ECEN 743: Reinforcement Learning

Markov Decision Process (MDP)

Dileep Kalathil
Assistant Professor
Department of Electrical and Computer Engineering
Texas A&M University

References

- [SB, Chapter 3-4]
- [DB, Chapter 2]
- [AJKS, Chapter 1]

- Acknowledgment: Some figures for this lecture are taken from UC Berkeley CS188 course, with permission.

Markov Decision Process (MDP)

Definition (Markov Decision Process (MDP))

A (discounted) Markov Decision Process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where,

- $\mathcal{S}$ is a set of states (state space)
- $\mathcal{A}$ is a set of actions (action space)
- P is a transition probability function, $P(s'|s,a) = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a), \forall s, s' \in \mathcal{S}, a \in \mathcal{A}$
- r is a reward function, $r(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}$
- γ is a discount factor, $\gamma \in (0,1)$

Markov Decision Process (MDP)

Definition (Markov Decision Process (MDP))

A (discounted) Markov Decision Process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where,

- $\mathcal{S}$ is a set of states (state space)
- $\mathcal{A}$ is a set of actions (action space)
- P is a transition probability function, $P(s'|s,a) = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a), \forall s, s' \in \mathcal{S}, a \in \mathcal{A}$
- r is a reward function, $r(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}$
- γ is a discount factor, $\gamma \in (0,1)$

- In the next few lectures, we will mainly focus on the case where state space and action space are finite

Markov Decision Process (MDP)

Definition (Markov Decision Process (MDP))

A (discounted) Markov Decision Process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where,

- $\mathcal{S}$ is a set of states (state space)
- $\mathcal{A}$ is a set of actions (action space)
- P is a transition probability function, $P(s'|s,a) = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a), \forall s, s' \in \mathcal{S}, a \in \mathcal{A}$
- r is a reward function, $r(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}$
- γ is a discount factor, $\gamma \in (0,1)$

- In the next few lectures, we will mainly focus on the case where state space and action space are finite
- A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the control action for each state

Policy Evaluation

MDP Questions

MDP Questions

- How do we compute the value of a policy π?

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, \pi\right], \text{ where, } s_{t+1} \sim P(\cdot|s_t, a_t),\ a_t \sim \pi(s_t, \cdot)$$

MDP Questions

- How do we compute the value of a policy π?

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, \pi\right], \text{ where, } s_{t+1} \sim P(\cdot|s_t, a_t), \ a_t \sim \pi(s_t, \cdot)$$

- How do we compute the optimal policy?

$$\pi^\star = \arg\max_{\pi \in \Pi} \ V_\pi$$

Policy Evaluation

- How do we compute the value of a policy π?

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)], \quad a_t \sim \pi(s_t, \cdot), \ s_{t+1} \sim P(\cdot|s_t, a_t), \ s_0 = s$$

Policy Evaluation

- How do we compute the value of a policy π?

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)], \quad a_t \sim \pi(s_t, \cdot), \ s_{t+1} \sim P(\cdot|s_t, a_t), \ s_0 = s$$

- V_π is also called the value function of policy π

Induced Markov Chain

- A policy π induces a Markov chain on the MDP

Induced Markov Chain

- A policy π induces a Markov chain on the MDP
- Given the transition probability function P of the MDP and the policy π, what is the probability of moving from state s to s' in one step?

Induced Markov Chain

- A policy π induces a Markov chain on the MDP
- Given the transition probability function P of the MDP and the policy π, what is the probability of moving from state s to s' in one step?

$$P_\pi(s, s') = \sum_a P(s'|s, a)\pi(s, a)$$

Induced Markov Chain

- A policy π induces a Markov chain on the MDP
- Given the transition probability function P of the MDP and the policy π, what is the probability of moving from state s to s' in one step?

$$P_\pi(s, s') = \sum_a P(s'|s, a)\pi(s, a)$$

- P_π is the transition probability matrix of the Markov chain induced by the policy π on the MDP

Induced Markov Chain

- A policy π induces a Markov chain on the MDP
- Given the transition probability function P of the MDP and the policy π, what is the probability of moving from state s to s' in one step?

$$P_\pi(s, s') = \sum_a P(s'|s, a)\pi(s, a)$$

- P_π is the transition probability matrix of the Markov chain induced by the policy π on the MDP
- What is the expected reward obtained from the state s under policy π?

Induced Markov Chain

- A policy π induces a Markov chain on the MDP
- Given the transition probability function P of the MDP and the policy π, what is the probability of moving from state s to s' in one step?

$$P_\pi(s, s') = \sum_a P(s'|s, a)\pi(s, a)$$

- P_π is the transition probability matrix of the Markov chain induced by the policy π on the MDP
- What is the expected reward obtained from the state s under policy π?

$$r_\pi(s) = \sum_a r(s, a)\pi(s, a)$$

Induced Markov Chain

- A policy π induces a Markov chain on the MDP
- Given the transition probability function P of the MDP and the policy π, what is the probability of moving from state s to s' in one step?

$$P_\pi(s, s') = \sum_a P(s'|s, a)\pi(s, a)$$

- P_π is the transition probability matrix of the Markov chain induced by the policy π on the MDP
- What is the expected reward obtained from the state s under policy π?

$$r_\pi(s) = \sum_a r(s, a)\pi(s, a)$$

- We will use the vector form V_π, r_π and the matrix form P_π

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- Intuition

$$V_\pi(s) = \underbrace{r_\pi(s)}_{\substack{\text{Expected reward for state } s \\ \text{under policy } \pi}} + \gamma \underbrace{\sum_{s'} P_\pi(s, s') V_\pi(s')}_{\substack{\text{Expected value of the next state} \\ \text{under policy } \pi}}$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- We can show this as follows:

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- We can show this as follows:

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] = \mathbb{E}[(r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)) \mid s_0 = s]$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- We can show this as follows:

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] = \mathbb{E}[(r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)) \mid s_0 = s]$$

$$= \mathbb{E}_{a \sim \pi(s,\cdot)}[r(s, a)] + \gamma \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s]$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- We can show this as follows:

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] = \mathbb{E}[(r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)) \mid s_0 = s]$$

$$= \mathbb{E}_{a \sim \pi(s,\cdot)}[r(s, a)] + \gamma \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s]$$

$$= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s, s_1 = s'] | s_0 = s]$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- We can show this as follows:

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] = \mathbb{E}[(r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)) \mid s_0 = s]$$

$$= \mathbb{E}_{a \sim \pi(s,\cdot)}[r(s, a)] + \gamma \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s]$$

$$= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s, s_1 = s']] | s_0 = s]$$

$$= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_1 = s']] | s_0 = s]$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- We can show this as follows:

$$\begin{aligned}
V_\pi(s) &= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] = \mathbb{E}[(r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)) \mid s_0 = s] \\
&= \mathbb{E}_{a \sim \pi(s,\cdot)}[r(s, a)] + \gamma \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s, s_1 = s'] | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_1 = s'] | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[V_\pi(s') | s_0 = s]
\end{aligned}$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- We can show this as follows:

$$\begin{aligned}
V_\pi(s) &= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] = \mathbb{E}[(r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)) \mid s_0 = s] \\
&= \mathbb{E}_{a \sim \pi(s,\cdot)}[r(s, a)] + \gamma \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s, s_1 = s'] | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_1 = s'] | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[V_\pi(s') | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}_{s' \sim P_\pi(s,\cdot)}[V_\pi(s')]
\end{aligned}$$

Consistency Equation

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

- We can show this as follows:

$$\begin{aligned}
V_\pi(s) &= \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] = \mathbb{E}[(r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)) \mid s_0 = s] \\
&= \mathbb{E}_{a \sim \pi(s,\cdot)}[r(s, a)] + \gamma \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_0 = s, s_1 = s'] | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) | s_1 = s'] | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}[V_\pi(s') | s_0 = s] \\
&= r_\pi(s) + \gamma \mathbb{E}_{s' \sim P_\pi(s,\cdot)}[V_\pi(s')] \\
&= r_\pi(s) + \gamma (P_\pi V_\pi)(s)
\end{aligned}$$

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

Proof: From the consistency equation, $(I - \gamma P_\pi)V_\pi = r_\pi$.
We can show that $(I - \gamma P_\pi)$ is invertible as follows.

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

Proof: From the consistency equation, $(I - \gamma P_\pi)V_\pi = r_\pi$.
We can show that $(I - \gamma P_\pi)$ is invertible as follows.
Let $V \in \mathbb{R}^{|\mathcal{S}|}$ be an arbitrary non-zero vector.

$$\begin{aligned}\|(I - \gamma P_\pi)V\|_\infty = \|(V - \gamma P_\pi V)\|_\infty &\overset{(a)}{\geq} \|V\|_\infty - \gamma \|P_\pi V\|_\infty \\ &\overset{(b)}{\geq} \|V\|_\infty - \gamma \|V\|_\infty = (1-\gamma)\|V\|_\infty > 0\end{aligned}$$

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

Proof: From the consistency equation, $(I - \gamma P_\pi)V_\pi = r_\pi$.
We can show that $(I - \gamma P_\pi)$ is invertible as follows.
Let $V \in \mathbb{R}^{|\mathcal{S}|}$ be an arbitrary non-zero vector.

$$\begin{aligned}\|(I - \gamma P_\pi)V\|_\infty = \|(V - \gamma P_\pi V)\|_\infty &\overset{(a)}{\geq} \|V\|_\infty - \gamma \|P_\pi V\|_\infty \\ &\overset{(b)}{\geq} \|V\|_\infty - \gamma \|V\|_\infty = (1-\gamma)\|V\|_\infty > 0\end{aligned}$$

So, $(I - \gamma P_\pi)$ does not have 0 as eigenvalue. □

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

- What is the computational complexity of policy evaluation?

