POLICIES AND VALUE FUNCTIONS

PROBLEM FORMULATION

OVERVIEW

An MDP describes the problem space.

- How the agent interacts with a world, i.e., state and action spaces
- How the world evolves, i.e., transition dynamics
- How the agent is evaluated, i.e., reward function, γ

What is the space of solutions?

PROBLEM FORMULATION

OVERVIEW

An MDP describes the problem space.

- How the agent interacts with a world, i.e., state and action spaces
- How the world evolves, i.e., transition dynamics
- How the agent is evaluated, i.e., reward function, γ

What is the space of solutions?

PROBLEM FORMULATION

SOLUTION SPACE

The search space of solutions for an MDP is the different ways an agent can select an action in each state.

We call the mechanism an agent uses to select actions a *policy* π.

$$\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$$

Describes a conditional probability distribution over action in each state.

$$\pi(a \mid s) = \Pr(A_t = a \mid S_t = s)$$

PROBLEM FORMULATION

SOLUTION SPACE

A policy is, in general, stochastic $\pi(a \mid s) = \Pr(A_t = a \mid S_t = s)$

Can be deterministic $\pi(a \mid s) = 1.0,\ \forall a' \neq a,\ \pi(a' \mid s) = 0$

Special notation $\pi\colon\ \mathcal{S} \to \mathcal{A}$, $a = \pi(s)$

PROBLEM FORMULATION

SOLUTION SPACE

A policy is, in general, stochastic $\pi(a \mid s) = \Pr(A_t = a \mid S_t = s)$

Can be deterministic $\pi(a \mid s) = 1.0,\ \forall a' \neq a,\ \pi(a' \mid s) = 0$

Special notation $\pi \colon \mathcal{S} \to \mathcal{A}, a = \pi(s)$

POLICIES AND TIME

STATIONARY VS NONSTATIONARY

Let π_t be the policy the agent uses at time t

A sequence of policies $\pi_0, \pi_1, \ldots, \pi_t, \ldots$ is ***stationary*** if

$$\forall (a, s),\ \forall t, t' \;\; \Pr(A_t = a \,|\, S_t = s) = \Pr(A_{t'} = a \,|\, S_{t'} = s)$$

Otherwise, the sequence of policies is ***nonstationary***

- The agent *will* change its policy as it learns
- It is challenging to reason about a changing policy

Instead, we will reason about how good one particular policy π is, assuming we won't change it

HOW GOOD IS A POLICY

VALUE FUNCTIONS

The value (quality) of a policy π in a state s is given by the *state value function*

$$v_\pi(s) \doteq \mathbb{E}[G_t \mid S_t = s]$$

$A_t, R_{t+1}, S_{t+1}, A_{t+1}, \dots$ all depend on π

RANDOM VARIABLES AND POLICIES

DEFINING A_t

A_t is the random variable for an action sampled from a policy π

What if we use a policy π'? Define a new random variable.

- A'_t is the random variable for an action sampled from a policy π'
- $\Pr(S_t = s), \Pr(R_t = r)$? All depend on A_t
- S'_t, R'_t are the states and rewards observed after taking action A'_{t-1} in state S'_{t-1}
- $p'(s', r \mid s, a) \doteq \Pr(S'_t = s', R'_t = r \mid S'_{t-1} = s, A'_{t-1} = a)$

RANDOM VARIABLES AND POLICIES

DEFINING A_t

$$G'_t = \sum_{k=0}^{\infty} \gamma^k R'_{t+1+k}$$

$$v_{\pi'}(s) \doteq \mathbb{E}[G'_t \,|\, S'_t = s]$$

π — represents a variable so we can represent specific policies with one notation

Let π_1 and π_2 be two specific policies.

$v_{\pi_1}(s) = v_\pi(s)$ for $\pi = \pi_1$ and $v_{\pi_2}(s) = v_\pi(s)$ for $\pi = \pi_2$

$$v_{\pi_1}(s) - v_{\pi_2}(s) = \mathbb{E}[G^1_t \,|\, S^1_t = s] - \mathbb{E}[G^2_t \,|\, S^2_t = s]$$

need to consider different probabilities for each policy when comparing them

RANDOM VARIABLES AND POLICIES

DEFINING A_t

$v_\pi(s) \doteq \mathbb{E}[G_t \mid S_t = s]$ — book, and this class

$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$ — Coursera, and many papers

Explicit, but not necessary sense A_t is defined to come from π

HOW GOOD IS A POLICY

VALUE FUNCTIONS

The value (quality) of a policy π in a state s is given by the *state value function*

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

The value of taking an action a in state s and then selecting actions according to π is given by the *action value function*

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

HOW GOOD IS A POLICY

VALUE FUNCTIONS

$v_\pi(s)$ says how good π is at making decisions in state s and the states that come after

$q_\pi(s, a)$ Use reason about how good one action is versus another when using π

$q_\pi(s, a_1) > q_\pi(s, a_2)?$

arg $\max_a q_\pi(s, a)$ — find best action(s) in the state and under the current policy

