

# Theory of Discounted Markov Decision Processes

Mohammad Sadegh Talebi  
m.shahi@di.ku.dk

Department of Computer Science

STARTING FROM SLIDE 9



# Markov Decision Process

An **infinite-horizon discounted** MDP is a tuple  $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ :

- **State-space**  $\mathcal{S}$  (finite, countably infinite, or continuous)
- **Action-space**  $\mathcal{A} = \cup_{s \in \mathcal{S}} \mathcal{A}_s$  (finite, countably infinite, or continuous)
  - $\mathcal{A}_s$  is the set of actions available in state  $s$
- **Transition function**  $P$ : Selecting  $a \in \mathcal{A}_s$  in  $s \in \mathcal{S}$  leads to a transition to  $s'$  with probability  $P(s'|s, a)$ .  $P(\cdot|s, a)$  is a probability distribution over  $\mathcal{S}$ , i.e.,

$$\sum_{s' \in \mathcal{S}} P(s'|s, a) = 1$$

- **Reward function**  $R$ : Selecting  $a \in \mathcal{A}_s$  in  $s \in \mathcal{S}$  yields a reward  $r \sim R(s, a)$ .
- **Discount factor**  $\gamma$ : Future rewards are discounted geometrically with a rate  $0 < \gamma < 1$ .

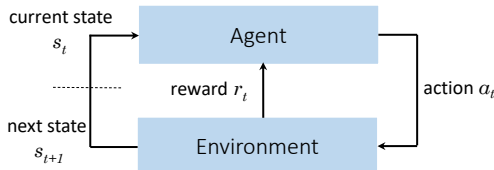


## Recap: Interaction with MDP

An **agent** interacts with the MDP for  $N$  rounds.

At each time step  $t$ :

- The agent observes the current state  $s_t$  and takes an action  $a_t \in \mathcal{A}_{s_t}$
- The environment (MDP) decides a reward  $r_t := r(s_t, a_t) \sim R(s_t, a_t)$  and a next state  $s_{t+1} \sim P(\cdot | s_t, a_t)$
- The agent receives  $r_t$  (any time in step  $t$  before start of  $t + 1$ )



This interaction produces a trajectory (or history)

$$h_t = (s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_{t-1}, a_{t-1}, r_{t-1}, s_t)$$



## Objective Function

**Infinite-Horizon Discounted MDPs:**  $N = \infty$ , and the goal is to maximize the total expected sum of **discounted** rewards

$$\max_{\text{all strategies}} \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \right]$$

Two views on discounting with a discount factor  $\gamma \in [0, 1)$ :

- Earlier rewards are more important. A unit reward at present will worth  $\gamma$  in the next slot.
- Problems with random horizon  $N$  and absorbing states



## Reward Function: Some Comments

- **Bounded Rewards Assumption:** We assume

$$R_{\max} := \sup_{s,a} |\mathbb{E}_{r \sim R(s,a)}[r]| < \infty$$

- For simplicity, we assume *deterministic rewards*
  - Hence,  $r \sim R(s,a)$  means  $r = R(s,a)$ .
  - Hence, we may use  $r(s,a)$  and  $R(s,a)$  interchangeably, but tend to keep  $r(s,a)$  for generality.
  - The results in this lecture will hold for stochastic rewards under mild assumptions (and often by replacing  $R(s,a)$  or  $r(s,a)$  with its mean).

**This lecture:** We consider deterministic and bounded rewards.



# Value Function



## Recap: Policy

When interacting with an MDP, actions are taken according to some **policy**:

|                   | deterministic                                                      | randomized                                                                 |
|-------------------|--------------------------------------------------------------------|----------------------------------------------------------------------------|
| stationary        | $\pi : \mathcal{S} \rightarrow \mathcal{A}, \quad \Pi^{\text{SD}}$ | $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A}), \quad \Pi^{\text{SD}}$ |
| history-dependent | $\pi : \mathcal{H} \rightarrow \mathcal{A}, \quad \Pi^{\text{SD}}$ | $\pi : \mathcal{H} \rightarrow \Delta(\mathcal{A}), \quad \Pi^{\text{SD}}$ |

- $\Delta(\mathcal{A})$  denotes the simplex of probability distributions over  $\mathcal{A}$ .
- $\mathcal{H}$  the set of all possible histories (trajectories).

