ECE276B: Planning & Learning in Robotics
Lecture 10: Bellman Equations

Instructor:
Nikolay Atanasov: natanasov@ucsd.edu

Teaching Assistants:
Zhichao Li: zhl355@eng.ucsd.edu
Jinzhao Li: jil016@eng.ucsd.edu

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering

First-Exit Problem

- The **first exit problem** is a slightly more general statement of the stochastic shortest path (SSP) problem
- **Terminal Set**: let $\mathcal{T} \subseteq \mathcal{X}$ be a set of terminal states with terminal cost $\mathfrak{q}(\mathbf{x})$ for $\mathbf{x} \in \mathcal{T}$
- **First-Exit Time**: trajectories terminate at $T := \inf\{t \geq 1 | \mathbf{x}_t \in \mathcal{T}\}$, the first passage time from an initial state $\mathbf{x}_0$ to a terminal state $\mathbf{x} \in \mathcal{T}$
- Note that T is a random variable unlike in the finite-horizon problem
- **First-Exit Problem**:

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

Discounted Problem

- **Discount factor** $\gamma \in [0, 1)$
- The optimal value function $V^*(\mathbf{x})$ and associated policy $\pi^*(\mathbf{x})$ are **stationary**
- The episodes $\rho_0 := \mathbf{x}_0, \mathbf{u}_0, \mathbf{x}_1, \mathbf{u}_1, \ldots \sim \pi$ continue forever but the costs are discounted by γ
- **Discounted Problem**:

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

Bellman Equation

- **First-Exit (SSP) Problem**: the optimal value function satisfies:

$$V^*(\mathbf{x}) = \mathfrak{q}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{T}$$
$$V^*(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \left(\ell(\mathbf{x}, \mathbf{u}) + \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}) V^*(\mathbf{x}') \right), \quad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

- **Discounted Problem**: the optimal value function satisfies (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 First-Exit problems:
  - Value Iteration (VI)
  - Policy Iteration (PI)
  - Linear Programming (LP)

Value Iteration (VI)

- **Value Iteration**: applies the Dynamic Programming recursion with an arbitrary initialization $V_0(\mathbf{x})$ to compute $V^*(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$
- The VI algorithm is the infinite-horizon equivalent of the DP algorithm
- VI requires infinite iterations for $V_k(\mathbf{x})$ to converge to $V^*(\mathbf{x})$. In practice, define a threshold for $|V_{k+1}(\mathbf{x}) - V_k(\mathbf{x})|$ for all $\mathbf{x} \in \mathcal{X}$
- **First-Exit Problem**:

$$V_k(\mathbf{x}) = \mathfrak{q}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{T}$$

$$V_{k+1}(\mathbf{x}) = \min_{\mathbf{u} \in \tilde{\mathcal{U}}(\mathbf{x})} \left[\ell(\mathbf{x}, \mathbf{u}) + \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}) V_k(\mathbf{x}') \right], \qquad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

- **Discounted Problem**:

$$V_{k+1}(\mathbf{x}) = \min_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \left[\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}) V_k(\mathbf{x}') \right], \qquad \forall \mathbf{x} \in \mathcal{X}$$

Gauss-Seidel Value Iteration

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

$$\hat{V}(\mathbf{x}) \leftarrow \min_{\mathbf{u}\in\mathcal{U}(\mathbf{x})} \left(\ell(\mathbf{x},\mathbf{u}) + \gamma \sum_{\mathbf{x}'\in\mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x},\mathbf{u}) V(\mathbf{x}') \right), \qquad \forall \mathbf{x} \in \mathcal{X}$$

$$V(\mathbf{x}) \leftarrow \hat{V}(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathcal{X}$$

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

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

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

Policy Evaluation

- The VI algorithm computes the optimal value function $V^*(\mathbf{x})$ for every state $\mathbf{x} \in \mathcal{X}$
- Instead of the optimal value function $V^*(\mathbf{x})$, is it possible to compute the value function $V^\pi(\mathbf{x})$ for a given policy π?