HOW GOOD IS A POLICY

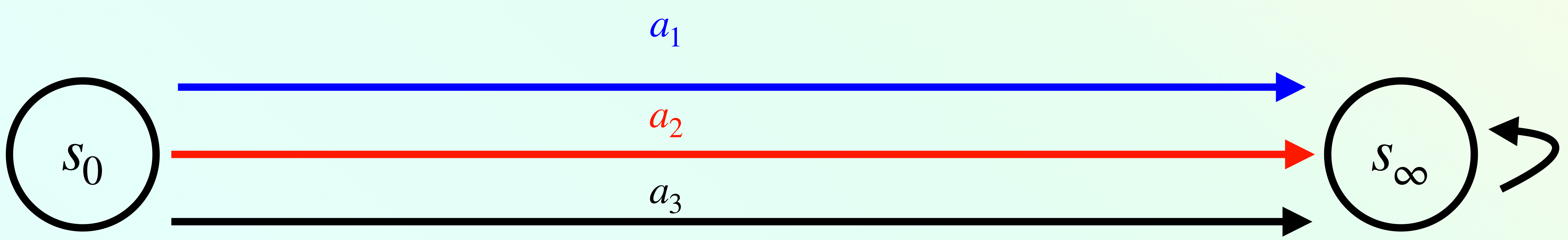
VALUE FUNCTIONS

Express v_π in terms of q_π

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}[G_t \mid S_t = s] \\
&= \sum_a \Pr(A_t = a \mid S_t = s)\mathbb{E}[G_t \mid S_t = s, A_t = a] \\
&= \sum_a \Pr(A_t = a \mid S_t = s) q_\pi(s, a) \\
&= \sum_a \pi(a \mid s) q_\pi(s, a)
\end{aligned}
$$

EXAMPLE

BANDIT



EXAMPLE

BANDIT

Let $Q_n(1) = 3,\ Q_n(2) = 2,\ Q_n(3) = 1$ be the estimates of $q_*(a)$

$\pi_{\text{greedy}}(1 \mid s_0) = 1.0$

$\pi_{\epsilon-\text{greedy}}(1 \mid s_0) = (1 - \epsilon) + \epsilon/3 \quad \pi_{\epsilon-\text{greedy}}(2 \mid s_0) = \pi_{\epsilon-\text{greedy}}(3 \mid s_0) = \epsilon/3$

EXAMPLE

BANDIT

For this problem $G_0 = R_1 + \gamma 0 + \gamma^2 0 + \ldots = R_1$, thus
$q_\pi(s_0, a) = \mathbb{E}[R_1 + \gamma 0 + \ldots \ | S_0 = s_0, A_0 = a] = \mathbb{E}[R_1 | S_0 = s_0, A_0 = a] = r(s_0, a) = q_*(a)$

$$v_{\pi_{\text{greedy}}}(s_0) = \sum_a \pi(a | s_0) q_{\pi_{\text{greedy}}}(s_0, a) = q_{\pi_{\text{greedy}}}(s_0, 1) = q_*(1)$$

$$v_{\pi_{\epsilon\text{-greedy}}}(s_0) = \sum_a \pi(a | s_0) q_{\pi_{\epsilon\text{-greedy}}}(s_0, a) = q_{\pi_{\epsilon\text{-greedy}}}(s_0, 1) = \begin{pmatrix} & \left((1-\epsilon) + \frac{\epsilon}{3}\right) q_*(1) \\ + & \frac{\epsilon}{3} q_*(2) \\ + & \frac{\epsilon}{3} q_*(3) \end{pmatrix}$$

$v_{\pi_{\text{greedy}}}(s_0) > v_{\pi_{\epsilon\text{-greedy}}}(s_0)$ ask if π_{greedy} better than $\pi_{\epsilon\text{-greedy}}$ in state s_0

EXAMPLE

BANDIT - OPTIMAL POLICY

What policy (or policies) are optimal for this problem?

Let $\mathscr{A}^* = \arg\max_a q_*(a)$

$\pi(s_0) \in \mathscr{A}^*$ any deterministic policy that chooses an action that maximizes q_*

Or any policy such that $\sum_{a \in \mathscr{A}^*} \pi(a \mid s_0) = 1$ only chooses optimal actions

EXAMPLE

BANDIT - OPTIMAL POLICY

At least one deterministic optimal policy

- n deterministic optimal policies if n optimal actions

Infinitely many optimal policies of $n > 1$

- $\{a_1, a_2\} = \mathscr{A}^* \ \ \forall p \in [0,1], \pi_p(a_1 | s_0) = p, \ \pi_p(a_2 | s_0) = 1 - p$
- All π_p are optimal

All bandit problems can be modeled as an MDP, so these are also true for MDPs

OPTIMAL POLICIES

DEFINITIONS

Set of all policies Π, set of all optimal policies Π_*

Definition #1

$\pi \geq \pi'$ if $\forall s,\ v_\pi(s) \geq v_{\pi'}(s)$ – at least as good in every state

$\pi \in \Pi_*$ if $\forall \pi' \in \Pi,\ \pi \geq \pi'$ – at least as good as any other policy in every state

