Policy Evaluation

Announcements

HW0 is out today (due March 2rd, 11:59 pm ET)

(Gradescope entry code: BP3B3N)

Office hours start this week:

Wen: Tuesday and Thursday, 10:55am - 11:30am
Wen-Ding: Friday 3pm-4pm
Hadi: Wednesday 2:30-3:30pm

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 $\pi : S \mapsto A$

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 $\pi : S \mapsto A$

Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| s_0 = s, a_h = \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h) \right]$

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 $\pi : S \mapsto A$

Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| s_0 = s, a_h = \pi(s_h), s_{h+1} \sim P(\cdot \,|\, s_h, a_h)\right]$

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

Recap: Optimal Policy

For Discounted infinite horizon MDP, $\exists$ a deterministic policy $\pi^\star : S \mapsto A$:

$$V^\star(s) \geq V^\pi(s), \forall s, \forall \pi$$

Recap: Optimal Policy

For Discounted infinite horizon MDP, $\exists$ a deterministic policy $\pi^\star : S \mapsto A$:

$$V^\star(s) \geq V^\pi(s), \forall s, \forall \pi$$

Bellman Optimality (DP):

1. For $V^\star$, we have $V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right], \forall s$

Recap: Optimal Policy

For Discounted infinite horizon MDP, $\exists$ a deterministic policy $\pi^\star : S \mapsto A$:

$$V^\star(s) \geq V^\pi(s), \forall s, \forall \pi$$

Bellman Optimality (DP):

1. For $V^\star$, we have $V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right], \forall s$

2. For V that satisfies $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right], \forall s$,
we have $V(s) = V^\star(s), \forall s$

Recap: Optimal Policy

For Discounted infinite horizon MDP, $\exists$ a deterministic policy $\pi^\star : S \mapsto A$:

$$V^\star(s) \geq V^\pi(s), \forall s, \forall \pi$$

Bellman Optimality (DP):

1. For $V^\star$, we have $V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right], \forall s$

2. For V that satisfies $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right], \forall s,$
we have $V(s) = V^\star(s), \forall s$

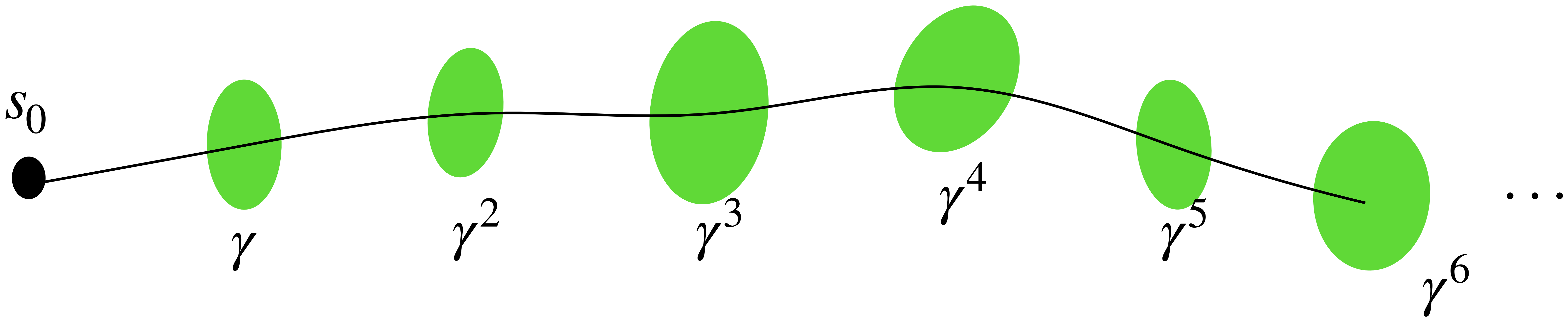
In HW0, we will study Bellman Optimality for $Q^\star / Q$