Policy Evaluation Theorem (Discounted Problem)

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

$$V^\pi(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \pi(\mathbf{x})) V^\pi(\mathbf{x}'), \qquad \forall \mathbf{x} \in \mathcal{X}$$

Furthermore, given any initial conditions $V_0(\mathbf{x})$, the sequence $V_k(\mathbf{x})$ generated by the recursion below converges to $V^\pi(\mathbf{x})$:

$$V_{k+1}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \pi(\mathbf{x})) V_k(\mathbf{x}'), \qquad \forall \mathbf{x} \in \mathcal{X}$$

Policy Evaluation

Policy Evaluation Theorem (First-Exit Problem)

The value function $V^{\pi}(\mathbf{x})$ at $\mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$ for policy π is the unique solution of:

$$V^{\pi}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \pi(\mathbf{x})) V^{\pi}(\mathbf{x}'). \qquad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

Furthermore, given any initial conditions $V_0(\mathbf{x})$, the sequence $V_k(\mathbf{x})$ generated by the recursion below converges to $V^{\pi}(\mathbf{x})$:

$$V_{k+1}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \pi(\mathbf{x})) V_k(\mathbf{x}'), \qquad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

- **Proof sketch**: This is a special case of the Bellman Equation Theorem (SSP). Consider a modified problem, where the only allowable control at state $\mathbf{x}$ is $\pi(\mathbf{x})$. Since the proper policy π is the only policy under consideration, the proper policy assumption is satisfied and the arg min over $\mathbf{u} \in \mathcal{U}(\mathbf{x})$ has to be $\pi(\mathbf{x})$.

Policy Evaluation as a Linear System

- Let $\mathcal{X} = \{1, \ldots, n\}$ for the Discounted Problem
- Let $\mathcal{X} = \mathcal{N} \cup \mathcal{T}$ for the First-Exit Problem with $\mathcal{N} = \{1, \ldots, n\}$
- Let $\mathbf{v}_i := V^\pi(i)$, $\boldsymbol{\ell}_i := \ell(i, \pi(i))$, $P_{ij} := p_f(j \mid i, \pi(i))$ for $i, j = 1, \ldots, n$
- Let $\mathfrak{q}_i := \mathfrak{q}(i)$ for $i \in \mathcal{T}$
- Policy evaluation requires solving a linear system:

$$\begin{aligned} &\textbf{Discounted:} \quad && \mathbf{v} = \boldsymbol{\ell} + \gamma P \mathbf{v} && \Rightarrow \quad (I - \gamma P)\mathbf{v} = \boldsymbol{\ell} \\ &\textbf{First-Exit:} \quad && \mathbf{v} = \boldsymbol{\ell} + P_{\mathcal{N}\mathcal{N}}\mathbf{v} + P_{\mathcal{N}\mathcal{T}}\mathfrak{q} && \Rightarrow \quad (I - P_{\mathcal{N}\mathcal{N}})\mathbf{v} = \boldsymbol{\ell} + P_{\mathcal{N}\mathcal{T}}\mathfrak{q} \end{aligned}$$

- **Existence of solution**:
  - **Discounted**: The matrix P has eigenvalues with modulus ≤ 1. All eigenvalues of γP have modulus < 1, so $(\gamma P)^T \to 0$ as $T \to \infty$ and $(I - \gamma P)^{-1}$ exists.
  - **First-Exit**: a unique solution for $\mathbf{v}$ exists as long as π is a proper policy. By the Chapman-Kolmogorov equation, $[P^k]_{ij} = \mathbb{P}(x_k = j \mid x_0 = i)$ and since π is proper, $[P^k]_{ij} \to 0$ as $k \to \infty$ for all $i, j \in \mathcal{X} \setminus \mathcal{T}$. Since $P^k_{\mathcal{N}\mathcal{N}}$ vanishes as $k \to \infty$, all eigenvalues of $P_{\mathcal{N}\mathcal{N}}$ must have modulus less than 1 and therefore $(I - P_{\mathcal{N}\mathcal{N}})^{-1}$ exists.

