

# CMPT 419/983: Theoretical Foundations of Reinforcement Learning

## Lecture 4

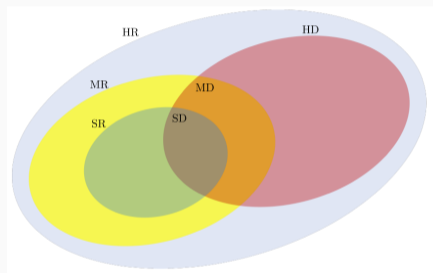
---

Sharan Vaswani

September 29, 2023

# Recap

- Given an MDP  $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, s_0)$ , interacting with  $M$  using a fixed policy  $\pi$  results in a stochastic process  $(S_0, A_0, S_1, \dots)$  over the state-action space and a corresponding reward process  $(R_0, R_1, \dots) = (r(S_0, A_0), r(S_1, A_1), \dots)$ .
- Objective:** Find policy  $\pi \in \Pi_{\text{HR}}$  that maximizes the value function  $v^\pi(s_0) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s_0]$ .
- For each  $s \in \mathcal{S}$ , for a given policy  $\pi = (\pi_0, \pi_1, \dots) \in \Pi_{\text{HR}}$ , there exists a policy  $\pi' = (\pi'_0, \pi'_1, \dots) \in \Pi_{\text{MR}}$  with the same value, conditioned on  $S_0 = s_0$ .
- Hence, considering the class  $\Pi_{\text{MR}}$  is sufficient when searching for the optimal policy.



# Infinite-horizon Discounted Setting

**Claim:** For  $\pi \in \Pi_{MR}$ , if we define

$$\mathbf{r}_\pi \in \mathbb{R}^S \quad \text{s.t.} \quad \mathbf{r}_\pi(s) := \sum_{a \in \mathcal{A}} r(s, a) \pi[a|s],$$

$$\mathbf{P}_\pi \in \mathbb{R}^{S \times S} \quad \text{s.t.} \quad \mathbf{P}_\pi[s, s'] = \Pr^\pi(s \rightarrow s') := \sum_{a \in \mathcal{A}} \Pr[s'|s, a] \pi(a|s),$$

then,  $v^\pi \in \mathbb{R}^S$  can be expressed as:

$$v^\pi = \sum_{t=0}^{\infty} \gamma^t \left[ \prod_{j=0}^{t-1} \mathbf{P}_{\pi_j} \right] \mathbf{r}_{\pi_t}.$$

Furthermore, for a policy  $\pi \in \Pi_{SR}$ ,  $v^\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^\pi$ . Examining each component,

$$v^\pi(s) = \mathbf{r}_\pi(s) + \gamma \sum_{s'} \mathbf{P}_\pi[s, s'] v^\pi(s') = \sum_{a \in \mathcal{A}} r(s, a) \pi[a|s] + \gamma \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{P}[s'|s, a] \pi[a|s] v^\pi(s')$$

This is the **Bellman equation** for a fixed policy  $\pi \in \Pi_{SR}$ .

# Infinite-horizon Discounted Setting

*Proof:* Starting from the definition of  $v^\pi(s_0)$ ,

$$v^\pi(s_0) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s_0 \right] = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \Pr[S_t = s, A_t = a \mid S_0 = s_0]$$

Let us evaluate the first three terms in this sum,

**For  $t = 0$ :**  $\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \Pr[S_0 = s, A_0 = a \mid S_0 = s_0] = \sum_{a \in \mathcal{A}} r(s_0, a) \pi_0(a \mid s_0) = \mathbf{r}_{\pi_0}(s_0)$

**For  $t = 1$ :**  $\gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \Pr[A_1 = a \mid S_1 = s, S_0 = s_0] \Pr[S_1 = s \mid S_0 = s_0]$   
 $= \gamma \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_1}(s) \Pr[S_1 = s \mid S_0 = s_0] = \gamma \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_1}(s) \sum_{a \in \mathcal{A}} \mathcal{P}[s \mid s_0, a] \pi_0(a \mid s_0) = \gamma \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_1}(s) \mathbf{P}_{\pi_0}[s_0, s]$

