Lecture 11 - Feb 17, 2023

- Dynamic programming
  - Policy Iteration
  - Value Iteration

  } Vector-Form

↓

  - Policy Iteration
  - Value Iteration

  } Matrix-Form

- Approximate Dynamic programming
  - Asynchronous DP
  - Generalized Policy Iteration

Exam 1 → Tuesday, Feb 21

HW2 → Due Feb 18

Project 2 → Due March 3

TA's office hour: Wendsdays, 2pm-3pm (in-person)

Fridays, 2pm-3pm (virtual)

Review:

DP
- Policy Iteration
- Value Iteration

① Policy Iteration
- Policy Evaluation
- Policy Improvement

$\pi^0 \xrightarrow{PE} V_{\pi^0} \xrightarrow{PI} \pi^1 \xrightarrow{PE} V_{\pi^1} \cdots$

PE: $V_\pi(s) = \sum_{s'} P(s'|s,\pi(s)) [R(s,\pi(s),s') + \gamma V_\pi(s')]$

$\hookrightarrow V_{k+1}(s) = \sum_{s'} P(s'|s,\pi(s)) [R(s,\pi(s),s') + \gamma V_k(s')]$

PI: $\pi'(s) = \underset{a \in A}{\text{argmax}} \sum_{s'} P(s'|s,a) [R(s,a,s') + \gamma V_\pi(s')]$