OPTIMAL POLICIES

DEFINITIONS

Definition #2

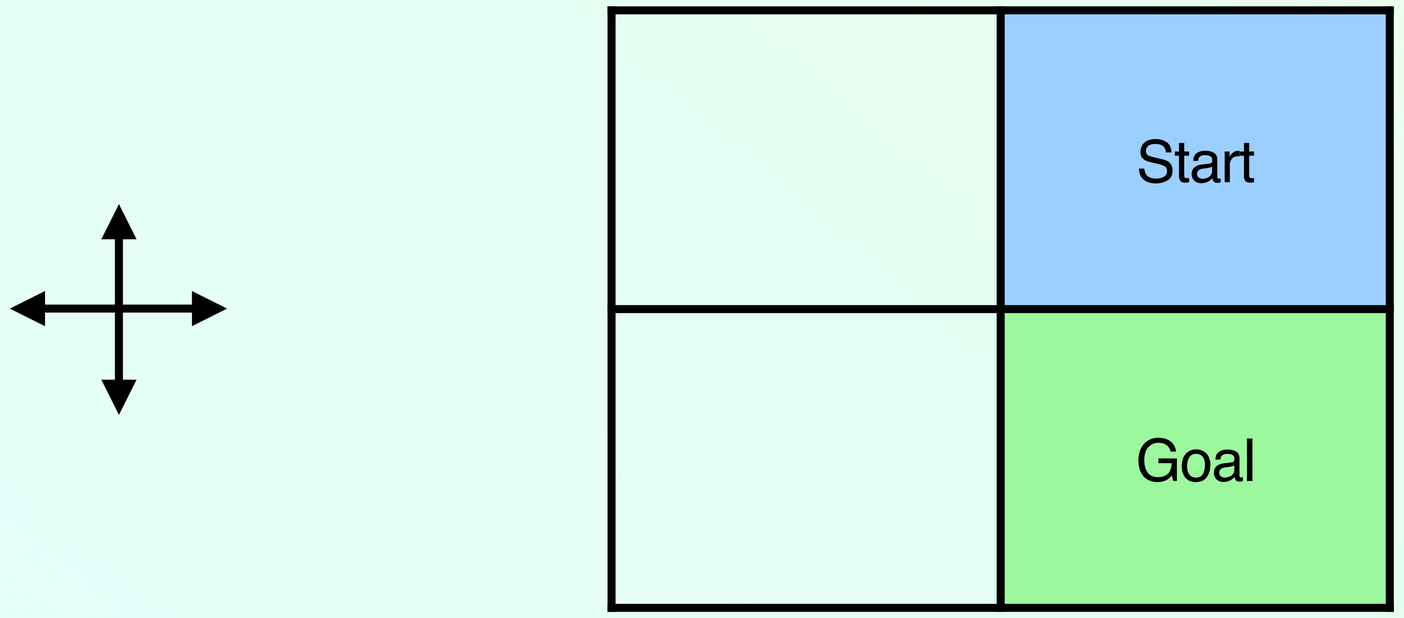
$J(\pi) \doteq \mathbb{E}[G_0] = \sum_s d_0(s) v_\pi(s)$ – policy's quality is determined by the start state distribution

$\Pi_* = \arg\max_\pi J(\pi)$ – all policies that optimize the value function in the first state

- Only has to be optimal in states that are reachable from the starting states
- The optimal policy can be terrible in states that are not reachable from the start state

