DYNAMIC PROGRAMMING

RECAP

WHERE WE ARE

Goal: find an optimal policy for an MDP

$\pi \in \Pi$ Search the space of policy for an optimal one

$v_\pi(s) \doteq \mathbb{E}[G_t \mid S_t = s]$ Quality of the policy in a state s

$$v_\pi(s) = \sum_a \left(\pi(a \mid s) r(s,a) + \gamma \sum_{s'} p(s,a,s') v^\pi(s') \right)$$ simpler expression of value

Compute v_π by solving a system of equations (becomes expensive if there are many states)

DIRECTION

THIS WEEK

1. An iterative method for computing v_π
2. Ways to use estimates of v_π to improve π

THE VALUE OF AN ACTION

DEPENDS ON THE POLICY

$q_\pi(s, a) \doteq \mathbb{E}[G_t \,|\, S_t = s, A_t = a]$

In Bandits, the optimal action was constant $q_*(a) = \mathbb{E}[R_t \,|\, A_t = a]$

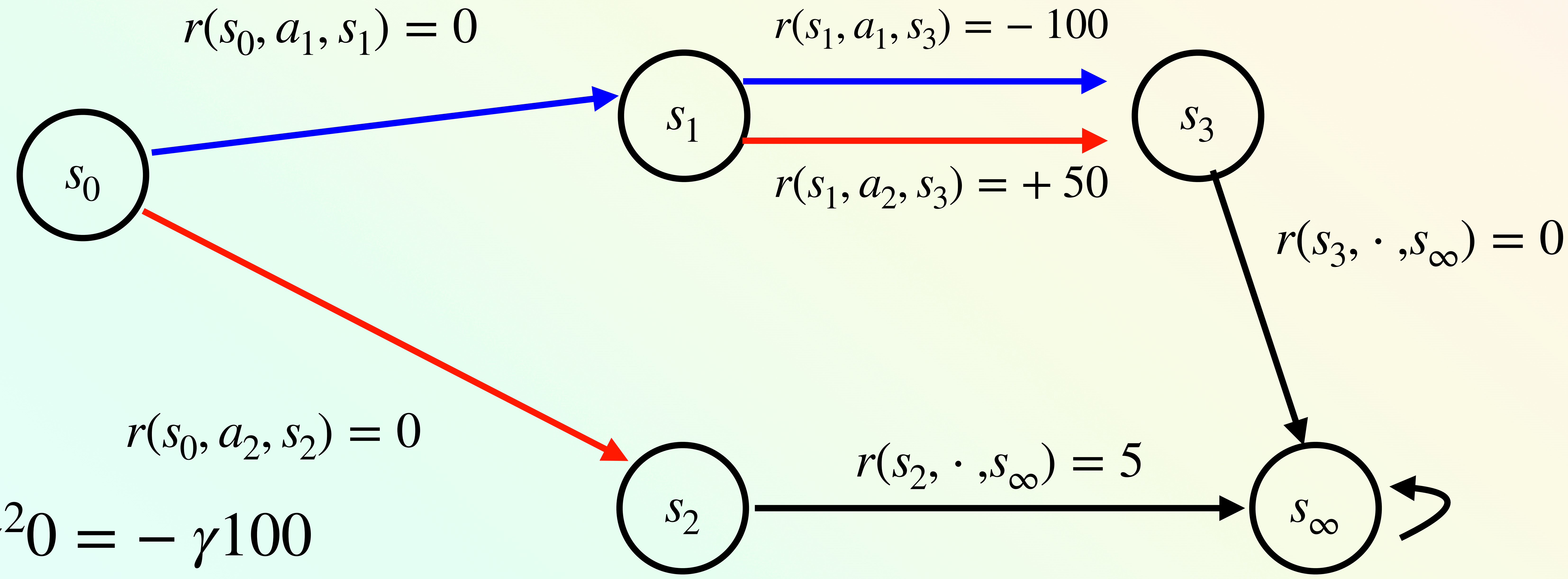
In MDPs an action can be both the best and worst choice

$a^* \in \arg\max_a q_*(s, a) \quad a^*$ is an optimal action

$\min_\pi q_\pi(s, a^*) = \min_\pi v_\pi(s) \quad a^*$ could also minimize the value of the state

THE VALUE OF AN ACTION

DEPENDS ON THE POLICY



$\pi(s_0) = a_1, \pi(s_1) = a_1$

$q_\pi(s_0, a_1) = 0 + \gamma(-100) + \gamma^2 0 = -\gamma 100$

$\pi(s_0) = a_1, \pi(s_1) = a_2$

$q_\pi(s_0, a_1) = 0 + \gamma 50 + \gamma^2 = \gamma 50$

TWO SETTINGS

CONTROL VS EVALUATION

Policy Evaluation — compute v_π or q_π

- Prediction settings: predict the value of state or action (supervised learning)

Policy Improvement — $\pi \rightarrow \pi'$ such that $\pi' > \pi$

- $\pi \rightarrow \pi' \rightarrow \ldots \rightarrow \pi_*$
- Control settings: searching for better (or optimal) policies
- *This is the settings we ultimately care about*

Policy evaluation is a useful step in policy improvement

TWO SETTINGS

CONTROL VS EVALUATION

Policy Evaluation — compute v_π or q_π

- Prediction settings: predict the value of state or action (supervised learning)

Policy Improvement — $\pi \to \pi'$ such that $\pi' > \pi$

- $\pi \to \pi' \to \ldots \to \pi_*$
- Control settings: searching for better (or optimal) policies
- *This is the settings we ultimately care about*

Policy evaluation is a useful step in policy improvement

RECALL: FUNCTION

SYSTEM OF EQUATIONS

$$
\begin{aligned}
v_1 = &\ r_1 + \gamma p_{1,1} v_1 + \gamma p_{1,2} v_2 + \ldots + \gamma p_{1,n} v_n \\
v_2 = &\ r_2 + \gamma p_{2,1} v_1 + \gamma p_{2,2} v_2 + \ldots + \gamma p_{2,n} v_n \\
\vdots & \\
v_n = &\ r_n + \gamma p_{n,1} v_1 + \gamma p_{n,2} v_2 + \ldots + \gamma p_{n,n} v_n
\end{aligned}
$$

Solve n equations for n unknown variables so that $v = v_\pi$

RECALL: FUNCTION

SYSTEM OF EQUATIONS

$$
\begin{aligned}
v_1 \leftarrow\ & r_1 + \gamma p_{1,1} v_1 + \gamma p_{1,2} v_2 + \ldots + \gamma p_{1,n} v_n \\
v_2 \leftarrow\ & r_2 + \gamma p_{2,1} v_1 + \gamma p_{2,2} v_2 + \ldots + \gamma p_{2,n} v_n \\
\vdots\ & \\
v_n \leftarrow\ & r_n + \gamma p_{n,1} v_1 + \gamma p_{n,2} v_2 + \ldots + \gamma p_{n,n} v_n
\end{aligned}
$$

Solve n equations for n unknown variables so that $v = v_\pi$

Idea: update the left-hand side with the right-hand side

ITERATIVE POLICY EVALUATION

DEFINITIONS

$$v^k = \begin{bmatrix} v_1^k \\ \vdots \\ v_n^k \end{bmatrix} = \begin{bmatrix} v^k(s_1) \\ \vdots \\ v^k(s_n) \end{bmatrix}$$

v^k vector of value estimates for each state at iteration k

ITERATIVE POLICY EVALUATION

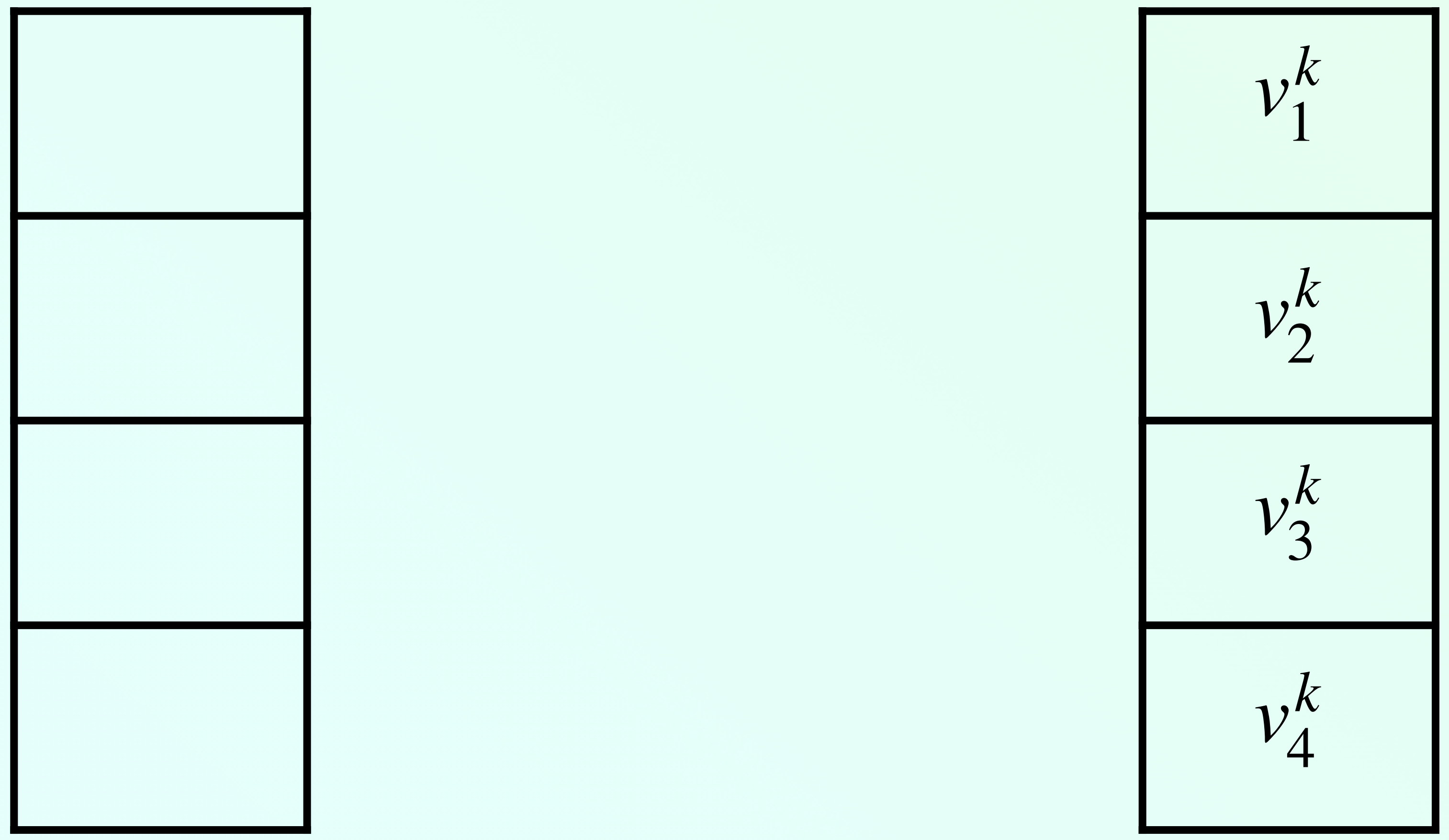
UPDATE

$$v_i^{\mathbf{k+1}} = \sum_a \pi(a \mid s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^{\mathbf{k}} \right)$$