---

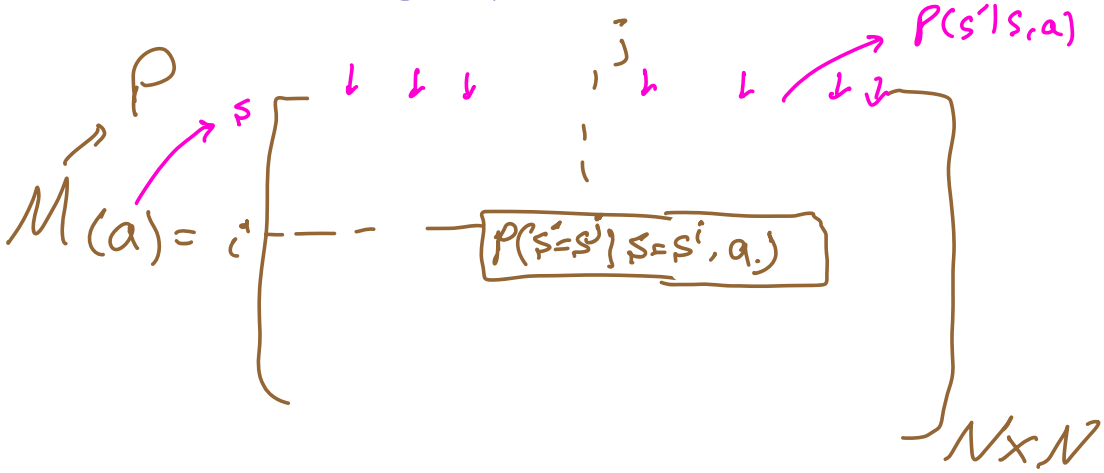
② Value Iteration

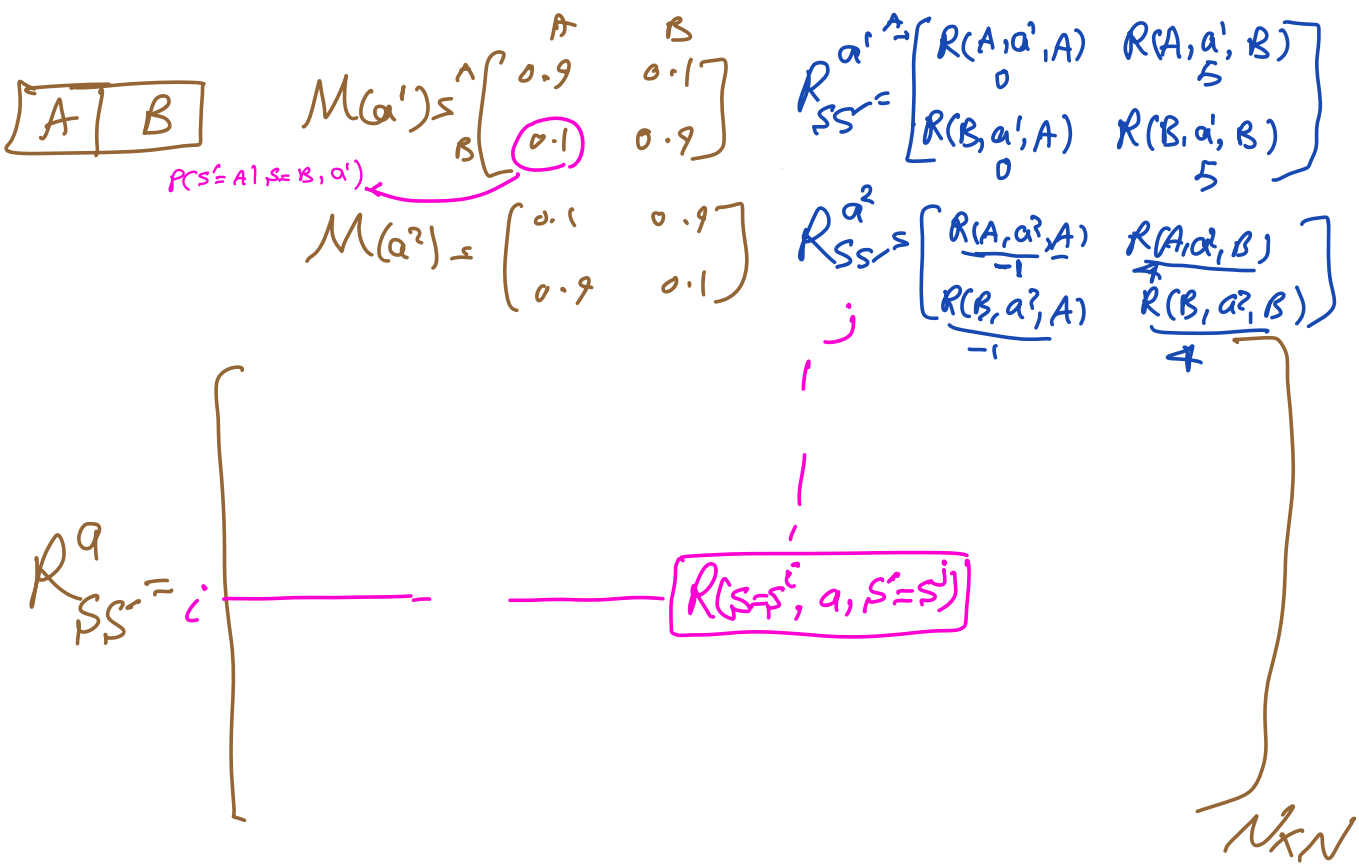
$V_{k+1}(s) = \max_{a \in A} \sum_{s'} P(s'|s,a) [R(s,a,s') + \gamma V_k(s')]$

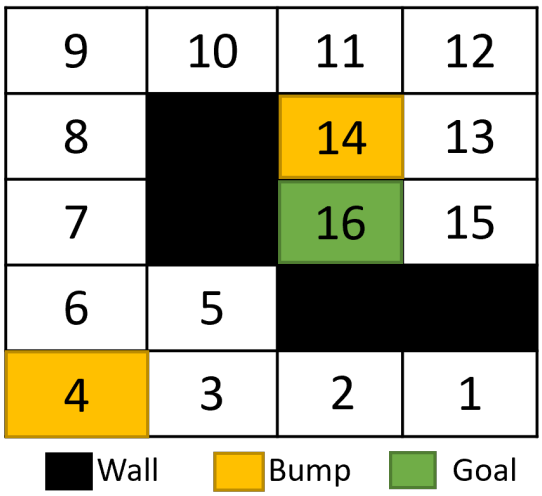
$V_0 - \cdots - V_T \quad V_{T+1} = V^*$

$$\pi^*(s) = \underset{a \in A}{\operatorname{argmax}} \sum_{s'} P(s' \mid s, a)\left[R(s, a, s') + \gamma V^*(s')\right]$$