Policy Evaluation as a Linear System

- The Policy Evaluation Thm. is an iterative solution to the linear system
- **Discounted**:

$$
\begin{aligned}
\mathbf{v}_1 &= \boldsymbol{\ell} + \gamma P \mathbf{v}_0 \\
\mathbf{v}_2 &= \boldsymbol{\ell} + \gamma P \mathbf{v}_1 = \boldsymbol{\ell} + \gamma P \boldsymbol{\ell} + (\gamma P)^2 \mathbf{v}_0 \\
&\vdots \\
\mathbf{v}_k &= (I + \gamma P + (\gamma P)^2 + \ldots + (\gamma P)^{k-1}) \boldsymbol{\ell} + (\gamma P)^k \mathbf{v}_0 \\
&\vdots \\
\mathbf{v}_\infty &\to (I - \gamma P)^{-1} \boldsymbol{\ell}
\end{aligned}
$$

- **First-Exit**:

$$
\begin{aligned}
\mathbf{v}_1 &= \boldsymbol{\ell} + P_{\mathcal{N}\mathcal{T}} \mathbf{q} + P_{\mathcal{N}\mathcal{N}} \mathbf{v}_0 \\
\mathbf{v}_2 &= \boldsymbol{\ell} + P_{\mathcal{N}\mathcal{T}} \mathbf{q} + P_{\mathcal{N}\mathcal{N}} \mathbf{v}_1 = \boldsymbol{\ell} + P_{\mathcal{N}\mathcal{T}} \mathbf{q} + P_{\mathcal{N}\mathcal{N}} (\boldsymbol{\ell} + P_{\mathcal{N}\mathcal{T}} \mathbf{q}) + P_{\mathcal{N}\mathcal{N}}^2 \mathbf{v}_0 \\
&\vdots \\
\mathbf{v}_\infty &\to (I - P_{\mathcal{N}\mathcal{N}})^{-1} (\boldsymbol{\ell} + P_{\mathcal{N}\mathcal{T}} \mathbf{q})
\end{aligned}
$$

Policy Iteration (PI)

- PI is an alternative algorithm to VI for computing $V^*(\mathbf{x})$
- PI iterates over policies instead of values
- **First-Exit Problem**: 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}:

$$V^{\pi}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \pi(\mathbf{x})) V^{\pi}(\mathbf{x}'), \qquad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

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

$$\pi'(\mathbf{x}) = \arg\min_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \left[\ell(\mathbf{x}, \mathbf{u}) + \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}) V^{\pi}(\mathbf{x}') \right], \qquad \forall \mathbf{x} \in \mathcal{X} \setminus \mathcal{T}$$

Policy Iteration (PI)

- PI is an alternative algorithm to VI for computing $V^*(\mathbf{x})$
- PI iterates over policies instead of values
- **Discounted Problem**: 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}:

$$V^{\pi}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \pi(\mathbf{x})) V^{\pi}(\mathbf{x}'), \qquad \forall \mathbf{x} \in \mathcal{X}$$

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

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

Policy Improvement Theorem

Let π and π' be deterministic policies 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**:

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

Theorem: Optimality of PI

Suppose that $\mathcal{X}$ is finite and:

- $\gamma \in [0, 1)$ (Discounted Problem)
- there exists a termination set $\mathcal{T}$ and 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 the Policy Evaluation step.
- Let π' be the policy obtained from the Policy Improvement step.
- By definition of the Policy Improvement step: $V^\pi(\mathbf{x}) \geq Q^\pi(\mathbf{x}, \pi'(\mathbf{x}))$ 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:

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

- Since this is the Bellman Equation for the First-Exit problem, we have converged to an optimal policy $\pi^* = \pi$ with optimal cost $V^* = V^\pi$.

Comparison between VI and PI

- PI and VI actually have a lot in common
- Rewrite VI as follows:
  2. **Policy Improvement**: Given $V_k(\mathbf{x})$ obtain a stationary policy:

$$\pi(\mathbf{x}) = \underset{\mathbf{u} \in \mathcal{U}(\mathbf{x})}{\arg\min} \left[\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}) V_k(\mathbf{x}') \right], \qquad \forall \mathbf{x} \in \mathcal{X}$$

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

$$V_{k+1}(\mathbf{x}) = \ell(\mathbf{x}, \pi(\mathbf{x})) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \pi(\mathbf{x})) V_k(\mathbf{x}'), \qquad \forall \mathbf{x} \in \mathcal{X}$$

- The Value Update step of VI is one step of an iterative solution to the linear system of equations in the Policy Evaluation Theorem
- PI solves the Policy Evaluation equation completely, which is equivalent to running the Value Update step of VI an infinite number of times!