For  $\pi \in \Pi^{\text{SR}}$ , we write  $a \sim \pi(\cdot|s)$  or  $a \sim \pi(s)$ . Also, given  $f : \mathcal{A}_s \rightarrow \mathbb{R}$ ,

$$\mathbb{E}_{a \sim \pi(s)}[f(a)] = \sum_{a \in \mathcal{A}_s} f(a) \pi(a|s)$$



## Value Function

The **value function** of policy  $\pi$  (or simply, **value of  $\pi$** ) is a mapping  $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$  defined as

$$V^\pi(s) := \mathbb{E}^\pi \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \middle| s_1 = s \right].$$

where  $\mathbb{E}^\pi$  indicates expectation over trajectories generated by  $\pi$ .

- Intuitively,  $V^\pi(s)$  measures the sum of future discounted rewards (in expectation) when the agent starts in  $s$  and follows  $\pi$ .
- A rough upper bound:

$$|V^\pi(s)| \leq \frac{R_{\max}}{1 - \gamma}, \quad \forall s \in \mathcal{S}$$





# Value Function

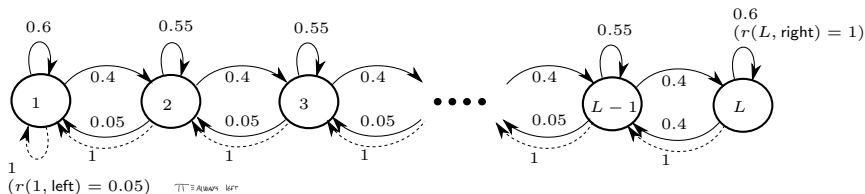
COMPUTING ALL POSSIBLE TRAJECTORIES MAY BE TOO COMPLEX, BUT WE CAN DO SIMULATIONS

The **value function** of policy  $\pi$  (or simply, **value of  $\pi$** ) is a mapping  $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$  defined as

$$V^\pi(s) := \mathbb{E}^\pi \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \mid s_1 = s \right].$$

where  $\mathbb{E}^\pi$  indicates expectation over trajectories generated by  $\pi$ .

**Example:** Value of  $\pi = \text{'always left'}$ ?



$\pi = \text{always left}$

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

$$V^\pi = \begin{bmatrix} V^\pi(1) \\ \vdots \\ V^\pi(L) \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix} + \gamma \begin{bmatrix} 0.05 & 0.4 & 0.55 & \dots & 0.4 & 0.6 \\ 0.05 & 0.4 & 0.55 & \dots & 0.4 & 0.6 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.05 & 0.4 & 0.55 & \dots & 0.4 & 0.6 \end{bmatrix} \begin{bmatrix} V^\pi(1) \\ \vdots \\ V^\pi(L) \end{bmatrix}$$

Block matrix equation



## Value Function



The **value function** of policy  $\pi$  (or simply, **value of  $\pi$** ) is a mapping  $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$  defined as

$$V^\pi(s) := \mathbb{E}^\pi \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \mid s_1 = s \right].$$

where  $\mathbb{E}^\pi$  indicates expectation over trajectories generated by  $\pi$ .

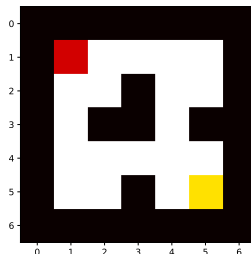
We may be interested in  $V^\pi(s_{\text{init}})$ .

### Example: 4-room Grid-World

- $s_{\text{init}}$  is .
- Terminal state .
- Our interest is to estimate  $V^\pi(\text{red square})$

ABSORBING STATE

IN THIS CASE WE DON'T  
CARE ABOUT ALL THE POSSIBLE  
INITIAL STATES



## Action-Value Function

The **action-value function of policy  $\pi$**  (or simply, **Q-value of  $\pi$** ) is a mapping  $Q^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  defined as (Under the bounded reward assumption)

Q LEARNING IS  
RELATED TO  
APPROXIMATING  
THIS  $\longrightarrow Q^\pi(s, a) := \mathbb{E}^\pi \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \mid s_1 = s, a_1 = a \right].$