Matrix-form DP:







$M(a=U) =$

$\begin{bmatrix} 1 & 0 & 0 & \cdots & & & & 0 \\ 0 & 1 & & & & & & \\ 0 & 0 & 0 & 0 & 1 & \cdots & \cdots & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \\ & & & & & & & \end{bmatrix}_{16 \times 16}$

$$R(s,a) = \sum_{s'} P(s'|s,a)\, R(s,a,s')$$

$R(s,a,s')$

s	a	s'	$P(s'\|s,a)$	$R(s,a,s')$
A	a^2	B	0.9	4
		A	0.1	-1

$$R(A,a^2) = 0.9 \times (4) + 0.1 \times (-1) = 3.5$$

$$R_s^a = \begin{bmatrix} R(s^1,a) \\ \\ \\ R(s^N,a) \end{bmatrix} = \left(M(a) \odot R_{ss'}^a\right) 1_{N\times 1}$$

component wise Hadamard product

$$\begin{bmatrix} P(s^1|s,a) & \cdots & P(s^N|s,a) \end{bmatrix} \odot \begin{bmatrix} R(s,a,s^1) & \cdots & R(s,a,s^N) \end{bmatrix}$$

$$\begin{bmatrix} P(s^1|s,a)R(s,a,s^1) & \cdots & P(s^N|s,a)R(s,a,s^N) \end{bmatrix}$$

$$M(a^1) = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \qquad R_{SS'}^{a^1} = \begin{bmatrix} 0 & 5 \\ 0 & 5 \end{bmatrix}$$

$$M(a^2) = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix} \qquad R_{SS'}^{a^2} = \begin{bmatrix} -1 & 4 \\ -1 & 4 \end{bmatrix}$$

$$R_S^{a^1} = \begin{bmatrix} R(A, a^1) \\ R(B, a^1) \end{bmatrix} = \underset{2\times 1}{\left(M(a^1) \odot R_{SS'}^{a^1}\right)} = \left(\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \odot \begin{bmatrix} 0 & 5 \\ 0 & 5 \end{bmatrix}\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0.5 \\ 0 & 4.5 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 4.5 \end{bmatrix}$$

$$R_S^{a^2} = \begin{bmatrix} R(A, a^2) \\ R(B, a^2) \end{bmatrix} = \left(M(a^2) \odot R_{SS'}^{a^2}\right) \underset{N\times 1}{=} \left(\begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix} \odot \begin{bmatrix} -1 & 4 \\ -1 & 4 \end{bmatrix}\right)\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} -0.1 & 3.6 \\ -0.9 & 0.4 \end{bmatrix} = \begin{bmatrix} 3.5 \\ -0.5 \end{bmatrix}$$

Value Iteration :

$$V_{k+1} = \max_{a \in A} R_s^a + \gamma M(a) V_k \quad (I)$$

$$V_{k+1}(s) = \max_{a \in A} \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V_k(s')\right]$$

$$= \max_{a \in A} \underbrace{\sum_{s'} P(s'|s,a) R(s,a,s')}_{R(s,a)} + \gamma \sum_{s'} P(s'|s,a) V_k(s')$$

$$P(s'=s^1|s,a) V_k(s^1) + P(s'=s^2|s,a) V_k(s^2) + \cdots + P(s'=s^N|s,a) V_k(s^N)$$

$$M(a) V_k$$

$V_0 \longrightarrow V_1 \longrightarrow V_2 \longrightarrow \cdots V_T$

$\|V_T - V_{T-1}\|_\infty < \theta \rightarrow V^*$ ✓

$$\pi^*(s) = \underset{a \in A}{\operatorname{argmax}} \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V^*(s')\right]$$

$$\begin{bmatrix} \pi^*(s^1) \\ \vdots \\ \pi^*(s^N) \end{bmatrix} = \pi^* = \underset{a \in A}{\operatorname{argmax}} \; R_s^a + \gamma M(a) V^* \qquad (II)$$

