# ECE276B: Planning & Learning in Robotics
# Lecture 11: Bellman Equations II


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 Formulation

- Trajectories terminate at $T < \infty$

$$\min_{\pi} V_{\tau}^{\pi}(x) = \mathbb{E}\left[\sum_{t=\tau}^{T-1} g_t(x_t, \pi_t(x_t)) + g_T(x_T) \middle| x_\tau = x\right]$$

- The optimal cost-to-go $V_t^*(x)$ can be found with a single backward pass through time, initialized from $V_T^*(x) = g_T(x)$ and following the recursion:

## Bellman Equations (Finite Horizon)

| | |
|---|---|
| Hamiltonian: | $H_t[x, u, V(\cdot)] = g_t(x, u) + \mathbb{E}_{x' \sim p_f(\cdot \mid x, u)} V(x')$ |
| Policy Evaluation: | $V_t^{\pi}(x) = H_t[x, \pi_t(x), V_{t+1}^{\pi}(\cdot)]$ |
| Bellman Equation: | $V_t^*(x) = \min_{u \in \mathcal{U}(x)} H_t[x, u, V_{t+1}^*(\cdot)]$ |
| Optimal policy: | $\pi_t^*(x) = \arg\min_{u \in \mathcal{U}(x)} H_t[x, u, V_{t+1}^*(\cdot)]$ |

# First Exit (SSP) Formulation

- Trajectories terminate at $T_{first}$, when a goal state $x \in \mathcal{T} \subseteq \mathcal{X}$ is reached:

$$\min_{\pi} V^{\pi}(x) = \mathbb{E}\left[\sum_{t=0}^{T_{first}-1} g(x_t, \pi(x_t)) + g_{\mathcal{T}}(x_{T_{first}}) \middle| x_0 = x\right]$$

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

## Bellman Equations (First Exit)

| | |
|---|---|
| Hamiltonian: | $H[x, u, V(\cdot)] = g(x, u) + \mathbb{E}_{x' \sim p_f(\cdot \mid x, u)} V(x')$ |
| Policy Evaluation: | $V^{\pi}(x) = H[x, \pi(x), V^{\pi}(\cdot)]$ |
| Bellman Equation: | $V^*(x) = \min_{u \in \mathcal{U}(x)} H[x, u, V^*(\cdot)]$ |
| Optimal policy: | $\pi^*(x) = \arg\min_{u \in \mathcal{U}(x)} H[x, u, V^*(\cdot)]$ |

# Discounted Formulation

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

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

## Bellman Equations (Discounted)

| | |
|---|---|
| Hamiltonian: | $H[x, u, V(\cdot)] = g(x, u) + \gamma \mathbb{E}_{x' \sim p_f(\cdot \mid x, u)} V(x')$ |
| Policy Evaluation: | $V^{\pi}(x) = H[x, \pi(x), V^{\pi}(\cdot)]$ |
| Bellman Equation: | $V^*(x) = \min_{u \in \mathcal{U}(x)} H[x, u, V^*(\cdot)]$ |
| Optimal policy: | $\pi^*(x) = \arg\min_{u \in \mathcal{U}(x)} H[x, u, V^*(\cdot)]$ |

- Every discounted problem can be converted to a first exist problem by scaling the transition probabilities by $\gamma$, introducing a terminal state with zero cost, and setting all transition probabilities to that state to $1 - \gamma$

# Value Function

- **Value Function**: the expected long-term cost of following policy $\pi$ starting from state $x$:

$$\begin{aligned} V^{\pi}(x) :=& \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g(x_t, \pi(x_t)) \middle| x_0 = x\right] \\ =& g(x, \pi(x)) + \gamma \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} g(x_t, \pi(x_t)) \middle| x_0 = x\right] \\ =& g(x, \pi(x)) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x, \pi(x))}\left[V^{\pi}(x')\right] \end{aligned}$$