- Intuitively,  $Q^\pi(s, a)$  measures the sum of future discounted rewards (in expectation) when the agent starts in  $s$  and takes action  $a$  in the first step (possibly  $a \neq \pi(s)$ ), and then follows  $\pi$  afterwards.
- A rough upper bound:

$$|Q^\pi(s, a)| \leq \frac{R_{\max}}{1 - \gamma}, \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

- For all  $s \in \mathcal{S}$ ,  $Q^\pi(s, \pi(s)) = V^\pi(s)$ .

$$Q^\pi = \begin{bmatrix} \text{LEFT} & \text{RIGHT} \\ V^\pi(s_1) & V^\pi(s_1) \\ V^\pi(s_2) & V^\pi(s_2) \\ \vdots & \vdots \\ V^\pi(s_L) & V^\pi(s_L) \end{bmatrix}$$

POSSIBLE FIRST MOVE (orange arrow pointing to the columns)

POSSIBLE INITIAL STATES (cyan arrow pointing to the rows)



# Policy Evaluation



## Recap: Induced Markov Chains

- Every  $\pi \in \Pi^{\text{SR}}$  induces a **Markov chain** on  $M$ , with transition probability matrix  $P^\pi$  given by:

$$P_{s,s'}^\pi = \sum_{a \in \mathcal{A}_s} P(s'|s, a) \pi(a|s), \quad s, s' \in \mathcal{S}.$$

- Every  $\pi \in \Pi^{\text{SR}}$  induces a reward vector  $r^\pi \in \mathbb{R}^{\mathcal{S}}$  on  $M$  defined by:

$$r^\pi(s) = \sum_{a \in \mathcal{A}_s} R(s, a) \pi(a|s), \quad s \in \mathcal{S}.$$

- If  $\pi \in \Pi^{\text{SD}}$ , then  $P_{s,s'}^\pi = P(s'|s, \pi(s))$  and  $r^\pi(s) = R(s, \pi(s))$ .

Every policy  $\pi \in \Pi^{\text{SR}}$  induces a **Markov Reward Process (MRP)** on  $M$ , specified by  $r^\pi$  and  $P^\pi$ .



## Bellman Equation for $\pi$

FIRST WE DEFINED  $V^\pi$ , BUT IN THE THEORETICAL DEFINITION A  $\sum$  IS INCLUDED, SO WE CAN'T USE IT TO ACTUALLY COMPUTE  $V^\pi$ .  
BELLMAN EQUATION SOLVES THIS PROBLEM

$$\begin{aligned} V^\pi &= r^\pi + \gamma P^\pi V^\pi \Leftrightarrow V^\pi - \gamma P^\pi V^\pi = r^\pi \\ (I - \gamma P^\pi) V^\pi &= r^\pi \\ \Leftrightarrow V^\pi &= (I - \gamma P^\pi)^{-1} r^\pi \end{aligned}$$

## Theorem (Bellman Equation for $\pi$ )

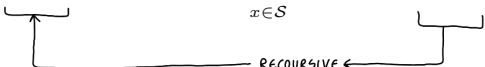
Let  $\pi \in \Pi^{SR}$ . For all  $s \in \mathcal{S}$ ,

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_{a \sim \pi(s)}[r(s, a)] + \gamma \mathbb{E}_{a \sim \pi(s)} \left[ \sum_{x \in \mathcal{S}} P(x|s, a) V^\pi(x) \right] \\ &= \sum_{a \in \mathcal{A}_s} \pi(a|s) r(s, a) + \gamma \sum_{a \in \mathcal{A}_s} \pi(a|s) \sum_{x \in \mathcal{S}} P(x|s, a) V^\pi(x) \end{aligned}$$

Equivalently,  $V^\pi = r^\pi + \gamma P^\pi V^\pi$ .