A	B

$$M(a^1) = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \quad M(a^2) = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}$$

$$R_{ss'}^{a^1} = \begin{bmatrix} 0 & 5 \\ 0 & 5 \end{bmatrix} \quad R_{ss'}^{a^2} = \begin{bmatrix} -1 & 4 \\ -1 & 4 \end{bmatrix}$$

$$R_s^{a^1} = \begin{bmatrix} 0.5 \\ 4.5 \end{bmatrix} \quad R_s^{a^2} = \begin{bmatrix} 3.5 \\ -0.5 \end{bmatrix}$$

$$V_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \xrightarrow{VIB} V_1$$

$$V_1 = \max_{a \in A} R_s^a + \gamma M(a) V_0$$

$$= \max \left\{ \underbrace{R_s^{a^1} + \gamma M(a) V_0}_{a^1} , \underbrace{R_s^{a^2} + \gamma M(a) V_0}_{a^2} \right\}$$

$$= \max \left\{ \begin{bmatrix} 0.5 \\ 4.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} , \begin{bmatrix} 3.5 \\ -0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

Row-wise

$$= \max \left\{ \begin{bmatrix} 0.5 \\ 4.5 \end{bmatrix} , \begin{bmatrix} 3.5 \\ -0.5 \end{bmatrix} \right\} = \begin{bmatrix} 3.5 \\ 4.5 \end{bmatrix}$$

$$V_2 = \max_{a \in A} R_s^a + \gamma M(a) V_1$$

$$= \max \left\{ \underbrace{R_s^{a1} + \gamma M(a) V_1}_{a1} , \underbrace{R_s^{a2} + \gamma M(a) V_1}_{a2} \right\}$$

$$= \max \left\{ \begin{bmatrix} 0.5 \\ 4.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 3.5 \\ 4.5 \end{bmatrix}, \begin{bmatrix} 3.5 \\ -0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix} \begin{bmatrix} 3.5 \\ 4.5 \end{bmatrix} \right\}$$

$$= \max \left\{ \begin{bmatrix} 3.74 \\ 8.46 \end{bmatrix}, \begin{bmatrix} 7.46 \\ 2.74 \end{bmatrix} \right\} = \begin{bmatrix} 7.46 \\ 8.46 \end{bmatrix}$$

$$\| V_2 - V_1 \|_\infty < \theta$$

$$\left\| \begin{bmatrix} 7.46 \\ 8.46 \end{bmatrix} - \begin{bmatrix} 3.5 \\ 4.5 \end{bmatrix} \right\| < 0.001 \quad \times \downarrow$$

V_3

V_4

$\vdots$

$$V_{1000} \cong V_{1001} = \begin{bmatrix} 43.1 \\ 44.1 \end{bmatrix} = V^*$$

$$\pi^* = \begin{bmatrix} \pi^*(A) \\ \pi^*(B) \end{bmatrix} = \underset{a}{\operatorname{argmax}}\ R_s^a + \gamma M(a) V^*$$

$$\pi^* = \underset{a \in A}{\operatorname{argmax}} \left\{ R_s^{a^1} + \gamma M(a^1) V^*,\ R_s^{a^2} + \gamma M(a^2) V^* \right\}$$

$$= \underset{a \in A}{\operatorname{argmax}} \left\{ \begin{bmatrix} 0.5 \\ 4.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 43.1 \\ 44.1 \end{bmatrix},\ \begin{bmatrix} 3.5 \\ -0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix} \begin{bmatrix} 43.1 \\ 44.1 \end{bmatrix} \right\}$$

$$= \underset{\text{row-wise}}{\operatorname{argmax}} \left\{ \underbrace{\begin{bmatrix} 39.38 \\ 44.1 \end{bmatrix}}_{a^1},\ \underbrace{\begin{bmatrix} 43.1 \\ 38.38 \end{bmatrix}}_{a^2} \right\} = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix} \checkmark$$

Value Iteration – Matrix Form

Transition matrices $M(a)$, $a \in A$, Reward $R^a_{ss'}$, Number of states N.

