CMPT 419/983: Theoretical Foundations of Reinforcement Learning

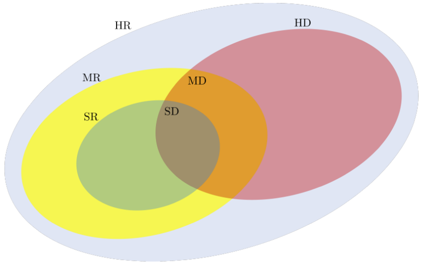
Lecture 4

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 π 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* $v^\pi(s_0) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s_0\right]$.
- 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 Π_{MR} is sufficient when searching for the optimal policy.



Infinite-horizon Discounted Setting

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

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

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

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 $\pi \in \Pi_{\text{SR}}$, $v^\pi = \mathbf{r}_\pi + \gamma\, \mathbf{P}_\pi\, v^\pi$. 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 $\pi \in \Pi_{\text{SR}}$.

Infinite-horizon Discounted Setting

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

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

Let us evaluate the first three terms in this sum,

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

For $t = 1$: $\gamma \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a) \Pr[A_1 = a | S_1 = s, S_0 = s_0] \Pr[S_1 = s | S_0 = s_0]$

$= \gamma \sum_{s\in\mathcal{S}} \mathbf{r}_{\pi_1}(s) \Pr[S_1 = s | S_0 = s_0] = \gamma \sum_{s\in\mathcal{S}} \mathbf{r}_{\pi_1}(s) \sum_{a\in\mathcal{A}} \mathcal{P}[s|s_0,a]\, \pi_0(a|s_0) = \gamma \sum_{s\in\mathcal{S}} \mathbf{r}_{\pi_1}(s)\, \mathbf{P}_{\pi_0}[s_0, s]$

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

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

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,

$$v^\pi(s_0) = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in \mathcal{S}} \mathbf{r}_{\pi_t}(s) \sum_{s_{t-1} \in \mathcal{S}} \ldots \sum_{s_1 \in \mathcal{S}} \mathbf{P}_{\pi_{t-1}}[s_{t-1}, s] \, \mathbf{P}_{\pi_{t-2}}[s_{t-2}, s_{t-1}] \cdots \mathbf{P}_{\pi_0}[s_0, s_1]$$

$$\implies v^\pi = \sum_{t=0}^{\infty} \gamma^t \left[\prod_{j=0}^{t-1} \mathbf{P}_{\pi_j} \right] \mathbf{r}_{\pi_t} \qquad (v^\pi(s_0) \text{ is the } s_0 \text{ component of the vector } v^\pi)$$

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

$$v^\pi = \sum_{t=0}^{\infty} \gamma^t \, [\mathbf{P}_\pi]^t \, \mathbf{r}_\pi = \mathbf{r}_\pi + \gamma \, \mathbf{P}_\pi \, \mathbf{r}_\pi + \gamma^2 \, [\mathbf{P}_\pi]^2 \, \mathbf{r}_\pi + \ldots$$

$$= \mathbf{r}_\pi + \gamma \, \mathbf{P}_\pi \, \left[\mathbf{r}_\pi + \gamma \, \mathbf{P}_\pi \, \mathbf{r}_\pi + \gamma^2 \, [\mathbf{P}_\pi]^2 \, \mathbf{r}_\pi + \ldots \right] = \mathbf{r}_\pi + \gamma \, \mathbf{P}_\pi \, v^\pi$$

$$\implies v^\pi = \mathbf{r}_\pi + \gamma \, \mathbf{P}_\pi \, v^\pi \quad \square$$

Infinite-horizon Discounted Setting

For $\pi \in \Pi_{\text{MR}}$, 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 sums 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^π from the previous slide.
- Q: For a vector $x \geq 0$, prove that $(I_S - \gamma \mathbf{P}_\pi)^{-1}\, x \geq x \geq 0$
- Q: For vectors $u \geq v$, prove that $(I_S - \gamma \mathbf{P}_\pi)^{-1}\, u \geq (I_S - \gamma \mathbf{P}_\pi)^{-1}\, v$

Bellman policy evaluation operator for π: $\mathcal{T}_\pi : \mathbb{R}^S \to \mathbb{R}^S$ 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** $\mathcal{T} : \mathbb{R}^S \to \mathbb{R}^S$. For a vector $u \in R^S$,

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

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

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

(Optimization over degenerate distributions)

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

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

Bellman Optimality Operator

Claim: $\mathcal{T}$ is a contraction mapping with modulus γ, i.e. for any 2 vectors $u, w \in \mathbb{R}^S$ $\left\|\mathcal{T}u - \mathcal{T}w\right\|_\infty \le \gamma \left\|u - w\right\|_\infty$.

