Policy Evaluation

Announcements

P1 is delayed, will be released on Monday

HW0 due today

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

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

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

Relationship between V and Q

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

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 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 $\mathcal{M} = (S, A, P, \gamma, r)$, 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 $s \in S$

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

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

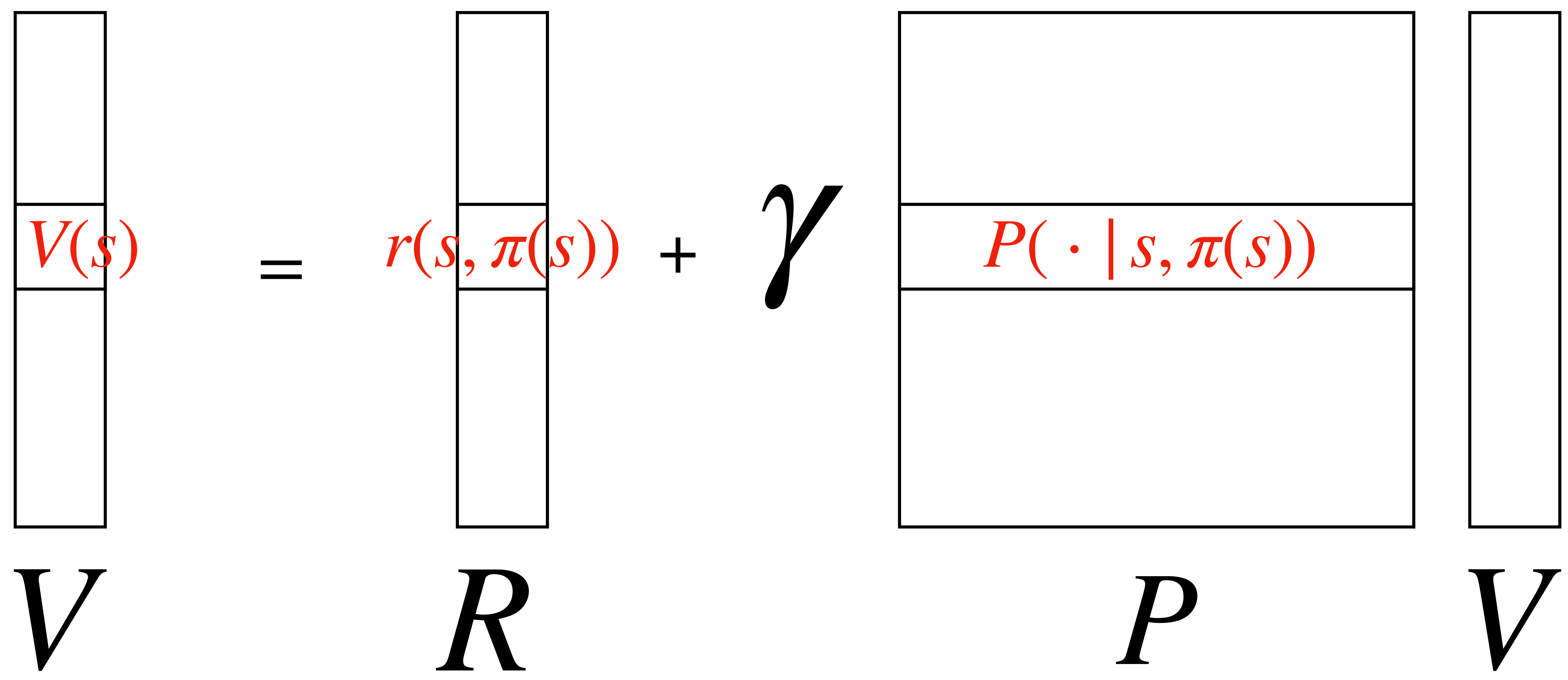
we can **combine all S many constraints together:**

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

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:

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



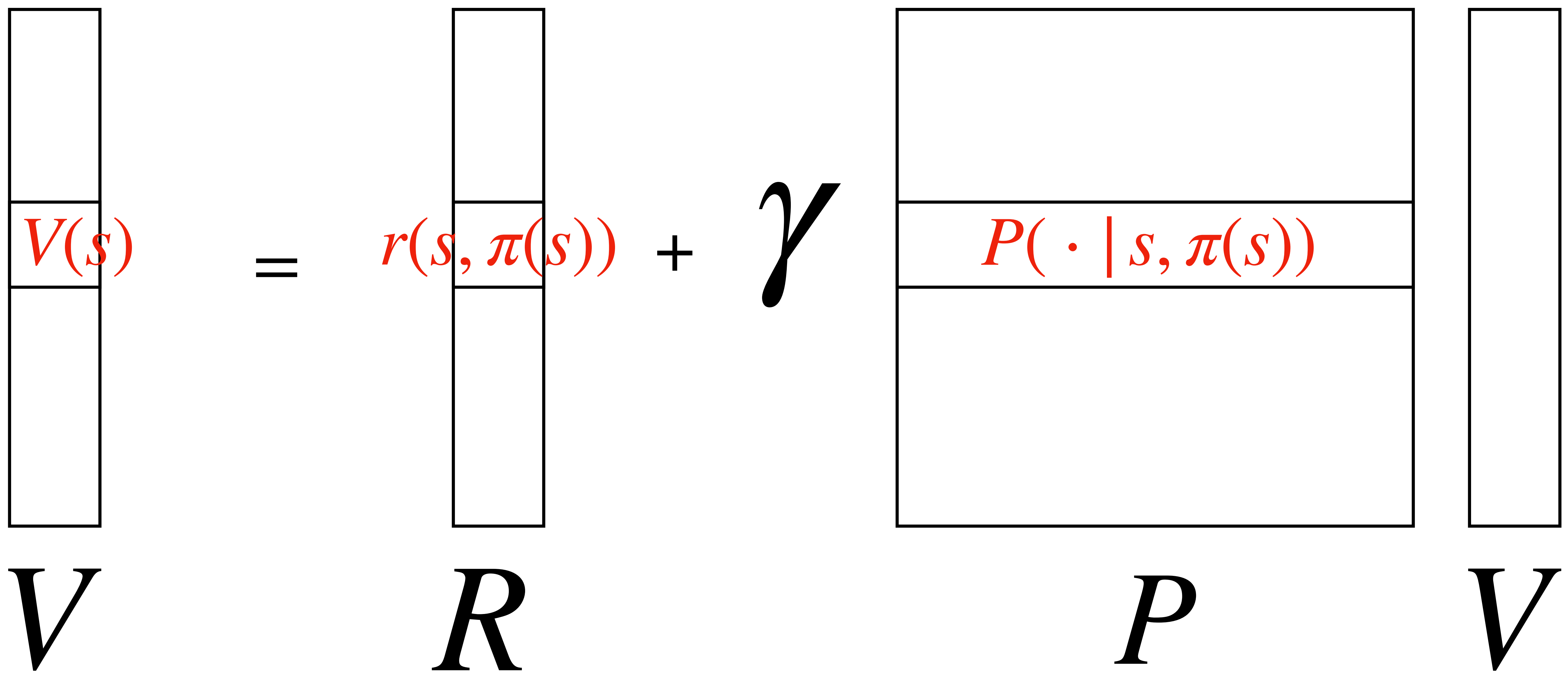
Exact Policy Evaluation

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:



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

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)