OPTIMAL POLICIES

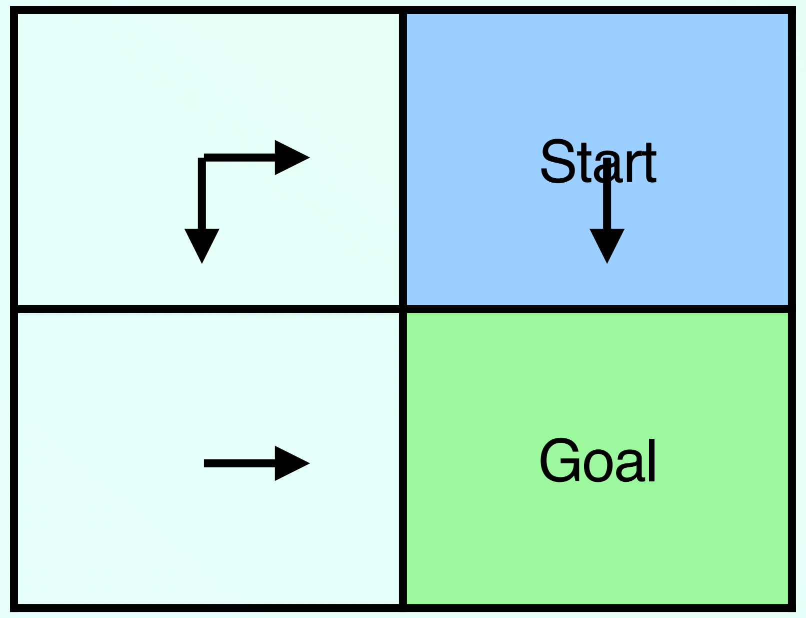
DEFINITIONS — EXAMPLE



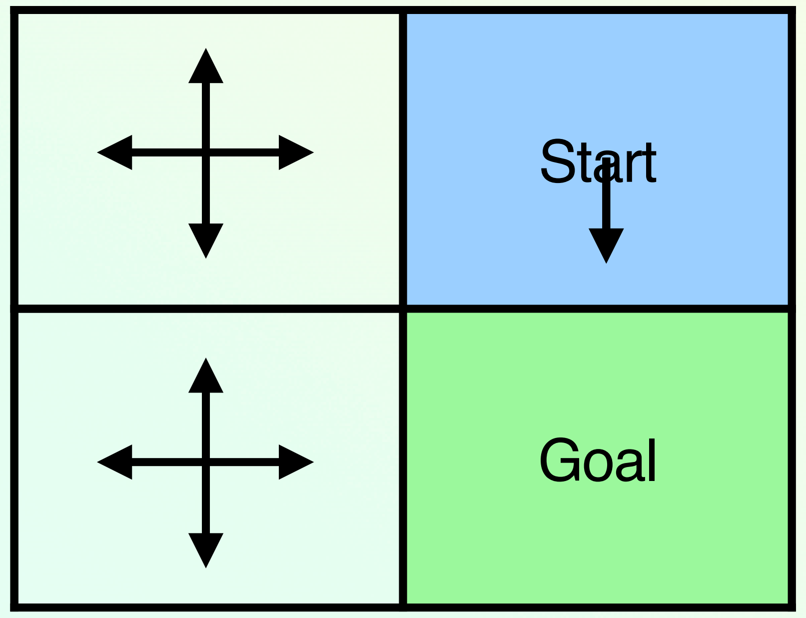
OPTIMAL POLICIES

DEFINITIONS — EXAMPLE

Definition #1



Definition #2



OPTIMAL POLICIES

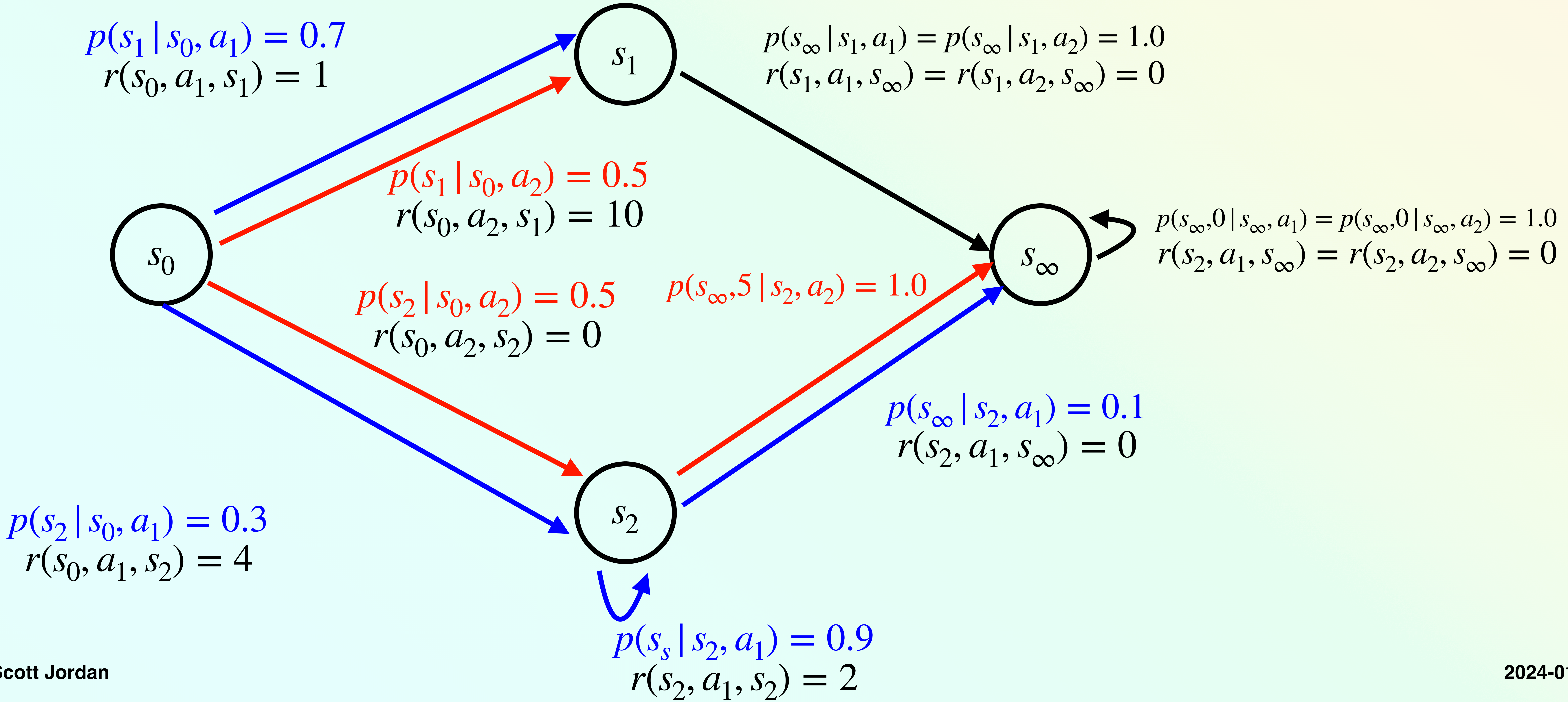