ITERATIVE POLICY EVALUATION

FULL SWEEP

$$v_i^{\mathbf{k+1}} = \sum_a \pi(a \mid s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^{\mathbf{k}} \right)$$

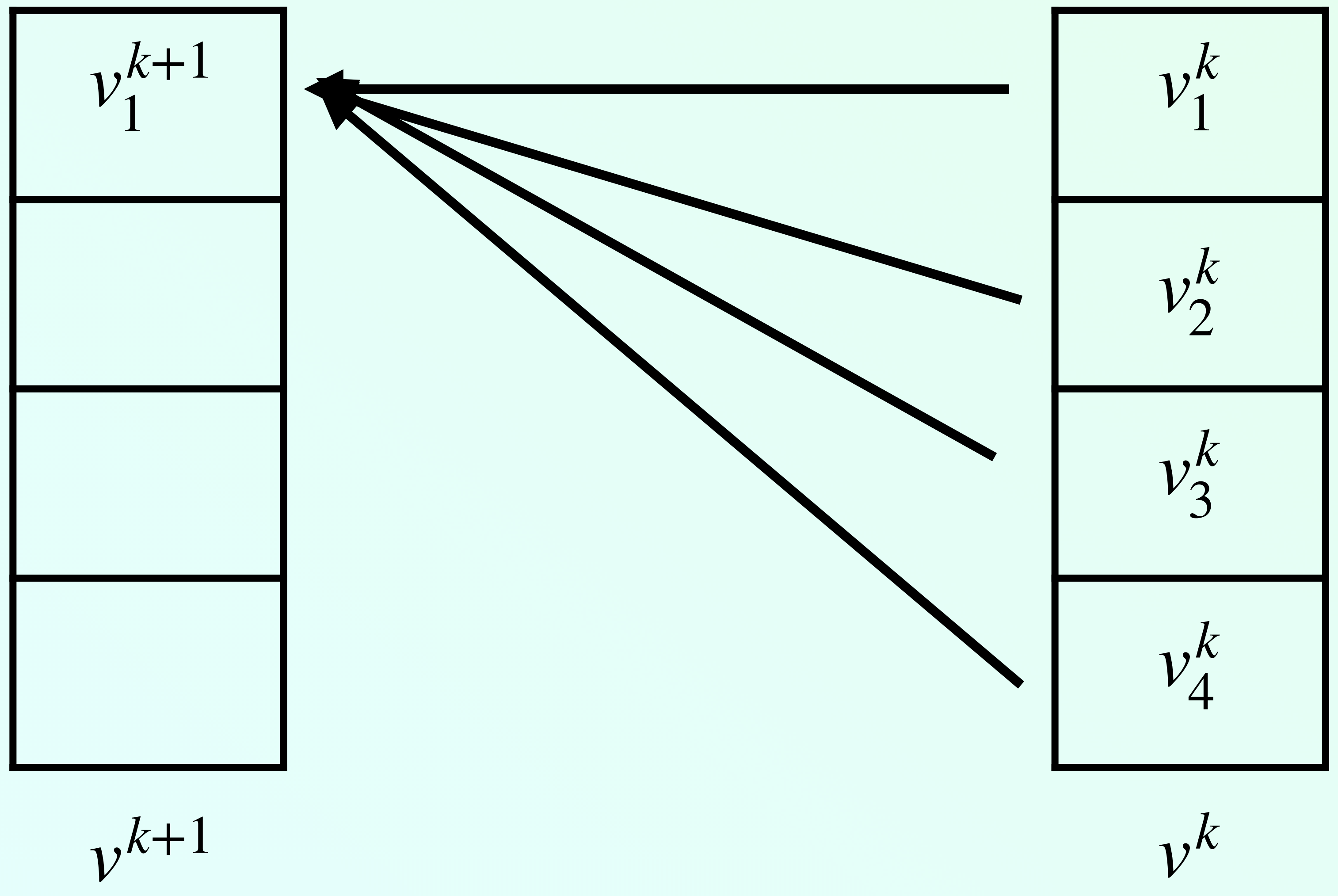


v^{k+1} v^k

ITERATIVE POLICY EVALUATION

FULL SWEEP

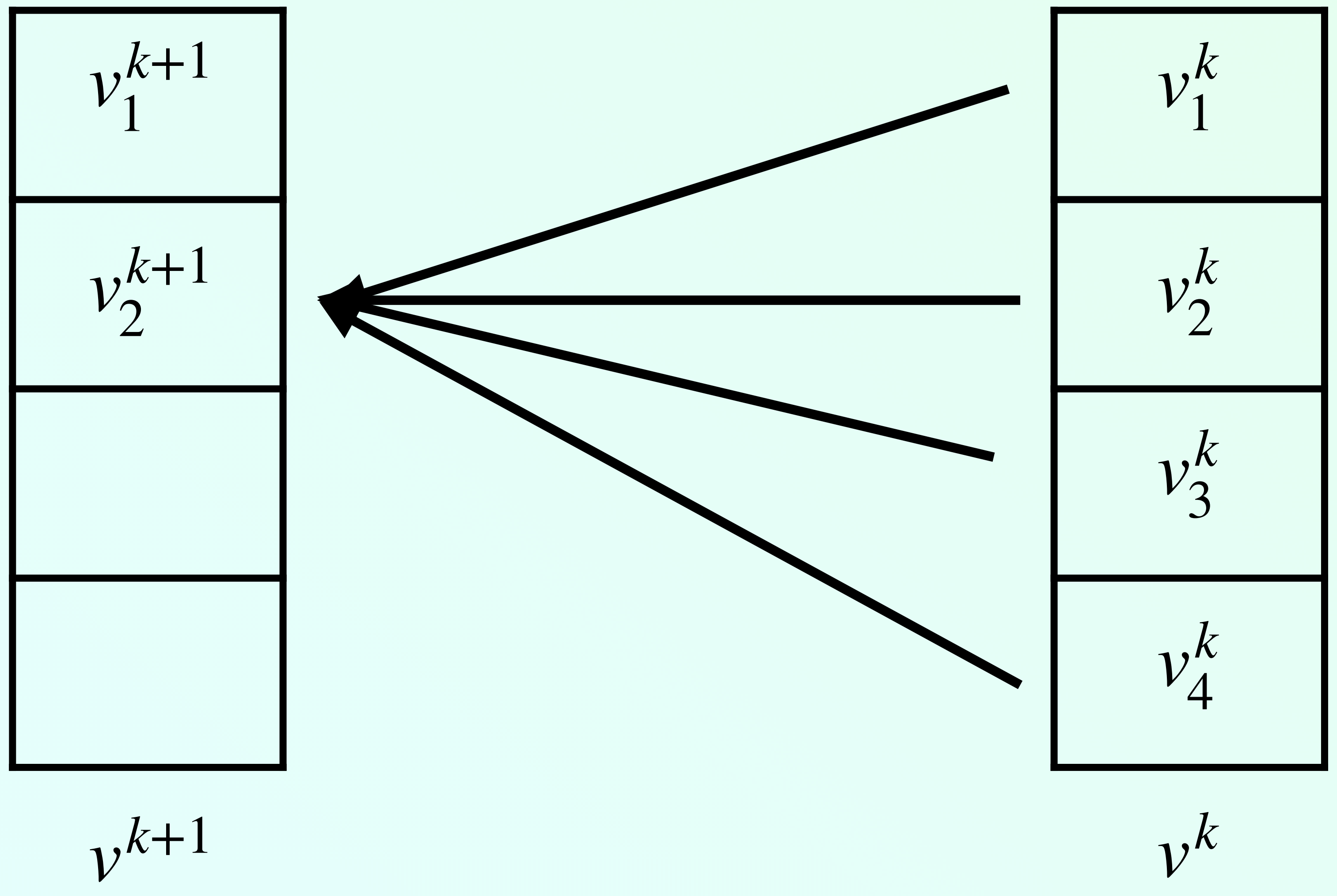
$$v_i^{\mathbf{k+1}} = \sum_a \pi(a \mid s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^{\mathbf{k}} \right)$$