- These relations are called the **Bellman equation**.
- The theorem tells us that for  $\pi \in \Pi^{SR}$ ,  $V^\pi$  satisfies the Bellman equation.
- For a deterministic **policy  $\pi \in \Pi^{SD}$** , the Bellman equation becomes:

$$V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{x \in \mathcal{S}} P(x|s, \pi(s)) V^\pi(x), \quad s \in \mathcal{S}.$$


 RECURSIVE



## Bellman Operator for $\pi$

The **Bellman operator** associated to  $\pi \in \Pi^{\text{SR}}$  is a mapping  $\mathcal{T}^\pi : \mathbb{R}^S \rightarrow \mathbb{R}^S$ , such that for any function  $f : S \rightarrow \mathbb{R}$ ,

$$\mathcal{T}^\pi f := r^\pi + \gamma P^\pi f.$$

- Intuitively,  $\mathcal{T}^\pi$  is the value of  $\pi$  for the same one-stage problem.
- $\mathcal{T}^\pi$  applies to (or *operates on*) a function defined on  $S$  and returns another function defined on  $S$ .
- The **Bellman equation**  $V^\pi = r^\pi + \gamma P^\pi V^\pi$  reads

$$V^\pi = \mathcal{T}^\pi V^\pi$$

In other words,  $V^\pi$  is the *unique* fixed-point of the operator  $\mathcal{T}^\pi$ .



# Bellman Equation for $\pi$

We prove the theorem for  $\pi \in \Pi^{\text{SD}}$ . (See Lecture Notes for  $\pi \in \Pi^{\text{SR}}$ .)

**Proof.** Let  $\pi \in \Pi^{\text{SD}}$  and  $s \in \mathcal{S}$ . We have

$$\begin{aligned}
 V^\pi(s) &= \mathbb{E}^\pi \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, \pi(s_t)) \middle| s_1 = s \right] \\
 &= r(s, \pi(s)) + \mathbb{E}^\pi \left[ \sum_{t=2}^{\infty} \gamma^{t-1} r(s_t, \pi(s_t)) \middle| s_1 = s \right] \\
 &= r(s, \pi(s)) + \gamma \sum_{x \in \mathcal{S}} \mathbb{P}(s_2 = x | s_1 = s, a_1 = \pi(s_1)) \underbrace{\mathbb{E}^\pi \left[ \sum_{t=2}^{\infty} \gamma^{t-2} r(s_t, \pi(s_t)) \middle| s_2 = x \right]}_{= V^\pi(x)} \\
 &= r(s, \pi(s)) + \gamma \sum_{x \in \mathcal{S}} \mathbb{P}(s_2 = x | s_1 = s, a_1 = \pi(s_1)) V^\pi(x) \\
 &= r(s, \pi(s)) + \gamma \sum_{x \in \mathcal{S}} P(x | s, \pi(s)) V^\pi(x).
 \end{aligned}$$





## Policy Evaluation

$$\begin{aligned}
 x &= f(x) & g(x) &= x e^x - \sqrt{x+1} = 0 \\
 x &= \sqrt{x+1} \cdot e^{-x} \\
 x &= f(x) \\
 &\uparrow \\
 &(\text{??}) \\
 &\text{I PROBABLY} \\
 &\text{WROTE IT WRONG}
 \end{aligned}$$

### Policy Evaluation: Computing $V^\pi$ for a given $\pi$

- **Direct Computation:** Using Bellman equation,

$$V^\pi = r^\pi + \gamma P^\pi V^\pi \implies I - \gamma P^\pi \text{ is invertible} \quad V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$$

- **Iterative Policy Evaluation:** Using  $V^\pi = \mathcal{T}^\pi V^\pi$ , the sequence

$$V_{n+1} = \mathcal{T}^\pi V_n = \underbrace{\mathcal{T}^\pi \cdots \mathcal{T}^\pi}_{n+1 \text{ times}} V_0$$

$$\begin{aligned}
 &V_0 \\
 &V_1 = \mathcal{T}^\pi V_0 \\
 &V_2 = \mathcal{T}^\pi \mathcal{T}^\pi V_0 \\
 &\vdots
 \end{aligned}$$

converges to  $V^\pi$  starting from any  $V_0$ .

- **Monte-Carlo Method:** Generate a number of trajectories of  $\pi$  and use the sample mean as an estimator to  $V^\pi$ .



So far:

- We defined policies and the value function.
- We characterized the value of stationary policies (via Bellman equations and operator).
- We developed ways to compute the value of a *fixed* stationary policy.