DEFINITIONS

We will use definition #1 in the first part of this course

We will use definition #2 when we switch to function approximation

COMPUTING VALUES FUNCTIONS

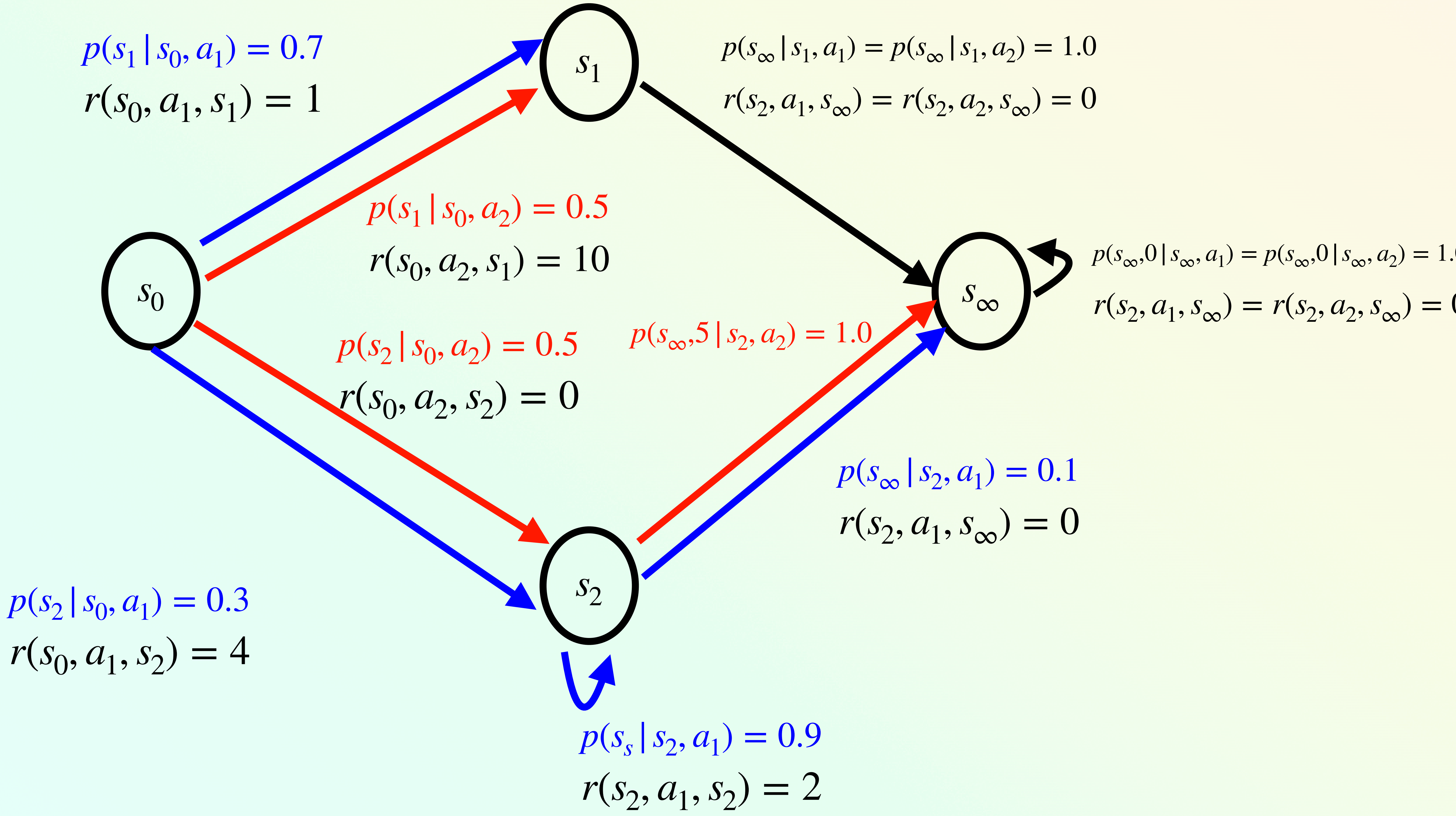
EXAMPLE



COMPUTING VALUES FUNCTIONS

EXAMPLE