ITERATIVE POLICY EVALUATION

FULL SWEEP

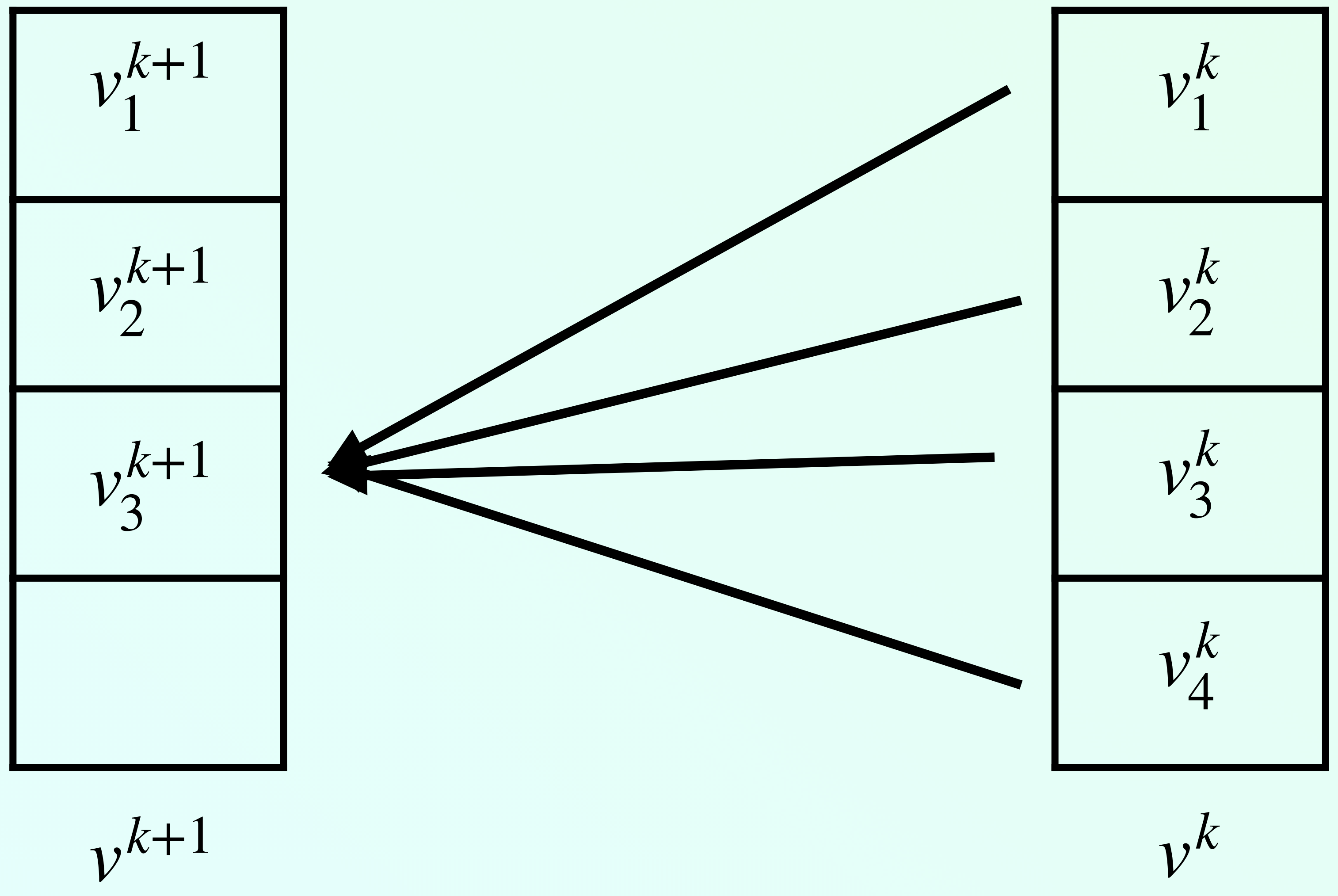
$$v_i^{\mathbf{k+1}} = \sum_a \pi(a \mid s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^{\mathbf{k}} \right)$$



ITERATIVE POLICY EVALUATION

FULL SWEEP

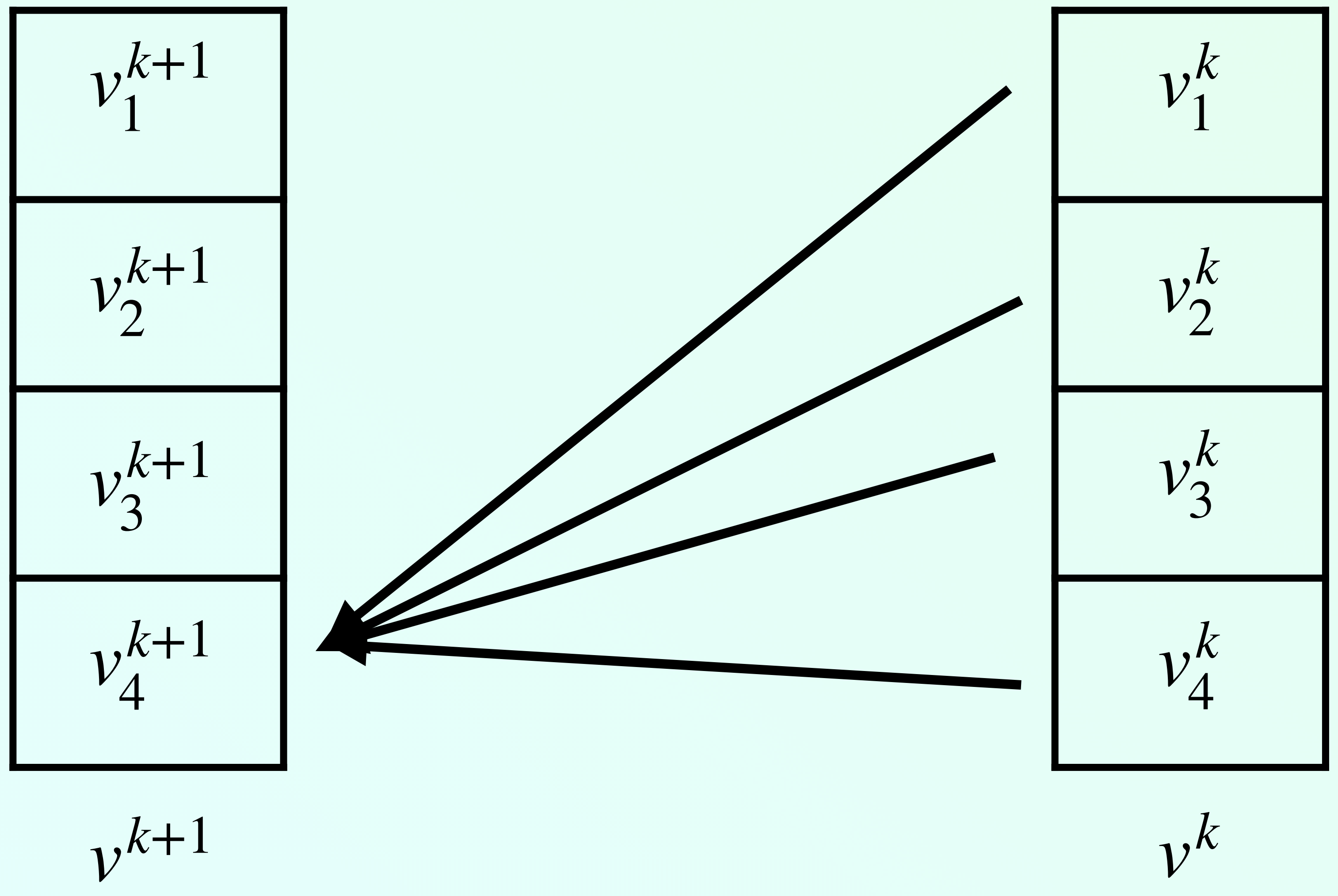
$$v_i^{\mathbf{k+1}} = \sum_a \pi(a \mid s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^{\mathbf{k}}\right)$$



ITERATIVE POLICY EVALUATION

FULL SWEEP

$$v_i^{\mathbf{k+1}} = \sum_a \pi(a \mid s_i) \left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^{\mathbf{k}} \right)$$



ITERATIVE POLICY EVALUATION

FULL SWEEP

$$v_i^{\mathbf{k+1}} = \sum_a \pi(a \mid s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^{\mathbf{k}} \right)$$

v^{k+2}

v_1^{k+1}
v_2^{k+1}
v_3^{k+1}
v_4^{k+1}

v^{k+1}

ITERATIVE POLICY EVALUATION

EXAMPLE

$$s_1 \xrightarrow{r(s_1,\,\cdot\,)=0} s_2 \xrightarrow{r(s_2,\,\cdot\,)=0} s_3 \xrightarrow{r(s_3,\,\cdot\,)=0} s_4 \xrightarrow{r(s_4,\,\cdot\,)=1} s_\infty$$

$$\underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{v^1} \qquad \underbrace{\begin{bmatrix} v_1^2 \\ v_2^2 \\ v_3^2 \\ v_4^2 \end{bmatrix}}_{v^2} \qquad \underbrace{\begin{bmatrix} \gamma^3 \\ \gamma^2 \\ \gamma \\ 1 \end{bmatrix}}_{v_\pi}$$

$$v_1^2 = r(s_1,\,\cdot\,) + \gamma v_2^1 = 0 + \gamma 0 = 0$$

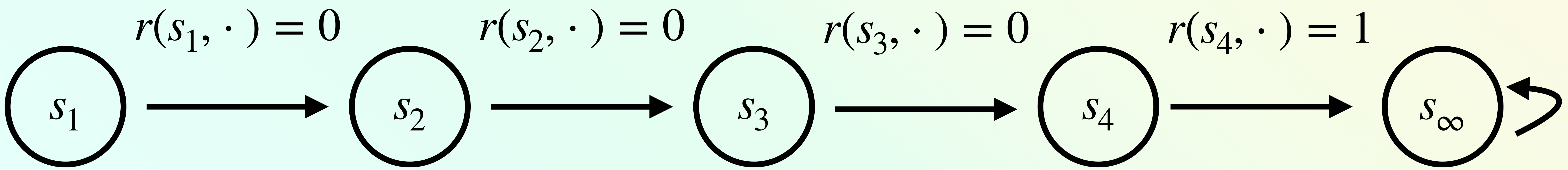
$$v_2^2 = r(s_2,\,\cdot\,) + \gamma v_3^1 = 0 + \gamma 0 = 0$$

