Policy Evaluation

Announcements

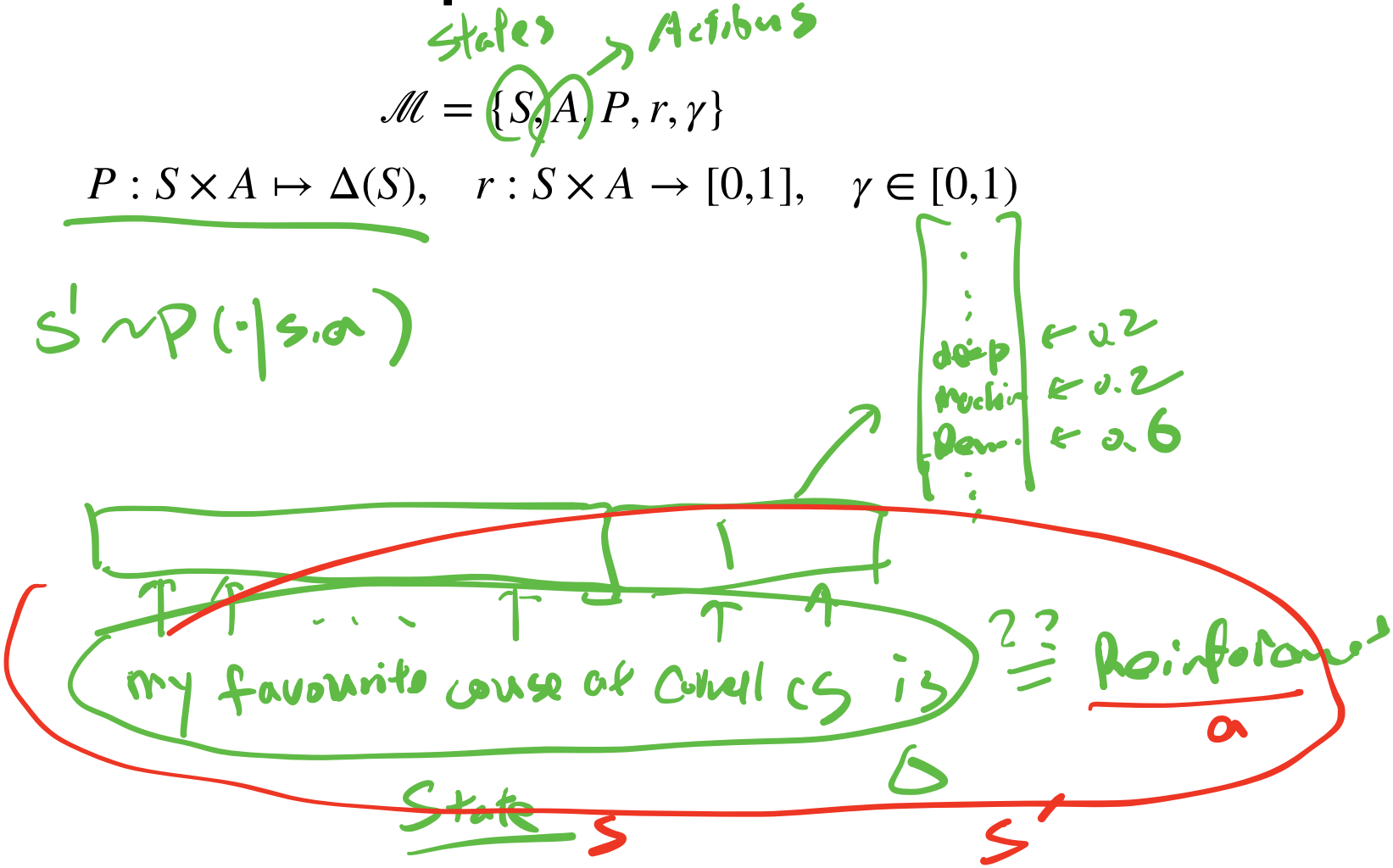
P1 is delayed, will be released on Monday