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

- What is the computational complexity of policy evaluation?
- We are solving a linear system of equations $(I - \gamma P_\pi) V_\pi = r_\pi$

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

- What is the computational complexity of policy evaluation?
- We are solving a linear system of equations $(I - \gamma P_\pi)V_\pi = r_\pi$
- Solving a system of linear equations has a complexity of $O(n^3)$, where n is the number of unknowns

Computing V_π: Linear System of Equations Approach

- Value function V_π satisfies the following consistency equation

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

Proposition

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

- What is the computational complexity of policy evaluation?
- We are solving a linear system of equations $(I - \gamma P_\pi)V_\pi = r_\pi$
- Solving a system of linear equations has a complexity of $O(n^3)$, where n is the number of unknowns
- Policy evaluation using the linear system of equations approach has $O(|\mathcal{S}|^3)$ complexity

Computing V_π: Iterative Approach

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0
- This will generate a sequence $V_0, V_1, V_2, \ldots$

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0
- This will generate a sequence $V_0, V_1, V_2, \ldots$
- Will this sequence converge?

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0
- This will generate a sequence $V_0, V_1, V_2, \ldots$
- Will this sequence converge?

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0
- This will generate a sequence $V_0, V_1, V_2, \ldots$
- Will this sequence converge?

Theorem (Convergence of Policy Evaluation Iteration)

Let $T_\pi V = r_\pi + \gamma P_\pi V$. Define the policy evaluation iteration $V_{k+1} = T_\pi V_k$. Then $\lim_{k\to\infty} V_k = V_\pi$

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0
- This will generate a sequence $V_0, V_1, V_2, \ldots$
- Will this sequence converge?

Theorem (Convergence of Policy Evaluation Iteration)

Let $T_\pi V = r_\pi + \gamma P_\pi V$. Define the policy evaluation iteration $V_{k+1} = T_\pi V_k$. Then $\lim_{k\to\infty} V_k = V_\pi$

- Proof relies on the contraction mapping theorem. We will discuss this in detail later

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0
- This will generate a sequence $V_0, V_1, V_2, \ldots$
- Will this sequence converge?

Theorem (Convergence of Policy Evaluation Iteration)

Let $T_\pi V = r_\pi + \gamma P_\pi V$. Define the policy evaluation iteration $V_{k+1} = T_\pi V_k$. Then $\lim_{k\to\infty} V_k = V_\pi$

- Proof relies on the contraction mapping theorem. We will discuss this in detail later
- How many computations per iteration?

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0
- This will generate a sequence $V_0, V_1, V_2, \ldots$
- Will this sequence converge?

Theorem (Convergence of Policy Evaluation Iteration)

Let $T_\pi V = r_\pi + \gamma P_\pi V$. Define the policy evaluation iteration $V_{k+1} = T_\pi V_k$. Then $\lim_{k\to\infty} V_k = V_\pi$

- Proof relies on the contraction mapping theorem. We will discuss this in detail later
- How many computations per iteration?
  - $O(|\mathcal{S}|^2)$

Computing V_π: Iterative Approach

- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = r_\pi + \gamma P_\pi V$$

- Define the iteration $V_{k+1} = T_\pi V_k$, starting with an initial (arbitrary) V_0
- This will generate a sequence $V_0, V_1, V_2, \ldots$
- Will this sequence converge?

Theorem (Convergence of Policy Evaluation Iteration)

Let $T_\pi V = r_\pi + \gamma P_\pi V$. Define the policy evaluation iteration $V_{k+1} = T_\pi V_k$. Then $\lim_{k\to\infty} V_k = V_\pi$

- Proof relies on the contraction mapping theorem. We will discuss this in detail later
- How many computations per iteration?
  - $O(|\mathcal{S}|^2)$
- How many iterations until convergence?

Action-Value (Q-value) function

- Recall that the value function of policy π is defined as

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$$

Action-Value (Q-value) function

- Recall that the value function of policy π is defined as

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$$

- Action-Value function (Q-value function) for policy π is defined as

$$Q_\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

Action-Value (Q-value) function

- Recall that the value function of policy π is defined as

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$$

- Action-Value function (Q-value function) for policy π is defined as

$$Q_\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- It is straight forward to show that

$$Q_\pi(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) V_\pi(s')$$

Action-Value (Q-value) function

- Recall that the value function of policy π is defined as

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$$

- Action-Value function (Q-value function) for policy π is defined as

$$Q_\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- It is straight forward to show that

$$Q_\pi(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) V_\pi(s')$$

Action-Value (Q-value) function

- Recall that the value function of policy π is defined as

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$$

- Action-Value function (Q-value function) for policy π is defined as

$$Q_\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- It is straight forward to show that

$$Q_\pi(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) V_\pi(s')$$

$$Q_\pi(s, a) = \underbrace{r(s, a)}_{\text{Instantaneous reward for } (x,a)} + \gamma \underbrace{\sum_{s'} P(s'|s, a) V_\pi(s')}_{\text{Expected value of the next state}}$$

Action-Value (Q-value) function

- Recall that the value function of policy π is defined as

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$$

- Action-Value function (Q-value function) for policy π is defined as

$$Q_\pi(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- It is straight forward to show that

$$Q_\pi(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) V_\pi(s')$$

$$Q_\pi(s, a) = \underbrace{r(s, a)}_{\text{Instantaneous reward for } (x,a)} + \gamma \underbrace{\sum_{s'} P(s'|s, a) V_\pi(s')}_{\text{Expected value of the next state}}$$

- It is also straight forward to show that

$$V_\pi = \sum_a Q_\pi(s, a) \pi(s, a)$$

Bellman Optimality Equation

MDP Questions

MDP Questions

- How do we find the value of a policy π?

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)], \quad a_t \sim \pi(s_t, \cdot), \; s_{t+1} \sim P(\cdot|s_t, a_t), \; s_0 = s$$

MDP Questions

- How do we find the value of a policy π?

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)], \quad a_t \sim \pi(s_t, \cdot), \; s_{t+1} \sim P(\cdot|s_t, a_t), \; s_0 = s$$

- How do we find the optimal policy?

$$\pi^*(s) = \arg\max_\pi V_\pi(s)$$

MDP Questions

- How do we find the value of a policy π?

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)], \quad a_t \sim \pi(s_t, \cdot), \; s_{t+1} \sim P(\cdot|s_t, a_t), \; s_0 = s$$

- How do we find the optimal policy?

$$\pi^*(s) = \arg\max_\pi V_\pi(s)$$

- How do we find the optimal value function V^*?

$$V^*(s) = \max_\pi \; V_\pi(s)$$

MDP Questions

- How do we find the value of a policy π?

$$V_\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)], \quad a_t \sim \pi(s_t, \cdot), \ s_{t+1} \sim P(\cdot|s_t, a_t), \ s_0 = s$$

- How do we find the optimal policy?

$$\pi^*(s) = \arg\max_\pi V_\pi(s)$$

- How do we find the optimal value function V^*?

$$V^*(s) = \max_\pi \ V_\pi(s)$$

  - Note that $V^* = V_{\pi^*}$

Bellman Optimality Equation

Theorem (Bellman Optimality Equation)

Let V^* be the optimal value function. Then V^* satisfies the equation

$$V^*(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right)$$

Bellman Optimality Equation

Theorem (Bellman Optimality Equation)

Let V^* be the optimal value function. Then V^* satisfies the equation

$$V^*(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right)$$

- **Reading assignment:** Proof of the above theorem

Bellman Optimality Equation

- Bellman Optimality Equation

$$V^*(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s')\right)$$

Bellman Optimality Equation

- Bellman Optimality Equation

$$V^*(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s')\right)$$

$$V^*(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$$

Bellman Optimality Equation

- Bellman Optimality Equation

$$V^*(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s')\right)$$

$$V^*(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$$

- Suppose the current state is s and we take action a. From the next state onwards, we follow the optimal policy π^*. What is the 'value' of state s (expected cumulative discounted reward) under this procedure?

Bellman Optimality Equation

- Bellman Optimality Equation

$$V^*(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s')\right)$$

$$V^*(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$$

- Suppose the current state is s and we take action a. From the next state onwards, we follow the optimal policy π^*. What is the 'value' of state s (expected cumulative discounted reward) under this procedure?

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

Bellman Optimality Equation

- Bellman Optimality Equation

$$V^*(s) = \max_a \left(r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s')\right)$$

$$V^*(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$$

- Suppose the current state is s and we take action a. From the next state onwards, we follow the optimal policy π^*. What is the 'value' of state s (expected cumulative discounted reward) under this procedure?