$$v_3^2 = r(s_3,\,\cdot\,) + \gamma v_4^1 = 0 + \gamma 0 = 0$$

$$v_4^2 = r(s_4,\,\cdot\,) + \gamma v_\pi(s_\infty) = 1 + \gamma 0 = 1$$

ITERATIVE POLICY EVALUATION

EXAMPLE



v^1	v^2	v_π
0	0	γ^3
0	0	γ^2
0	0	γ
0	1	1

$$v_1^2 = r(s_1, \cdot\,) + \gamma v_2^1 = 0 + \gamma 0 = 0$$

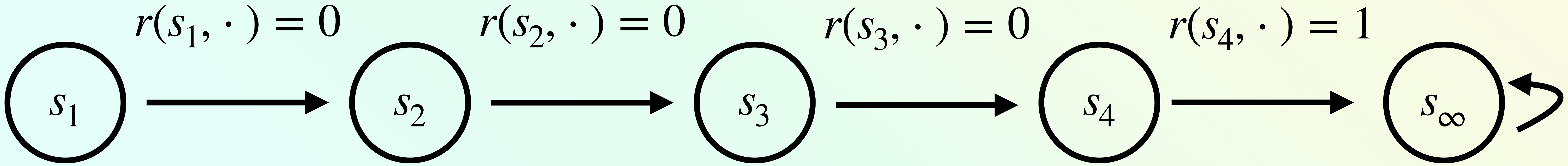
$$v_2^2 = r(s_2, \cdot\,) + \gamma v_3^1 = 0 + \gamma 0 = 0$$

$$v_3^2 = r(s_3, \cdot\,) + \gamma v_4^1 = 0 + \gamma 0 = 0$$

$$v_4^2 = r(s_4, \cdot\,) + \gamma v_\pi(s_\infty) = 1 + \gamma 0 = 1$$

ITERATIVE POLICY EVALUATION

EXAMPLE



v^1	v^2	v^3	v_π
0	0	v_1^3	γ^3
0	0	v_3^3	γ^2
0	0	v_4^3	γ
0	1	v_4^3	1

$$v_1^3 = 0 + \gamma v_2^2 = 0$$

$$v_2^3 = 0 + \gamma v_3^2 = 0$$

$$v_3^3 = 0 + \gamma v_4^2 = 0 + \gamma$$

$$v_4^3 = 1 + \gamma 0 = 1$$

ITERATIVE POLICY EVALUATION

EXAMPLE

$$s_1 \xrightarrow{r(s_1,\,\cdot\,)=0} s_2 \xrightarrow{r(s_2,\,\cdot\,)=0} s_3 \xrightarrow{r(s_3,\,\cdot\,)=0} s_4 \xrightarrow{r(s_4,\,\cdot\,)=1} s_\infty \circlearrowleft$$

$$v^1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad v^2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad v^3 = \begin{bmatrix} 0 \\ 0 \\ \gamma \\ 1 \end{bmatrix} \qquad v_\pi = \begin{bmatrix} \gamma^3 \\ \gamma^2 \\ \gamma \\ 1 \end{bmatrix}$$

$$v_1^3 = 0 + \gamma v_2^2 = 0$$

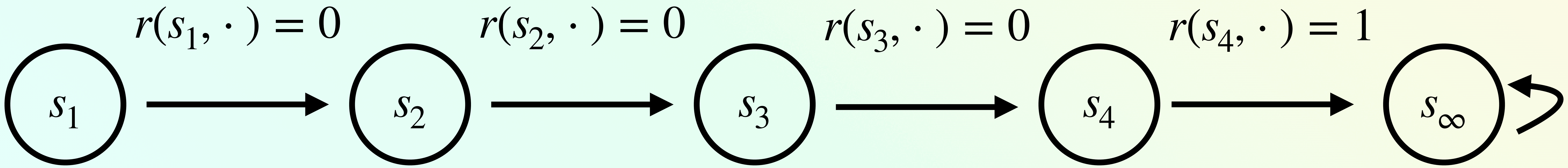
$$v_2^3 = 0 + \gamma v_3^2 = 0$$

$$v_3^3 = 0 + \gamma v_4^2 = 0 + \gamma$$

$$v_4^3 = 1 + \gamma 0 = 1$$

ITERATIVE POLICY EVALUATION

EXAMPLE



$$\underset{v^1}{\begin{bmatrix}0\\0\\0\\0\end{bmatrix}} \quad \underset{v^2}{\begin{bmatrix}0\\0\\0\\1\end{bmatrix}} \quad \underset{v^3}{\begin{bmatrix}0\\0\\\gamma\\1\end{bmatrix}} \quad \underset{v^4}{\begin{bmatrix}v_1^4\\v_2^4\\v_3^4\\v_4^4\end{bmatrix}} \quad \begin{aligned} v_1^4 &= 0 + \gamma v_2^3 = 0\\ v_2^4 &= 0 + \gamma v_3^3 = \gamma^2\\ v_3^4 &= 0 + v_4^3 = \gamma\\ v_4^4 &= 1 + \gamma 0 = 1 \end{aligned} \quad \underset{v_\pi}{\begin{bmatrix}\gamma^3\\\gamma^2\\\gamma\\1\end{bmatrix}}$$

ITERATIVE POLICY EVALUATION

EXAMPLE

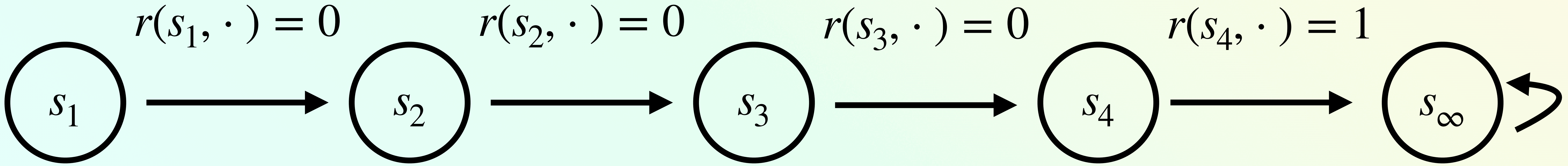
$$s_1 \xrightarrow{r(s_1,\,\cdot\,)=0} s_2 \xrightarrow{r(s_2,\,\cdot\,)=0} s_3 \xrightarrow{r(s_3,\,\cdot\,)=0} s_4 \xrightarrow{r(s_4,\,\cdot\,)=1} s_\infty$$

v^1	v^2	v^3	v^4	v_π
0	0	0	0	γ^3
0	0	0	γ^2	γ^2
0	0	γ	γ	γ
0	1	1	1	1

$$v_1^4 = 0 + \gamma v_2^3 = 0$$
$$v_2^4 = 0 + \gamma v_3^3 = \gamma^2$$
$$v_3^4 = 0 + v_4^3 = \gamma$$
$$v_4^4 = 1 + \gamma 0 = 1$$

ITERATIVE POLICY EVALUATION

EXAMPLE

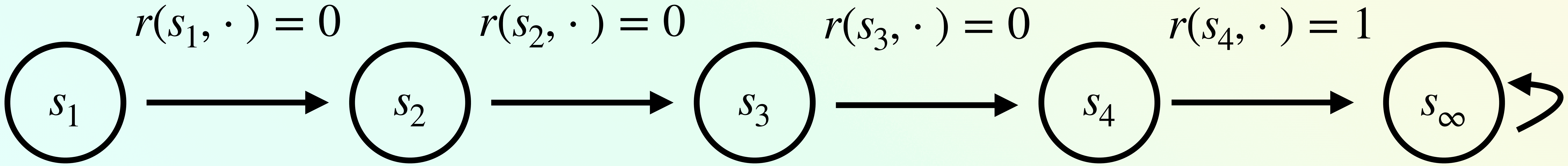


v^1	v^2	v^3	v^4	v^5	v_π
0	0	0	0	v_1^5	γ^3
0	0	0	γ^2	v_1^5	γ^2
0	0	γ	γ	v_1^5	γ
0	1	1	1	v_1^5	1

$$v_1^4 = 0 + \gamma v_2^4 = \gamma^3$$
$$v_2^4 = 0 + \gamma v_3^3 = \gamma^2$$
$$v_3^4 = 0 + v_4^3 = \gamma$$
$$v_4^4 = 1 + \gamma 0 = 1$$

ITERATIVE POLICY EVALUATION

EXAMPLE

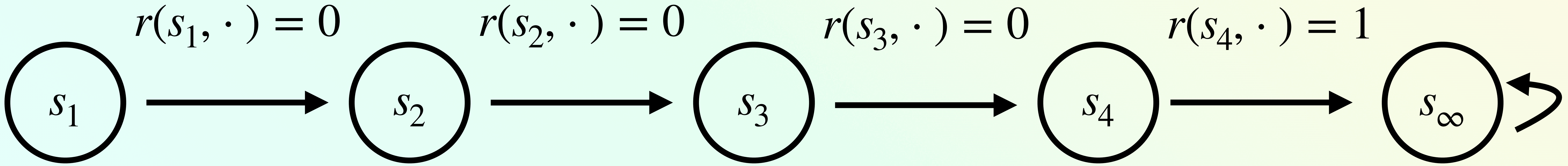


v^1	v^2	v^3	v^4	v^5	v_π
0	0	0	0	γ	γ^3
0	0	0	γ^2	γ^2	γ^2
0	0	γ	γ	γ	γ
0	1	1	1	1	1

$$v_1^4 = 0 + \gamma v_2^4 = \gamma^3$$
$$v_2^4 = 0 + \gamma v_3^3 = \gamma^2$$
$$v_3^4 = 0 + v_4^3 = \gamma$$
$$v_4^4 = 1 + \gamma 0 = 1$$