- **Value Iteration**: computes the optimal value function

$$V^*(x) := \min_{\pi} V^{\pi}(x) = \min_{u \in \mathcal{U}(x)} \left\{ g(x, u) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x,u)}\left[V^*(x')\right] \right\}$$

# Action-Value (Q) Function

- **Q Function**: the expected long-term cost of taking action $u$ in state $x$ and following policy $\pi$ afterwards:

$$\begin{aligned} Q^{\pi}(x,u) :=& g(x,u) + \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t g(x_t, \pi(x_t)) \middle| x_0 = x\right] \\ =& g(x,u) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x,u)}\left[V^{\pi}(x')\right] \\ =& g(x,u) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x,u)}\left[Q^{\pi}(x', \pi(x'))\right] \end{aligned}$$

- **Q-Value Iteration**: computes the optimal Q function

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

- $Q^*(x,u)$ allows us to choose optimal actions **without having to know anything about the dynamics** $p_f(x' \mid x, u)$:

$$\pi^*(x) = \arg\min_{u \in \mathcal{U}(x)} \left\{ g(x,u) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x,u)}\left[V^*(x')\right] \right\} = \arg\min_{u \in \mathcal{U}(x)} Q^*(x,u)$$

# Backup Operators

- **Policy Evaluation Backup Operator**:

$$\mathcal{T}_\pi[V](x) := H[x, \pi(x), V] = g(x, \pi(x)) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x,\pi(x))} \left[ V(x') \right]$$

- **Value Iteration Backup Operator**:

$$\mathcal{T}_*[V](x) := \min_{u \in \mathcal{U}(x)} H[x, u, V] = \min_{u \in \mathcal{U}(x)} \left\{ g(x, u) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x,u)} \left[ V(x') \right] \right\}$$

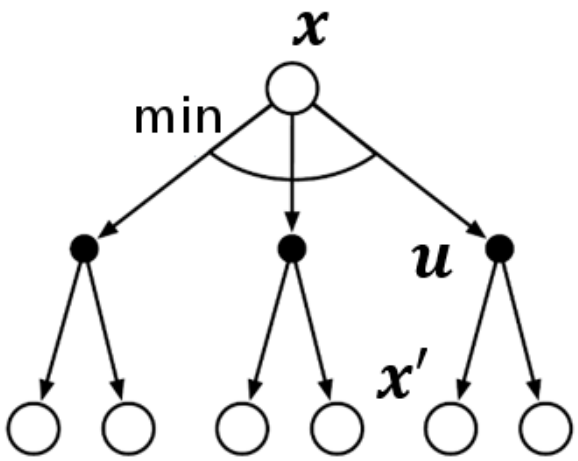
- **Policy Q-Evaluation Backup Operator**:

$$\mathcal{T}_\pi[Q](x, u) := g(x, u) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x,\pi(x))} \left[ Q(x', \pi(x')) \right]$$

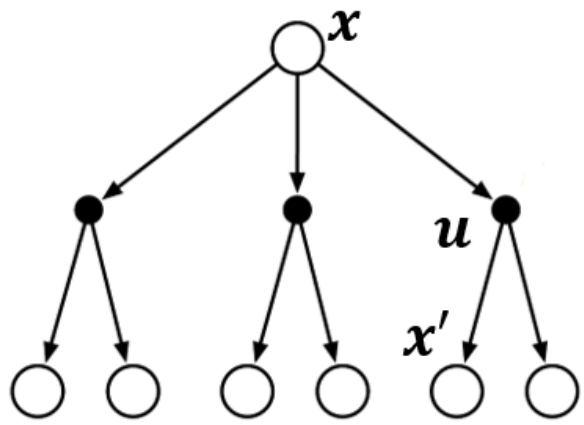
- **Q-Value Iteration Backup Operator**:

$$\mathcal{T}_*[Q](x, u) := g(x, u) + \gamma \mathbb{E}_{x' \sim p_f(\cdot|x,u)} \left[ \min_{u' \in \mathcal{U}(x')} Q(x', u') \right]$$