Recap: State-action distribution

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

Recap: State-action distribution

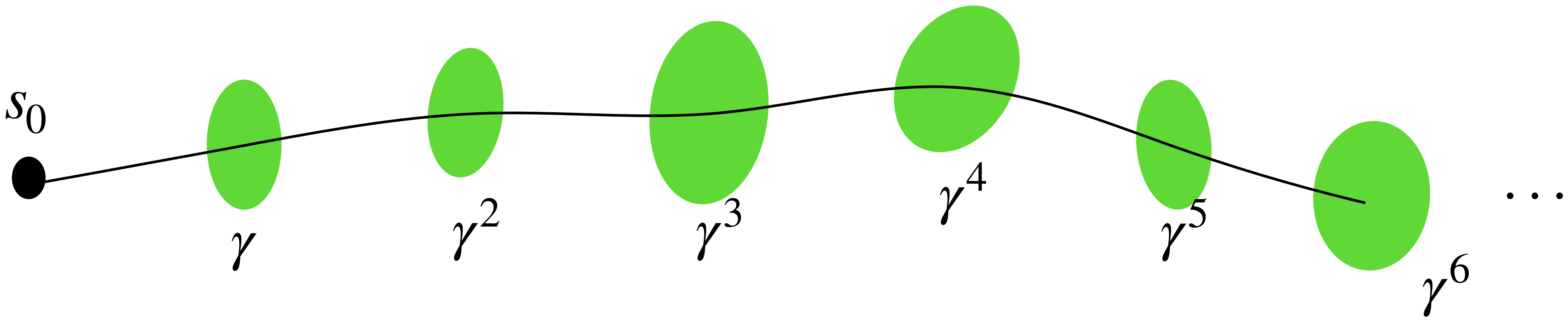
$$\mathbb{P}_h^{\pi}(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^{\pi}(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$



Recap: State-action distribution

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

$$d_{s_0}^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a; s_0)$$



Today: Policy Evaluation

Key Question:

Given MDP $\mathscr{M} = (S, A, r, P, \gamma)$ **& 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 opponent
(we can abstract our strategy as policy π)

Motivation for Policy Evaluation

We want to **evaluate** our strategy against some opponent
(we can abstract our strategy as policy π)

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 $\mathcal{M} = (S, A, P, \gamma, r)$, and policy π, we want to compute V^π

Exact Policy Evaluation

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

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

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

Exact Policy Evaluation

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

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

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

This gives us S many linear constraints

Exact Policy Evaluation

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

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

Exact Policy Evaluation

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

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

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 $P_{s,s'} = P(s' | s, \pi(s))$,