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

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

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

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

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

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

$$\left\|\mathcal{T}u - \mathcal{T}w\right\|_\infty \le \gamma \left\|u - w\right\|_\infty \quad \square$$

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 $\|u_n - u^*\|_\infty \to 0$ as $n \to \infty$.

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?*

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

Fundamental Theorem

Claim: There exists a policy $\pi^* \in \Pi_{\text{SD}}$ s.t. $v^{\pi^*}(s) = \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}(s)$ 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, $\max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}(s) = \max_{\pi \in \Pi_{\mathbf{MR}}} v^{\pi}(s)$.

(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,

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

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

We will prove that for a v s.t. $v = \mathcal{T} v$, $v = \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$. Together with (b), this implies that $v^* = \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$ 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 $v \geq \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$.
(ii) If $v \leq \mathcal{T} v$, then $v \leq \max_{\pi \in \Pi_{\mathbf{HR}}} v^{\pi}$.

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

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

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

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

(Recursing)

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

Fundamental Theorem

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

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

$$\implies v \leq v^\pi \leq \max_{\pi \in \Pi_{\mathbf{SD}}} v^\pi \tag{By def of v^π for $\pi \in \Pi_{\text{SD}}$}$$

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

The fundamental theorem immediately suggests a way to calculate π^*:

- 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

Input: MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho)$, $v_0 = 0$.
for $k = 1 \to K$ **do**
  $\forall s \in \mathcal{S}$, $v_k(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_{k-1}(s') \right\} = (\mathcal{T} v_{k-1})(s)$
end for
$\forall s \in \mathcal{S}$, 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?

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 $\left\| v^* - v^{\hat{\pi}} \right\|_\infty$.
- 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) π 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^π is the value function corresponding to policy π i.e. $v^\pi = \mathcal{T}_\pi v^\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^\pi$, then,

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

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

Policy Error Bound

Claim: For an arbitrary $v \in \mathbb{R}^S$ if (i) π 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^π is the value function corresponding to policy π i.e. $v^\pi = \mathcal{T}_\pi v^\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^\pi$, then,

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

Proof: For the proof, we need the following properties of the $\mathcal{T}$ and $\mathcal{T}_\pi$ operators.

$$\mathcal{T} v^* = v^* \quad ; \quad \mathcal{T} v = \mathcal{T}_\pi v \quad ; \quad v^\pi = \mathcal{T}_\pi v^\pi$$

We will also need the following properties: for $u, w \in \mathbb{R}^S$ s.t. $u \leq w$ (element-wise) and a constant c,

$$\mathcal{T}(u) \leq \mathcal{T}(w) \quad ; \quad \mathcal{T}_\pi(u) \leq \mathcal{T}_\pi(w) \qquad \text{(Monotonicity)}$$

$$\mathcal{T}(u + c\mathbf{1}) = \mathcal{T}(u) + c\gamma\,\mathbf{1} \quad ; \quad \mathcal{T}_\pi(u + c\mathbf{1}) = \mathcal{T}_\pi(u) + c\gamma\,\mathbf{1} \qquad \text{(Additivity)}$$

Prove in Assignment 2!

Policy Error Bound

Define $\epsilon := \|v^* - v\|_\infty \implies -\epsilon\mathbf{1} \leq v^* - v \leq \epsilon\mathbf{1}$ and define $\delta := v^* - v^\pi$.