*How to find an optimal strategy/policy? Alternatively, how to find policies with good values? (WITHOUT HAVING TO COMPUTE  $V^\pi$  FOR ALL OF THEM)*



# Optimization in Discounted MDPs: Optimal Policy and Value



## Optimal Value and Policy

Solving a discounted MDP  $M$  amounts to solving the following optimization problem:

$$V^*(s) = \sup_{\pi \in \Pi^{\text{HR}}} V^\pi(s), \quad \forall s \in \mathcal{S}.$$

- (i)  $V^* : \mathcal{S} \rightarrow \mathbb{R}$  is called the **optimal value function**.
- (ii) If there exists  $\pi^*$  such that  $V^{\pi^*}(s) = V^*(s)$  for all  $s \in \mathcal{S}$ , then  $\pi^*$  is called an **optimal policy**.
- (iii)  $\pi$  is  **$\varepsilon$ -optimal** for  $\varepsilon > 0$  if

$$V^\pi(s) \geq V^*(s) - \varepsilon, \quad \forall s \in \mathcal{S}$$



# Bellman Optimality Equation

## Theorem

$V^*$  satisfies the optimal Bellman equation:

$$V^*(s) = \max_{a \in \mathcal{A}_s} \left( r(s, a) + \gamma \sum_{x \in \mathcal{S}} P(x|s, a) V^*(x) \right), \quad s \in \mathcal{S}$$

↑ THERE IS NO POLICY HERE!

The optimal Bellman operator is a mapping  $\mathcal{T} : \mathbb{R}^{\mathcal{S}} \rightarrow \mathbb{R}^{\mathcal{S}}$ , such that for any function  $f : \mathcal{S} \rightarrow \mathbb{R}$ ,

$$(\mathcal{T}f)(s) := \max_{a \in \mathcal{A}_s} \left( r(s, a) + \gamma \sum_{x \in \mathcal{S}} P(x|s, a) f(x) \right), \quad s \in \mathcal{S}$$

- $V^*$  satisfies  $\mathcal{T}V^* = V^*$ .
- We can define  $\mathcal{T}$  and optimal Bellman equation for the optimal Q function (next lecture).



## Optimality Theorems

### Theorem

*Suppose the state space  $\mathcal{S}$  is finite. Then there exists a policy  $\pi^* \in \Pi^{SD}$ .*

- Thus, when seeking  $\pi^*$  in a discounted MDP with a finite  $\mathcal{S}$ , we can restrict our attention to  $\Pi^{SD}$ .
- In other words, for finite  $\mathcal{S}$ ,

$$\sup_{\pi \in \Pi^{HR}} V^\pi = \sup_{\pi \in \Pi^{SD}} V^\pi = \max_{\pi \in \Pi^{SD}} V^\pi$$



# Optimality Theorems

A fundamental result in the theory of discounted MDPs:

## Theorem

*A stationary deterministic policy  $\pi$  is optimal if and only if*

$$\mathcal{T}^\pi V^* = \mathcal{T}V^*$$

*Equivalently,  $\pi$  is optimal if and only if it attains the maximum in the Bellman optimality equations: For all  $s \in \mathcal{S}$ ,*

$$\pi(s) \in \arg \max_{a \in \mathcal{A}_s} \left( r(s, a) + \sum_{x \in \mathcal{S}} P(x|s, a) V^*(x) \right).$$



So far:

- We defined policies and the value function.
- We characterized the value of stationary policies (via Bellman equations and operator).
- We developed ways to compute the value of a *fixed* stationary policy.
- We defined the notion of optimality and showed that *there exists*  $\pi^* \in \Pi^{\text{SD}}$  when  $\mathcal{S}$  is finite.
- We characterized the optimal value function  $V^*$  (via *optimal Bellman equation*).

*How to actually compute  $\pi^*$ ?*





# Algorithms for Solving Discounted MDPs



# Major Solution Methods

Three major classes of algorithms for solving discounted MDPs:

- Value Iteration  $\rightarrow$  GIVES A NEAR OPTIMAL POLICY  $\in \Pi^{SD}$
- Policy Iteration  $\rightarrow$  GIVES AN OPTIMAL POLICY  $\in \Pi^{SD}$
- Linear Programming



THEY ARE NOT THE ONLY OPTIMAL  
POLICIES, THERE MAY BE OTHER MORE  
COMPLEX POLICIES, MAYBE ALSO IN  
 $\Pi^{SR}$  ETC...