ITERATIVE POLICY EVALUATION

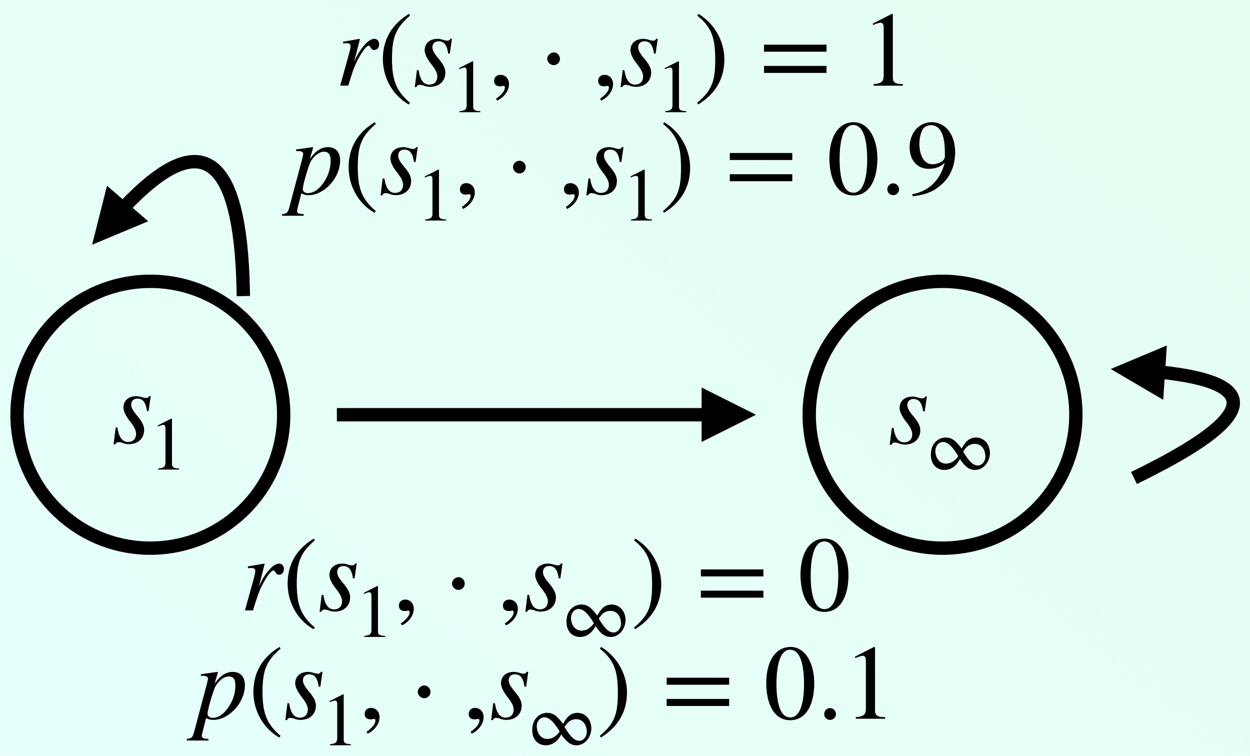
EXAMPLE



v^1	v^2	v^3	v^4	v^5	v^6	v_π
0	0	0	0	γ	γ	γ^3
0	0	0	γ^2	γ^2	γ^2	γ^2
0	0	γ	γ	γ	γ	γ
0	1	1	1	1	1	1

ITERATIVE POLICY EVALUATION

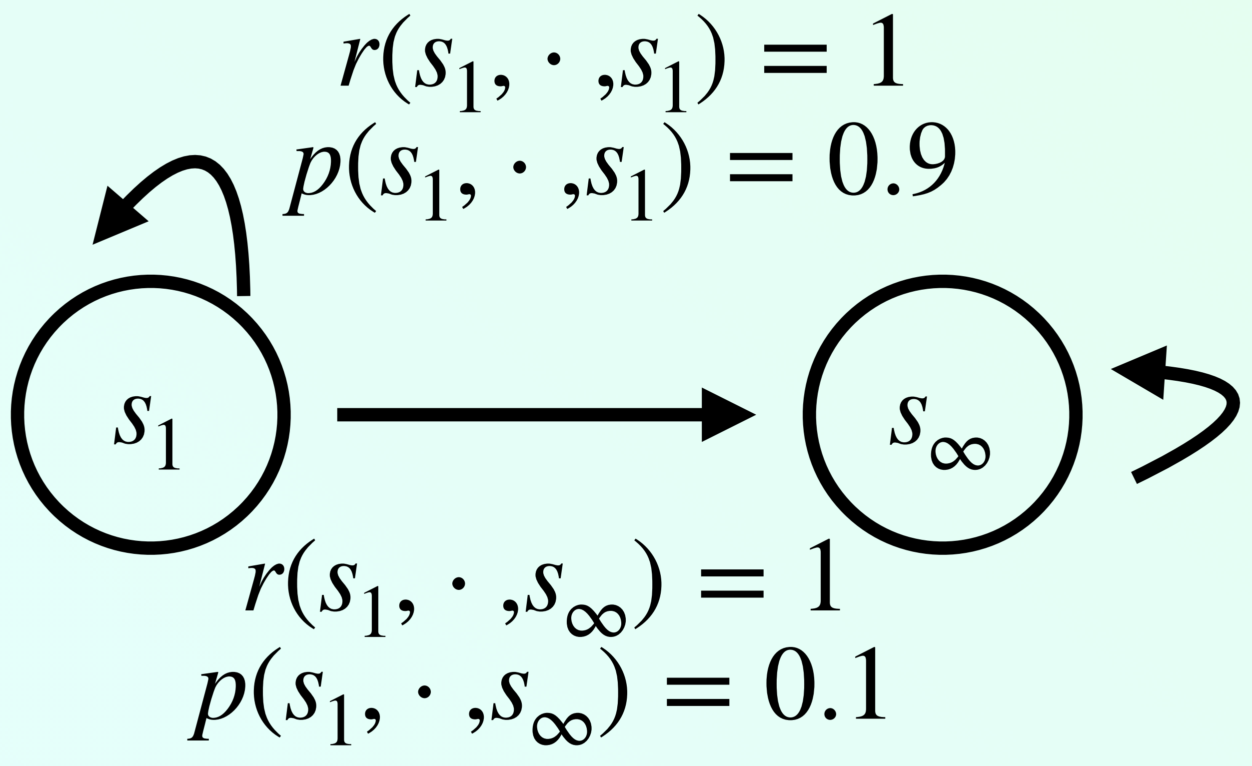
EXAMPLE 2



$$
\begin{aligned}
v_1^{k+1} &= p(s_1, \cdot, s_1)\left(r(s_1, \cdot, s_1) + \gamma v_1^k\right) + p(s_1, \cdot, s_\infty)\left(r(s_1, \cdot, s_\infty) + \gamma v_\pi(s_\infty)\right) \\
&= 0.9(1 + \gamma v_1^k) + 0.1(0 + \gamma 0) = 0.9 + 0.9\gamma v_1^k
\end{aligned}
$$

ITERATIVE POLICY EVALUATION

EXAMPLE 2



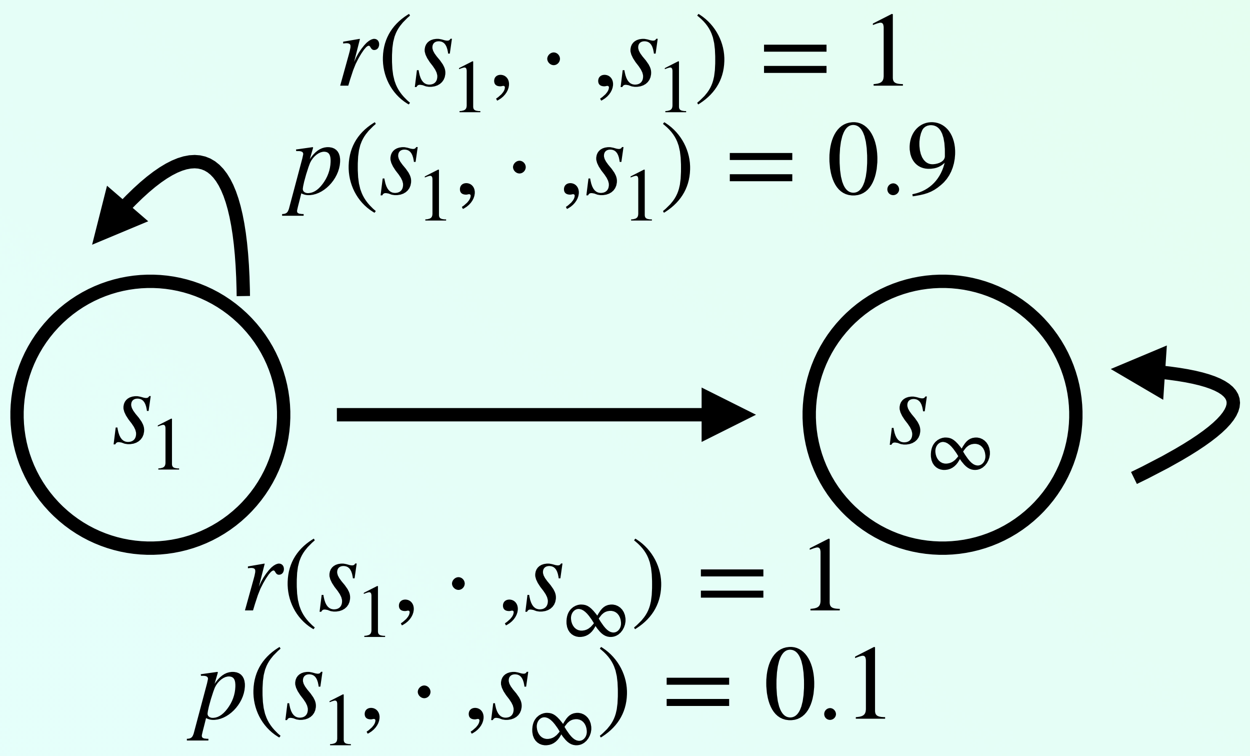
$\gamma = 0.5$

0							
v_1^1	v_1^2	v_1^3	v_1^4	v_1^5	v_1^6		

$$
\begin{aligned}
v_1^{k+1} &= p(s_1, \cdot, s_1)\big(r(s_1, \cdot, s_1) + \gamma v_1^k\big) + p(s_1, \cdot, s_\infty)\big(r(s_1, \cdot, s_\infty) + \gamma v_\pi(s_\infty)\big) \\
&= 0.9(1 + \gamma v_1^k) + 0.1(0 + \gamma 0) = 0.9 + 0.9\gamma v_1^k
\end{aligned}
$$