**For  $t = 2$ :**  $\gamma^2 \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_2}(s) \Pr[S_2 = s \mid S_0 = s_0] = \gamma^2 \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_2}(s) \sum_{s_1 \in \mathcal{S}} \mathbf{P}_{\pi_1}[s_1, s] \mathbf{P}_{\pi_0}[s_0, s_1]$

**For a general  $t$ :**  $\gamma^t \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_t}(s) \sum_{s_{t-1} \in \mathcal{S}} \dots \sum_{s_1 \in \mathcal{S}} \mathbf{P}_{\pi_{t-1}}[s_{t-1}, s] \mathbf{P}_{\pi_{t-2}}[s_{t-2}, s_{t-1}] \dots \mathbf{P}_{\pi_0}[s_0, s_1]$

# Infinite-horizon Discounted Setting

Recall that,  $v^\pi(s_0) = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \Pr[S_t = s, A_t = a | S_0 = s_0]$ , and that term  $t$  in the above sum is equal to  $\gamma^t \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_t}(s) \sum_{s_{t-1} \in \mathcal{S}} \cdots \sum_{s_1 \in \mathcal{S}} \mathbf{P}_{\pi_{t-1}}[s_{t-1}, s] \mathbf{P}_{\pi_{t-2}}[s_{t-2}, s_{t-1}] \cdots \mathbf{P}_{\pi_0}[s_0, s_1]$ . Hence,

$$\begin{aligned} v^\pi(s_0) &= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_t}(s) \sum_{s_{t-1} \in \mathcal{S}} \cdots \sum_{s_1 \in \mathcal{S}} \mathbf{P}_{\pi_{t-1}}[s_{t-1}, s] \mathbf{P}_{\pi_{t-2}}[s_{t-2}, s_{t-1}] \cdots \mathbf{P}_{\pi_0}[s_0, s_1] \\ \implies v^\pi &= \sum_{t=0}^{\infty} \gamma^t \left[ \prod_{j=0}^{t-1} \mathbf{P}_{\pi_j} \right] \mathbf{r}_{\pi_t} \quad (v^\pi(s_0) \text{ is the } s_0 \text{ component of the vector } v^\pi) \end{aligned}$$

For a policy  $\pi \in \Pi_{\text{SR}}$ ,  $\mathbf{P}_{\pi_t} = \mathbf{P}_\pi$  and  $\mathbf{r}_{\pi_t} = \mathbf{r}_\pi$  for all  $t$ . Hence,

$$\begin{aligned} v^\pi &= \sum_{t=0}^{\infty} \gamma^t [\mathbf{P}_\pi]^t \mathbf{r}_\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{r}_\pi + \gamma^2 [\mathbf{P}_\pi]^2 \mathbf{r}_\pi + \dots \\ &= \mathbf{r}_\pi + \gamma \mathbf{P}_\pi [\mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{r}_\pi + \gamma^2 [\mathbf{P}_\pi]^2 \mathbf{r}_\pi + \dots] = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^\pi \\ \implies v^\pi &= \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^\pi \quad \square \end{aligned}$$

# Infinite-horizon Discounted Setting

For  $\pi \in \Pi_{SR}$ , we have seen that  $v^\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^\pi$ . This corresponds to a system of linear equations, and can be solved in closed form. Since  $\gamma < 1$ , and  $\mathbf{P}_\pi$  is a stochastic matrix (i.e. its elements correspond to probabilities, and rows and columns add up to one), the eigenvalues of  $I_S - \gamma \mathbf{P}_\pi$  are strictly positive and hence it is invertible.

$$v^\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^\pi \implies (I_S - \gamma \mathbf{P}_\pi) v^\pi = \mathbf{r}_\pi \implies v^\pi = (I_S - \gamma \mathbf{P}_\pi)^{-1} \mathbf{r}_\pi.$$