# Value Iteration

## Value Iteration (VI)

- The most well-known, and perhaps the simplest, algorithm for solving discounted MDPs
- Around since the early days of MDPs
- Also known as successive approximation, backward induction, etc.

**Idea:** The optimal Bellman operator  $\mathcal{T}$  is *contracting*. Iterate  $\mathcal{T}$  until convergence:

$$V_{n+1} = \mathcal{T}V_n, \quad n = 0, 1, 2, \dots$$

Indeed, VI is an algorithm for approximating the fixed point of  $\mathcal{T}$ .



## Value Iteration (VI)

**input:**  $\varepsilon$

- **initialization:** Select a value function  $V_0 \in \mathbb{R}^S$ ,  $V_1 = R_{\max}/(1 - \gamma)\mathbf{1}$ , and set  $n = 0$
- **while** ( $\|V_{n+1} - V_n\|_\infty \geq \frac{\varepsilon(1-\gamma)}{2\gamma}$ )
  - (i) **Update**, for each  $s \in S$ ,

$$V_{n+1}(s) = \max_{a \in \mathcal{A}_s} \left( r(s, a) + \gamma \sum_{x \in S} P(x|s, a) V_n(x) \right)$$

- (ii) **Increment**  $n$ .

**output:**

$$\pi^{\text{VI}}(s) \in \arg \max_{a \in \mathcal{A}_s} \left( r(s, a) + \gamma \sum_{x \in S} P(x|s, a) V_n(x) \right), \quad s \in S$$



# Why VI works?

*Why does VI work?*

⇒ Because of contraction properties of Bellman operators.



# Contraction Mapping

An operator (or mapping)  $\mathcal{L} : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is called a  $\kappa$ -contraction mapping (with respect to  $\|\cdot\|$ ) if there exists  $\kappa \in [0, 1)$  such that for all  $v, v' \in \mathbb{R}^n$ ,

$$\|\mathcal{L}v - \mathcal{L}v'\| \leq \kappa \|v - v'\|.$$

↳ LIPSCHITZ CONSTANT

## Theorem (Banach Fixed-Point Theorem)

Suppose  $\mathcal{L}$  is a contraction mapping. Then

- (i) there exists a unique  $v^* \in \mathbb{R}^n$  such that  $\mathcal{L}v^* = v^*$ ;
- (ii) for any  $v_0 \in \mathbb{R}^n$ , the sequence  $(v_n)_{n \geq 0}$  with  $v_{n+1} = \mathcal{L}v_n = \mathcal{L}^{n+1}v_0$  for  $n \geq 0$  converges to  $v^*$ .



# $\mathcal{T}^\pi$ and $\mathcal{T}$ Are Contraction Mapping

## Lemma

For any  $v, v' \in \mathbb{R}^S$ , and any  $\pi$ ,

$$\|\mathcal{T}^\pi v - \mathcal{T}^\pi v'\|_\infty \leq \gamma \|v - v'\|_\infty,$$

$$\|\mathcal{T}v - \mathcal{T}v'\|_\infty \leq \gamma \|v - v'\|_\infty.$$

Hence,  $\mathcal{T}^\pi$  and  $\mathcal{T}$  are  $\gamma$ -contraction mappings w.r.t.  $\|\cdot\|_\infty$ .

**Proof.** First statement is easy to prove. For the second, we have:

$$\begin{aligned} & \|\mathcal{T}v - \mathcal{T}v'\|_\infty \\ &= \max_s \left| \max_{a \in \mathcal{A}_s} \left( r(s, a) + \gamma \sum_j P(j|s, a) v(j) \right) - \max_{a \in \mathcal{A}_s} \left( r(s, a) + \gamma \sum_j P(j|s, a) v'(j) \right) \right| \\ &\leq \max_s \max_{a \in \mathcal{A}_s} \left| \gamma \sum_j P(j|s, a) (v(j) - v'(j)) \right| \\ &\quad \text{(Using inequality } |\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|) \\ &\leq \gamma \max_s \max_{a \in \mathcal{A}_s} \max_j |v(j) - v'(j)| \sum_j P(j|s, a) = \gamma \|v - v'\|_\infty \end{aligned}$$