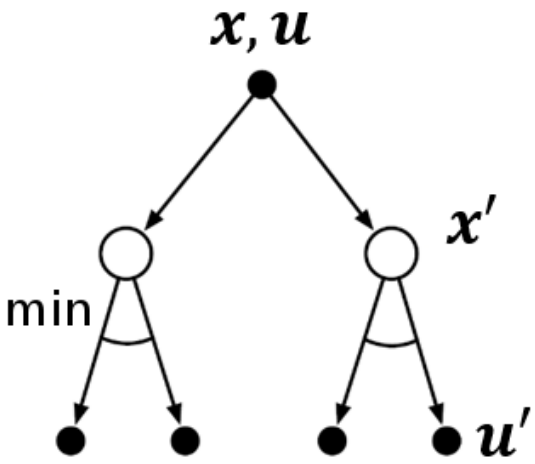
# Backup Operators (Stochastic Policy)



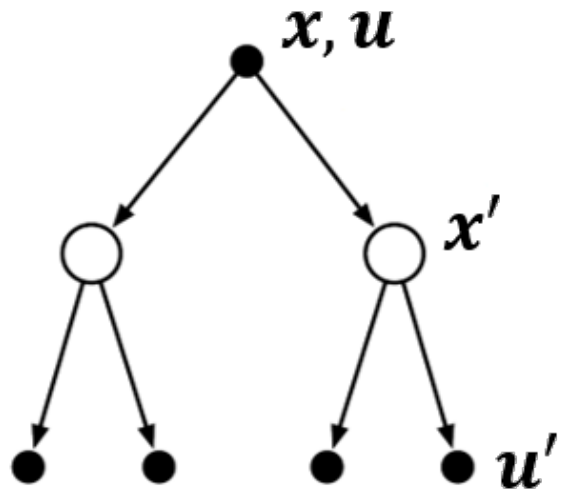

(a) $\mathcal{T}_*[V](x)$




(b) $\mathcal{T}_\pi[V](x)$




(c) $\mathcal{T}_*[Q](x,u)$




(d) $\mathcal{T}_\pi[Q](x,u)$

# Contraction in Discounted Problems

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

1. Monotonicity: $V(x) \leq V'(x) \quad \Rightarrow \quad T_*[V](x) \leq T_*[V'](x)$
2. Additivity: $T_*[V + d](x) = T_*[V](x) + \gamma d$
3. Contraction: $\| T_*[V](x) - T_*[V'](x) \|_\infty \leq \gamma \| V(x) - V'(x) \|_\infty$

- **Proof of Contraction**: Let $d = \max_x |V(x) - V'(x)|$. Then:

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

Apply $\mathcal{T}_*$ to both sides and use monotonicity and additivity:

$$\mathcal{T}_*[V](x) - \gamma d \leq \mathcal{T}_*[V'](x) \leq \mathcal{T}_*[V](x) + \gamma d, \quad \forall x \in \mathcal{X}$$

# VI and PI Revisited

- **Value Iteration**:
  - $V^*$ is the solution to $V = \mathcal{T}_*[V]$ (Bellman Equation)
  - Since $\mathcal{T}_*$ is a contraction, the fixed-point equation has a unique solution (Contraction Mapping Theorem), which can be determined iteratively:

$$V_{k+1} = \mathcal{T}_*[V_k] \quad \text{(Value Iteration)}$$

  - **Initialization**:
    - Discounted: arbitrary
    - First exit: $V_k(x) = g_{\mathcal{T}}(x)$ for all $k$ and all terminal $x \in \mathcal{T}$

- **Policy Iteration**:
  - **Policy Evaluation**: Given $\pi$ compute $V^\pi$ via

$$\mathbf{v} = (I - \gamma P)^{-1}\mathbf{g} \quad \text{OR} \quad V_{k+1} = \mathcal{T}_\pi[V_k] \quad \text{(Policy Evaluation Thm)}$$