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

$$V(s) = \underbrace{r(s,a)}_{\text{Reward for } (x,a)} + \gamma \underbrace{\mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]}_{\substack{\text{Expected value of the next state} \\ \text{under the optimal policy } \pi^*}}$$

Bellman Optimality Equation

- Bellman Optimality Equation

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- Suppose the current state is s and we take action a. From the next state onwards, we follow the optimal policy π^*. What is the 'value' of state s (expected cumulative discounted reward) under this procedure?

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

Bellman Optimality Equation

- Bellman Optimality Equation

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- Suppose the current state is s and we take action a. From the next state onwards, we follow the optimal policy π^*. What is the 'value' of state s (expected cumulative discounted reward) under this procedure?

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- What is the best action that we can take in the current state s?

Bellman Optimality Equation

- Bellman Optimality Equation

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- Suppose the current state is s and we take action a. From the next state onwards, we follow the optimal policy π^*. What is the 'value' of state s (expected cumulative discounted reward) under this procedure?

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- What is the best action that we can take in the current state s?
  - Take the action which maximizes the value $V(s)$

Bellman Optimality Equation

- Bellman Optimality Equation

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- Suppose the current state is s and we take action a. From the next state onwards, we follow the optimal policy π^*. What is the 'value' of state s (expected cumulative discounted reward) under this procedure?

$$V(s) = r(s,a) + +\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

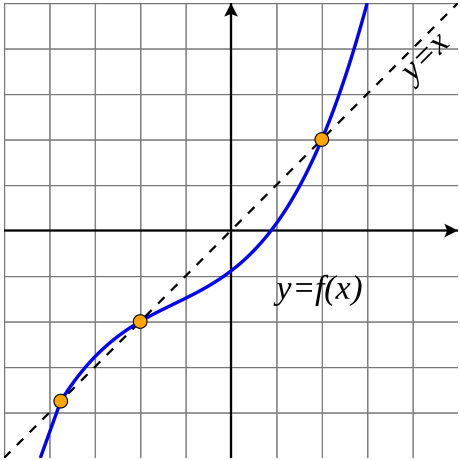
- What is the best action that we can take in the current state s?
  - Take the action which maximizes the value $V(s)$
- So, we are taking the optimal action in the current state and in all following states. This should give the optimal value corresponding to the current state

Fixed Point and Contraction Mapping

Fixed Point of a Function

Definition (Fixed Point)

Let $f : \mathcal{U} \to \mathcal{U}$. u^* is a fixed point of $f(\cdot)$ if $f(u^*) = u^*$.



Contraction Mapping

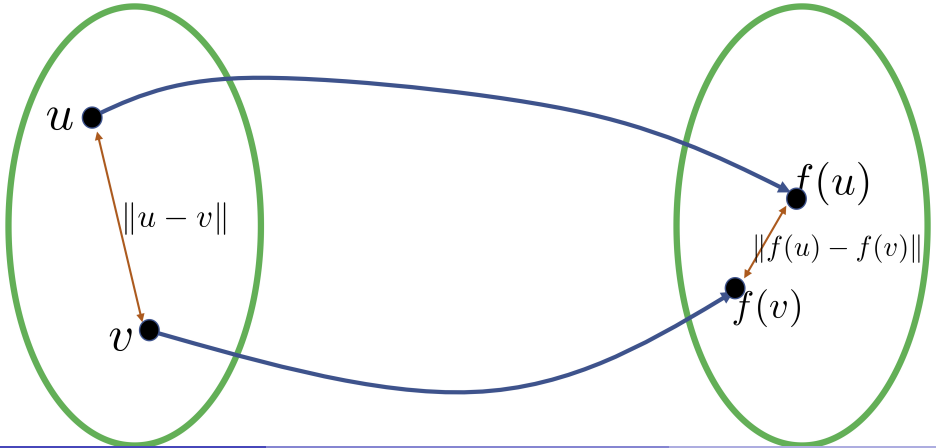
Definition

Let $f : \mathcal{U} \to \mathcal{U}$ and let $\|\cdot\|$ be a norm defined on $\mathcal{U}$. We say $f(\cdot)$ is a contraction mapping (or simply, contraction) if $\|f(u) - f(v)\| \leq \alpha\|u - v\|$ for an $\alpha \in (0, 1)$ and for all $u, v \in \mathcal{U}$.

Contraction Mapping

Definition

Let $f : \mathcal{U} \to \mathcal{U}$ and let $\|\cdot\|$ be a norm defined on $\mathcal{U}$. We say $f(\cdot)$ is a contraction mapping (or simply, contraction) if $\|f(u) - f(v)\| \leq \alpha \|u - v\|$ for an $\alpha \in (0, 1)$ and for all $u, v \in \mathcal{U}$.



Contraction Mapping

Definition

Let $f : \mathcal{U} \to \mathcal{U}$ and let $\|\cdot\|$ be a norm defined on $\mathcal{U}$. We say $f(\cdot)$ is a contraction mapping (or simply, contraction) if $\|f(u) - f(v)\| \leq \alpha \|u - v\|$ for an $\alpha \in (0, 1)$ for all $u, v \in \mathcal{U}$.

- Example: $f : [\sqrt{a}, \infty] \to [\sqrt{a}, \infty], f(u) = \frac{1}{2}\left(u + \frac{a}{u}\right)$. Then $f(\cdot)$ is a contraction mapping.

Contraction Mapping

Definition

Let $f : \mathcal{U} \to \mathcal{U}$ and let $\|\cdot\|$ be a norm defined on $\mathcal{U}$. We say $f(\cdot)$ is a contraction mapping (or simply, contraction) if $\|f(u) - f(v)\| \leq \alpha \|u - v\|$ for an $\alpha \in (0, 1)$ for all $u, v \in \mathcal{U}$.

- Example: $f : [\sqrt{a}, \infty] \to [\sqrt{a}, \infty], f(u) = \frac{1}{2}\left(u + \frac{a}{u}\right)$. Then $f(\cdot)$ is a contraction mapping.

$$|f(u) - f(v)| = \frac{1}{2}\left|1 - \frac{a}{uv}\right| |u - v| < \frac{1}{2}|u - v|$$

Contraction Mapping

Definition

Let $f : \mathcal{U} \to \mathcal{U}$ and let $\|\cdot\|$ be a norm defined on $\mathcal{U}$. We say $f(\cdot)$ is a contraction mapping (or simply, contraction) if $\|f(u) - f(v)\| \leq \alpha\|u - v\|$ for an $\alpha \in (0, 1)$ for all $u, v \in \mathcal{U}$.

- Example: $f : [\sqrt{a}, \infty] \to [\sqrt{a}, \infty], f(u) = \frac{1}{2}\left(u + \frac{a}{u}\right)$. Then $f(\cdot)$ is a contraction mapping.

$$|f(u) - f(v)| = \frac{1}{2}\left|1 - \frac{a}{uv}\right| |u - v| < \frac{1}{2}|u - v|$$

- Example: $f : \mathbb{R}^n \to \mathbb{R}^n, f(u) = Au$, where $A \in \mathbb{R}^n \times \mathbb{R}^n$. Assume that the row sum of A is strictly less than 1, i.e., $\sum_j |a_{ij}| \leq \alpha < 1$. Then $f(\cdot)$ is a contraction mapping with respect to $\|\cdot\|_\infty$

Banach Fixed Point Theorem

Theorem (Banach Fixed Point Theorem)

Let $\mathcal{U}$ be a closed and bounded subset of $\mathbb{R}^n$. Let $f : \mathcal{U} \to \mathcal{U}$ be a contraction mapping, i.e, $\|f(u) - f(v)\| \leq \alpha \|u - v\|$, $\alpha \in (0, 1)$, for all $u, v \in \mathcal{U}$. Then,

1. *$f(\cdot)$ has a unique fixed point u^**
2. *The iteration $u_{k+1} = f(u_k)$ with an arbitrary initial element u_0 will converge to the unique fixed point, i.e., $\lim_{k \to \infty} u_k = u^*$*

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$.

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$
$$\|u_3 - u_2\| = \|f(u_2) - f(u_1)\| \leq \alpha \|u_2 - u_1\| \leq \alpha^2 \|u_1 - u_0\|$$

□

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$
$$\|u_3 - u_2\| = \|f(u_2) - f(u_1)\| \leq \alpha \|u_2 - u_1\| \leq \alpha^2 \|u_1 - u_0\|$$
$$\|u_4 - u_3\| = \|f(u_3) - f(u_2)\| \leq \alpha \|u_3 - u_2\| \leq \alpha^3 \|u_1 - u_0\|$$

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$
$$\|u_3 - u_2\| = \|f(u_2) - f(u_1)\| \leq \alpha \|u_2 - u_1\| \leq \alpha^2 \|u_1 - u_0\|$$
$$\|u_4 - u_3\| = \|f(u_3) - f(u_2)\| \leq \alpha \|u_3 - u_2\| \leq \alpha^3 \|u_1 - u_0\|$$

We can show that $\|u_{n+1} - u_n\| \leq \alpha^n \|u_1 - u_0\|, \forall n \geq 1$

□

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$
$$\|u_3 - u_2\| = \|f(u_2) - f(u_1)\| \leq \alpha \|u_2 - u_1\| \leq \alpha^2 \|u_1 - u_0\|$$
$$\|u_4 - u_3\| = \|f(u_3) - f(u_2)\| \leq \alpha \|u_3 - u_2\| \leq \alpha^3 \|u_1 - u_0\|$$

We can show that $\|u_{n+1} - u_n\| \leq \alpha^n \|u_1 - u_0\|, \forall n \geq 1$

We can now show that the sequence $\{u_0, u_1, u_2, \ldots\}$ is a Cauchy sequence.