- By the Neumann series,  $(I - A)^{-1} = \sum_{t=0}^{\infty} A^t$ . Hence,  $(I_S - \gamma \mathbf{P}_\pi)^{-1} \mathbf{r}_\pi = \sum_{t=0}^{\infty} \gamma^t [\mathbf{P}_\pi]^t \mathbf{r}_\pi$  which recovers the expression for  $v^\pi$  from the previous slide.
- **Q:** For a vector  $x \geq 0$ , prove that  $(I_S - \gamma \mathbf{P}_\pi)^{-1} x \geq x \geq 0$  **Ans:** Use the Neumann series
- **Q:** For vectors  $u \geq v$ , prove that  $(I_S - \gamma \mathbf{P}_\pi)^{-1} u \geq (I_S - \gamma \mathbf{P}_\pi)^{-1} v$  **Ans:**  $x = u - v$  above.

**Bellman policy evaluation operator for  $\pi$ :**  $\mathcal{T}_\pi : \mathbb{R}^S \rightarrow \mathbb{R}^S$  s.t. for vector  $u \in \mathbb{R}^S$   
 $\mathcal{T}_\pi u = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi u$  and  $(\mathcal{T}_\pi u)(s) = \mathbf{r}_\pi(s) + \gamma \sum_{s'} \mathbf{P}_\pi[s, s'] u(s')$ .

# Bellman Optimality Operator

Define the **Bellman optimality operator**  $\mathcal{T} : \mathbb{R}^S \rightarrow \mathbb{R}^S$ . For a vector  $u \in \mathbb{R}^S$ ,

$$(\mathcal{T}u)(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) u(s') \right\}$$

Consider  $w := \max_{\pi \in \Pi_{\text{SD}}} \{ \mathbf{r}_\pi + \gamma \mathbf{P}_\pi u \}$ ,

$$\begin{aligned} w(s) &= \max_{\pi \in \Pi_{\text{SD}}} \left\{ \mathbf{r}_\pi(s) + \gamma \sum_{s'} \mathbf{P}_\pi[s, s'] u(s') \right\} \\ &= \max_{\substack{\pi(\cdot|s) \\ \exists a^* \text{ s.t. } \pi(a^*|s)=1}} \left\{ \sum_a \pi(a|s) \left[ r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) u(s') \right] \right\} \\ &\quad \text{(Optimization over degenerate distributions)} \\ &= \max_a \left\{ r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) u(s') \right\} = (\mathcal{T}u)(s) \end{aligned}$$

$$\implies \mathcal{T}u = \max_{\pi \in \Pi_{\text{SD}}} \{ \mathbf{r}_\pi + \gamma \mathbf{P}_\pi u \}$$

# Bellman Optimality Operator

**Claim:**  $\mathcal{T}$  is a contraction mapping with modulus  $\gamma$ , i.e. for any 2 vectors  $u, w \in \mathbb{R}^S$

$$\|\mathcal{T}u - \mathcal{T}w\|_\infty \leq \gamma \|u - w\|_\infty.$$

*Proof:* For a fixed  $s$ , without loss of generality, consider the case when  $(\mathcal{T}w)(s) \geq (\mathcal{T}u)(s)$ . By the definition of  $\mathcal{T}$ , if  $a^*(s) = \arg \max \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a)w(s')\}$ , then,

$$(\mathcal{T}w)(s) = r(s, a^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s))w(s')$$

$$r(s, a^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s))u(s') \leq \max_a \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a)u(s')\} = (\mathcal{T}u)(s)$$

$$\implies (\mathcal{T}w)(s) - (\mathcal{T}u)(s) \leq \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s)) [w(s') - u(s')]$$

$$\leq \gamma \|\mathcal{P}(\cdot|s, a^*(s))\|_1 \|w - u\|_\infty = \gamma \|w - u\|_\infty$$

Similarly,  $(\mathcal{T}w)(s) - (\mathcal{T}u)(s) \leq \gamma \|w - u\|_\infty$ . Since this result is true for an arbitrary  $s$ ,

$$\|\mathcal{T}u - \mathcal{T}w\|_\infty \leq \gamma \|u - w\|_\infty \quad \square$$



# Banach's Fixed Point Theorem

**Fact:** Under certain technical assumptions, if  $L$  is a contraction mapping, then,

- There exists a unique fixed point  $u^*$  such that  $Lu^* = u^*$ .
- For any vector  $u_0$ ,  $u_{n+1} = Lu_n = L^{n+1}u_0$  converges to  $u^*$  i.e.  $\|u_n - u^*\|_\infty \rightarrow 0$  as  $n \rightarrow \infty$ .