  - **Policy Improvement**: choose the action that minimizes the Hamiltonian:

$$\pi'(x) = \arg\min_{u \in \mathcal{U}(x)} H[x, u, V^\pi(\cdot)]$$

  - **Initialization**: arbitrary as long as $V^\pi$ is finite

# Value Iteration

- $V^*$ is the fixed point of $\mathcal{T}_*$:
  $V^{(0)}, \mathcal{T}_*[V^{(0)}], \mathcal{T}_*^2[V^{(0)}], \mathcal{T}_*^3[V^{(0)}], \ldots \quad \to V^*$

**Algorithm 1** Value Iteration

1: Initialize $V^{(0)}$
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: $\quad V^{(n+1)} = \mathcal{T}_* \left[ V^{(n)} \right]$

- $Q^*$ is the fixed point of $\mathcal{T}_*$:
  $Q^{(0)}, \mathcal{T}_*[Q^{(0)}], \mathcal{T}_*^2[Q^{(0)}], \mathcal{T}_*^3[Q^{(0)}], \ldots \quad \to Q^*$

**Algorithm 2** Q-Value Iteration

1: Initialize $Q^{(0)}$
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: $\quad Q^{(n+1)} = \mathcal{T}_* \left[ Q^{(n)} \right]$

# Policy Iteration

- Policy Evaluation: $V^{(0)},\ \mathcal{T}_\pi[V^{(0)}],\ \mathcal{T}_\pi^2[V^{(0)}],\ \mathcal{T}_\pi^3[V^{(0)}], \ldots \quad \to V^\pi$

**Algorithm 3** Policy Iteration

1: Initialize $V^{(0)}$
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: $\quad \pi^{(n+1)}(x) = \arg\min_{u \in \mathcal{U}(x)} H[x, u, V^{(n)}(\cdot)]$ ▷ Policy Improvement
4: $\quad V^{(n+1)} = \mathcal{T}^\infty_{\pi^{(n+1)}}\left[V^{(n)}\right]$ ▷ Policy Evaluation

- Policy Q-Evaluation: $Q^{(0)},\ \mathcal{T}_\pi[Q^{(0)}],\ \mathcal{T}_\pi^2[Q^{(0)}],\ \mathcal{T}_\pi^3[Q^{(0)}], \ldots \quad \to Q^\pi$

**Algorithm 4** Q-Policy Iteration

1: Initialize $Q^{(0)}$
2: **for** $n = 0, 1, 2 \ldots$ **do**
3: $\quad \pi^{(n+1)}(x) = \arg\min_{u \in \mathcal{U}(x)} Q^{(n)}(x, u)$ ▷ Policy Improvement
4: $\quad Q^{(n+1)} = \mathcal{T}^\infty_{\pi^{(n+1)}}\left[Q^{(n)}\right]$ ▷ Policy Evaluation

# Generalized Policy Iteration

**Algorithm 5** Generalized Policy Iteration

1: Initialize $V^{(0)}$
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: $\quad \pi^{(n+1)}(x) = \arg\min_{u \in \mathcal{U}(x)} H[x, u, V^{(n)}(\cdot)]$ ▷ Policy Improvement
4: $\quad V^{(n+1)} = \mathcal{T}^k_{\pi^{(n+1)}}\left[V^{(n)}\right], \quad \text{for } k \geq 1$ ▷ Policy Evaluation

**Algorithm 6** Generalized Q-Policy Iteration

1: Initialize $Q^{(0)}$
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: $\quad \pi^{(n+1)}(x) = \arg\min_{u \in \mathcal{U}(x)} Q^{(n)}(x, u)$ ▷ Policy Improvement
4: $\quad Q^{(n+1)} = \mathcal{T}^k_{\pi^{(n+1)}}\left[Q^{(n)}\right], \quad \text{for } k \geq 1$ ▷ Policy Evaluation