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

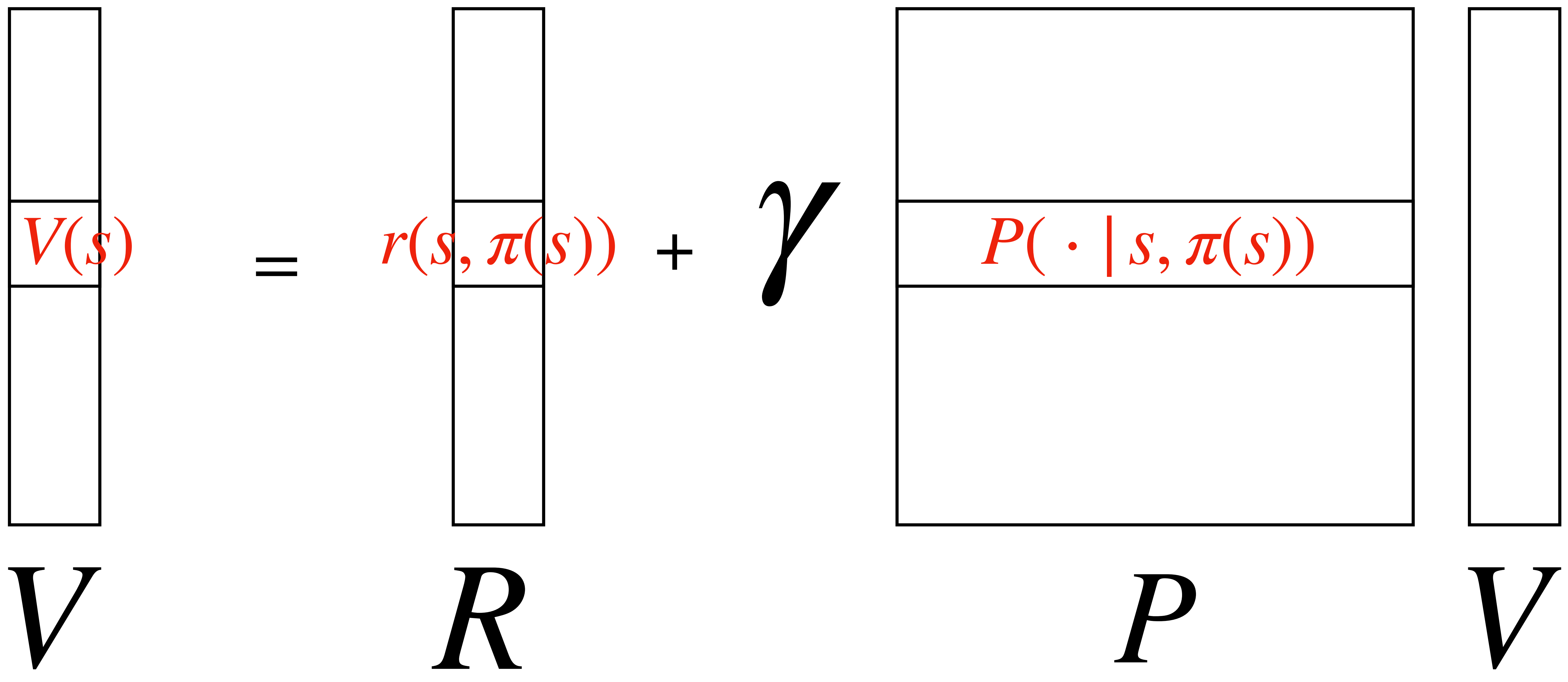
$$V = R + \gamma P V$$

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:

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



Exact Policy Evaluation

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

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

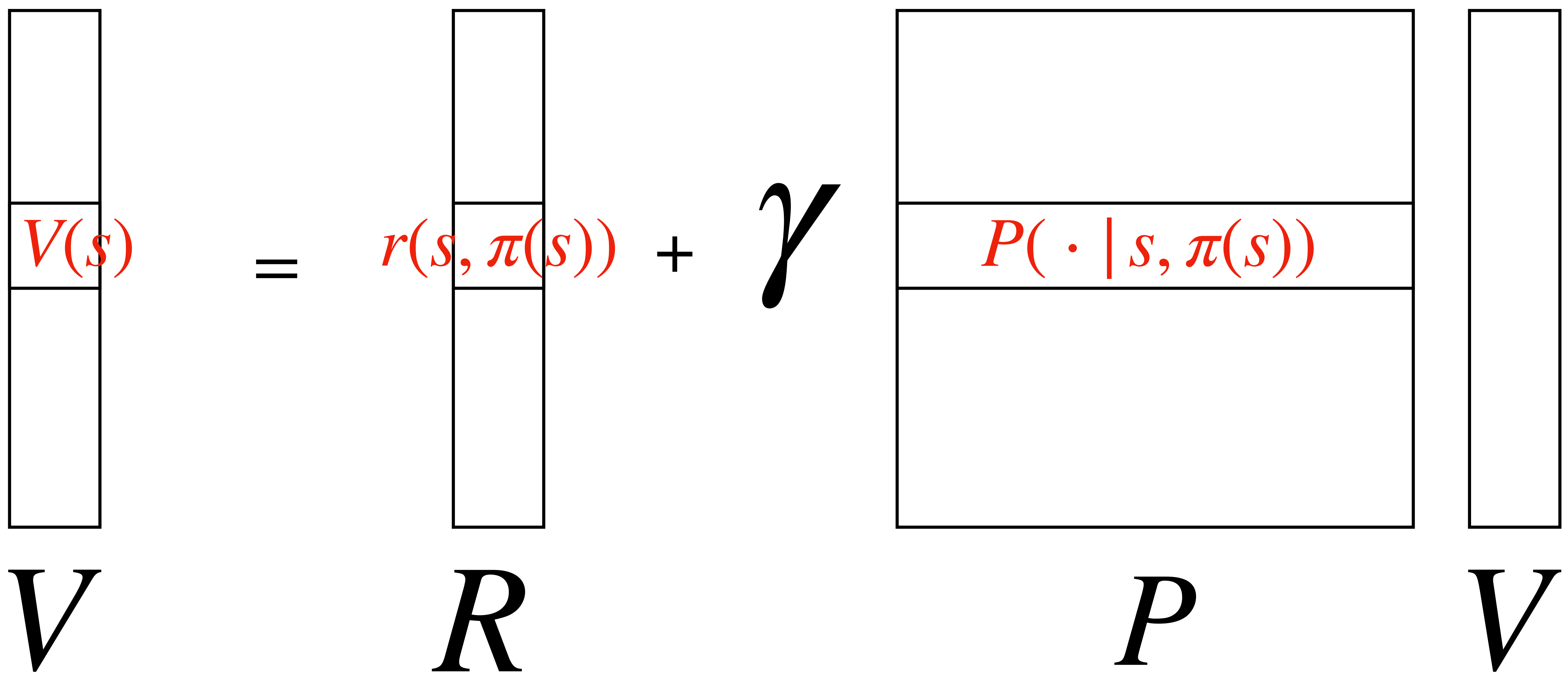