ITERATIVE POLICY EVALUATION

EXAMPLE 2



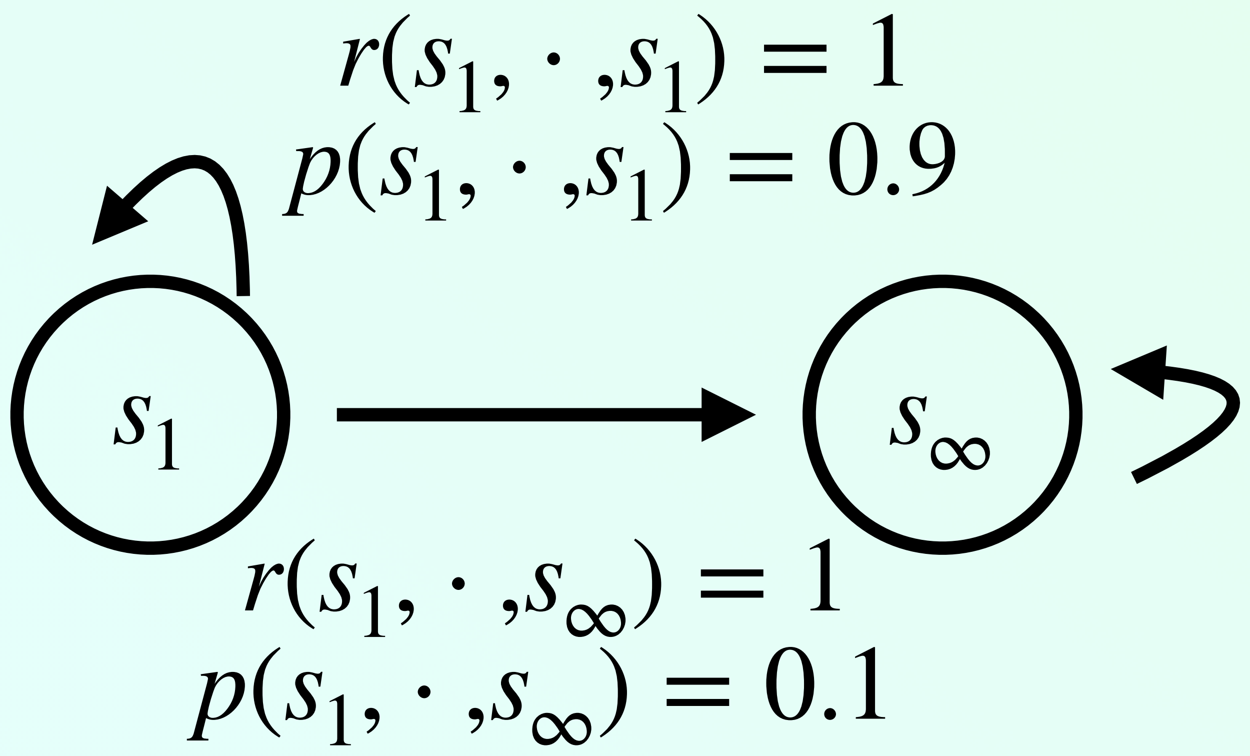
$\gamma = 0.5$

0	0.9						
v_1^1	v_1^2	v_1^3	v_1^4	v_1^5	v_1^6		

$$
\begin{aligned}
v_1^{k+1} &= p(s_1, \cdot, s_1)\big(r(s_1, \cdot, s_1) + \gamma v_1^k\big) + p(s_1, \cdot, s_\infty)\big(r(s_1, \cdot, s_\infty) + \gamma v_\pi(s_\infty)\big) \\
&= 0.9(1 + \gamma v_1^k) + 0.1(0 + \gamma 0) = 0.9 + 0.9\gamma v_1^k
\end{aligned}
$$

ITERATIVE POLICY EVALUATION

EXAMPLE 2



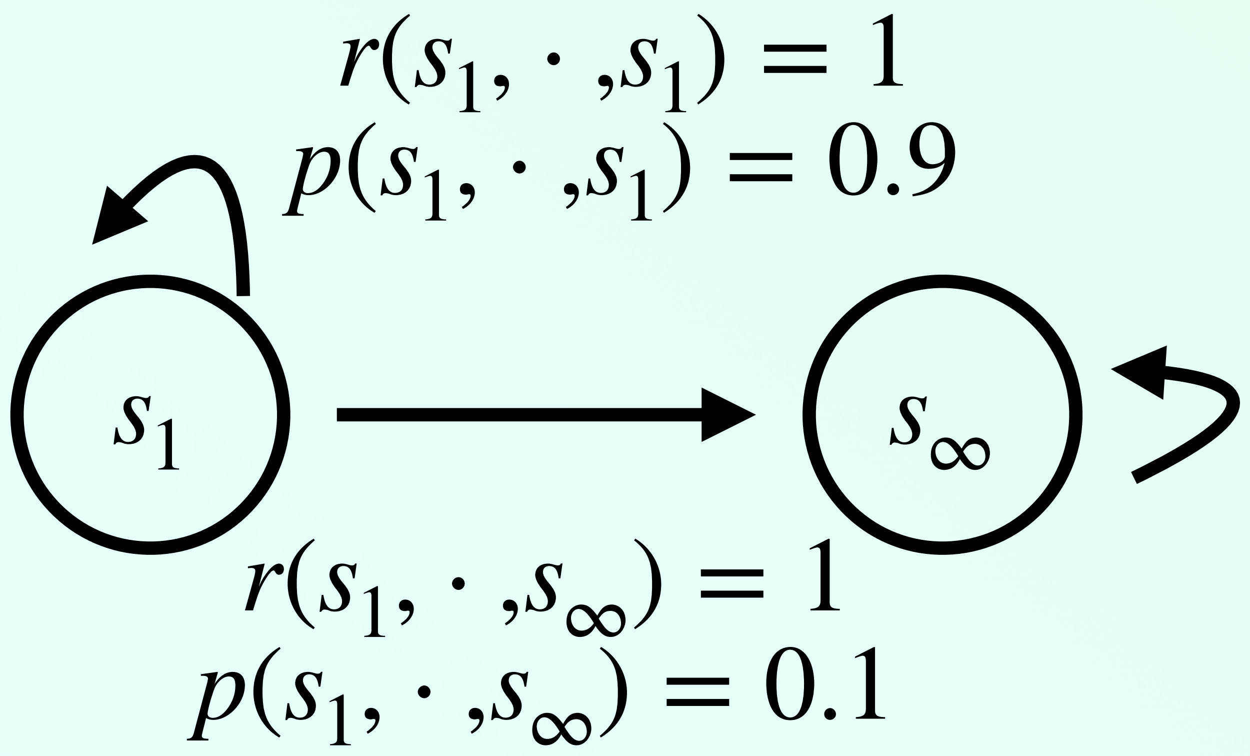
$\gamma = 0.5$

0	0.9	1.31					
v_1^1	v_1^2	v_1^3	v_1^4	v_1^5	v_1^6		

$$
\begin{aligned}
v_1^{k+1} &= p(s_1, \cdot, s_1)\left(r(s_1, \cdot, s_1) + \gamma v_1^k\right) + p(s_1, \cdot, s_\infty)\left(r(s_1, \cdot, s_\infty) + \gamma v_\pi(s_\infty)\right) \\
&= 0.9(1 + \gamma v_1^k) + 0.1(0 + \gamma 0) = 0.9 + 0.9\gamma v_1^k
\end{aligned}
$$