Since the Bellman optimality operator,  $\mathcal{T}$  is a contraction mapping, using Banach's Fixed Point Theorem above, there exists a fixed point  $u^* \in \mathbb{R}^S$  s.t.  $\mathcal{T}u^* = u^*$ .

**Claim:** For  $u_0 \in \mathbb{R}^S$ ,  $\|u^* - \mathcal{T}^n u_0\|_\infty \leq \gamma^n \|u^* - u_0\|_\infty$  i.e.  $u_n := \mathcal{T}^n u_0$  converges to  $u^*$  at a linear rate.

**Q: Proof?** **Ans:** For any  $s \leq n$ ,

$$\begin{aligned}\|u^* - u_s\|_\infty &= \|\mathcal{T}u^* - \mathcal{T}u_{s-1}\|_\infty \leq \gamma \|u^* - u_{s-1}\|_\infty \\ \implies \|u^* - u_n\|_\infty &\leq \gamma \|u^* - u_{n-1}\|_\infty \leq \gamma^n \|u^* - u_0\|_\infty \quad \square\end{aligned}$$

Similarly,  $\mathcal{T}_\pi$  is a  $\gamma$ -contraction, and converges to a unique fixed point equal to  $v^\pi$  at a linear rate. Prove in Assignment 2!

# Fundamental Theorem

**Claim:** There exists a policy  $\pi^* \in \Pi_{\text{SD}}$  s.t.  $v^{\pi^*}(s) = \max_{\pi \in \Pi_{\text{HR}}} v^{\pi}(s)$  for all  $s \in \mathcal{S}$ .

- Hence, for MDPs, it is sufficient to only consider the class of stationary, deterministic policies in order to compute the optimal policy.

*Proof:* We know the following:

(a) From Slide 19 in Lecture 3,  $\max_{\pi \in \Pi_{\text{HR}}} v^{\pi}(s) = \max_{\pi \in \Pi_{\text{MR}}} v^{\pi}(s)$ .

(b) If  $v^*$  is the fixed point of  $\mathcal{T}$  and  $\pi^* \in \Pi_{\text{SD}}$  is the *greedy* policy s.t.

$\pi^*(s) = \arg \max_a \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) v^*(s')\}$ , then,

$$v^* = \mathcal{T}v^* = \max_{\pi \in \Pi_{\text{SD}}} \{\mathbf{r}_{\pi} + \gamma \mathbf{P}_{\pi} v^*\} = \mathcal{T}_{\pi^*} v^* = \mathbf{r}_{\pi^*} + \gamma \mathbf{P}_{\pi^*} v^*$$

(c)  $\max_{\pi \in \Pi_{\text{SD}}} \{\mathbf{r}_{\pi} + \gamma \mathbf{P}_{\pi} v^*\} = \max_{\pi \in \Pi_{\text{SR}}} \{\mathbf{r}_{\pi} + \gamma \mathbf{P}_{\pi} v^*\}$  i.e. randomized policies cannot increase the value. (Prove in Assignment 2!)

We will prove that for a  $v$  s.t.  $v = \mathcal{T}v$ ,  $v = \max_{\pi \in \Pi_{\text{HR}}} v^{\pi}$ . Together with (b), this implies that  $v^* = \max_{\pi \in \Pi_{\text{HR}}} v^{\pi}$  and that this value function corresponds to the policy  $\pi^* \in \Pi_{\text{SD}}$ .

# Fundamental Theorem

We will now prove that:

(i) If  $v \geq \mathcal{T}v$ , then  $v \geq \max_{\pi \in \Pi_{\text{HR}}} v^\pi$ .

(ii) If  $v \leq \mathcal{T}v$ , then  $v \leq \max_{\pi \in \Pi_{\text{HR}}} v^\pi$ .