HW0 due today

Recap: Definitions

$$\mathcal{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$



Recap: Definitions

$$\mathscr{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto \Delta(A)$

$a \sim \pi(\cdot|s)$

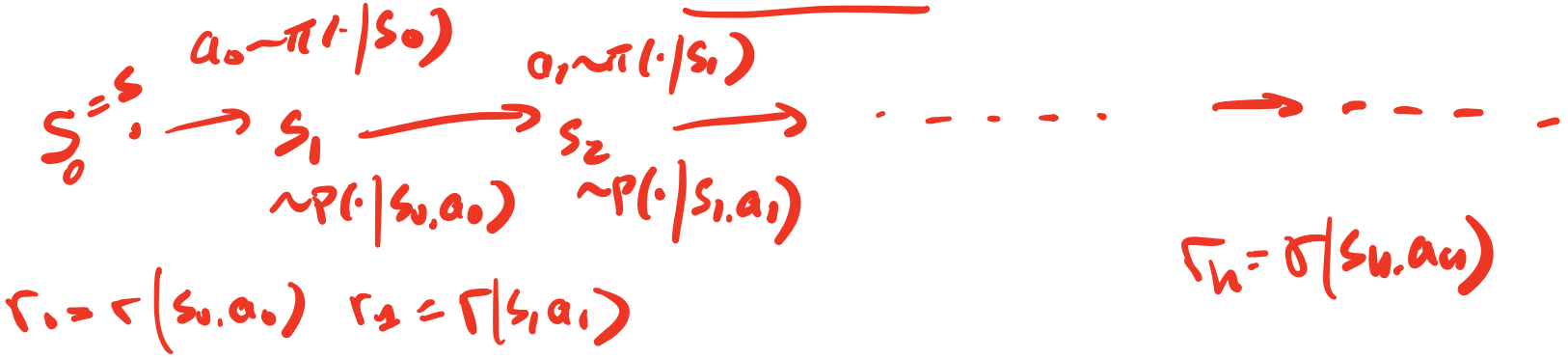
Recap: Definitions

$$\mathcal{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto \Delta(A)$

Value function $V^{\pi}(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, \pi\right]$



Recap: Definitions

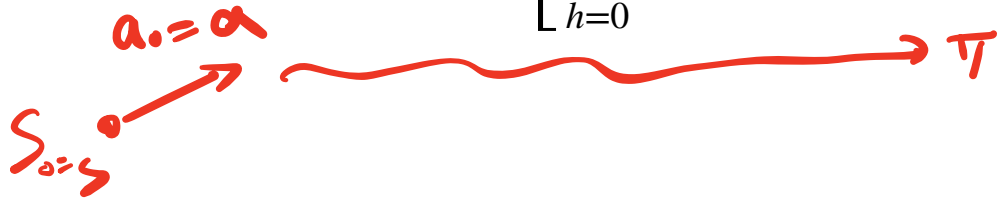
$$\mathcal{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto \Delta(A)$

Value function $V^{\pi}(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, \pi\right]$

Q function $Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s,a), a_h \sim \pi(\cdot \mid s_h) \text{ for } h \geq 1\right]$



Recap: Bellman equation (consistency)

Bellman Eq

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

$a \sim \pi(\cdot|s)$

$s \cdot \longrightarrow s' \sim P(\cdot|s,a)$

$V^{\pi}(s')$

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

Recap: Bellman equation (consistency)

Bellman Eq

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

Relationship between V and Q

Exercise: can you write V^{π} using Q^{π}, and then Q^{π} using V^{π}

$V^{\pi}(s)$?? $\mathbb{E}_{a\sim\pi(\cdot|s)} Q^{\pi}(s,a) = V^{\pi}(s)$

$Q^{\pi}(s,a)$?? $= r(s,a) + \gamma \mathbb{E}_{s'\sim p(s,a)} V^{\pi}(s')$

Today: Policy Evaluation

Key Question:

Given MDP $\mathcal{M} = (S, A, r, P, \gamma)$ & a $\pi : S \mapsto A$,
how good is π?

i.e., how to compute $V^{\pi}(s), \forall s$?

Motivation for Policy Evaluation

We want to **evaluate** our strategy against some fixed opponent (AlphaGo constantly estimates the current probability of winning)

Motivation for Policy Evaluation

We want to **evaluate** our strategy against some fixed opponent (AlphaGo constantly estimates the current probability of winning)

We want to **evaluate** our recommendation strategy before we release it to users

A more fundamental motivation…

Recall that we have A^S many policies.
To select the optimal policy, we need to do evaluation

Outline:

1. **Exact** Policy Evaluation

2. **Approximate** Policy Evaluation via an Iterative Algorithm

Exact Policy Evaluation

Setup: we have MDP $\mathscr{M} = (S, A, P, \gamma, r)$, and ***deterministic*** π, we want to compute V^π

Exact Policy Evaluation

Setup: we have MDP $\mathscr{M} = (S, A, P, \gamma, r)$, and ***deterministic*** π, we want to compute V^π