ITERATIVE POLICY EVALUATION

EXAMPLE 2



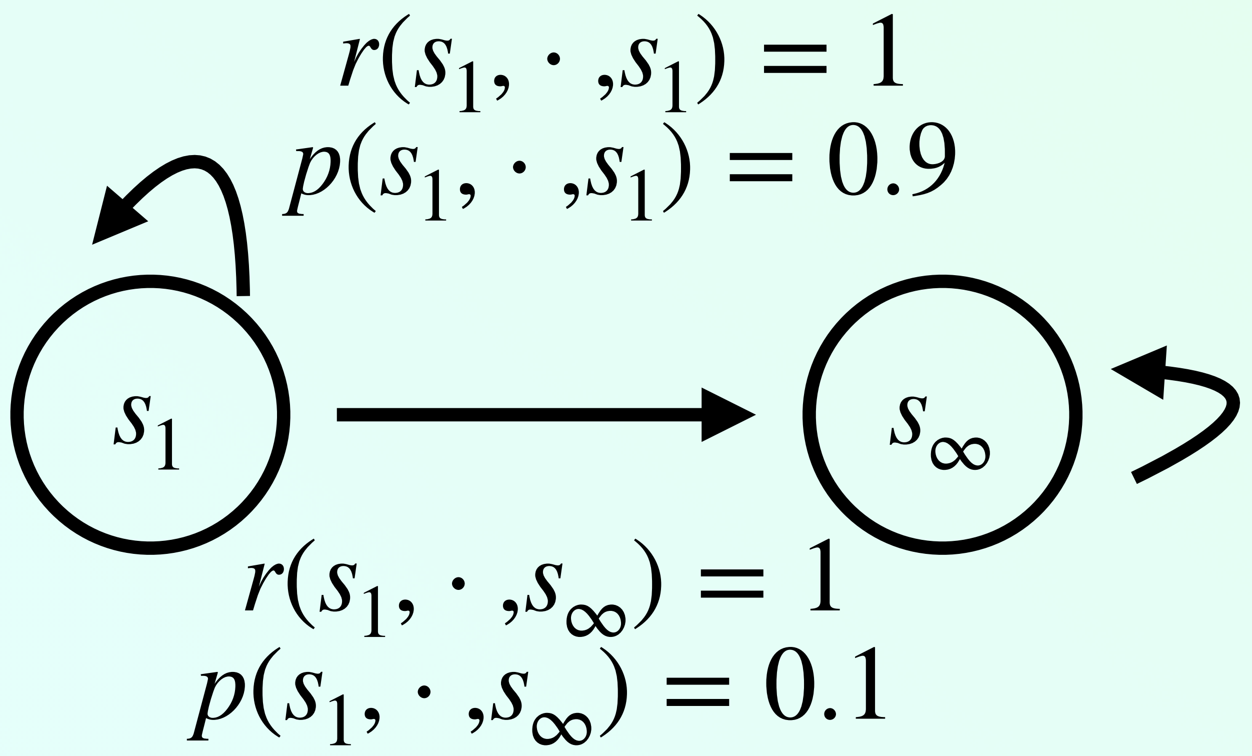
$\gamma = 0.5$

0	0.9	1.31	1.49				
v_1^1	v_1^2	v_1^3	v_1^4	v_1^5	v_1^6		

$$
\begin{aligned}
v_1^{k+1} &= p(s_1, \cdot, s_1)\big(r(s_1, \cdot, s_1) + \gamma v_1^k\big) + p(s_1, \cdot, s_\infty)\big(r(s_1, \cdot, s_\infty) + \gamma v_\pi(s_\infty)\big) \\
&= 0.9(1 + \gamma v_1^k) + 0.1(0 + \gamma 0) = 0.9 + 0.9\gamma v_1^k
\end{aligned}
$$

ITERATIVE POLICY EVALUATION

EXAMPLE 2



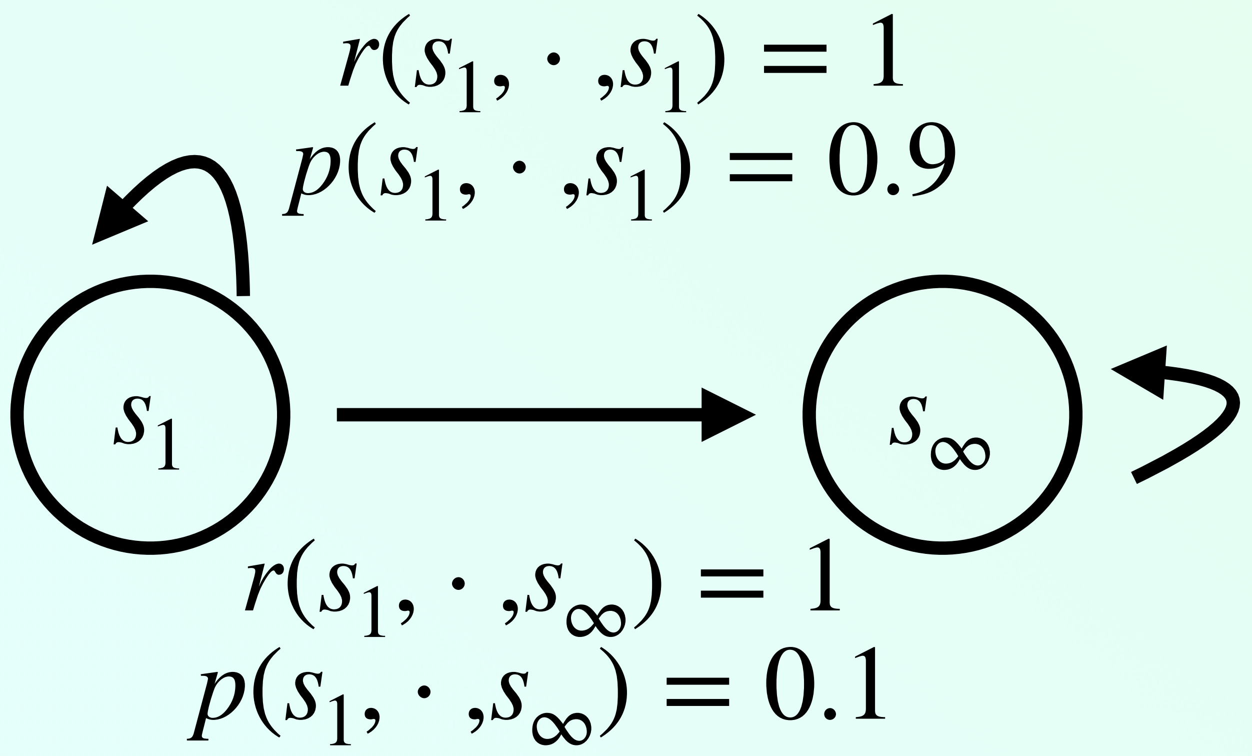
$\gamma = 0.5$

0	0.9	1.31	1.49	1.57			
v_1^1	v_1^2	v_1^3	v_1^4	v_1^5	v_1^6		

$$\begin{aligned} v_1^{k+1} &= p(s_1, \cdot, s_1)\left(r(s_1, \cdot, s_1) + \gamma v_1^k\right) + p(s_1, \cdot, s_\infty)\left(r(s_1, \cdot, s_\infty) + \gamma v_\pi(s_\infty)\right) \\ &= 0.9(1 + \gamma v_1^k) + 0.1(0 + \gamma 0) = 0.9 + 0.9\gamma v_1^k \end{aligned}$$

ITERATIVE POLICY EVALUATION

EXAMPLE 2



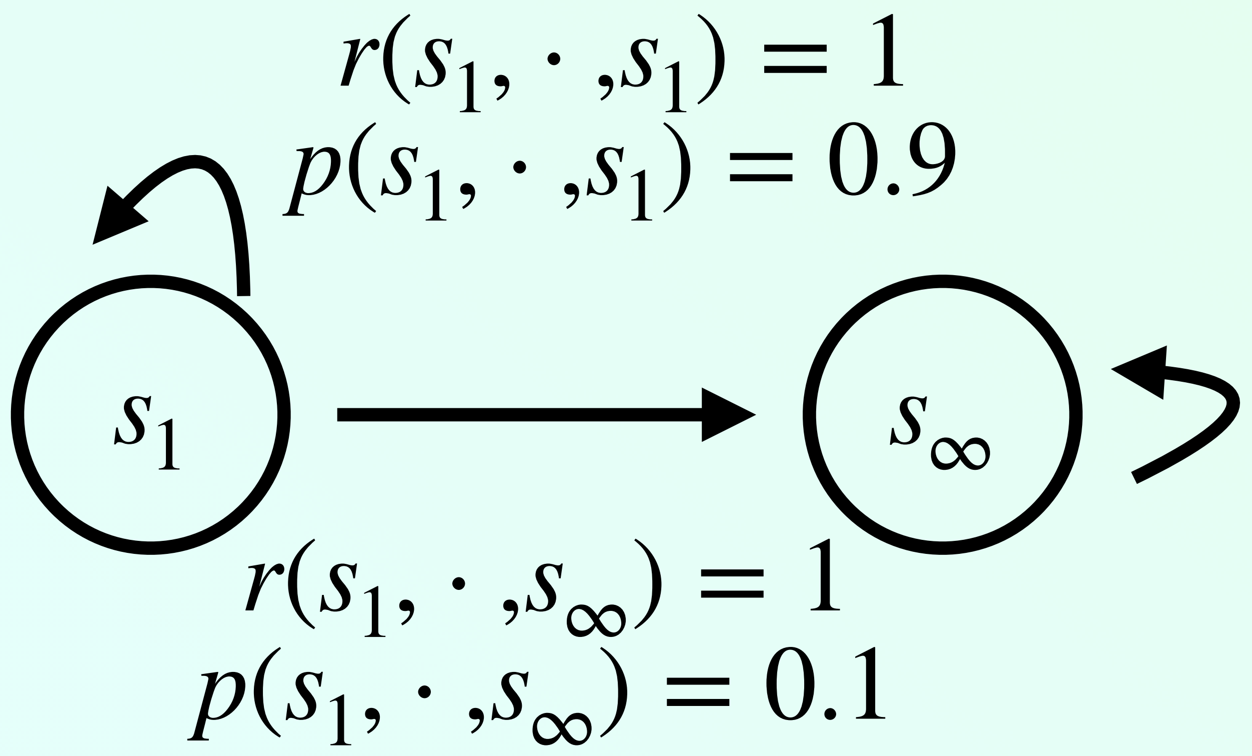
$\gamma = 0.5$

0	0.9	1.31	1.49	1.57	1.61		
v_1^1	v_1^2	v_1^3	v_1^4	v_1^5	v_1^6		

$$\begin{aligned} v_1^{k+1} &= p(s_1, \cdot, s_1)\big(r(s_1, \cdot, s_1) + \gamma v_1^k\big) + p(s_1, \cdot, s_\infty)\big(r(s_1, \cdot, s_\infty) + \gamma v_\pi(s_\infty)\big) \\ &= 0.9(1 + \gamma v_1^k) + 0.1(0 + \gamma 0) = 0.9 + 0.9\gamma v_1^k \end{aligned}$$

ITERATIVE POLICY EVALUATION

EXAMPLE 2



$\gamma = 0.5$

0	0.9	1.31	1.49	1.57	1.61	$\cdots$	1.636
v_1^1	v_1^2	v_1^3	v_1^4	v_1^5	v_1^6	$\cdots$	v_1^∞

$$\begin{aligned} v_1^{k+1} &= p(s_1, \cdot, s_1)\big(r(s_1, \cdot, s_1) + \gamma v_1^k\big) + p(s_1, \cdot, s_\infty)\big(r(s_1, \cdot, s_\infty) + \gamma v_\pi(s_\infty)\big) \\ &= 0.9(1 + \gamma v_1^k) + 0.1(0 + \gamma 0) = 0.9 + 0.9\gamma v_1^k \end{aligned}$$

ITERATIVE POLICY EVALUATION

PROPERTIES

$$\lim_{k\to\infty} v^k \to v_\pi$$

Can require an infinite number of updates!