Exact Policy Evaluation

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

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

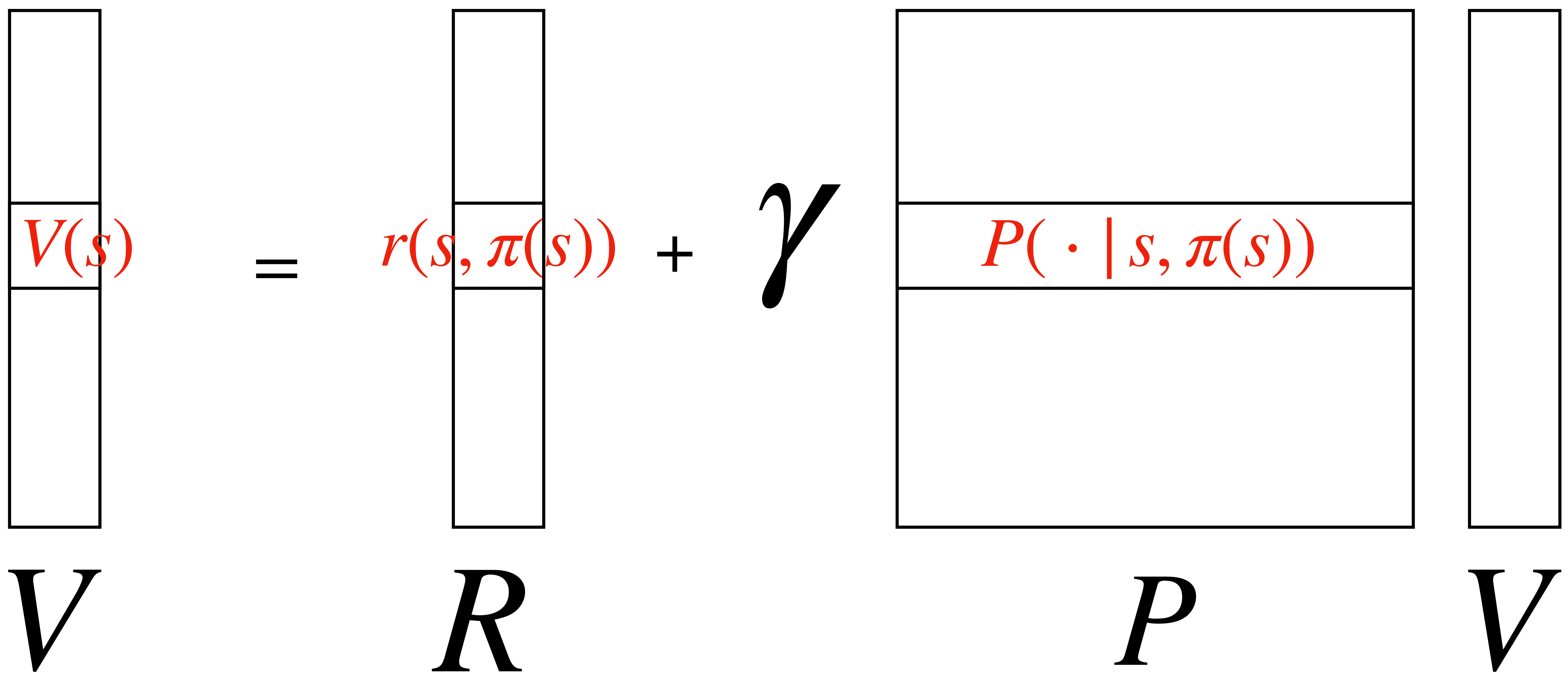
In HW0, we will show that $(I - \gamma P)$ is full rank (thus invertible)