Hence, if  $v = \mathcal{T}v$ , then  $v = \max_{\pi \in \Pi_{\text{HR}}} v^\pi$ .

Let us first prove (i). Define an arbitrary  $\pi' := \{\pi'_1, \pi'_2, \dots\} \in \Pi_{\text{MR}}$ . For an arbitrary  $i$ , define  $\pi_i := \{\pi'_i, \pi'_i, \dots\} \in \Pi_{\text{SR}}$ .

$$v \geq \mathcal{T}v = \max_{\pi \in \Pi_{\text{SD}}} \{\mathbf{r}_\pi + \gamma \mathbf{P}_\pi v\} = \max_{\pi \in \Pi_{\text{SR}}} \{\mathbf{r}_\pi + \gamma \mathbf{P}_\pi v\} \geq \mathbf{r}_{\pi_i} + \gamma \mathbf{P}_{\pi_i} v \quad (\text{Using (c)})$$

$$\implies v \geq \mathbf{r}_{\pi_0} + \gamma \mathbf{P}_{\pi_0} v \geq \mathbf{r}_{\pi_0} + \gamma \mathbf{P}_{\pi_0} [\mathbf{r}_{\pi_1} + \gamma \mathbf{P}_{\pi_1} v] \implies v \geq \sum_{t=0}^{\infty} \gamma^t \left[ \prod_{j=0}^{t-1} \mathbf{P}_{\pi_j} \right] \mathbf{r}_{\pi_t} \quad (\text{Recurring})$$

$$\implies v \geq v^{\pi'} \implies v \geq \max_{\pi \in \Pi_{\text{MR}}} v^\pi = \max_{\pi \in \Pi_{\text{HR}}} v^\pi \quad (\text{Using def of } v^{\pi'} \text{ for } \pi' \in \Pi_{\text{MR}}, \text{ and then (a)})$$

# Fundamental Theorem

Now let us prove (ii): if  $v \leq \mathcal{T}v$ , then  $v \leq \max_{\pi \in \Pi_{\text{HR}}} v^\pi$ . For a specific  $\pi \in \Pi_{\text{SD}}$ ,

$$v \leq \mathcal{T}v = \mathcal{T}_\pi v = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v \leq \mathbf{r}_\pi + \gamma \mathbf{P}_\pi [\mathbf{r}_\pi + \gamma \mathbf{P}_\pi v] \implies v \leq \sum_{t=0}^{\infty} \gamma^t [\mathbf{P}_\pi]^t \mathbf{r}_\pi$$

(Recurring)

$$\implies v \leq v^\pi \leq \max_{\pi \in \Pi_{\text{SD}}} v^\pi \quad (\text{By def of } v^\pi \text{ for } \pi \in \Pi_{\text{SD}})$$

$$= \max_{\pi \in \Pi_{\text{SR}}} v^\pi \leq \max_{\pi \in \Pi_{\text{MR}}} v^\pi \implies v \leq \max_{\pi \in \Pi_{\text{HR}}} v^\pi \quad \square \quad (\text{Using (c) and then (a)})$$

The fundamental theorem immediately suggests a way to calculate  $\pi^*$ :

- Starting from an arbitrary vector  $v_0 \in \mathbb{R}^S$ , iterate  $v = \mathcal{T}v$  to converge to a fixed point  $v^*$ .
- Once we have computed  $v^*$ , compute the greedy policy in each state  $s \in \mathcal{S}$ :  
$$\pi^*(s) = \arg \max_a \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) v^*(s')\}.$$

This is value iteration!