## VI: Convergence

VI is a *globally convergent* method for finding an  $\varepsilon$ -optimal policy. Formally:

### Theorem

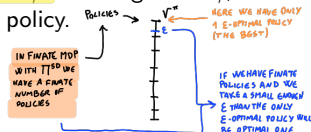
Let  $(V_n)_{n \geq 0}$  a sequence of value functions generated by VI with some  $\varepsilon > 0$  starting from an arbitrary initial point  $V_0 \in \mathbb{R}^S$ . Then,

- (i)  $V_n$  converges to  $V^*$  in norm;
- (ii) the algorithm stops after finitely many iterations;
- (iii)  $\pi^{VI}$  is  $\varepsilon$ -optimal;
- (iv) when convergence criterion is satisfied,  $\|V_{n+1} - V^*\|_\infty < \varepsilon/2$ .

- Each iteration of VI involves  $O(S^2A)$  arithmetic calculations.
- The *iteration complexity* of VI depends on both  $\varepsilon$  and  $\gamma$ . The larger the  $\gamma$ , the more iteration until the algorithm finds an  $\varepsilon$ -optimal policy.

$$\sqrt{s} \geq \sqrt{s} - \varepsilon, \quad \forall s \leftarrow \text{At worst is } \varepsilon\text{-OPTIMAL}$$

! Nice!





# Policy Iteration

## Policy Iteration (PI)

- A popular algorithm for solving discounted MDPs
- Around since early days of MDPs
- Like VI, it is an iterative algorithm but directly searches in the space of policies.

**Idea:** Starting from an initial policy, at each iterate  $n$ ,

- Find  $V^{\pi_n}$  (policy evaluation)
- Improve  $\pi_n$  to  $\pi_{n+1}$  using  $V^{\pi_n}$  (policy improvement)



# Policy Iteration (PI)

- **initialization:** Select  $\pi_0$  and  $\pi_1$  arbitrarily ( $\pi_0 \neq \pi_1$ ), and set  $n = 0$

- **while** ( $\pi_{n+1} \neq \pi_n$ )

- (i) **Policy Evaluation:** Find  $V_n$ , the value of  $\pi_n$  by solving

$$(I - \gamma P^{\pi_n})V_n = r^{\pi_n} \rightarrow \text{HERE WE CAN COMPUTE A LESS COMPLEX APPROXIMATION}$$

- (ii) **Policy Improvement:** Choose  $\pi_{n+1}$  such that

$$\pi_{n+1}(s) \in \arg \max_{a \in \mathcal{A}_s} \left( r(s, a) + \gamma \sum_{x \in \mathcal{S}} P(x|s, a) V_n(x) \right)$$

and if possible, set  $\pi_{n+1} = \pi_n$ .

- (iii) **Increment  $n$ .**

- **output:**  $\pi^{\text{PI}} = \pi_n$



WATCH OUT FOR LOOPS! THERE CAN BE MORE OPTIMAL POLICIES AND THE ALGORITHM CAN END UP UPDATING BETWEEN THEM

→ SOLUTION: WE CAN DECIDE TO STOP AS SOON AS WE REACH A CERTAIN VALUE



# PI: Convergence

## Theorem

Suppose  $M$  has a finite state-action space. Then,

(i) *PI terminates in at most*

$$O\left(\max\left\{\frac{SA}{1-\gamma} \log \frac{1}{1-\gamma}, \frac{A^S}{S}\right\}\right) \text{ iterations};$$

(ii)  $\pi^{PI} = \pi^*$ .

IT'S A SUPER  
CONSERVATIVE BOUND  
(IN PRACTICE WE NEED  
MANY LESS ITERATIONS)

- Under PI,  $V_{n+1} \geq V_n$  for any  $n$ . Further, the number of policies is finite  $A^S$ .
- Each iteration in PI involves solving a linear system with  $S$  equations and  $S$  unknowns. Hence, per iteration complexity of PI is  $O(S^3 + S^2 A)$ .
- In practice, PI converges within, at most, a few tens of iterations.

MORE COMPLEX ITERATION  
THAN VI, BUT HERE WE  
USUALLY NEED LESS OF THEM