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:

✓ 1. Exact Policy Evaluation

2. Approximate Policy Evaluation via an Iterative Algorithm

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

Detour: fix-point solution

Consider $x^\star = f(x^\star), \quad f : [a, b] \mapsto [a, b]$

Detour: fix-point solution

Consider $x^\star = f(x^\star), \quad f : [a, b] \mapsto [a, b]$

Common approach to find $x^\star$:

Initialize $x^0 \in [a, b]$, repeat: $x^{t+1} = f(x^t)$

Detour: fix-point solution

Consider $x^\star = f(x^\star), \quad f: [a,b] \mapsto [a,b]$

Common approach to find $x^\star$:

Initialize $x^0 \in [a,b]$, repeat: $x^{t+1} = f(x^t)$

If f is a contraction mapping,

i.e., $\forall x, x', \; |f(x) - f(x')| \leq \gamma |x - x'|$, for some $\gamma \in [0,1)$, then:

$x^t \to x^\star$, as $t \to \infty$

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^π is a fix-point solution:

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

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

Iterative Policy Evaluation:

Algorithm (Iterative PE):

Start with some initialization $V^0 \in \left[0, 1/(1-\gamma)\right]^{|S|}$, **repeat for** $t = 0\ldots$:

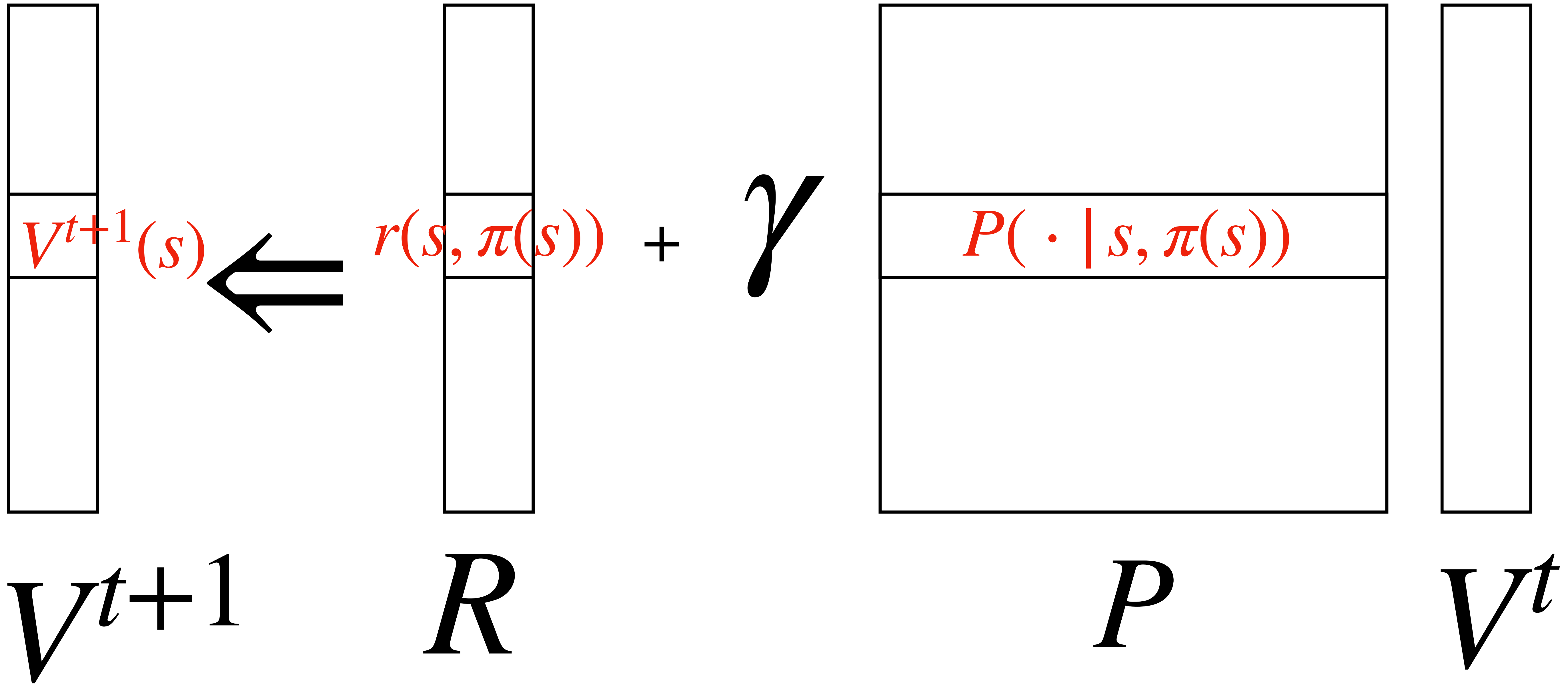
$$V^{t+1} \Leftarrow R + \gamma P V^t$$

Iterative Policy Evaluation:

Algorithm (Iterative PE):

Start with some initialization $V^0 \in \left[0,\ 1/(1-\gamma)\right]^{|S|}$, **repeat for** $t = 0\ldots$:

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

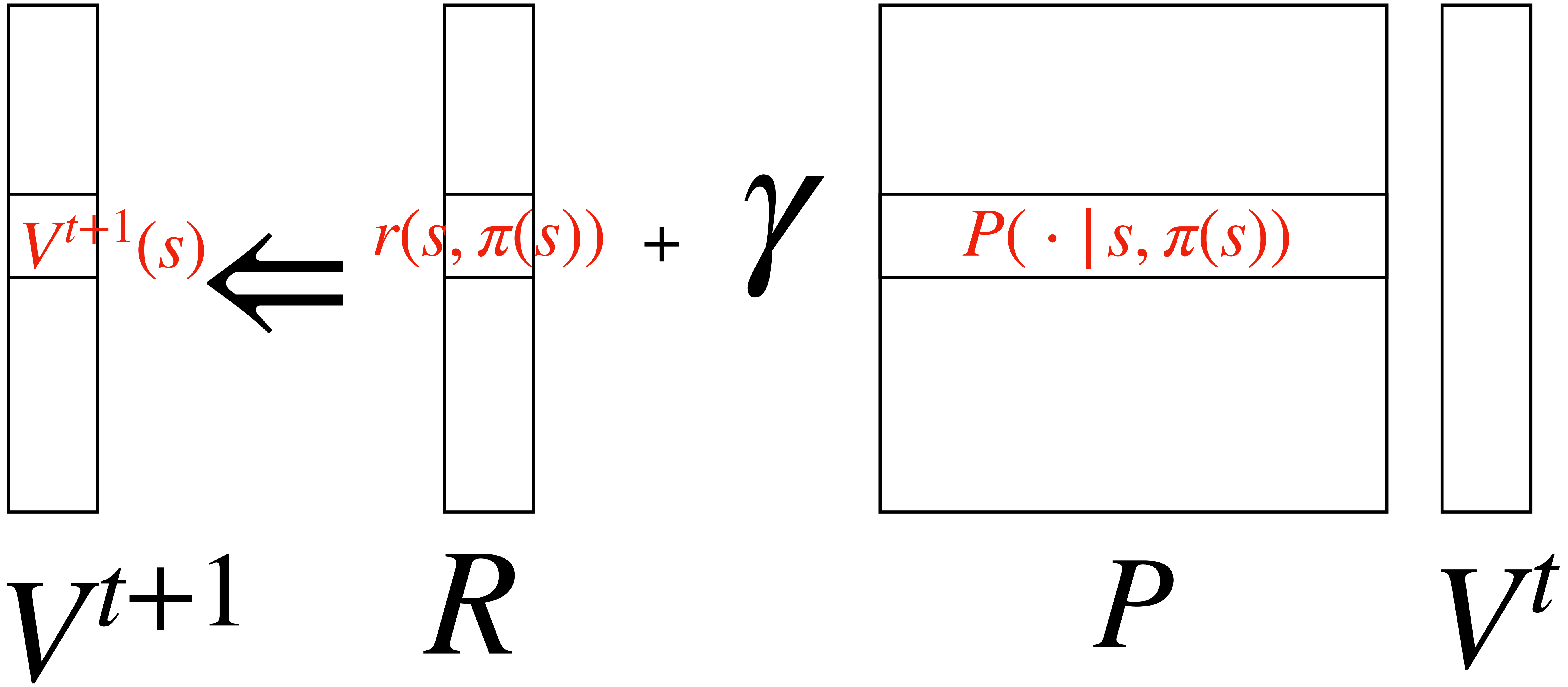


Iterative Policy Evaluation:

Algorithm (Iterative PE):

Start with some initialization $V^0 \in \left[0,\, 1/(1-\gamma)\right]^{|S|}$, **repeat for** $t = 0\ldots$:

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



Q: What's computation complexity per iteration?

Iterative Policy Evaluation:

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

$$V^{t+1}(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi(s))} V^{t}(s')$$

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

Convergence of Iterative PE

Theorem:

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

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

Convergence of Iterative PE

Theorem:

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

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

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

Convergence of Iterative PE

Theorem:

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

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

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

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

Convergence of Iterative PE

Theorem:

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

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

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

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

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

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:

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

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

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

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

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

$$\leq \gamma \parallel V^t - V^\pi \parallel_\infty$$

Convergence of Iterative PE

Theorem:

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

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

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

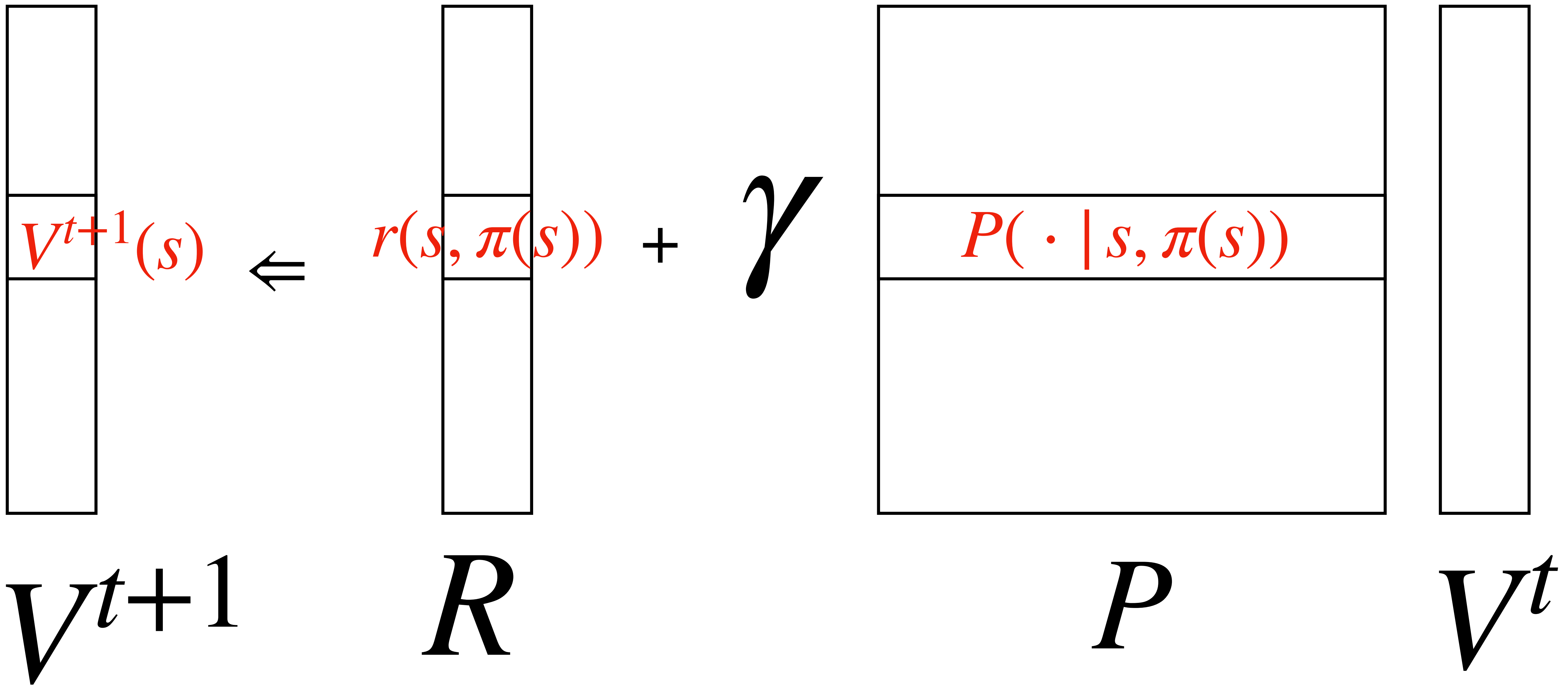
$$= \left| r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^t(s') - \left(r(s,\pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s') \right) \right|$$

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

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

$$\leq \gamma \parallel V^t - V^\pi \parallel_\infty \quad \Rightarrow \parallel V^{t+1} - V^\pi \parallel_\infty \leq \gamma \parallel V^t - V^\pi \parallel_\infty$$

Summary so far:



Convergence:

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

Outline:

1. Exact Policy Evaluation

2. Approximate Policy Evaluation via an Iterative Algorithm

Summary

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

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 $O(S^3)$

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 $O(S^3)$

1. For iterative PE algorithm, to find a ϵ accurate value function, we need # of iterations:

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 $O(S^3)$

1. For iterative PE algorithm, to find a ϵ accurate value function, we need # of iterations:

$$\ln\left(\frac{\| V^0 - V^\star \|_\infty}{\epsilon}\right) \Big/ \ln(1/\gamma)$$

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 $O(S^3)$

1. For iterative PE algorithm, to find a ϵ accurate value function, we need # of iterations:

$$\ln\left(\frac{\| V^0 - V^\star \|_\infty}{\epsilon}\right) \Big/ \ln(1/\gamma)$$

Computation wise, we need $O\left(S^2 \ln\left(\frac{1}{\epsilon}\right)\right)$

Summary

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

Bellman Equation

↓

A fix-point equation:

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

Summary

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

Bellman Equation

↓

A fix-point equation:

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

Fix-point iteration framework

Summary

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

Bellman Equation

↓

A fix-point equation:

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

→

Alg: Iterative PE

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

←

Fix-point iteration framework

Summary

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

Bellman Equation

↓

A fix-point equation:

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

Contraction

Alg: Iterative PE

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

Fix-point iteration framework

Summary

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

Bellman Equation

A fix-point equation:

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

Contraction

Theorem

$$\|V^t - V^\pi\|_\infty \leq \gamma^t \|V^0 - V^\pi\|_\infty$$

Alg: Iterative PE

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

Fix-point iteration framework

Next two lectures:

Given MDP $\mathscr{M}$, how to compute the optimal policy $\pi^\star$, and $V^\star$