We know that for V^π, we have **Bellman equation:**

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(\cdot|s,\pi(s))} V^\pi(s')$$

$a \in \pi(s)$

Exact Policy Evaluation

Setup: we have MDP $\mathscr{M} = (S, A, P, \gamma, r)$, and ***deterministic*** π, we want to compute V^π

We know that for V^π, we have **Bellman equation:**

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(\cdot|s,\pi(s))} V^\pi(s')$$

This gives us S many **linear** constraints

Exact Policy Evaluation

Let's form linear constraints. Denote $V(s)$ as our estimate for $s \in S$

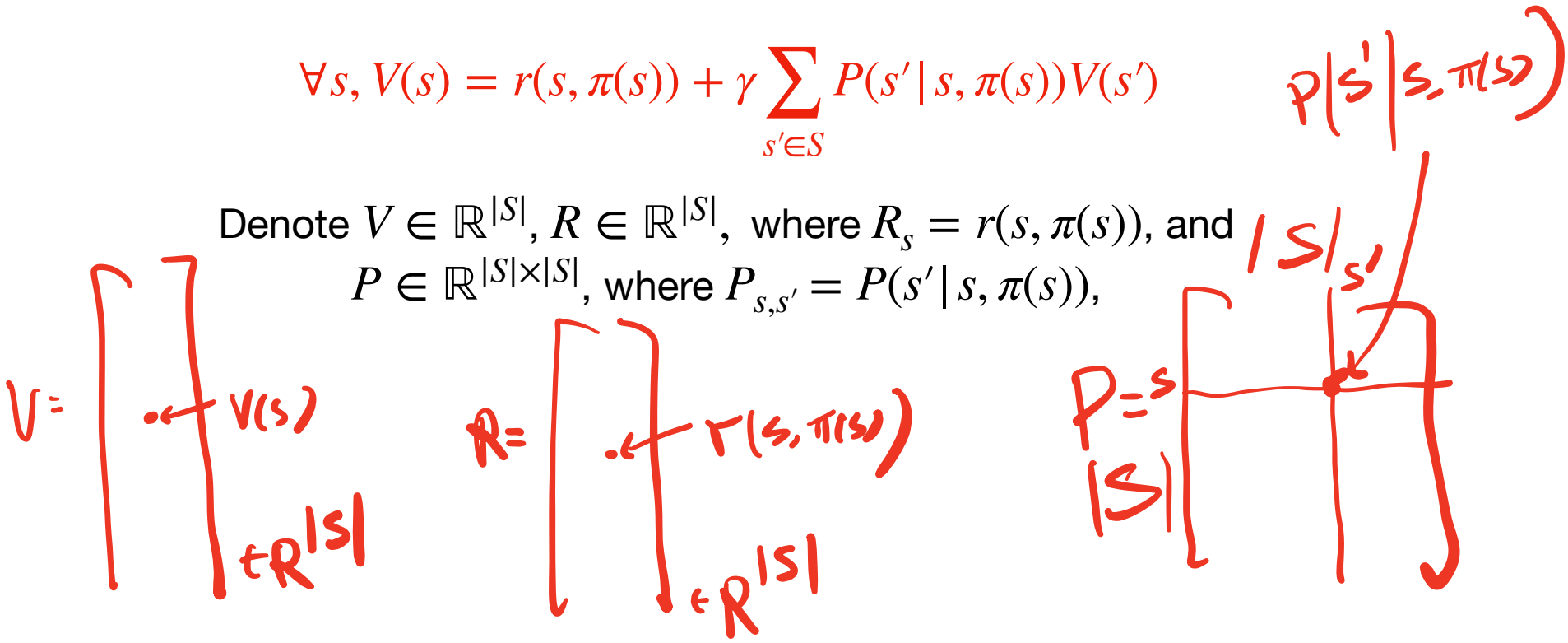
$$\forall s, V(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V(s')$$

Exact Policy Evaluation

Let's form linear constraints. Denote $V(s)$ as our estimate for $s \in S$

$$\forall s, V(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V(s')$$

Denote $V \in \mathbb{R}^{|S|}$, $R \in \mathbb{R}^{|S|}$, where $R_s = r(s, \pi(s))$, and $P \in \mathbb{R}^{|S| \times |S|}$, where $P_{s,s'} = P(s' \mid s, \pi(s))$,



Exact Policy Evaluation

Let's form linear constraints. Denote $V(s)$ as our estimate for $s \in S$

$$\forall s, V(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s' \mid s, \pi(s)) V(s')$$