Comparison between VI and PI

- **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

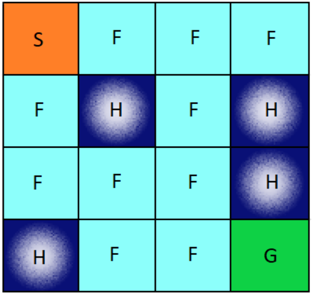
Generalized Policy Iteration

- Assuming that 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

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'll fall into the freezing water.
- At this time, there's an international frisbee shortage, so it's absolutely imperative that you navigate across the lake and retrieve the disc.
- However, the ice is slippery, so you won't always move in the direction you intend.

Example: Frozen Lake Problem

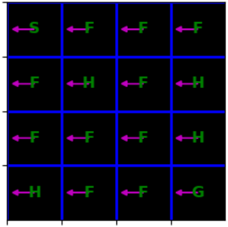


- S : starting point, safe
- F : frozen surface, safe
- H : hole, fall to your doom
- G : goal, where the frisbee is located
- $\mathcal{X} = \{0, 1, \dots, 15\}$
- $\mathcal{U}(x) = \{\text{Left}(0), \text{Down}(1), \text{Right}(2), \text{Up}(3)\}$
- You receive a reward of 1 if you reach the goal, and zero otherwise
- A requested action $u \in \mathcal{U}(x)$ succeeds 80% of the time. A neighboring action is executed in the other 50% of the time due to slip:

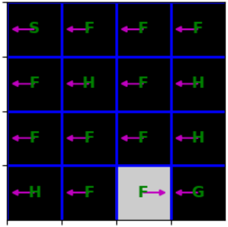
$$x' \mid x = 9, u = 1 = \begin{cases} 13, & \text{with prob. } 0.8 \\ 8, & \text{with prob. } 0.1 \\ 10, & \text{with prob. } 0.1 \end{cases}$$

- 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.

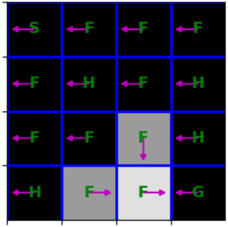
Value Iteration on Frozen Lake



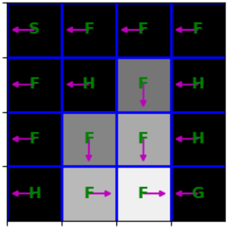
(a) $t = 0$



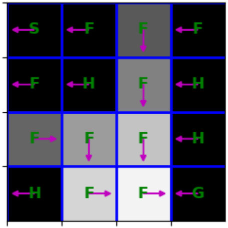
(b) $t = 1$



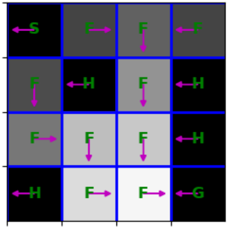
(c) $t = 2$



(d) $t = 3$



(e) $t = 4$

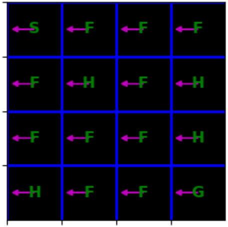


(f) $t = 5$

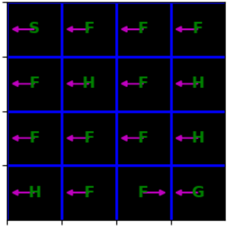
Value Iteration on Frozen Lake

Iteration	$\max_x \lvert V_{t+1}(x) - V_t(x) \rvert$	**# 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