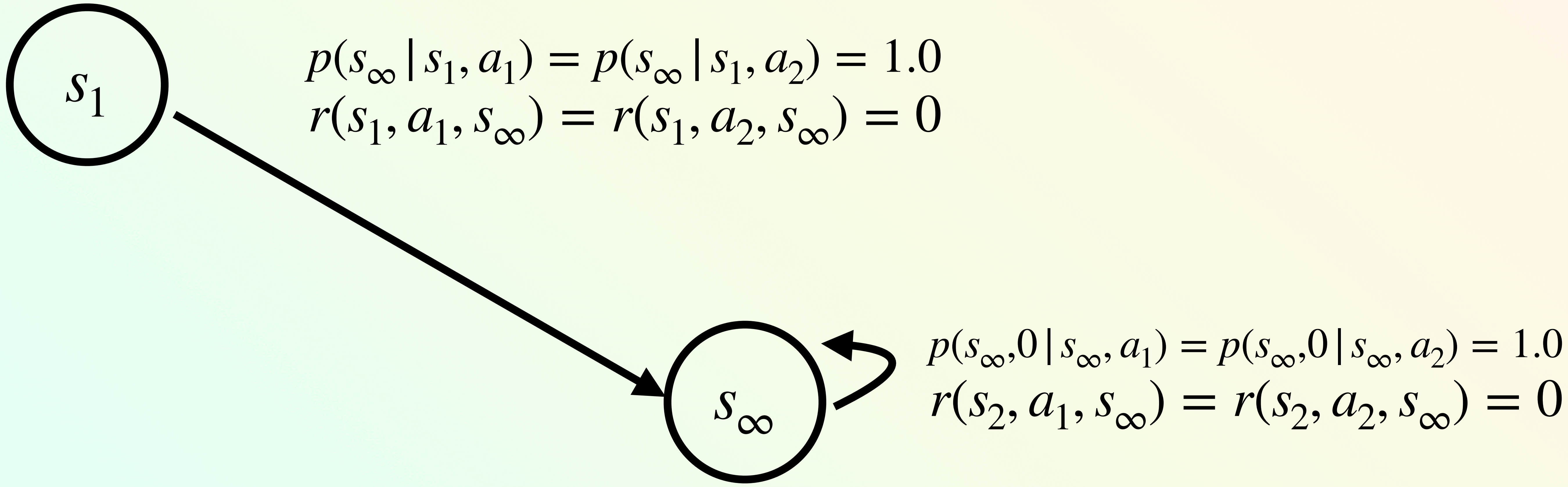
Compute q_π, v_π for s_0, s_1, s_2



COMPUTING VALUES FUNCTIONS

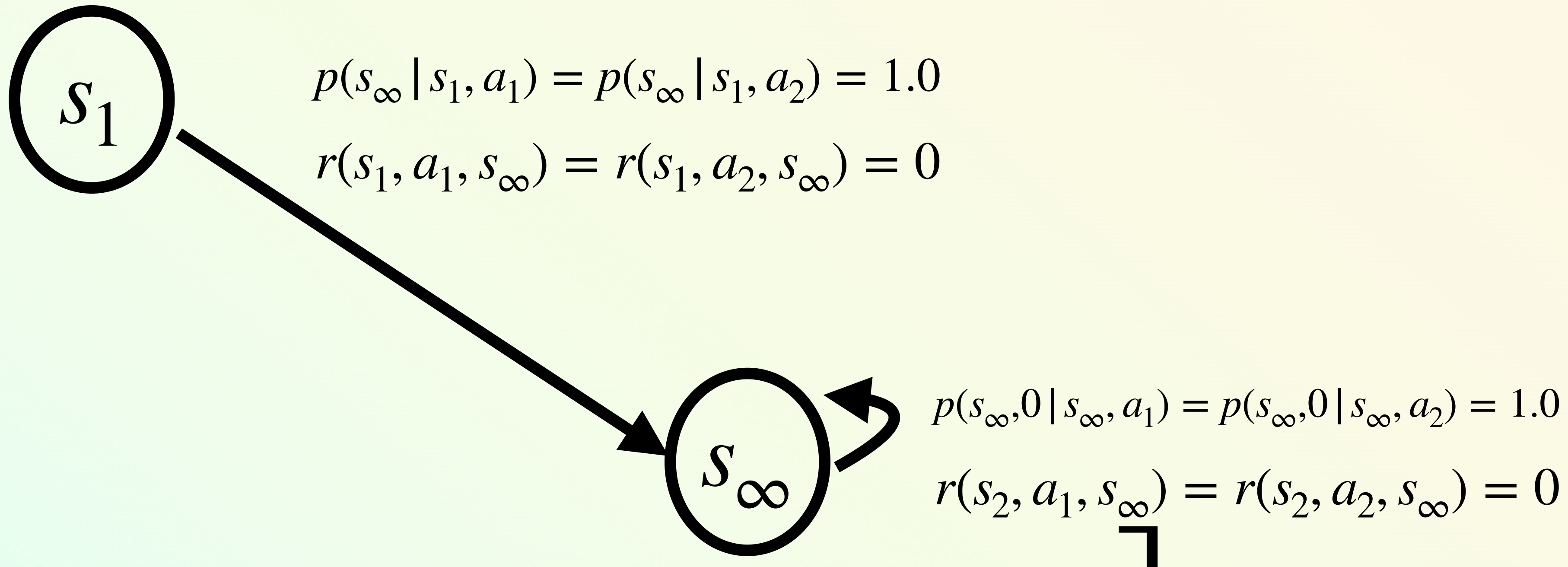
EXAMPLE

$q_\pi(s_1, a_1) = ?$



COMPUTING VALUES FUNCTIONS

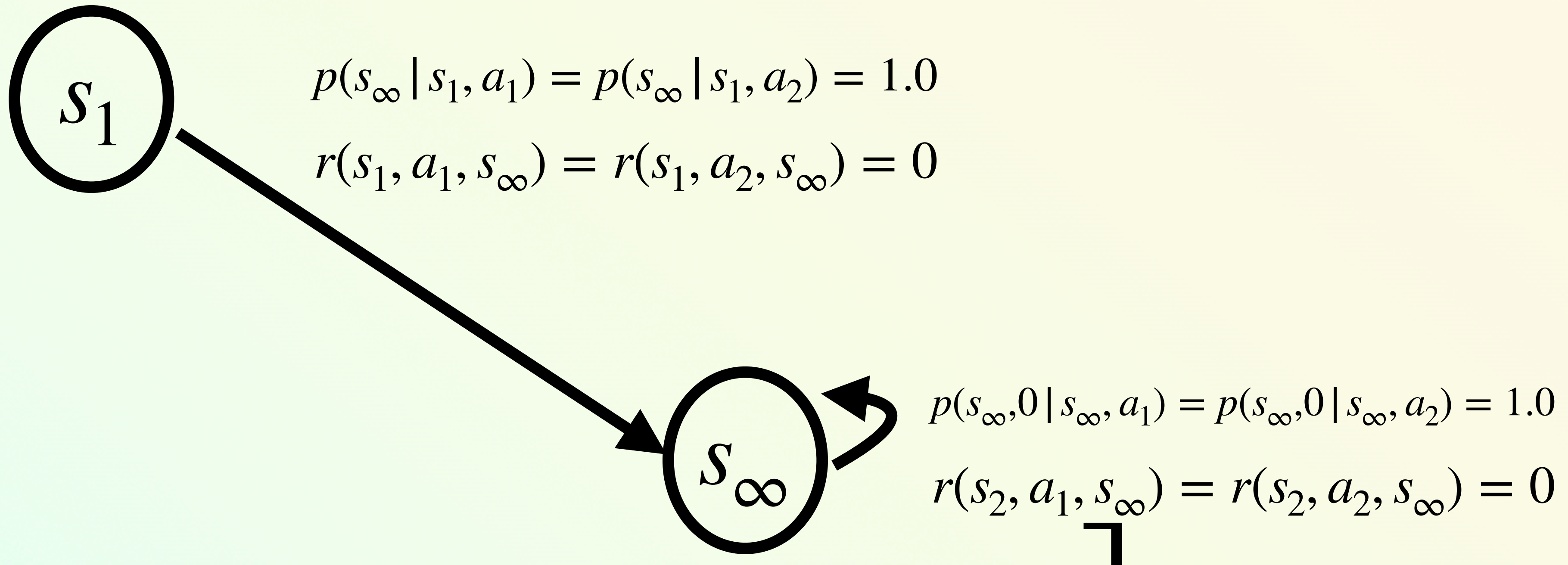
EXAMPLE



$$
\begin{aligned}
q_\pi(s_1, a_1) &= \mathbb{E}[G_t \,|\, S_t = s_2, A_t = a_1] \\
&= \mathbb{E}[R_{t+1} \,|\, S_t = s, A_t = a_1] + \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k R_{t+1+k} \,|\, S_t = s_1, A_t = a_1\right]
\end{aligned}
$$