□

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$
$$\|u_3 - u_2\| = \|f(u_2) - f(u_1)\| \leq \alpha \|u_2 - u_1\| \leq \alpha^2 \|u_1 - u_0\|$$
$$\|u_4 - u_3\| = \|f(u_3) - f(u_2)\| \leq \alpha \|u_3 - u_2\| \leq \alpha^3 \|u_1 - u_0\|$$

We can show that $\|u_{n+1} - u_n\| \leq \alpha^n \|u_1 - u_0\|, \forall n \geq 1$
We can now show that the sequence $\{u_0, u_1, u_2, \ldots\}$ is a Cauchy sequence.
Since $\mathcal{U}$ is closed and bounded, the sequence converges, i.e., there exists a limit $u^* = \lim_{k \to \infty} u_k$.

□

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$
$$\|u_3 - u_2\| = \|f(u_2) - f(u_1)\| \leq \alpha \|u_2 - u_1\| \leq \alpha^2 \|u_1 - u_0\|$$
$$\|u_4 - u_3\| = \|f(u_3) - f(u_2)\| \leq \alpha \|u_3 - u_2\| \leq \alpha^3 \|u_1 - u_0\|$$

We can show that $\|u_{n+1} - u_n\| \leq \alpha^n \|u_1 - u_0\|, \forall n \geq 1$
We can now show that the sequence $\{u_0, u_1, u_2, \ldots\}$ is a Cauchy sequence.
Since $\mathcal{U}$ is closed and bounded, the sequence converges, i.e., there exists a limit $u^* = \lim_{k \to \infty} u_k$.

We will now show that u^* is a fixed point of $f(\cdot)$.

□

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$
$$\|u_3 - u_2\| = \|f(u_2) - f(u_1)\| \leq \alpha \|u_2 - u_1\| \leq \alpha^2 \|u_1 - u_0\|$$
$$\|u_4 - u_3\| = \|f(u_3) - f(u_2)\| \leq \alpha \|u_3 - u_2\| \leq \alpha^3 \|u_1 - u_0\|$$

We can show that $\|u_{n+1} - u_n\| \leq \alpha^n \|u_1 - u_0\|, \forall n \geq 1$
We can now show that the sequence $\{u_0, u_1, u_2, \ldots\}$ is a Cauchy sequence.
Since $\mathcal{U}$ is closed and bounded, the sequence converges, i.e., there exists a limit $u^* = \lim_{k \to \infty} u_k$.

We will now show that u^* is a fixed point of $f(\cdot)$.
Since $f(\cdot)$ is a contraction, it is also a continuous function.

□

Banach Fixed Point Theorem: Proof

Proof: **Existence:** Let $u_0 \in \mathcal{U}$ be arbitrary. Define the iteration $u_{k+1} = f(u_k)$. Now, repeatedly applying the contraction property,

$$\|u_2 - u_1\| = \|f(u_1) - f(u_0)\| \leq \alpha \|u_1 - u_0\|$$
$$\|u_3 - u_2\| = \|f(u_2) - f(u_1)\| \leq \alpha \|u_2 - u_1\| \leq \alpha^2 \|u_1 - u_0\|$$
$$\|u_4 - u_3\| = \|f(u_3) - f(u_2)\| \leq \alpha \|u_3 - u_2\| \leq \alpha^3 \|u_1 - u_0\|$$

We can show that $\|u_{n+1} - u_n\| \leq \alpha^n \|u_1 - u_0\|, \forall n \geq 1$
We can now show that the sequence $\{u_0, u_1, u_2, \ldots\}$ is a Cauchy sequence.
Since $\mathcal{U}$ is closed and bounded, the sequence converges, i.e., there exists a limit $u^* = \lim_{k \to \infty} u_k$.

We will now show that u^* is a fixed point of $f(\cdot)$.
Since $f(\cdot)$ is a contraction, it is also a continuous function.
Then, $u^* = \lim_{k \to \infty} u_k = \lim_{k \to \infty} u_{k+1} = \lim_{k \to \infty} f(u_k) = f(\lim_{k \to \infty} u_k) = f(u^*)$

□

Banach Fixed Point Theorem: Proof

Proof: **Uniqueness:** Suppose there exists another fixed point $\bar{u}^* \neq u^*$. Then,

Banach Fixed Point Theorem: Proof

Proof: **Uniqueness:** Suppose there exists another fixed point $\bar{u}^* \neq u^*$. Then,

$$\|u^* - \bar{u}^*\| = \|f(u^*) - f(\bar{u}^*)\| \leq \alpha \|u^* - \bar{u}^*\|.$$

Banach Fixed Point Theorem: Proof

Proof: **Uniqueness:** Suppose there exists another fixed point $\bar{u}^* \neq u^*$. Then,

$$\|u^* - \bar{u}^*\| = \|f(u^*) - f(\bar{u}^*)\| \leq \alpha \|u^* - \bar{u}^*\|.$$

This is a contradiction.

Banach Fixed Point Theorem: Proof

Proof: **Uniqueness:** Suppose there exists another fixed point $\bar{u}^* \neq u^*$. Then,

$$\|u^* - \bar{u}^*\| = \|f(u^*) - f(\bar{u}^*)\| \leq \alpha \|u^* - \bar{u}^*\|.$$

This is a contradiction.
So, $\bar{u}^* = u^*$.

□

Value Iteration

Bellman Operator

- Bellman operator $T : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$$

Bellman Operator

- Bellman operator $T : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$$

Proposition (Contraction property of Bellman operator)

Bellman operator T is a contraction with respect to $\|\cdot\|_\infty$, i.e, for any $V_1, V_2 \in \mathcal{R}^{|\mathcal{S}|}$, $\|TV_1 - TV_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$.

Bellman Operator

- Bellman operator $T : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$$

Proposition (Contraction property of Bellman operator)

Bellman operator T is a contraction with respect to $\|\cdot\|_\infty$, i.e, for any $V_1, V_2 \in \mathcal{R}^{|\mathcal{S}|}$, $\|TV_1 - TV_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$.

Lemma

Let g, h be any two real-valued function. Then,

$$\left|\max_a \ g(x,a) - \max_a \ h(x,a)\right| \leq \max_a \ |g(x,a) - h(x,a)|$$

Bellman Operator

Proof: (of contraction property)

Bellman Operator

Proof: (of contraction property)

$$
\begin{aligned}
&|(TV_1)(s) - (TV_2)(s)| \\
&= |\max_{a \in \mathcal{A}} (r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_1(s')]) - \max_{a \in \mathcal{A}} (r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_2(s')])|
\end{aligned}
$$

Bellman Operator

Proof: (of contraction property)

$$
\begin{aligned}
&|(TV_1)(s) - (TV_2)(s)| \\
&= |\max_{a\in\mathcal{A}}(r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_1(s')]) - \max_{a\in\mathcal{A}}(r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_2(s')])| \\
&\overset{(a)}{\leq} \max_{a\in\mathcal{A}} |(r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_1(s')]) - (r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_2(s')])|
\end{aligned}
$$

Bellman Operator

Proof: (of contraction property)

$$
\begin{aligned}
&|(TV_1)(s) - (TV_2)(s)| \\
&= |\max_{a \in \mathcal{A}} (r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_1(s')]) - \max_{a \in \mathcal{A}} (r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_2(s')])| \\
&\overset{(a)}{\leq} \max_{a \in \mathcal{A}} |(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_1(s')]) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_2(s')])| \\
&= \max_{a \in \mathcal{A}} |\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_1(s')] - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_2(s')]|
\end{aligned}
$$

□

Bellman Operator

Proof: (of contraction property)

$$
\begin{aligned}
&|(TV_1)(s) - (TV_2)(s)| \\
&= |\max_{a \in \mathcal{A}} (r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_1(s')]) - \max_{a \in \mathcal{A}} (r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_2(s')])| \\
&\overset{(a)}{\leq} \max_{a \in \mathcal{A}} |(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_1(s')]) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_2(s')])| \\
&= \max_{a \in \mathcal{A}} |\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_1(s')] - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_2(s')]| \\
&\leq \max_{a \in \mathcal{A}} \ \gamma \ \mathbb{E}_{s' \sim P(\cdot|s,a)}[|V_1(s') - V_2(s')|]
\end{aligned}
$$