V^π is a fix-point solution

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

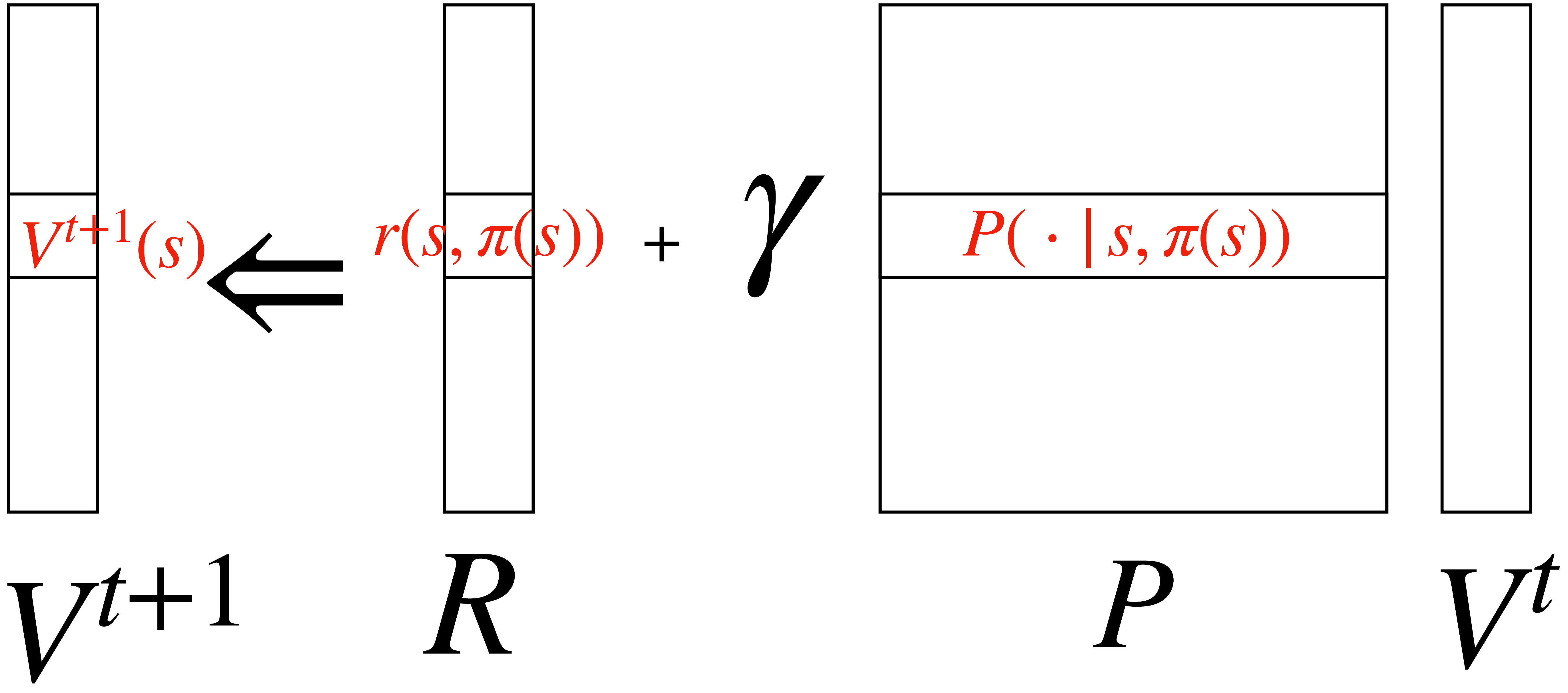
$$V^\pi = R + \gamma P V^\pi$$

Iterative Policy Evaluation:

Algorithm (Iterative PE):

Start with some initialization $V^0 \in \mathbb{R}^{|S|}$, **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:

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

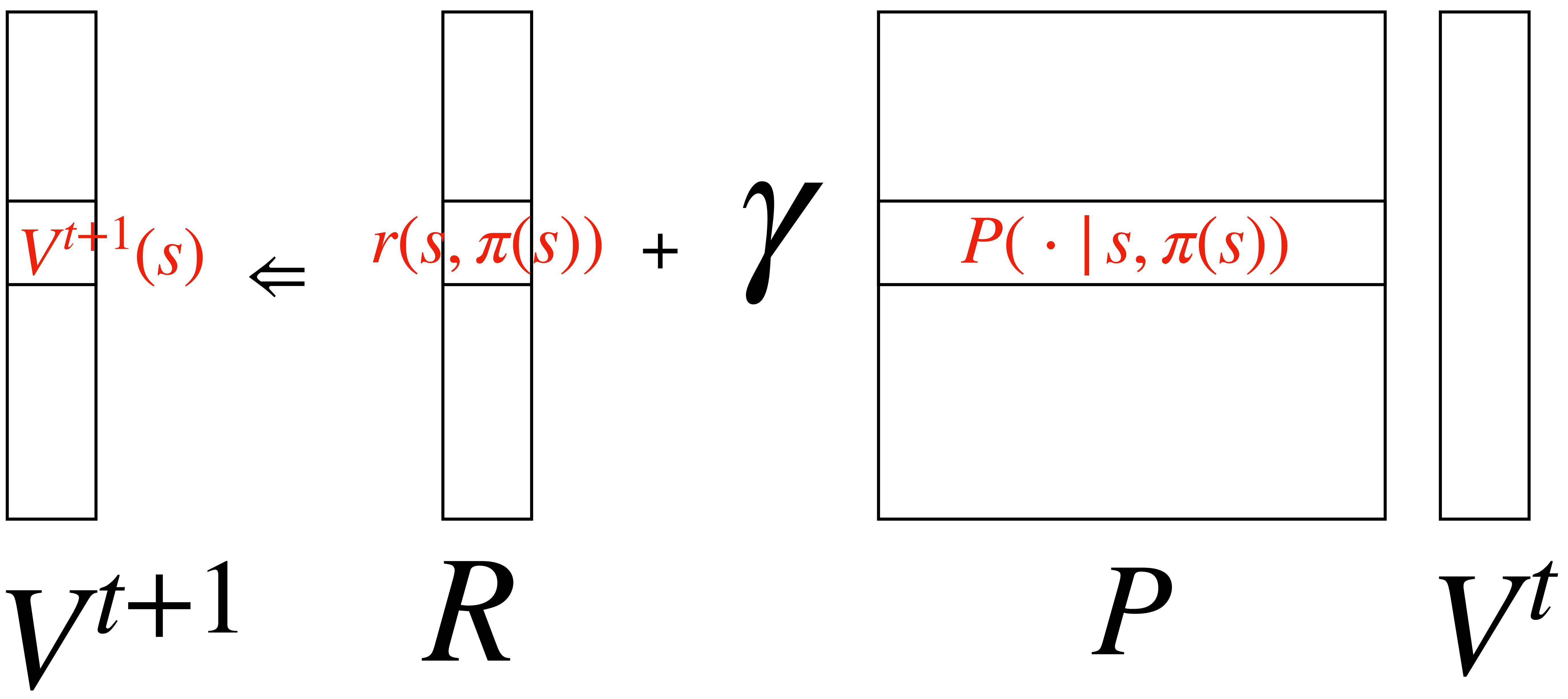
$$\forall s, \left| V^{t+1}(s) - V^\pi(s) \right|$$

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

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

$$\leq \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} \left| V^t(s') - V^\pi(s') \right| \leq \gamma \parallel V^t - V^\pi \parallel_\infty \Rightarrow \parallel V^{t+1} - V^\pi \parallel_\infty \leq \gamma \parallel V^t - V^\pi \parallel_\infty$$

Summary so far:



Convergence:

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

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

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)