Denote $V \in \mathbb{R}^{|S|}$, $R \in \mathbb{R}^{|S|}$, where $R_s = r(s, \pi(s))$, and
$P \in \mathbb{R}^{|S| \times |S|}$, where $P_{s,s'} = P(s' \mid s, \pi(s))$,

we can **combine all S many constraints together:**

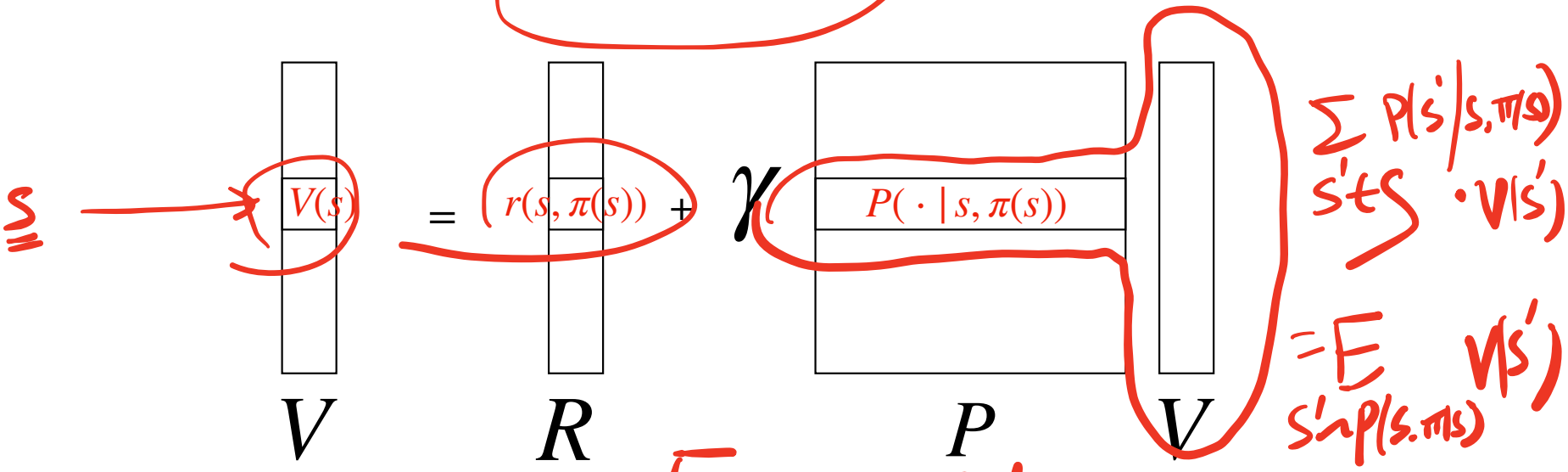
$$V = R + \gamma P V$$

Exact Policy Evaluation

$V \in \mathbb{R}^{|S|}$, $R \in \mathbb{R}^{|S|}$, where $R_s = r(s, \pi(s))$, and $P \in \mathbb{R}^{|S| \times |S|}$, where $P_{s',s} = P(s' \mid s, \pi(s))$,

we can combine all constraints together:

$$V = R + \gamma P V$$



$$V(s) = r(s, \pi(s)) + \gamma \mathop{\mathbb{E}}_{s' \sim P(s, \pi(s))} V(s')$$

Exact Policy Evaluation

$(I - \gamma P)V = R$

Since $V = r + \gamma PV$, we can obtain V as:

$$V = (I - \gamma P)^{-1} R$$

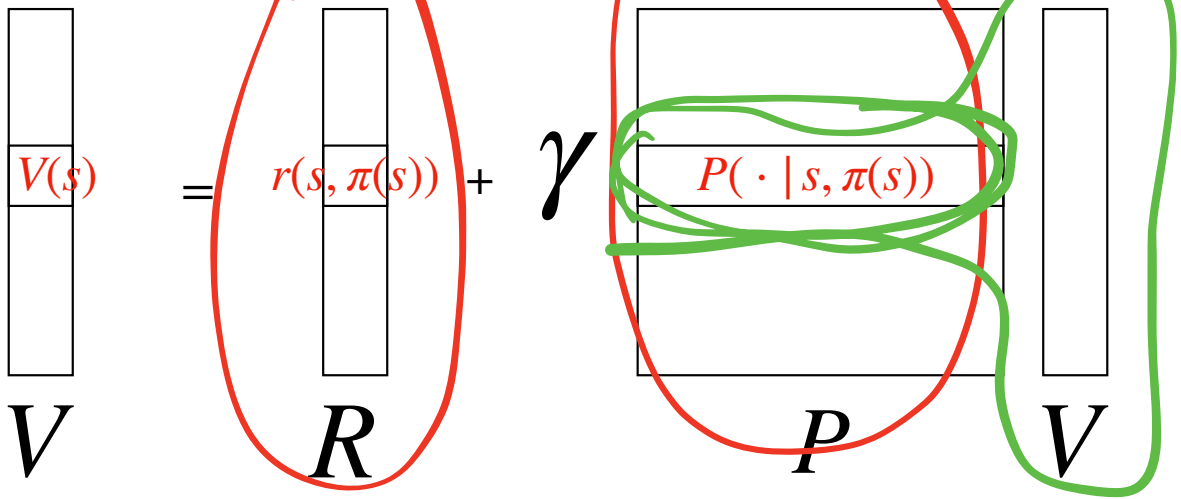
Exact Policy Evaluation

Since $V = r + \gamma PV$, we can obtain V as:

$$V = (I - \gamma P)^{-1} R$$

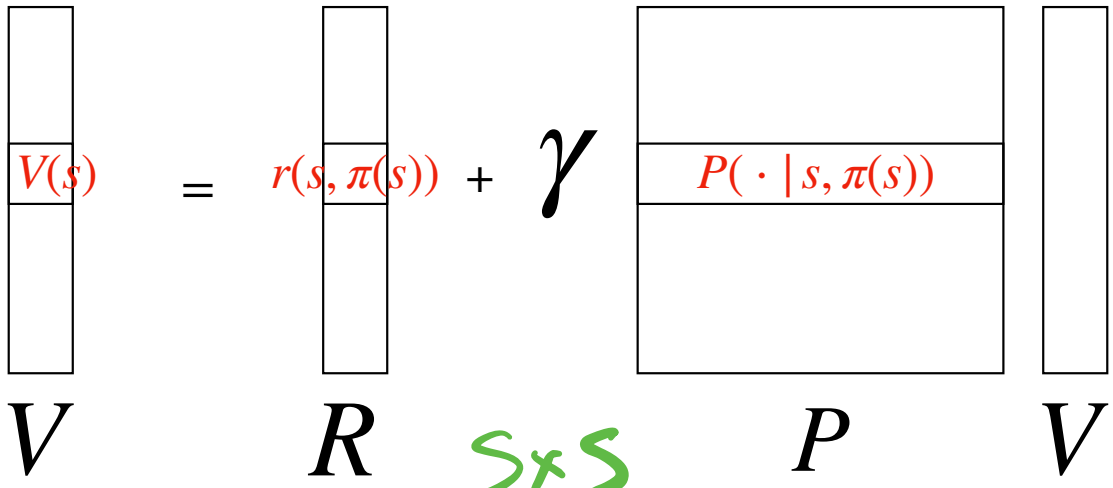
See the assigned AJKS section for proving $(I - \gamma P)$ is full rank (thus invertible)

Summary so far:



$$V = (I - \gamma P)^{-1} R$$

Summary so far:



$$V = (I - \gamma P)^{-1} R$$

Downside: computation expensive: matrix inverse is $O(S^3)$

Outline:

$V \approx V^{\pi}$

1. Exact Policy Evaluation

2. Approximate Policy Evaluation via an Iterative Algorithm

(An approximation solution could be enough, i.e., trade accuracy for computation)

V^π is a fix-point solution

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(s,\pi(s))} V^\pi(s')$$

$$V^\pi = R + \gamma P V^\pi$$

$$R + \gamma P \cdot V^\pi \Rightarrow V^\pi$$

V^π is a fix-point solution

$$\forall s, V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s \sim P(s,\pi(s))} V^\pi(s')$$

$$V^\pi = R + \gamma P V^\pi$$

V^0 $\quad V' = R + \gamma P V^0 \rightarrow V^2 = R + \gamma P V^1$

$\| V^{t+1} - V^t \|_2 \approx 0$

Iterative Policy Evaluation:

Algorithm (Iterative PE):

Start with some initialization $V^0 \in \mathbb{R}^{|S|}$, **repeat for** $t = 0 \ldots$:

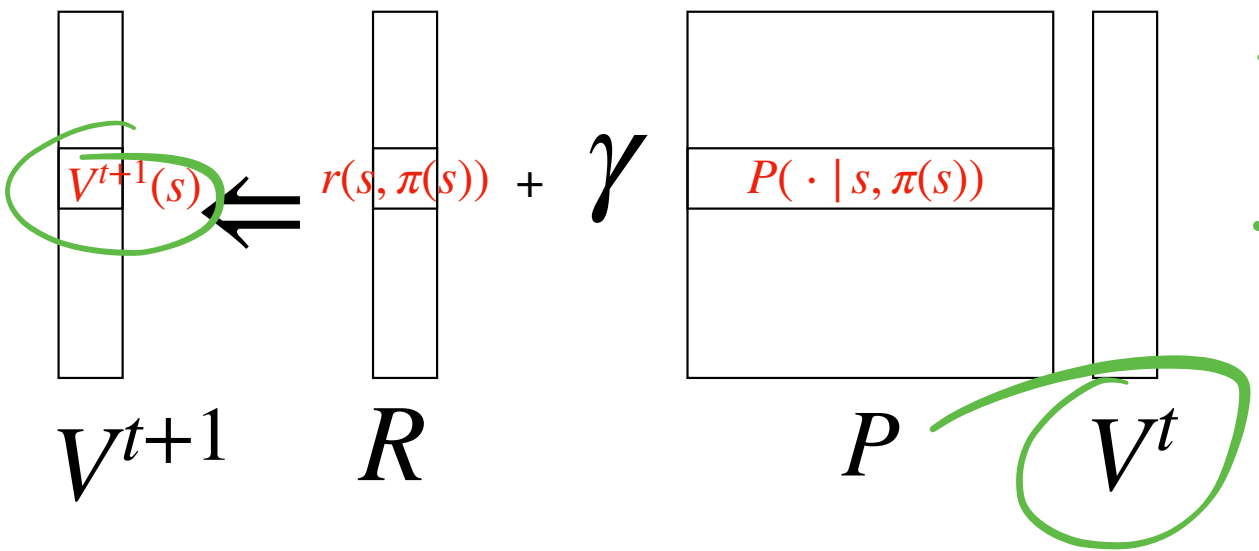
$$V^{t+1} \Leftarrow R + \gamma P V^t$$

Iterative Policy Evaluation:

Algorithm (Iterative PE):

Start with some initialization $V^0 \in \mathbb{R}^{|S|}$, **repeat for** $t = 0 \ldots$:

$$V^{t+1} \Leftarrow R + \gamma P V^t$$



$V^{t+1}(s)$

$= r(s, \pi(s))$

$+ \gamma \sum_{s' \sim P(\cdot \mid s, \pi(s))} V^t(s')$

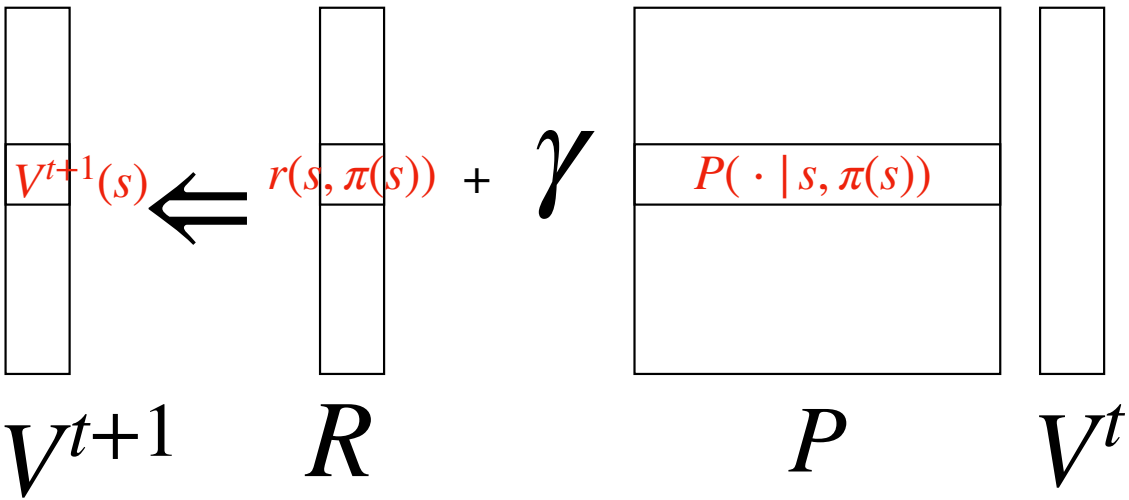
$V^t \approx V^\pi$

Iterative Policy Evaluation:

Algorithm (Iterative PE):

Start with some initialization $V^0 \in \mathbb{R}^{|S|}$, **repeat for** $t = 0 \ldots$:

$$V^{t+1} \Leftarrow R + \gamma P V^t$$



Q: What's computation complexity per iteration?