□

Bellman Operator

Proof: (of contraction property)

$$
\begin{aligned}
&|(TV_1)(s) - (TV_2)(s)| \\
&= |\max_{a\in\mathcal{A}}(r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_1(s')]) - \max_{a\in\mathcal{A}}(r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_2(s')])| \\
&\overset{(a)}{\leq} \max_{a\in\mathcal{A}} |(r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_1(s')]) - (r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_2(s')])| \\
&= \max_{a\in\mathcal{A}} |\gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_1(s')] - \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_2(s')]| \\
&\leq \max_{a\in\mathcal{A}} \ \gamma \ \mathbb{E}_{s'\sim P(\cdot|s,a)}[|V_1(s') - V_2(s')|] \\
&\leq \max_{a\in\mathcal{A}} \ \gamma \ \mathbb{E}_{s'\sim P(\cdot|s,a)}[\|V_1 - V_2\|_\infty] = \gamma \|V_1 - V_2\|_\infty
\end{aligned}
$$

□

Bellman Operator

Proof: (of contraction property)

$$
\begin{aligned}
&|(TV_1)(s) - (TV_2)(s)| \\
&= |\max_{a\in\mathcal{A}}(r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_1(s')]) - \max_{a\in\mathcal{A}}(r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_2(s')])| \\
&\overset{(a)}{\leq} \max_{a\in\mathcal{A}} |(r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_1(s')]) - (r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_2(s')])| \\
&= \max_{a\in\mathcal{A}} |\gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_1(s')] - \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_2(s')]| \\
&\leq \max_{a\in\mathcal{A}} \ \gamma \ \mathbb{E}_{s'\sim P(\cdot|s,a)}[|V_1(s') - V_2(s')|] \\
&\leq \max_{a\in\mathcal{A}} \ \gamma \ \mathbb{E}_{s'\sim P(\cdot|s,a)}[\|V_1 - V_2\|_\infty] = \gamma \|V_1 - V_2\|_\infty
\end{aligned}
$$

This implies

$$\|TV_1 - TV_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$$

□

Optimal Value Function

- Bellman operator: $(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$

Optimal Value Function

- Bellman operator: $(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$

Theorem

1. *Optimal value function V^* is the unique fixed point of the Bellman operator T, i.e., $TV^* = V^*$*

Optimal Value Function

- Bellman operator: $(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$

Theorem

1. *Optimal value function V^* is the unique fixed point of the Bellman operator T, i.e., $TV^* = V^*$*
2. *Value Iteration, defined as $V_{k+1} = TV_k$, with an arbitrary initial value V_0 will converge to the optimal value function V^*, i.e, $\lim_{k \to \infty} V_k = V^*$*

Optimal Value Function

- Bellman operator: $(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$

Theorem

1. *Optimal value function V^* is the unique fixed point of the Bellman operator T, i.e., $TV^* = V^*$*
2. *Value Iteration, defined as $V_{k+1} = TV_k$, with an arbitrary initial value V_0 will converge to the optimal value function V^*, i.e, $\lim_{k\to\infty} V_k = V^*$*
3. *Optimal policy π^* can be computed as*

$$\pi^*(s) = \arg\max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$$

Optimal Value Function

- Bellman operator: $(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$

Theorem

1. *Optimal value function V^* is the unique fixed point of the Bellman operator T, i.e., $TV^* = V^*$*
2. *Value Iteration, defined as $V_{k+1} = TV_k$, with an arbitrary initial value V_0 will converge to the optimal value function V^*, i.e, $\lim_{k\to\infty} V_k = V^*$*
3. *Optimal policy π^* can be computed as*

$$\pi^*(s) = \arg\max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$$

- **Reading assignment:** Proof of the above theorem

Optimal Value Function

- Bellman operator: $(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$

Theorem

1. *Optimal value function V^* is the unique fixed point of the Bellman operator T, i.e., $TV^* = V^*$*
2. *Value Iteration, defined as $V_{k+1} = TV_k$, with an arbitrary initial value V_0 will converge to the optimal value function V^*, i.e, $\lim_{k\to\infty} V_k = V^*$*
3. *Optimal policy π^* can be computed as*

$$\pi^*(s) = \arg\max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$$

- **Reading assignment:** Proof of the above theorem
- Part (1) of the above theorem is essentially stating the Bellman optimality equation, $V^*(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$

Optimal Value Function

- Bellman operator: $(TV)(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]\right)$

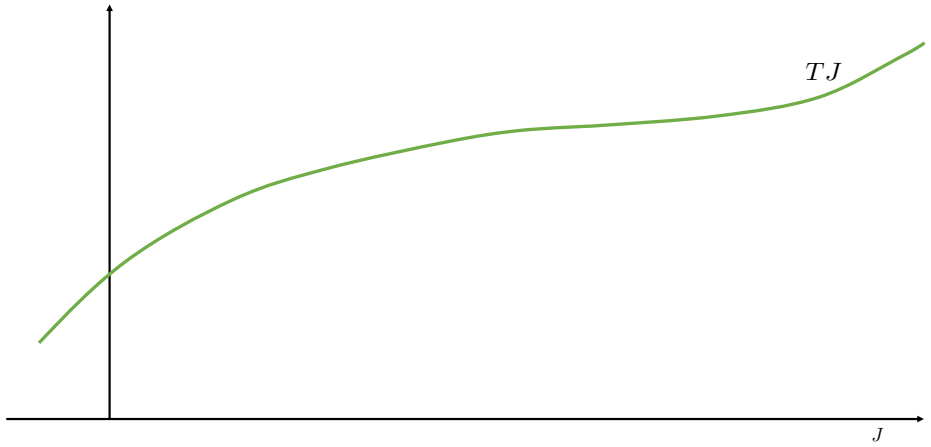
Theorem

1. *Optimal value function V^* is the unique fixed point of the Bellman operator T, i.e., $TV^* = V^*$*
2. *Value Iteration, defined as $V_{k+1} = TV_k$, with an arbitrary initial value V_0 will converge to the optimal value function V^*, i.e, $\lim_{k\to\infty} V_k = V^*$*
3. *Optimal policy π^* can be computed as*

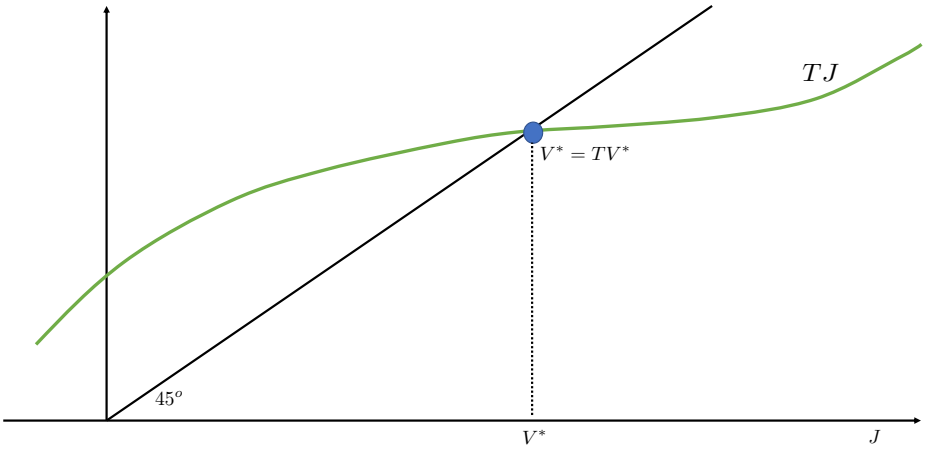
$$\pi^*(s) = \arg\max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$$

- **Reading assignment:** Proof of the above theorem
- Part (1) of the above theorem is essentially stating the Bellman optimality equation, $V^*(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$
- Part (2) follows from the fact that T is a contraction mapping and V^* is its unique fixed point

