}') \right) \geq V_0(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}$$

$$\begin{aligned} V_2(\mathbf{x}) &= \min_{\mathbf{u} \in \mathcal{U}} \left(\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) V_1(\mathbf{x}') \right) \\ &\geq \min_{\mathbf{u} \in \mathcal{U}} \left(\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) V_0(\mathbf{x}') \right) = V_1(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X} \end{aligned}$$

Linear Programming Solution to the Bellman Equation

- ▶ The above shows that $V_{k+1}(\mathbf{x}) \geq V_k(\mathbf{x})$ for all k and $\mathbf{x} \in \mathcal{X}$
- ▶ Since VI guarantees that $V_k(\mathbf{x}) \rightarrow V^*(\mathbf{x})$ as $k \rightarrow \infty$, we also have:

$$V^*(\mathbf{x}) \geq V_0(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X} \quad \Rightarrow \quad \sum_{\mathbf{x} \in \mathcal{X}} w(\mathbf{x}) V^*(\mathbf{x}) \geq \sum_{\mathbf{x} \in \mathcal{X}} w(\mathbf{x}) V_0(\mathbf{x})$$

for any $w(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$.

- ▶ The above holds for **any** V_0 that satisfies:

$$V_0(\mathbf{x}) \leq \min_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \left(\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) V_0(\mathbf{x}') \right), \quad \forall \mathbf{x} \in \mathcal{X}$$

- ▶ Since V^* satisfies this condition with equality (Bellman Equation), it is the maximal V_0 that satisfies the condition

Linear Programming Solution to the Bellman Equation

LP Solution to Bellman Equation (Discounted Problem)

For finite \mathcal{X} , the solution $V^*(\mathbf{x})$ to the linear program with $w(\mathbf{x}) > 0$:

$$\begin{aligned} \max_V \quad & \sum_{\mathbf{x} \in \mathcal{X}} w(\mathbf{x}) V(\mathbf{x}) \\ \text{s.t.} \quad & V(\mathbf{x}) \leq \left(\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) V(\mathbf{x}') \right), \quad \forall \mathbf{u} \in \mathcal{U}, \forall \mathbf{x} \in \mathcal{X} \end{aligned}$$

also solves the Bellman Equation to yield the optimal value function of an infinite-horizon finite-state discounted stochastic optimal control problem.

- ▶ An equivalent result holds for the First-Exit Problem

LP Solution to Bellman Equation (Proof)

- ▶ Let J^* be the solution to the linear program so that:

$$J^*(\mathbf{x}) \leq \left(\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}) J^*(\mathbf{x}') \right), \quad \forall \mathbf{u} \in \mathcal{U}, \forall \mathbf{x} \in \mathcal{X}$$

- ▶ Since J^* is feasible, it satisfies $J^*(\mathbf{x}) \leq V^*(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$
- ▶ By contradiction, suppose that $J^* \neq V^*$
- ▶ Then, there exists a state $\mathbf{y} \in \mathcal{X}$ such that:

$$J^*(\mathbf{y}) < V^*(\mathbf{y}) \quad \Rightarrow \quad \sum_{\mathbf{x} \in \mathcal{X}} w(\mathbf{x}) J^*(\mathbf{x}) < \sum_{\mathbf{x} \in \mathcal{X}} w(\mathbf{x}) V^*(\mathbf{x})$$

for any positive $w(\mathbf{x})$ but since V^* solves the Bellman Equation:

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

V^* is feasible and has higher value than J^* , which is a contradiction.

Dual Linear Program

- ▶ Dual linear program:

$$\min_{\lambda \geq 0} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{u} \in \mathcal{U}} \ell(\mathbf{x}, \mathbf{u}) \lambda(\mathbf{x}, \mathbf{u})$$

$$\text{s.t. } \sum_{\mathbf{u}' \in \mathcal{U}} \lambda(\mathbf{x}', \mathbf{u}') = w(\mathbf{x}) + \gamma \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{u} \in \mathcal{U}} \lambda(\mathbf{x}, \mathbf{u}) p_f(\mathbf{x}' | \mathbf{x}, \mathbf{u}), \quad \forall \mathbf{x}' \in \mathcal{X}$$

- ▶ If $\sum_{\mathbf{x} \in \mathcal{X}} w(\mathbf{x}) = 1$, the constraint ensures that $\lambda(\mathbf{x}, \mathbf{u})$ is a probability measure on $\mathcal{X} \times \mathcal{U}$ induced by an optimal policy π :

$$\lambda(\mathbf{x}, \mathbf{u}) = \sum_{\mathbf{x}_0 \in \mathcal{X}} w(\mathbf{x}_0) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^{\pi}(\mathbf{x}_t = \mathbf{x}, \mathbf{u}_t = \mathbf{u} | \mathbf{x}_0)$$

- ▶ Optimal policy:

$$\pi^*(\mathbf{x}) \in \arg \min_{\mathbf{u} \in \mathcal{U}} \lambda(\mathbf{x}, \mathbf{u})$$