In practice, stop when $\|v^{k+1} - v^k\|_\infty < \Delta$

- Δ is a small user-specified value
- $\|v^{k+1} - v^k\|_\infty = \max_i |v_i^{k+1} - v_i^k|$

CONVERGENCE OF v^k

PROOF

$$\lim_{k \to \infty} v^k \to v_\pi$$

Great! But we need to prove it.

BELLMAN OPERATOR (EVALUATION)

DEFINITION

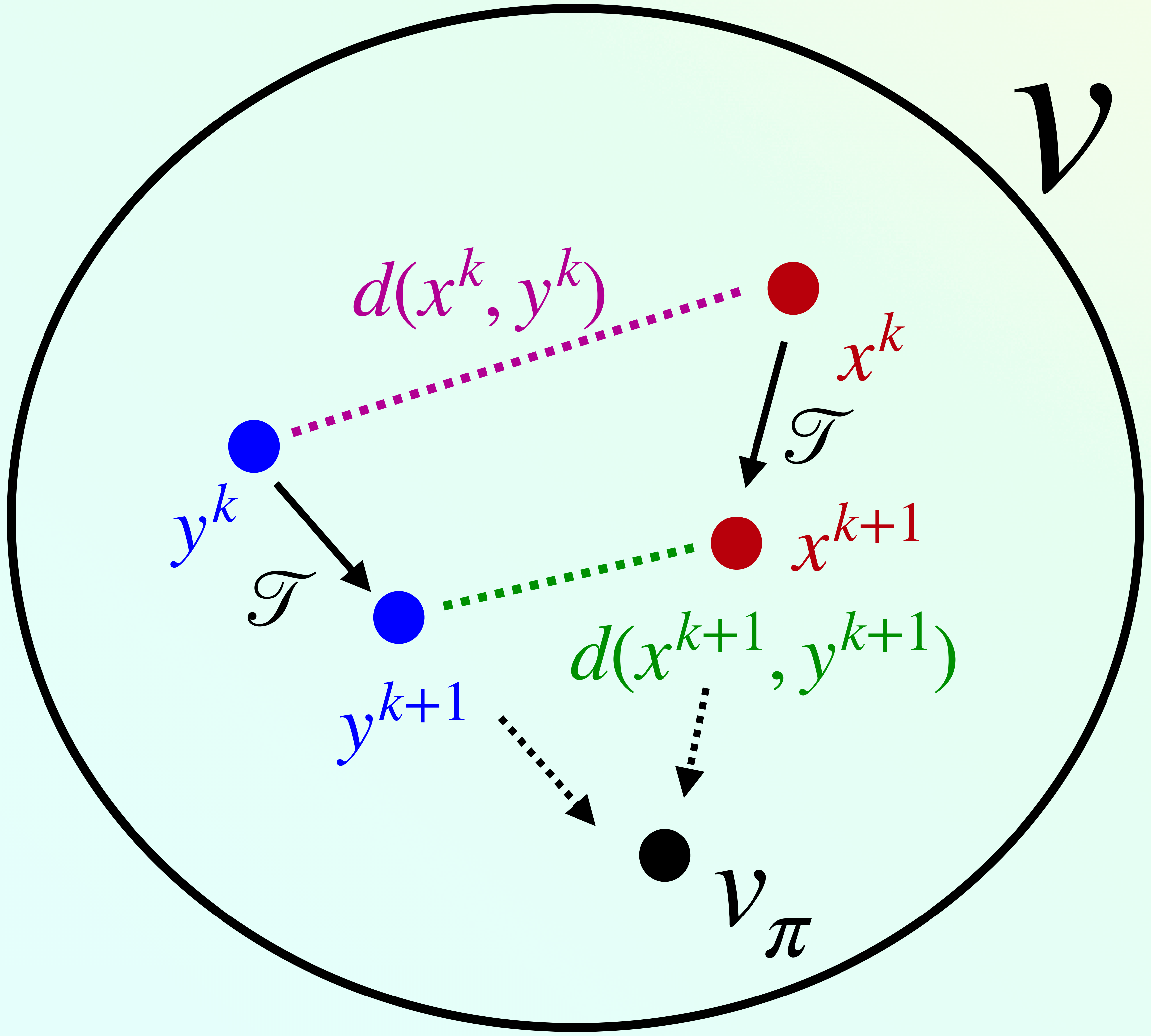
$\mathcal{T} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is a Bellman evaluation operator that iteratively updates v

$$\mathcal{T}(v^k) \doteq v^{k+1}$$

Want to show:

$$\forall x, y \text{ and } \lambda \in [0,1)$$

$$d(\mathcal{T}(x), \mathcal{T}(y)) \leq \lambda \, d(x, y)$$



BELLMAN OPERATOR (EVALUATION)

DEFINITION CONTRACTION MAPPING

$f\colon \mathbb{R}^n \to \mathbb{R}^n$ is a contraction mapping if $\forall x, y \in \mathbb{R}^n$ there exist some $\lambda \in [0,1)$ such that

$$d\left(f(x), f(y)\right) \leq \lambda d(x, y)$$

Where $d\colon \mathbb{R}^n \to \mathbb{R}$ is a distance function

BELLMAN OPERATOR (EVALUATION)

DEFINITION CONTRACTION MAPPING

Banach Fixed Point Theorem: (paraphrased, see Wikipedia for precise statement)

If f is a contraction mapping, then f has a *unique fixed point* x^* and

the sequence defined by $x^{k+1} = f(x^k)$ with x^1 chosen arbitrarily converges to x^*

BELLMAN OPERATOR (EVALUATION)

DEFINITION CONTRACTION MAPPING

Banach Fixed Point Theorem: (paraphrased; see Wikipedia for precise statement)

If f is a contraction mapping, then f has a *unique fixed point* x^* and

the sequence defined by $x^{k+1} = f(x^k)$ with x^1 chosen arbitrarily converges to x^*

For us:

1. Prove $\mathscr{T}$ is a contraction mapping
2. Prove the fixed point is v_π

BELLMAN OPERATOR (EVALUATION)

PROOF CONTRACTION MAPPING

$$d(v, v') = \|v - v'\|_\infty = \max_i |v_i - v_i'|$$

$$\begin{aligned}
\|\mathcal{T}(v) - \mathcal{T}(v')\|_\infty &= \max_i \left| \mathcal{T}(v)_i - \mathcal{T}(v')_i \right| \\
&= \max_i \left| \sum_a \pi(a \mid s_i) \left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j \right) - \sum_a \pi(a \mid s_i) \left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j' \right) \right| \\
&= \max_i \left| \sum_a \pi(a \mid s_i) \left(r(s_i, a) - r(s_i, a) + \gamma \sum_j p(s_i, a, s_j)(v_j - v_j') \right) \right| \\
&= \max_i \left| \sum_a \pi(a \mid s_i) \gamma \sum_j p(s_i, a, s_j)(v_j - v_j') \right|
\end{aligned}$$

BELLMAN OPERATOR (EVALUATION)

PROOF CONTRACTION MAPPING

We know that $\left| \sum_i (x_i - y_i) \right| \leq \sum_i \left| (x_i - y_i) \right|$

$$\|\mathcal{T}(v) - \mathcal{T}(v')\|_\infty = \max_i \left| \sum_a \pi(a \mid s_i) \gamma \sum_j p(s_i, a, s_j)(v_j - v'_j) \right|$$

$$\leq \max_i \sum_a \pi(a \mid s_i) \gamma \sum_j p(s_i, a, s_j) \left| v_j - v'_j \right|$$

BELLMAN OPERATOR (EVALUATION)

PROOF CONTRACTION MAPPING

$$
\begin{aligned}
\|\mathcal{T}(v) - \mathcal{T}(v')\|_\infty &= \max_i \left| \sum_a \pi(a \mid s_i)\gamma \sum_j p(s_i, a, s_j)(v_j - v'_j) \right| \\
&\le \max_i \sum_a \pi(a \mid s_i)\gamma \sum_j p(s_i, a, s_j) \left| v_j - v'_j \right| \\
&\le \max_i \max_a \underbrace{\pi(a \mid s_i)}_{\le 1} \gamma \sum_j p(s_i, a, s_j) \left| v_j - v'_j \right| \\
&\le \max_i \max_a \gamma \sum_j p(s_i, a, s_j) \left| v_j - v'_j \right| \\
&\le \max_i \max_a \gamma \max_j \underbrace{p(s_i, a, s_j)}_{\le 1} \left| v_j - v'_j \right| \\
&\le \max_i \max_a \gamma \max_j \left| v_j - v'_j \right| = \gamma \max_j \left| v_j - v'_j \right|
\end{aligned}
$$

BELLMAN OPERATOR (EVALUATION)

PROOF CONTRACTION MAPPING

$$\|\mathcal{T}(v) - \mathcal{T}(v')\|_\infty \leq \gamma \max_j \left| v_j - v'_j \right| = \gamma d(v, v')$$

For $\gamma \in [0,1)$, since $d\left(\mathcal{T}(v), \mathcal{T}(v')\right) \leq \gamma d(v, v')$ then $\mathcal{T}$ is a contraction mapping

BELLMAN OPERATOR (EVALUATION)

PROOF FIXED POINT IS v_π

Let v^* be the fixed point of $\mathcal{T}$ (not the optimal value function)

We have that $v^* = \mathcal{T}(v^*)$, so $\forall i$

$$v_i^* = \mathcal{T}(v^*)_i$$

$$= \sum_a \pi(a \mid s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^* \right)$$

This is the Bellman equation, and it is only true for $v = v^\pi$, so $v^* = v_\pi$.

NEXT CLASS

WHAT YOU SHOULD DO

1. Quiz due Tonight night: Dynamic Programming

Monday: Policy Iteration and Value Iteration