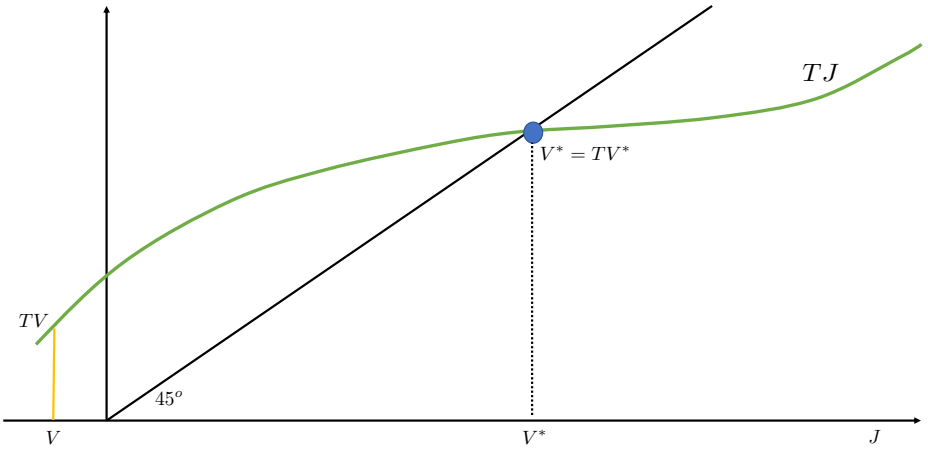
Contraction, Fixed Point, Convergence



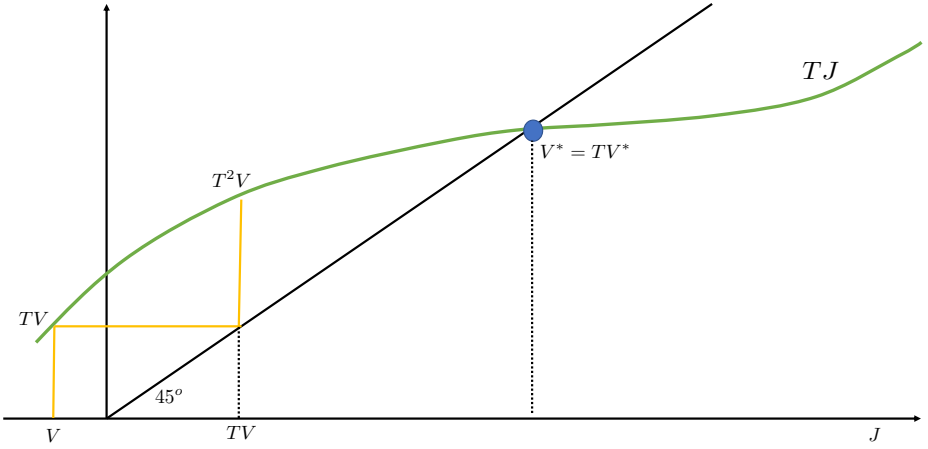
Contraction, Fixed Point, Convergence



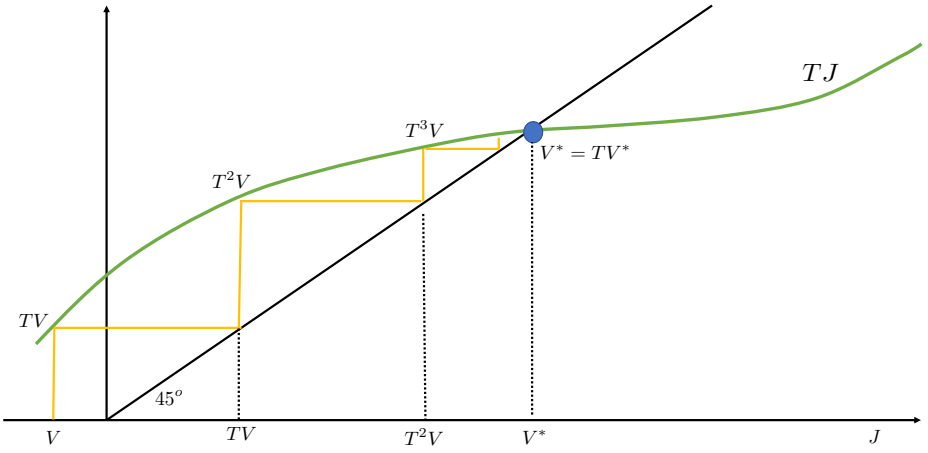
Contraction, Fixed Point, Convergence



Contraction, Fixed Point, Convergence



Contraction, Fixed Point, Convergence



Value Iteration

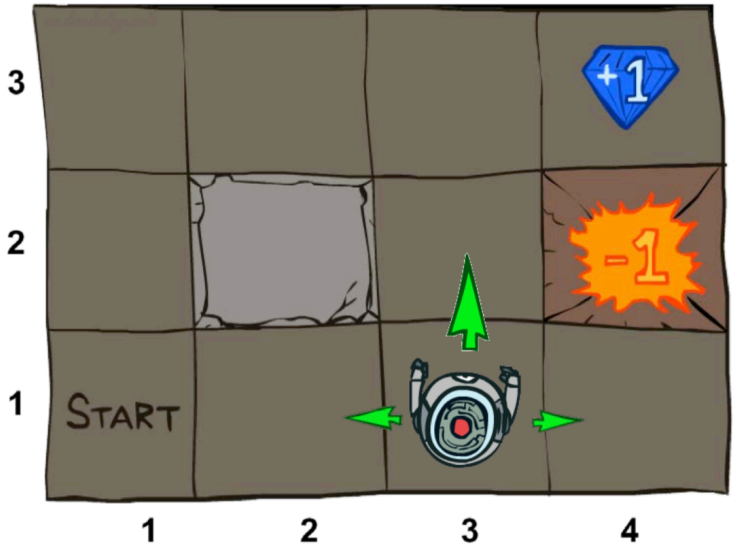
- How do we find the optimal value function V^*?

$$V^*(s) = \max_{\pi} V_\pi(s)$$

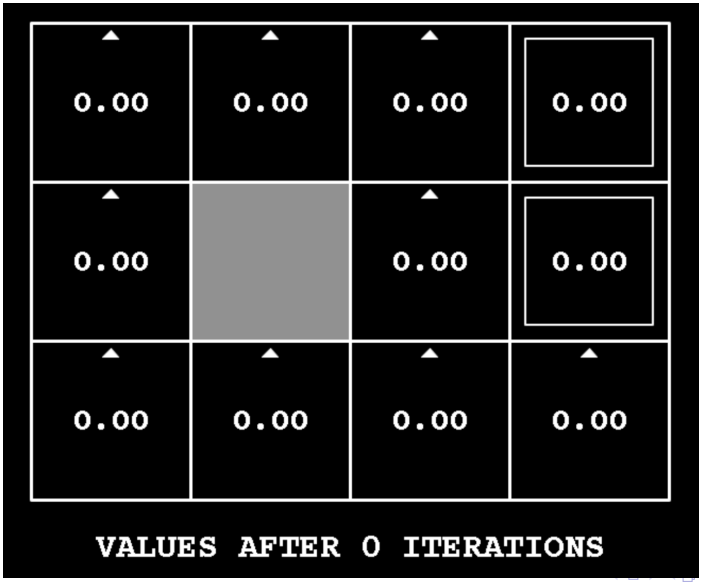
- Value Iteration: $V_{k+1} = TV_k$
- How do we find the optimal policy?

$$\pi^*(s) = \arg\max_{a} \left(r(x,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^*(s')\right)$$