$$
\begin{aligned}
\delta = v^* - v^\pi = \mathcal{T}v^* - v^\pi &= \mathcal{T}v^* - \mathcal{T}_\pi v^\pi && \text{(By definitions of } \mathcal{T}, \mathcal{T}_\pi\text{)} \\
&\leq \mathcal{T}(v + \epsilon\mathbf{1}) - \mathcal{T}_\pi v^\pi = \mathcal{T}v + \epsilon\gamma\mathbf{1} - \mathcal{T}_\pi v^\pi && \text{(By monotonicity, additivity of } \mathcal{T}\text{)} \\
&= \mathcal{T}_\pi v + \epsilon\gamma\mathbf{1} - \mathcal{T}_\pi v^\pi && \text{(Since } \mathcal{T}v = \mathcal{T}_\pi v\text{)} \\
&\leq \mathcal{T}_\pi(v^* + \epsilon\mathbf{1}) + \epsilon\gamma\mathbf{1} - \mathcal{T}_\pi v^\pi = \mathcal{T}_\pi v^* + \gamma\epsilon\mathbf{1} + \epsilon\gamma\mathbf{1} - \mathcal{T}_\pi v^\pi \\
& && \text{(By monotonicity, additivity of } \mathcal{T}_\pi\text{)} \\
&= \mathcal{T}_\pi v^* - \mathcal{T}_\pi v^\pi + 2\gamma\epsilon\mathbf{1} \\
&= [\mathbf{r}_\pi + \gamma\mathbf{P}_\pi v^*] - [\mathbf{r}_\pi + \gamma\mathbf{P}_\pi v^\pi] + 2\gamma\epsilon\mathbf{1} && \text{(By definition of } \mathcal{T}_\pi\text{)} \\
&= \gamma\mathbf{P}_\pi(v^* - v^\pi) + 2\gamma\epsilon\mathbf{1} \\
\implies \delta &\leq \gamma\mathbf{P}_\pi\delta + 2\gamma\epsilon\mathbf{1} \\
\implies |\delta| &\leq \gamma|\mathbf{P}_\pi\delta| + 2\gamma\epsilon\mathbf{1} \\
& && \text{(Taking an element-wise absolute value and using the triangle inequality)}
\end{aligned}
$$

Policy Error Bound

Recall that $\epsilon = \|v^* - v\|_\infty$, $\delta := v^* - v^\pi$ and $|\delta| \leq \gamma |\mathbf{P}_\pi \delta| + 2\gamma\epsilon\mathbf{1}$ Let us simplify $|\mathbf{P}_\pi \delta|$. For an arbitrary s,

$$
\begin{aligned}
|\mathbf{P}_\pi \delta|(s) &= \left| \sum_{s'} \mathbf{P}_\pi(s, s')\delta(s') \right| \leq \sum_{s'} |\mathbf{P}_\pi(s, s')\delta(s')| = \sum_{s'} \mathbf{P}_\pi(s, s')|\delta(s')| \\
&\leq \|\delta\|_\infty \sum_{s'} \mathbf{P}_\pi(s, s') = \|\delta\|_\infty \\
\implies |\mathbf{P}_\pi \delta| &\leq \|\delta\|_\infty \mathbf{1} \implies |\delta| \leq \gamma \|\delta\|_\infty \mathbf{1} + 2\gamma\epsilon\mathbf{1} \\
\implies \|\delta\|_\infty &\leq \gamma \|\delta\|_\infty + 2\gamma\epsilon \implies \|\delta\|_\infty \leq \frac{2\gamma\epsilon}{1-\gamma}
\end{aligned}
$$

(By taking the element-wise maximum on both sides)

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

Value Iteration

- We have seen that VI requires $O\left(\frac{S^2 A \log(1/\epsilon)}{1-\gamma}\right)$ operations to produce an ϵ-optimal policy π that guarantees $v^\pi \geq v^* - \epsilon\mathbf{1}$.
- **Lower Bound**: For $\epsilon \in [0, \gamma/1-\gamma)$, any algorithm guaranteed to produce ϵ-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 ϵ) (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).
- Next, we will study Policy Iteration (PI) which can converge to the optimal policy with finite operations.

Policy Iteration

Policy Iteration

Algorithm Policy Iteration

1: **Input**: MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho)$, π_0.
2: **for** $k = 0 \to K$ **do**
3: **Policy Evaluation**: Calculate v^{π_k} as the solution to $(I - \gamma \mathbf{P}_{\pi_k})v = \mathbf{r}_{\pi_k}$.
4: **Policy Improvement**: $\forall s$, $\pi_{k+1}(s) = \arg\max_a \{r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\, v^{\pi_k}(s')\}$
5: **end for**

- Computational Complexity: $O((S^3 + S^2 A)\, K)$
- We will prove that $K = O\left(\frac{SA}{1-\gamma}\right)$ iterations of PI are sufficient to ensure exact convergence to the optimal policy. Hence, PI requires $O\left(\frac{S^4 A + S^3 A^2}{1-\gamma}\right)$ operations.

We will do the proof in two steps:

(i) Show that the sequence of v^{π_k} converges to v^* at a linear rate (similar to VI).
(ii) Relate v^{π_k} to the greedy policy chosen by PI at each iteration.