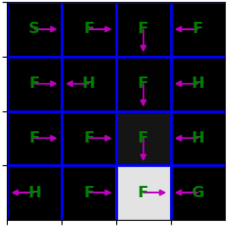
Policy Iteration on Frozen Lake



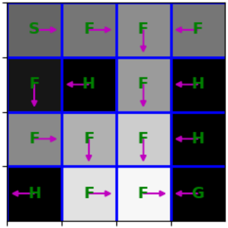
(a) $t = 0$



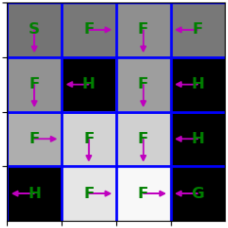
(b) $t = 1$



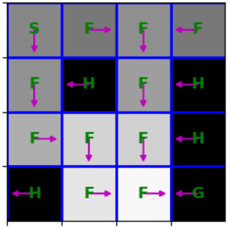
(c) $t = 2$



(d) $t = 3$



(e) $t = 4$

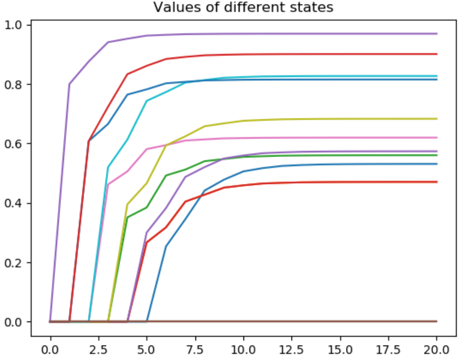


(f) $t = 5$

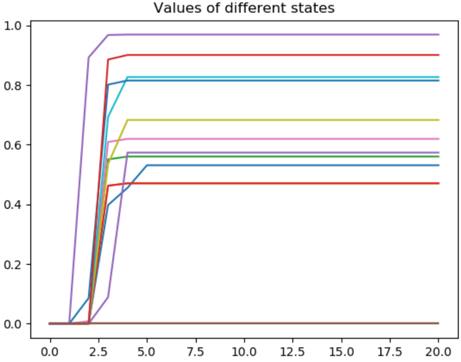
Policy Iteration on Frozen Lake

Iteration	$\max_x \lvert V_{t+1}(x) - V_t(x) \rvert$	**# 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

Value Iteration vs Policy Iteration

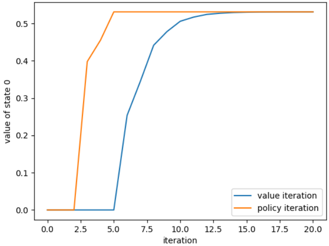


(a) VI

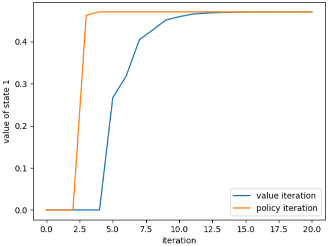


(b) PI

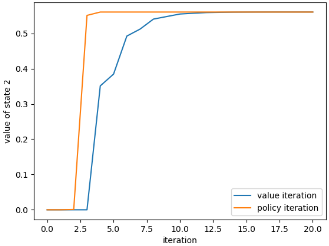
Value Iteration vs Policy Iteration



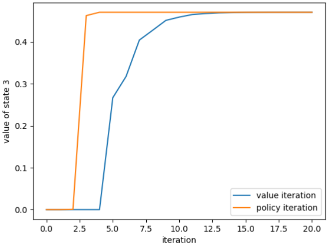
(a) State 0



(b) State 1



(c) State 2



(d) State 3

Linear Programming Solution to the Bellman Equation

- Suppose we initialize VI with V_0 that satisfies a relaxed Bellman Equation:

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

- Applying VI to V_0 leads to:

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

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 $V_k(\mathbf{x}) \to V^*(\mathbf{x})$ as $k \to \infty$ 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:

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

- Note that V^* also satisfies this condition with equality (Bellman Equation) and hence is the maximal V_0 (at each state) that satisfies the condition.

Linear Programming Solution to the Bellman Equation

LP Solution to the Bellman Equation

The solution V^* to the linear program with $w(\mathbf{x}) > 0$:

$$\max_{V} \sum_{\mathbf{x} \in \mathcal{X}} w(\mathbf{x}) V(\mathbf{x})$$

$$\text{s.t. } V(\mathbf{x}) \leq \left(\ell(\mathbf{x}, \mathbf{u}) + \gamma \sum_{\mathbf{x}' \in \mathcal{X}} p_f(\mathbf{x}' \mid \mathbf{x}, \mathbf{u}) V(\mathbf{x}') \right), \qquad \forall \mathbf{u} \in \mathcal{U}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$$

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

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

LP Solution to the BE (Proof)

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

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

- 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:

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

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

Bellman Equations (Summary)

Finite-Horizon Problem

- Trajectories terminate at fixed $T < \infty$

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

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

Bellman Equations (Finite-Horizon Problem)

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

First-Exit Problem

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

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

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

Bellman Equations (First-Exit Problem)

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

Discounted Problem

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

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

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

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

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

- Every discounted problem can be converted to a first exit problem by scaling the transition probabilities by γ, 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 π starting from state $\mathbf{x}$:

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

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

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

Action-Value (Q) Function

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

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

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

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

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

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

Bellman Backup Operators

- **Policy Evaluation Backup Operator**:

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

- **Value Iteration Backup Operator**:

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

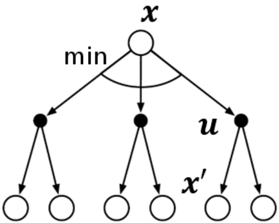
- **Policy Q-Evaluation Backup Operator**:

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

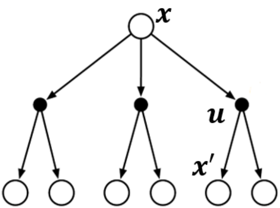
- **Q-Value Iteration Backup Operator**:

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

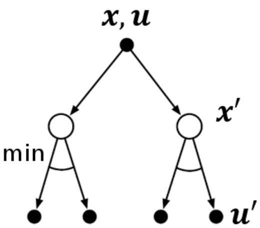
Bellman Backup Operators (Stochastic Policy)



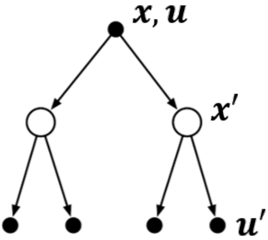
(a) $\mathcal{T}_*[V](\mathbf{x})$



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



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



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

Contraction in Discounted Problems

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

1. Monotonicity: $V(\mathbf{x}) \leq V'(\mathbf{x}) \quad \Rightarrow \quad \mathcal{T}_*[V](\mathbf{x}) \leq \mathcal{T}_*[V'](\mathbf{x})$
2. γ-Additivity: $\mathcal{T}_*[V(\cdot) + d](\mathbf{x}) = \mathcal{T}_*[V](\mathbf{x}) + \gamma d$
3. Contraction: $\|\mathcal{T}_*[V](\mathbf{x}) - \mathcal{T}_*[V'](\mathbf{x})\|_\infty \leq \gamma \|V(\mathbf{x}) - V'(\mathbf{x})\|_\infty$

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

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

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

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

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:

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

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

- **Policy Iteration**:
  - **Policy Evaluation**: Given π compute V^π via

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

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

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

  - **Initialization**: arbitrary as long as V^π is finite

Value Iteration

- V^* **is a 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** $k = 0, 1, 2, \ldots$ **do**
3: $\quad V_{k+1} = \mathcal{T}_*\left[V_k\right]$

- Q^* **is a 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** $k = 0, 1, 2, \ldots$ **do**
3: $\quad Q_{k+1} = \mathcal{T}_*\left[Q_k\right]$

Policy Iteration

- Policy Evaluation: $V_0,\ \mathcal{T}_\pi[V_0],\ \mathcal{T}_\pi^2[V_0],\ \mathcal{T}_\pi^3[V_0], \dots \quad \to V^\pi$

Algorithm 3 Policy Iteration

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

- Policy Q-Evaluation: $Q_0,\ \mathcal{T}_\pi[Q_0],\ \mathcal{T}_\pi^2[Q_0],\ \mathcal{T}_\pi^3[Q_0], \dots \quad \to Q^\pi$

Algorithm 4 Q-Policy Iteration

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

Generalized Policy Iteration

Algorithm 5 Generalized Policy Iteration

Initialize V_0

for $k = 0, 1, 2, \ldots$ **do**

$\pi_{k+1}(\mathbf{x}) = \arg\min_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} H[\mathbf{x}, \mathbf{u}, V_k(\cdot)]$ ▷ Policy Improvement

$V_{k+1} = \mathcal{T}^n_{\pi_{k+1}}[V_k], \quad \text{for } n \geq 1$ ▷ Policy Evaluation

Algorithm 6 Generalized Q-Policy Iteration

Initialize Q_0

for $k = 0, 1, 2, \ldots$ **do**

$\pi_{k+1}(\mathbf{x}) = \arg\min_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} Q_k(\mathbf{x}, \mathbf{u})$ ▷ Policy Improvement

$Q_{k+1} = \mathcal{T}^n_{\pi_{k+1}}[Q_k], \quad \text{for } n \geq 1$ ▷ Policy Evaluation