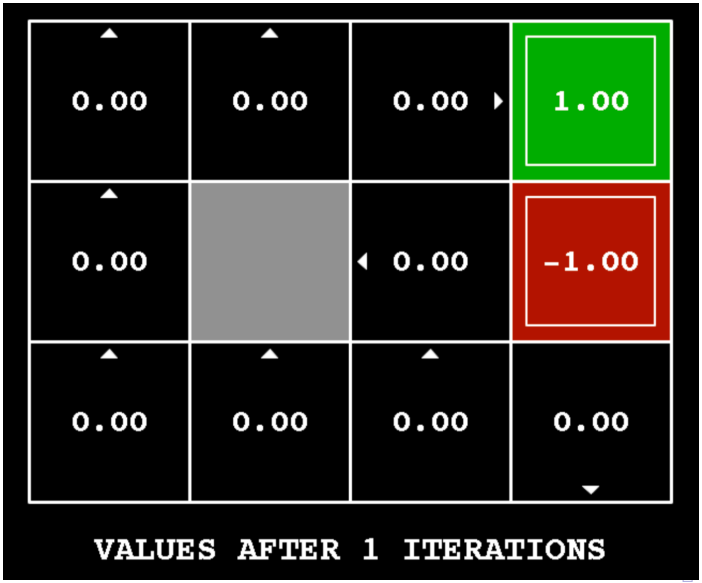
Value Iteration: Example



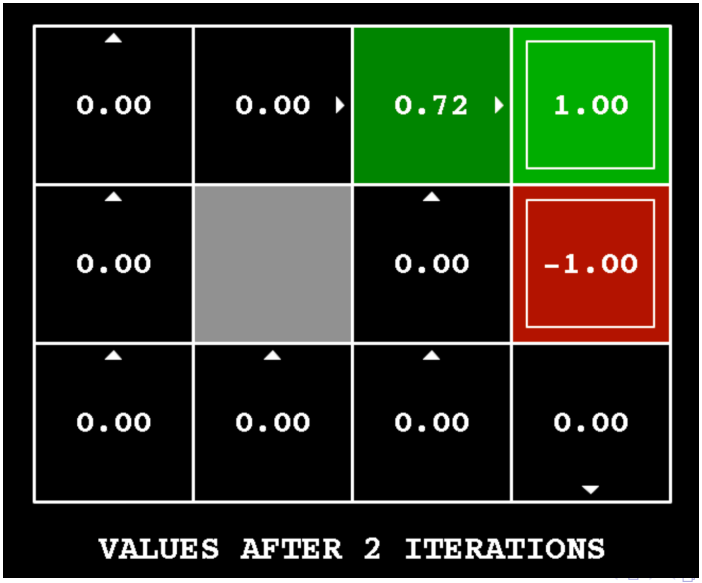
Value Iteration: Example



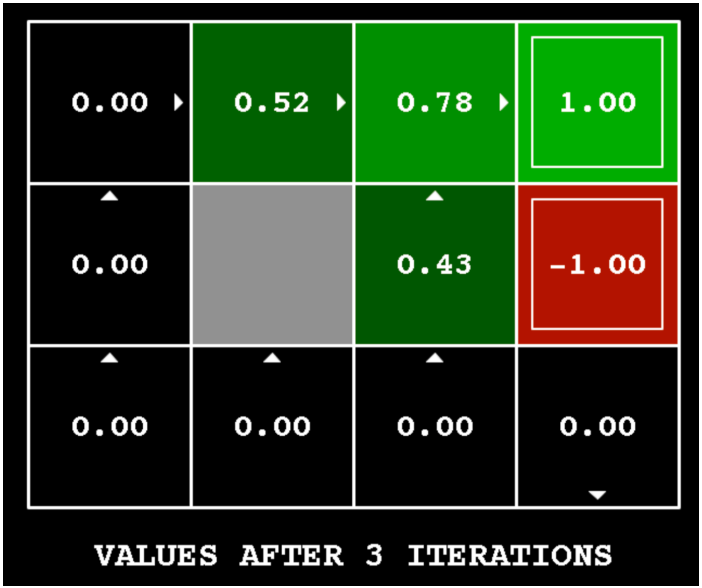
Value Iteration: Example



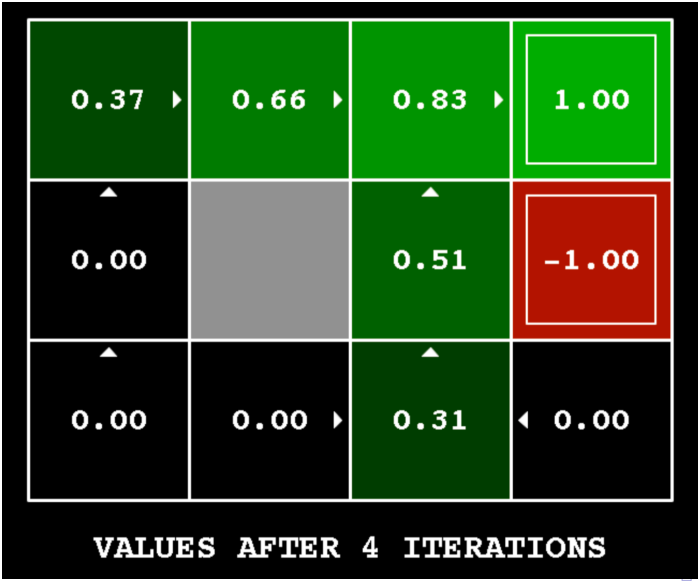
Value Iteration: Example



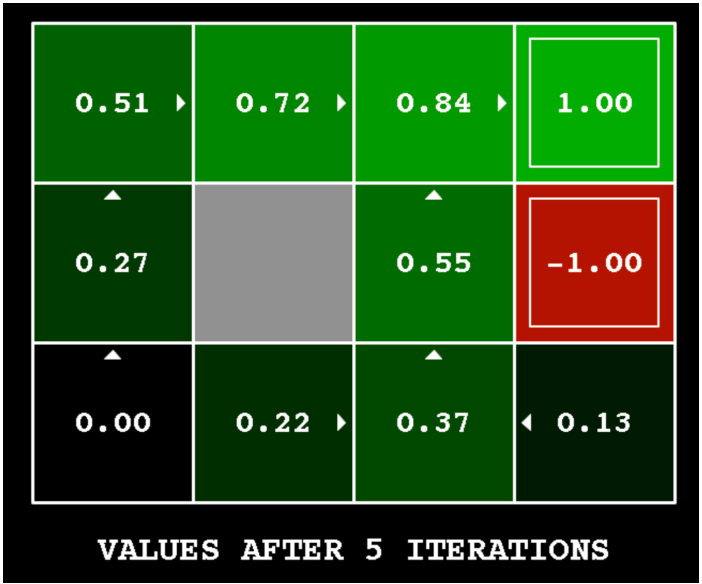
Value Iteration: Example



Value Iteration: Example



Value Iteration: Example



Value Iteration: Example

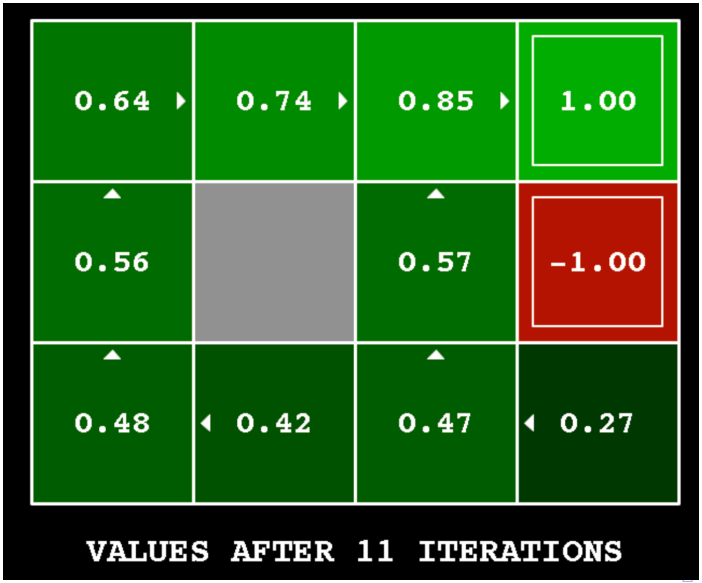
Value Iteration: Example

Value Iteration: Example

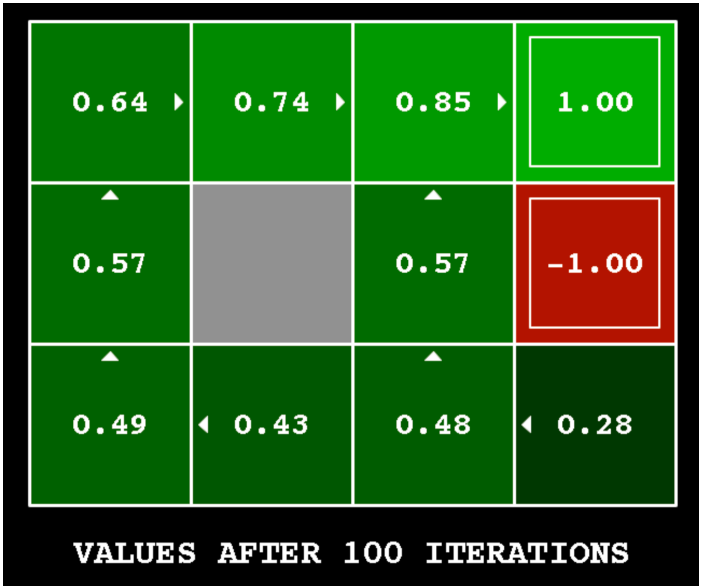
Value Iteration: Example

Value Iteration: Example

Value Iteration: Example



Value Iteration: Example



Optimal Q-value Function

- Q-value function of a policy π

$$Q_\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

Optimal Q-value Function

- Q-value function of a policy π

$$Q_\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- Optimal Q-value function

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

Optimal Q-value Function

- Q-value function of a policy π

$$Q_\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- Optimal Q-value function

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

- V^* to Q^*

Optimal Q-value Function

- Q-value function of a policy π

$$Q_\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- Optimal Q-value function

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

- V^* to Q^*

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

Optimal Q-value Function

- Q-value function of a policy π

$$Q_\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- Optimal Q-value function

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

- V^* to Q^*

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- Q^* to V^*

Optimal Q-value Function

- Q-value function of a policy π

$$Q_\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- Optimal Q-value function

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

- V^* to Q^*

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- Q^* to V^*

$$V^*(s) = \max_a \; Q^*(s,a)$$

Optimal Q-value Function

- Q-value function of a policy π

$$Q_\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- Optimal Q-value function

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

- V^* to Q^*

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

- Q^* to V^*

$$V^*(s) = \max_a \ Q^*(s,a)$$

- π^* from Q^*

Optimal Q-value Function

- Q-value function of a policy π

$$Q_\pi(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a]$$

- Optimal Q-value function

$$Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

- V^* to Q^*

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

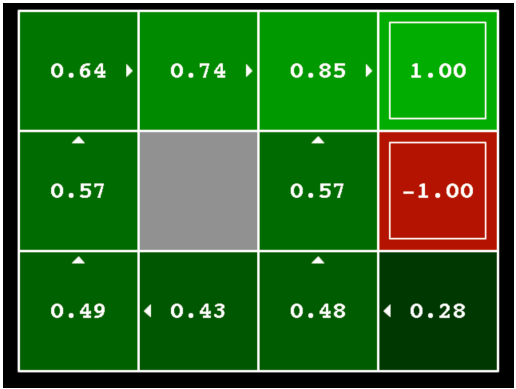
- Q^* to V^*

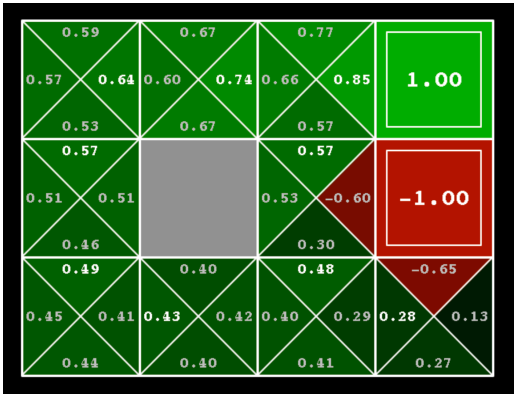
$$V^*(s) = \max_a \; Q^*(s,a)$$

- π^* from Q^*

$$\pi^*(s) = \arg\max_a \; Q^*(s,a)$$

Policy Computation from V^*/Q^*





$$V^*(s) = \max_a Q^*(s,a), \qquad Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^*(s'),$$