Convergence of Iterative PE

Theorem:

Recall $\gamma \in [0,1)$. After t iterations, we have:

$$\forall s, \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty$$

$\max_s |V^0(s) - V^\pi(s)|$

$\gamma < 1$

$\gamma = 0.99$

$(0.99)^{500} \approx 0$

$V^{t+1} = R + \gamma P \cdot V^t$

Convergence of Iterative PE

Theorem:

Recall $\gamma \in [0,1)$. After t iterations, we have:

$$\forall s, \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty$$

$\forall s, \left| V^{t+1}(s) - V^\pi(s) \right|$

Convergence of Iterative PE

Theorem:

Recall $\gamma \in [0,1)$. After t iterations, we have:

$$\forall s, \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty$$

$$\forall s, \left| V^{t+1}(s) - V^\pi(s) \right|$$

$$= \left| r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \left(r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right) \right|$$

$V^\pi(s) =$

Alg

Convergence of Iterative PE

Theorem:

Recall $\gamma \in [0,1)$. After t iterations, we have:

$$\forall s, \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty$$

$$\begin{aligned}
&\forall s, \left| V^{t+1}(s) - V^\pi(s) \right| \\
&= \left| r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \left(r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right) \right| \\
&= \gamma \left| \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right|
\end{aligned}$$

Convergence of Iterative PE

Theorem:

Recall $\gamma \in [0,1)$. After t iterations, we have:

$$\forall s, \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty$$

$$\forall s, \left| V^{t+1}(s) - V^\pi(s) \right|$$

$$= \left| r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \left(r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right) \right|$$

$$= \gamma \left| \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right|$$

$$\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} \left| V^t(s') - V^\pi(s') \right|$$

Convergence of Iterative PE

Theorem:

Recall $\gamma \in [0,1)$. After t iterations, we have:

$$\forall s, \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty$$

$$\begin{aligned}
&\forall s, \left| V^{t+1}(s) - V^\pi(s) \right| \\
&= \left| r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \left(r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right) \right| \\
&= \gamma \left| \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right| \\
&\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} \left| V^t(s') - V^\pi(s') \right| \leq \gamma \left\| V^t - V^\pi \right\|_\infty
\end{aligned}$$

$\mathbb{E}|f(x)| \leq \max_x |f(x)|$

Convergence of Iterative PE

Theorem:

Recall $\gamma \in [0,1)$. After t iterations, we have:

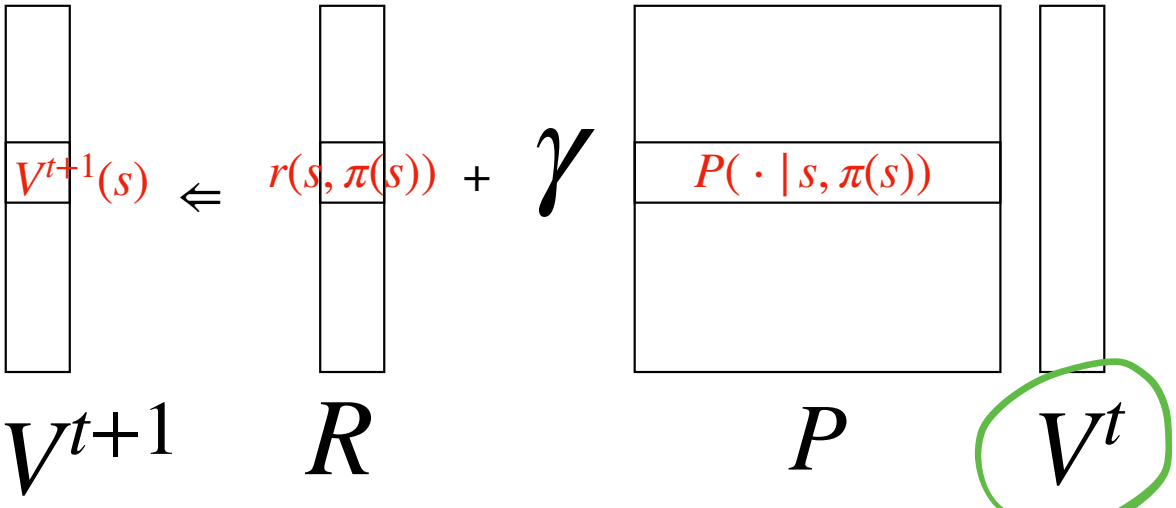
$$\forall s, \left| V^t(s) - V^\pi(s) \right| \leq \gamma^t \left\| V^0 - V^\pi \right\|_\infty$$