- Reward: $R^a_s = \left(M(a) \odot R^a_{ss'}\right) 1_{N\times 1}$, for $a \in A$. ($\odot$: Hadamard Product)
- $V_0 = 0_{N\times 1}$, $k = 0$

Repeat

- Value Iteration Backup: $V_{k+1} = \max\limits_{a\in A} \left\{ R^a_s + \gamma M(a) V_k \right\}$ (max: Row-wise)
- $k = k+1$

Until $\max\limits_{i\in\{1,\dots,N\}} |V_k(i) - V_{k-1}(i)| < \theta$

- optimal state values: $V^* = V_k$
- optimal policy: $\pi^* = \arg\max\limits_{a\in A} R^a_s + \gamma M(a) V^*$ (max: Row-wise)

Policy Iteration:

$\pi^0 \xrightarrow{PE} V_{\pi^0} \xrightarrow{PI} \pi^1 \xrightarrow{PE} V_{\pi^1} \cdots$

PE

① $V_\pi(s) = \sum_{s'} P(s'|s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V_\pi(s') \right]$

② $V_{k+1}(s) = \sum_{s'} P(s'|s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V_k(s') \right]$

Matrix_form

$V_\pi(s) = \underbrace{\sum_{s'} P(s'|s, \pi(s)) R(s, \pi(s), s')}_{R(s, \pi(s))} + \gamma \sum_{s'} P(s'|s, \pi(s)) V_\pi(s')$

$P(s'=s^1|s, \pi(s)) V_\pi(s^1) + P(s'=s^2|s, \pi(s)) V_\pi(s^2) + \cdots + P(s'=s^N|s, \pi(s)) V_\pi(s^N)$

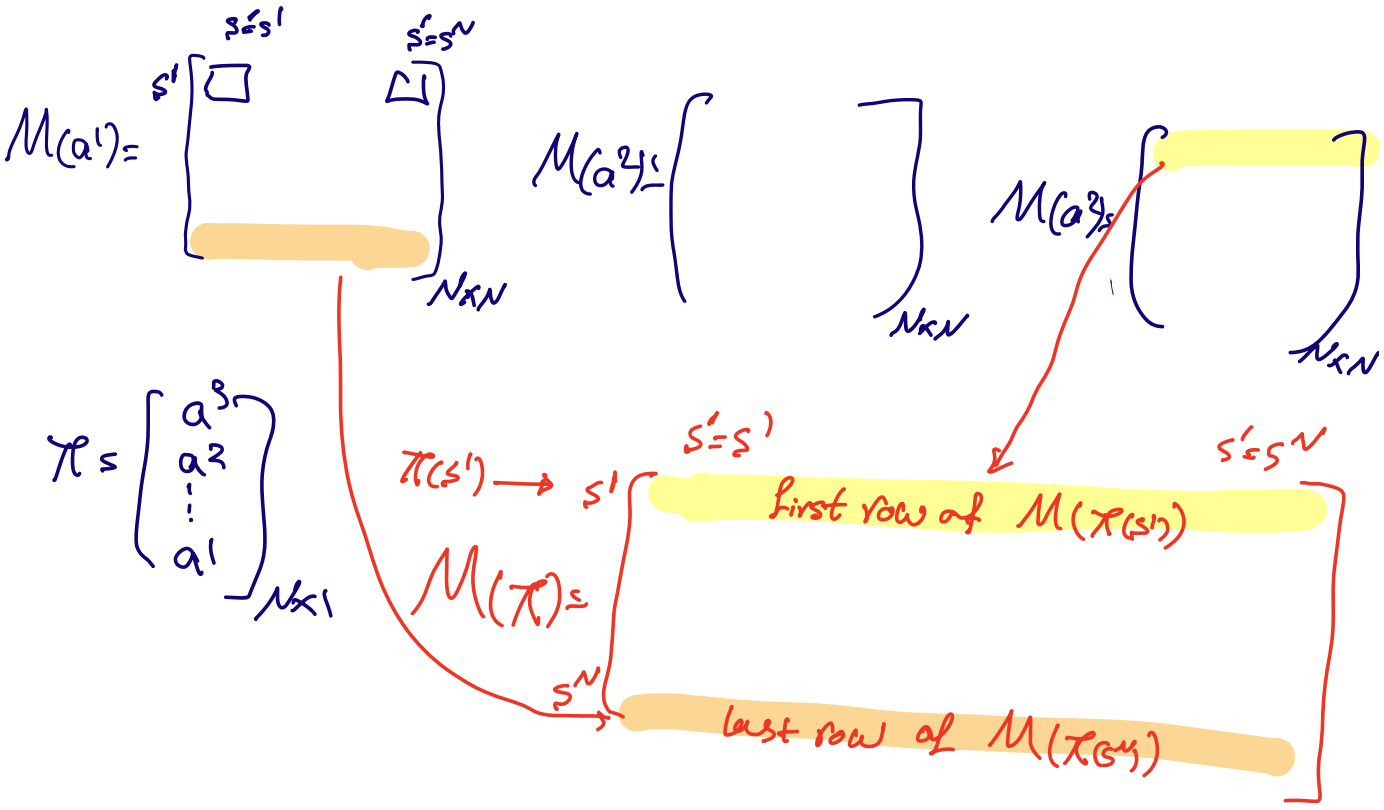
$s \rightarrow \left[\quad M(\pi(s)) \quad \right]$