Policy Iteration

(i) Claim: For PI, $\|v^{\pi_K} - v^*\|_\infty \leq \gamma^K \|v^{\pi_0} - v^*\|_\infty$.
Proof: We will first prove a more general result: for any π, π', if π' is the greedy policy w.r.t v^π, then, $v^\pi \leq \mathcal{T} v^\pi \leq v^{\pi'}$. To see this, note that,

$$\mathcal{T} v^\pi = \mathcal{T}_{\pi'} v^\pi \quad ; \quad v^\pi = \mathcal{T}_\pi v^\pi \leq \mathcal{T} v^\pi \quad \text{(By definition of } \pi' \text{ and by definitions of } \mathcal{T} \text{ and } \mathcal{T}_\pi\text{)}$$

We will use induction to show that $v^\pi \leq \mathcal{T} v^\pi \leq \mathcal{T}^n_{\pi'} v^\pi$ for all n. As $n \to \infty$, $v^\pi \leq \mathcal{T} v^\pi \leq v^{\pi'}$.
Base Case: For $n = 1$, from the above definition, we know that $v^\pi \leq \mathcal{T} v^\pi = \mathcal{T}_{\pi'} v^\pi$.
Inductive Hypothesis: Assume that $v^\pi \leq \mathcal{T} v^\pi \leq \mathcal{T}^{n-1}_{\pi'} v^\pi$. Let us prove it for n,

$$v^\pi \leq \mathcal{T}^{n-1}_{\pi'} v^\pi \implies \mathcal{T}_{\pi'} v^\pi \leq \mathcal{T}^n_{\pi'} v^\pi \implies \mathcal{T} v^\pi \leq \mathcal{T}^n_{\pi'} v^\pi \implies v^\pi \leq \mathcal{T} v^\pi \leq \mathcal{T}^n_{\pi'} v^\pi$$

Using this result for PI, we get that $v^{\pi_k} \leq \mathcal{T} v^{\pi_k} \leq v^{\pi_{k+1}}$. Using this result recursively,

$$\mathcal{T} v^{\pi_0} \leq v^{\pi_1} \implies \mathcal{T}^2 v^{\pi_0} \leq \mathcal{T} v^{\pi_1} \leq v^{\pi_2} \implies \mathcal{T}^K v^{\pi_0} \leq v^{\pi_K}$$

Policy Iteration

Recall we have proved that $\mathcal{T}^K v^{\pi_0} \leq v^{\pi_K}$. Since v^* is the optimal value function,

$$\mathcal{T}^K v^{\pi_0} \leq v^{\pi_K} \leq v^* \implies v^* - v^{\pi_K} \leq v^* - \mathcal{T}^K v^{\pi_0}$$
$$\implies \left\|v^* - v^{\pi_K}\right\|_\infty \leq \left\|v^* - \mathcal{T}^K v^{\pi_0}\right\|_\infty$$
$$\implies \left\|v^* - v^{\pi_K}\right\|_\infty \leq \left\|\mathcal{T}^K v^* - \mathcal{T}^K v^{\pi_0}\right\|_\infty \leq \gamma^K \left\|v^* - v^{\pi_0}\right\|_\infty \quad \square$$

For proving **(ii)**, we will require an intermediate result – the *value difference lemma*.

Claim: For any $\pi, \pi' \in \Pi_{\text{MR}}$, $v^{\pi'} - v^\pi = (I - \gamma \mathbf{P}_{\pi'})^{-1} g(\pi', \pi)$ where $g(\pi', \pi) := \mathcal{T}_{\pi'} v^\pi - v^\pi$.
Proof: Recall that $v^{\pi'} = (I - \gamma \mathbf{P}_{\pi'})^{-1} \mathbf{r}_{\pi'}$.

$$v^{\pi'} - v^\pi = (I - \gamma \mathbf{P}_{\pi'})^{-1} \mathbf{r}_{\pi'} - v^\pi = (I - \gamma \mathbf{P}_{\pi'})^{-1} [\mathbf{r}_{\pi'} - (I - \gamma \mathbf{P}_{\pi'})\ v^\pi]$$
$$= (I - \gamma \mathbf{P}_{\pi'})^{-1} \left[(\mathbf{r}_{\pi'} + \gamma \mathbf{P}_{\pi'} v^\pi) - v^\pi\right] = (I - \gamma \mathbf{P}_{\pi'})^{-1} \left[\mathcal{T}_{\pi'} v^\pi - v^\pi\right]$$
$$= (I - \gamma \mathbf{P}_{\pi'})^{-1}\ g(\pi', \pi) \quad \square$$

Policy Iteration

Claim: Consider an arbitrary sub-optimal stationary deterministic policy π'_0 and define π'_K to be the policy returned by PI after K iterations starting from policy π'_0. For all $K \geq K^* := \lceil \frac{\log(1/1-\gamma)}{\log(1/\gamma)} \rceil + 1$, there exists a state s' such that $\pi'_K[s'] \neq \pi'_0[s']$. This means that for all $K \geq K^*$, the action corresponding to $\pi'_0[s']$ is *eliminated* for state s'.