$$\begin{aligned}
&\forall s, \left| V^{t+1}(s) - V^\pi(s) \right| \\
&= \left| r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \left(r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right) \right| \\
&= \gamma \left| \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right| \\
&\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} \left| V^t(s') - V^\pi(s') \right| \leq \gamma \left\| V^t - V^\pi \right\|_\infty
\end{aligned}$$

$$\Rightarrow \left\| V^{t+1} - V^\pi \right\|_\infty \leq \gamma \left\| V^t - V^\pi \right\|_\infty$$

$$\leq \gamma^2 \| V^{t-1} - V^\pi \|_\infty$$

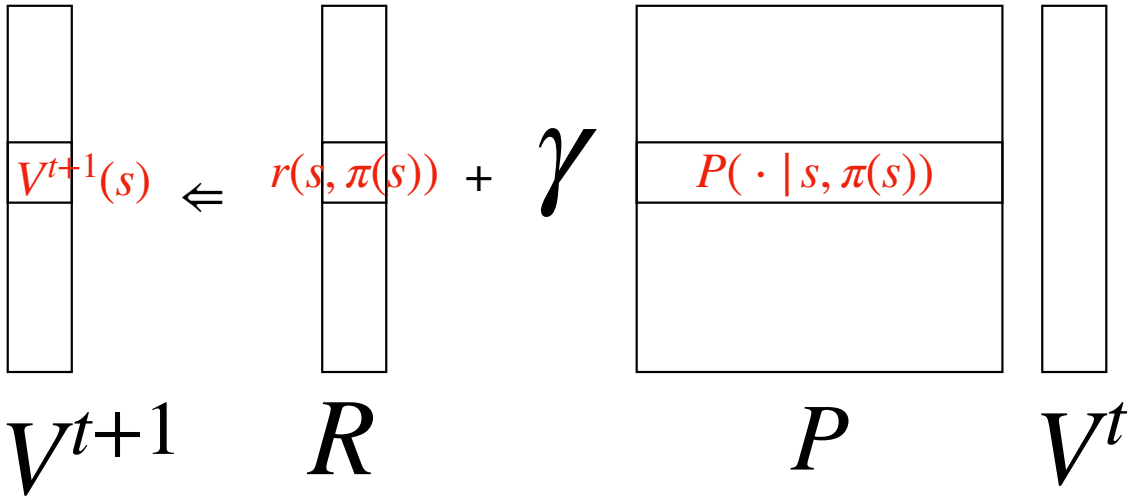
Summary so far:



V^0

$\| V^{t+1} - V^t \|_\infty \approx 0$

Summary so far:



Convergence:

$$\| V^{t+1} - V^{\pi} \|_{\infty} \leq \gamma \| V^{t} - V^{\pi} \|_{\infty} \leq \gamma^{t+1} \| V^{0} - V^{\pi} \|_{\infty}$$

Extension to stochastic policy and Q functions

Your homework:

How to modify the two algorithms so that it can handle stochastic policy and learn Q functions

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathop{E}_{s' \sim p(s,a)} V^{\pi}(s')$$

Summary

Key Question today: Given MDP $\mathscr{M}$, and a policy π, How to compute $V^{\pi}(s), \forall s$?

Summary

Key Question today: Given MDP $\mathscr{M}$, and a policy π, How to compute $V^{\pi}(s), \forall s$?

1. The **exact** algorithm $V = (I - \gamma P)^{-1} R$ requires matrix inverse (computation complexity at least $O(S^3)$)

Summary

Key Question today: Given MDP $\mathcal{M}$, and a policy π, How to compute $V^{\pi}(s), \forall s$?

1. The **exact** algorithm $V = (I - \gamma P)^{-1} R$ requires matrix inverse (computation complexity at least $O(S^3)$)

2. Iterative algorithm can quickly find an **approximate** solution (error shrinks in the rate of γ^t)

Summary

Key Question today: Given MDP $\mathscr{M}$, and a policy π, How to compute $V^{\pi}(s), \forall s$?

1. The **exact** algorithm $V = (I - \gamma P)^{-1} R$ requires matrix inverse (computation complexity at least $O(S^3)$)

2. Iterative algorithm can quickly find an **approximate** solution (error shrinks in the rate of γ^t)

(We will see many similar iterative algorithms later)