$$V^{\pi} = R_s^{\pi} + \gamma M(\pi) V^{\pi} \Rightarrow \text{Calculate } V^{\pi}$$

$$(I - \gamma M(\pi)) V^{\pi} = R_s^{\pi}$$

$$V^{\pi} = (I - \gamma M(\pi))^{-1} R_s^{\pi}$$

Policy Evaluation



$$R_s^a = (M(a) \odot R_{ss'}^a) 1_N$$

$$R_S^{a^1} = \begin{bmatrix} R(s^1, a^1) \\ \vdots \\ R(s^N, a^1) \end{bmatrix} \qquad R_S^{a^2} = \begin{bmatrix} R(s^1, a^2) \\ \vdots \\ R(s^N, a^2) \end{bmatrix} \cdots \quad R_S^{a^3} = \begin{bmatrix} R(s^1, a^3) \\ \\ R(s^N, a^3) \end{bmatrix}$$

$$R_S^{\pi} = \begin{bmatrix} R(s^1, \overrightarrow{\pi(s^1)}^{a^3}) \\ \\ R(s^N, \overrightarrow{\pi(s^N)}^{a^1}) \end{bmatrix}$$

first row of $R_S^{\pi(s^1)}$

last row of $R_S^{\pi(s^N)}$

$$V_{k+1}(s) = \sum_{s'} P(s'|s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V_k(s') \right]$$

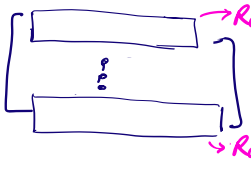
$$V_{k+1} = R_S^{\pi} + \gamma M(\pi) V_k$$

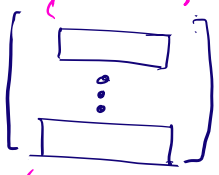
Policy Evaluation: Matrix-form

Input: Policy π

- Number of states: N, Transition Matrices: $M(a^1), \dots M(a^L)$.

- Reward: $R_s^a = \left(M(a) \overset{\text{Hadamard product}}{\odot} R_{ss'}^a\right) 1_{N\times 1}$, for $a \in A$.

- $M(\pi) =$ 

, $R_s^\pi =$ 

- $V^\pi = \left(I - \gamma M(\pi)\right)^{-1} R_s^\pi$

output: V^π

Policy Improvement

$$\pi'(s) = \underset{a \in A}{\operatorname{argmax}} \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V_{\pi}(s')\right]$$

Policy Improvement

$$\pi' = \underset{a \in A}{\operatorname{argmax}} \; R_s^a + \gamma M(a) V^{\pi}$$