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

Computing Q^* Directly

- Optimal Q-value function Q^* satisfies the Bellman equation

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_b Q^*(s',b)]$$

Computing Q^* Directly

- Optimal Q-value function Q^* satisfies the Bellman equation

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_b Q^*(s',b)]$$

- Define the mapping $F : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as

$$(FQ)(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_b Q(s',b)]$$

Computing Q^* Directly

- Optimal Q-value function Q^* satisfies the Bellman equation

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_b Q^*(s',b)]$$

- Define the mapping $F : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as

$$(FQ)(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_b Q(s',b)]$$

Computing Q^* Directly

- Optimal Q-value function Q^* satisfies the Bellman equation

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_b Q^*(s',b)]$$

- Define the mapping $F : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as

$$(FQ)(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_b Q(s',b)]$$

Proposition

1. *F is a contraction mapping w.r.t $\|\cdot\|_\infty$*
2. *Q^* is the unique fixed point of F*
3. *Define Q-value iteration, $Q_{k+1} = FQ_k$. Then, $\lim_{k\to\infty} Q_k = Q^*$*

Policy Iteration

Policy Evaluation: Recap

- Value function V_π satisfies the equation $V_\pi = R_\pi + \gamma P_\pi V_\pi$

Policy Evaluation: Recap

- Value function V_π satisfies the equation $V_\pi = R_\pi + \gamma P_\pi V_\pi$
- How do we find V_π?

Policy Evaluation: Recap

- Value function V_π satisfies the equation $V_\pi = R_\pi + \gamma P_\pi V_\pi$
- How do we find V_π?
- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = R_\pi + \gamma P_\pi V$$

Policy Evaluation: Recap

- Value function V_π satisfies the equation $V_\pi = R_\pi + \gamma P_\pi V_\pi$
- How do we find V_π?
- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = R_\pi + \gamma P_\pi V$$

Policy Evaluation: Recap

- Value function V_π satisfies the equation $V_\pi = R_\pi + \gamma P_\pi V_\pi$
- How do we find V_π?
- Policy evaluation operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is defined as

$$T_\pi V = R_\pi + \gamma P_\pi V$$

Proposition

1. T_π *is a contraction mapping w.r.t* $\|\cdot\|_\infty$
2. V_π *is the unique fixed point of* T_π
3. *Define the policy evaluation iteration,* $V_{k+1} = T_\pi V_k$. *Then* $\lim_{k\to\infty} V_k = V_\pi$

Policy Iteration

- Can we find the optimal policy by 'cleverly' searching over the policy space?

Policy Iteration

- Can we find the optimal policy by 'cleverly' searching over the policy space?
- Policy Iteration: For $k = 0, 1, 2, \ldots$

Policy Iteration

- Can we find the optimal policy by 'cleverly' searching over the policy space?
- Policy Iteration: For $k = 0, 1, 2, \ldots$
  1. Policy evaluation: compute the value of the policy to get V_{π_k}:

$$\text{Solve} \quad T_{\pi_k} V_{\pi_k} = V_{\pi_k}$$

Policy Iteration

- Can we find the optimal policy by 'cleverly' searching over the policy space?
- Policy Iteration: For $k = 0, 1, 2, \ldots$
  1. Policy evaluation: compute the value of the policy to get V_{π_k}:

$$\text{Solve} \quad T_{\pi_k} V_{\pi_k} = V_{\pi_k}$$

  2. Policy improvement: perform greedy update to get the next policy π_{k+1}:

$$\pi_{k+1}(x) = \arg\max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_{\pi_k}(s')]\right)$$

Policy Iteration

- Can we find the optimal policy by 'cleverly' searching over the policy space?
- Policy Iteration: For $k = 0, 1, 2, \ldots$
  1. Policy evaluation: compute the value of the policy to get V_{π_k}:

$$\text{Solve} \quad T_{\pi_k} V_{\pi_k} = V_{\pi_k}$$

  2. Policy improvement: perform greedy update to get the next policy π_{k+1}:

$$\pi_{k+1}(x) = \arg\max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_{\pi_k}(s')]\right)$$

Proposition (Convergence of Policy Iteration)

1. $V_{\pi_{k+1}} \geq V_{\pi_k}$
2. $\|V_{\pi_{k+1}} - V^*\|_\infty \leq \gamma \|V_{\pi_k} - V^*\|_\infty$

Policy Iteration

Lemma (Monotone Property of T_π)

Let $V_1 \geq V_2$. *Then,* $T_\pi V_1 \geq T_\pi V_2$.

Policy Iteration

Lemma (Monotone Property of T_π)

Let $V_1 \geq V_2$. Then, $T_\pi V_1 \geq T_\pi V_2$.

Proof: (of convergence of policy iteration):
Monotone property: Note that

$$V_{\pi_k} = T_{\pi_k} V_{\pi_k} \leq T V_{\pi_k} = T_{\pi_{k+1}} V_{\pi_k}$$

Policy Iteration

Lemma (Monotone Property of T_π)

Let $V_1 \geq V_2$. Then, $T_\pi V_1 \geq T_\pi V_2$.

Proof: (of convergence of policy iteration):
Monotone property: Note that

$$V_{\pi_k} = T_{\pi_k} V_{\pi_k} \leq T V_{\pi_k} = T_{\pi_{k+1}} V_{\pi_k}$$

Using the monotone property of T_π,

$$V_{\pi_k} \leq T_{\pi_{k+1}} V_{\pi_k} \leq T^{(2)}_{\pi_{k+1}} V_{\pi_k} \leq \cdots \leq T^{(m)}_{\pi_{k+1}} V_{\pi_k} \leq \cdots \leq \lim_{m \to \infty} T^{(m)}_{\pi_{k+1}} V_{\pi_k} = V_{\pi_{k+1}}$$

Policy Iteration

Lemma (Monotone Property of T_π)

Let $V_1 \geq V_2$. Then, $T_\pi V_1 \geq T_\pi V_2$.

Proof: (of convergence of policy iteration):
Monotone property: Note that

$$V_{\pi_k} = T_{\pi_k} V_{\pi_k} \leq T V_{\pi_k} = T_{\pi_{k+1}} V_{\pi_k}$$

Using the monotone property of T_π,

$$V_{\pi_k} \leq T_{\pi_{k+1}} V_{\pi_k} \leq T^{(2)}_{\pi_{k+1}} V_{\pi_k} \leq \cdots \leq T^{(m)}_{\pi_{k+1}} V_{\pi_k} \leq \cdots \leq \lim_{m \to \infty} T^{(m)}_{\pi_{k+1}} V_{\pi_k} = V_{\pi_{k+1}}$$

Since there are only finite number of deterministic policies, policy iteration will converge in finite number of steps.

Policy Iteration

Lemma (Monotone Property of T_π)

Let $V_1 \geq V_2$. Then, $T_\pi V_1 \geq T_\pi V_2$.

Proof: (of convergence of policy iteration):
Monotone property: Note that

$$V_{\pi_k} = T_{\pi_k} V_{\pi_k} \leq T V_{\pi_k} = T_{\pi_{k+1}} V_{\pi_k}$$

Using the monotone property of T_π,

$$V_{\pi_k} \leq T_{\pi_{k+1}} V_{\pi_k} \leq T^{(2)}_{\pi_{k+1}} V_{\pi_k} \leq \cdots \leq T^{(m)}_{\pi_{k+1}} V_{\pi_k} \leq \cdots \leq \lim_{m \to \infty} T^{(m)}_{\pi_{k+1}} V_{\pi_k} = V_{\pi_{k+1}}$$

Since there are only finite number of deterministic policies, policy iteration will converge in finite number of steps.
Rate of convergence:

$$\|V^* - V_{\pi_{k+1}}\|_\infty \leq \|V^* - T V_{\pi_k}\|_\infty \leq \gamma \|V^* - V_{\pi_k}\|_\infty$$

□

Dynamic Programming

Dynamic Programming

- How do we find the value of a policy π?

Dynamic Programming

- How do we find the value of a policy π?
  - Policy evaluation iteration

Dynamic Programming

- How do we find the value of a policy π?
  - Policy evaluation iteration
- How do we find the optimal value function V^*?

Dynamic Programming

- How do we find the value of a policy π?
  - Policy evaluation iteration
- How do we find the optimal value function V^*?
  - Value iteration

Dynamic Programming

- How do we find the value of a policy π?
  - Policy evaluation iteration
- How do we find the optimal value function V^*?
  - Value iteration
- How do we find the optimal policy?

Dynamic Programming

- How do we find the value of a policy π?
  - Policy evaluation iteration
- How do we find the optimal value function V^*?
  - Value iteration
- How do we find the optimal policy?
  - Value iteration, policy iteration

Reinforcement Learning Questions

- How do we find the value of a policy π?
- How do we find the optimal value function V^*?
- How do we find the optimal policy?

Reinforcement Learning Questions

- How do we find the value of a policy π?
- How do we find the optimal value function V^*?
- How do we find the optimal policy?

... without the knowledge of the model P