# Value Iteration

# Value Iteration

---

## Algorithm Value Iteration

---

- 1: **Input:** MDP  $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho)$ ,  $v_0 = 0$ .
  - 2: **for**  $k = 1 \rightarrow K$  **do**
  - 3:    $\forall s \in \mathcal{S}$ ,  $v_k(s) = \max_{a \in \mathcal{A}} \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) v_{k-1}(s')\} = (\mathcal{T} v_{k-1})(s)$
  - 4: **end for**
  - 5:  $\forall s \in \mathcal{S}$ , return  $\hat{\pi}(s) = \arg \max_a \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) v_K(s')\}$
- 

**Q:** What is the computational complexity of VI? **Ans:**  $O(S^2AK)$

**Claim:** After  $K \geq \frac{\log(1/\epsilon(1-\gamma))}{1-\gamma}$  iterations, value iteration returns a  $v_K$  s.t.  $\|v_K - v^*\|_\infty \leq \epsilon$ .

*Proof:* By using the contraction property of  $\mathcal{T}$ ,

$$\|v_K - v^*\|_\infty \leq \gamma^K \|v_0 - v^*\|_\infty = \gamma^K \|v^*\|_\infty \leq \gamma^K \frac{1}{1-\gamma}$$

Setting  $K \geq \frac{\log(1/\epsilon(1-\gamma))}{1-\gamma} \geq \frac{\log(1/\epsilon(1-\gamma))}{\log(1/\gamma)}$  ensures that  $\|v_K - v^*\|_\infty \leq \epsilon$ . ( $\because 1 - \gamma \leq \log(1/\gamma)$ )

Recall that the greedy step w.r.t  $v_K$  can also be written as:  $\mathcal{T} v_K = \mathcal{T}_{\hat{\pi}} v_K$ .

# Value Iteration

- The previous result gives a bound on the quality of  $v_K$ .
- Since  $\hat{\pi}$  is the policy returned by VI, we want a bound on  $\|v^* - v^{\hat{\pi}}\|_{\infty}$ .
- We will prove a general result bounding the error for the greedy policy inferred from  $v$ .

**Claim:** For an arbitrary  $v \in \mathbb{R}^S$  if (i)  $\pi$  is the greedy policy w.r.t  $v$ , i.e.

$\pi(s) = \arg \max_a \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) v(s')\}$ , (ii)  $v^{\pi}$  is the value function corresponding to policy  $\pi$  i.e.  $v^{\pi} = \mathcal{T}_{\pi} v^{\pi} = \mathbf{r}_{\pi} + \gamma \mathbf{P}_{\pi} v^{\pi}$ , then,

$$v^{\pi} \geq v^* - \frac{2\gamma \|v - v^*\|_{\infty}}{1 - \gamma} \mathbf{1}$$

- Hence, the error in  $\|v - v^*\|_{\infty}$  “blows up” when inferring policy  $\pi$ .
- This result is sharp meaning that the constant  $\frac{2\gamma}{1-\gamma}$  cannot be improved.
- Using this result, we conclude that VI requires  $K \geq \frac{\log(2\gamma/\epsilon(1-\gamma)^2)}{1-\gamma}$  iterations to obtain a greedy policy  $\hat{\pi}$  s.t.  $v^* - v^{\hat{\pi}} \leq \epsilon \mathbf{1}$ .

# Value Iteration

- We have seen that VI requires  $O\left(\frac{S^2 A \log(1/\epsilon)}{1-\gamma}\right)$  operations to produce an  $\epsilon$ -optimal policy  $\pi$  that guarantees  $v^\pi \geq v^* - \epsilon \mathbf{1}$ .
- **Lower Bound:** For  $\epsilon \in [0, \gamma/(1-\gamma))$ , any algorithm guaranteed to produce  $\epsilon$ -optimal policies in an MDP with finite state-action spaces (with sizes  $S$  and  $A$  respectively) and bounded (in  $[0, 1]$ ) rewards requires  $\Omega(S^2 A)$  operations (no dependence on  $\epsilon$ ) (see Csaba's notes, Lecture 3 for details).
- Is our VI analysis loose or is the  $O(\log(1/\epsilon))$  dependence necessary?
- There exists a family of MDPs with deterministic transitions, three states, two actions and bounded (in  $[0, 1]$ ) rewards such that the worst-case iteration complexity of VI to find an *exactly* optimal policy is infinite. (see Csaba's notes, Lecture 4 for details).
- In the next class, we will study Policy Iteration (PI) which can converge to the optimal policy with finite operations.