We will use this claim multiple times starting from $\pi'_0 = \pi_0$. In particular,

- After $K \geq K^*$ iterations of PI, we know there exists a state s' for which the action corresponding to $\pi_0[s']$ is eliminated.
- If we continue running PI, after a further K^* iterations, another action would be eliminated. Specifically, for $\pi'_0 = \pi_{K^*}$, there exists a state s'' for which the action corresponding to $\pi_{K^*}[s'']$ is eliminated.
- Since we are considering deterministic policies, we need to eliminate at most $SA - S$ actions, and need to run PI for at most $(SA - S)\, K^*$ iterations. Hence, PI will converge to the optimal policy in $O\left(\frac{S\,A\,\log(1/1-\gamma)}{1-\gamma}\right)$ iterations.

Policy Iteration

Proof: We will make use of the value difference lemma to bound $g(\pi, \pi^*)$. Note that $g(\pi, \pi^*) = \mathcal{T}_\pi v^* - v^* < 0$ for all sub-optimal policies π.

$$-g(\pi'_K, \pi^*) = \left(I - \gamma \mathbf{P}_{\pi'_K}\right)[v^* - v^{\pi'_K}] = [v^* - v^{\pi'_K}] - \gamma \mathbf{P}_{\pi'_K} \underbrace{[v^* - v^{\pi'_K}]}_{\text{Non-negative}} \leq [v^* - v^{\pi'_K}]$$

$$\implies \|g(\pi'_K, \pi^*)\|_\infty \leq \left\| v^* - v^{\pi'_K} \right\|_\infty$$

(Taking element-wise absolute value and max over the states)

$$\leq \gamma^K \left\| v^{\pi'_0} - v^* \right\|_\infty \qquad \text{(From the claim in } \textbf{(i)}\text{)}$$

$$= \gamma^K \left\| \left(I - \gamma \mathbf{P}_{\pi'_0}\right)^{-1} g(\pi'_0, \pi^*) \right\|_\infty \qquad \text{(Value Difference Lemma)}$$

$$\leq \frac{\gamma^K}{1-\gamma} \|g(\pi'_0, \pi^*)\|_\infty \qquad \text{(Using the Neumann series)}$$

$$\implies \|g(\pi'_K, \pi^*)\|_\infty < \|g(\pi'_0, \pi^*)\|_\infty \qquad \left(K \geq K^* = \left\lceil \frac{\log(1/1-\gamma)}{\log(1/\gamma)} \right\rceil + 1\right)$$

Policy Iteration

Recall that $\|g(\pi'_K, \pi^*)\|_\infty < \|g(\pi'_0, \pi^*)\|_\infty$.
If $s' := \arg\max_s |g(\pi'_0, \pi^*)(s)| \implies \|g(\pi'_0, \pi^*)\|_\infty = -g(\pi'_0, \pi^*)(s')$, then,

$$
\begin{aligned}
\|g(\pi'_K, \pi^*)\|_\infty < -g(\pi'_0, \pi^*)(s') &\implies \max_s |g(\pi'_K, \pi^*)| \le -g(\pi'_0, \pi^*)(s') \\
&\implies -g(\pi'_K, \pi^*)(s') < -g(\pi'_0, \pi^*)(s') \\
&\implies v^*(s') - (\mathcal{T}_{\pi'_K} v^*)(s') < v^*(s') - (\mathcal{T}_{\pi'_\mathbf{0}} v^*)(s') \quad \text{(Recall that } -g(\pi', \pi^*) = v^* - \mathcal{T}_{\pi'} v^*) \\
&\implies \mathbf{r}_{\pi'_K}(s') + (\mathbf{P}_{\pi'_K} v^*)(s') > \mathbf{r}_{\pi'_\mathbf{0}}(s') + (\mathbf{P}_{\pi'_\mathbf{0}} v^*)(s') \qquad \text{(Recall that } \mathcal{T}_{\pi'} v^* = \mathbf{r}_{\pi'} + \mathbf{P}_{\pi'} v^*) \\
&\implies \pi'_K(s') \neq \pi'_0(s') \quad \square
\end{aligned}
$$