Example:

$$R_s^{a1} = \left(M(a1) \odot R_{ss'}^{a'}\right)\mathbf{1}_{2\times 1} = \begin{bmatrix} 0.5 \\ 4.5 \end{bmatrix}$$

$$R_s^{a2} = \begin{bmatrix} 3.5 \\ -0.5 \end{bmatrix}$$

$$M(a1) = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$$

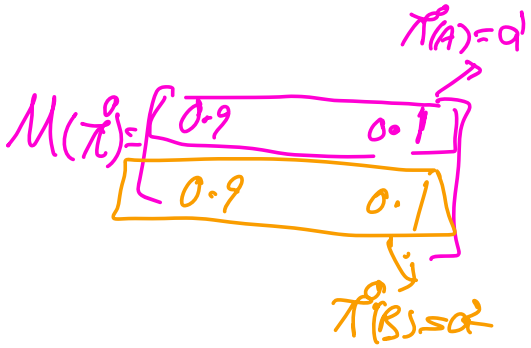
$$M(a2) = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}$$

$$R_{ss'}^{a'} = \begin{bmatrix} 0 & 5 \\ * & 5 \end{bmatrix}$$

$$R_{ss'}^{a2} = \begin{bmatrix} -1 & 4 \\ -1 & 4 \end{bmatrix}$$

Policy Iteration:

$\pi^0 = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$ ← Random $\xrightarrow{PE} V^{\pi^0}$



$$R_S^{\pi^0} = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} \leftarrow R(A, a^1), \quad \leftarrow R(B, a^2)$$

$$V^{\pi^0} = (I - \gamma M(\pi^0))^{-1} R_S^{\pi^0}$$

$$V^{\pi^0} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 4.1 \\ 3.1 \end{bmatrix}$$

$$\pi' = \underset{a \in A}{\operatorname{argmax}} \; R_S^a + \gamma M(a) V^{\pi^0}$$

$$= \underset{a \in A}{\operatorname{argmax}} \left\{ R_S^{a^1} + \gamma M(a^1) V^{\pi^0}, \; R_S^{a^2} + \gamma M(a^2) V^{\pi^0} \right\}$$

$$= \operatorname{argmax} \left\{ \begin{bmatrix} 0.5 \\ 4.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 4.1 \\ 3.1 \end{bmatrix} \right.$$

$$\left[\begin{bmatrix}3.5\\-0.5\end{bmatrix}+0.9\begin{bmatrix}0.1 & 0.1\\0.9 & 0.1\end{bmatrix}\begin{bmatrix}4.1\\5.1\end{bmatrix}\right]$$

$$= \text{argmax}\left\{\underbrace{\begin{bmatrix}4.1\\7.38\end{bmatrix}}_{a^1},\underbrace{\begin{bmatrix}6.38\\3.1\end{bmatrix}}_{a^2}\right\} = \begin{bmatrix}a^2\\a^1\end{bmatrix}$$

$\rightarrow M(\pi^2)\ R_S^{\pi^2} \rightarrow V^{\pi^2}$

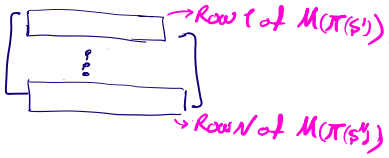
π^3

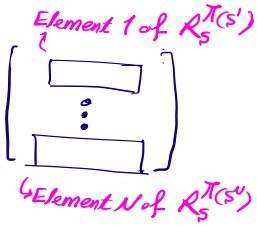
Policy Iteration: Matrix Form

Initialization: arbitrary policy π, transition matrices $M(a)$, Reward $R^a_{ss'}$

- Reward: $R^a_s = \left(M(a) \overset{\text{Hadamard Product}}{\odot} R^a_{ss'}\right) 1_{N\times 1}$, for $a \in A$.

Policy Evaluation

- $M(\pi) =$ 

, $R^\pi_s =$ 

- $V^\pi = \left(I - \gamma M(\pi)\right)^{-1} R^\pi_s$

Policy Improvement

- $\pi' = \underset{a\in A}{\text{argmax}}$ (Row-wise) $R^a_s + \gamma M(a) V^\pi$

If $\pi \neq \pi'$, $\pi = \pi'$ and go back to policy Evaluation step.

Else

- optimal policy: $\pi^* = \pi'$
- optimal state values: $V^* = V^\pi$