COMPUTING VALUES FUNCTIONS

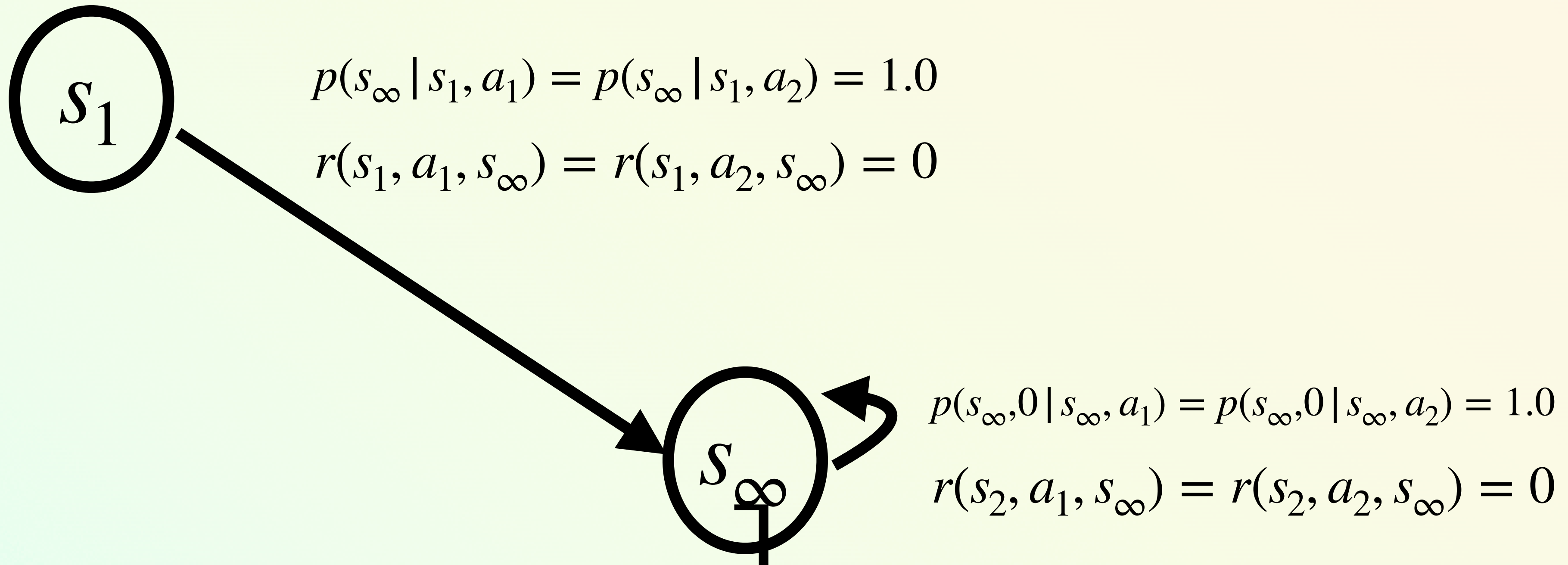
EXAMPLE



$$
\begin{aligned}
q_\pi(s_1, a_1) &= \mathbb{E}[G_t \mid S_t = s_2, A_t = a_1] \\
&= \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a_1] + \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k R_{t+1+k} \mid S_t = s_1, A_t = a_1\right] \\
&= \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a_1] + \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k R_{t+1+k} \mid S_{t+1} = s_\infty\right] \\
&= \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a_1] + 0
\end{aligned}
$$

COMPUTING VALUES FUNCTIONS

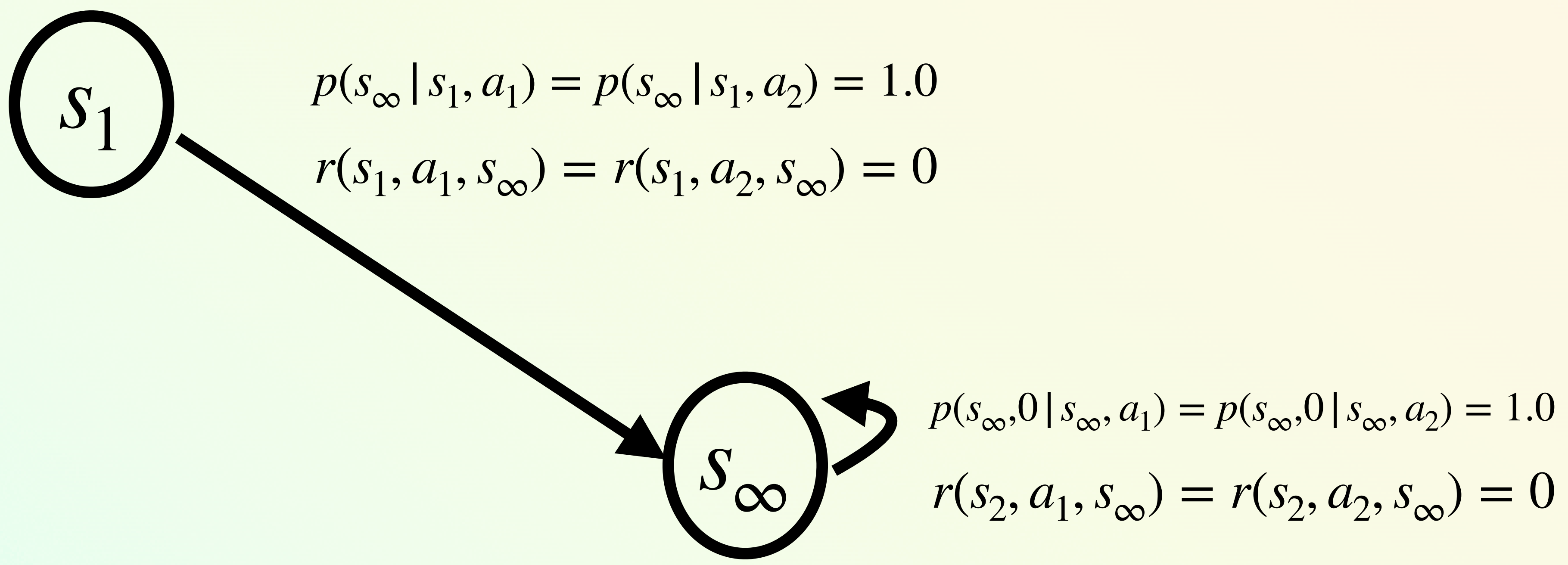
EXAMPLE



$$
\begin{aligned}
q_\pi(s_1, a_1) = &\ \mathbb{E}[G_t \,|\, S_t = s_2, A_t = a_1] \\
= &\ \mathbb{E}[R_{t+1} \,|\, S_t = s, A_t = a_1] + \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k R_{t+1+k} \,|\, S_t = s_1, A_t = a_1\right] \\
= &\ \mathbb{E}[R_{t+1} \,|\, S_t = s, A_t = a_1] + \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k R_{t+1+k} \,|\, S_{t+1} = s_\infty\right] \\
= &\ \mathbb{E}[R_{t+1} \,|\, S_t = s, A_t = a_1] + 0 \\
= &\ r(s_1, a_1, s_\infty) = 0
\end{aligned}
$$

COMPUTING VALUES FUNCTIONS

EXAMPLE



$q_\pi(s_1, a_1) = q_\pi(s_1, a_2) = 0$

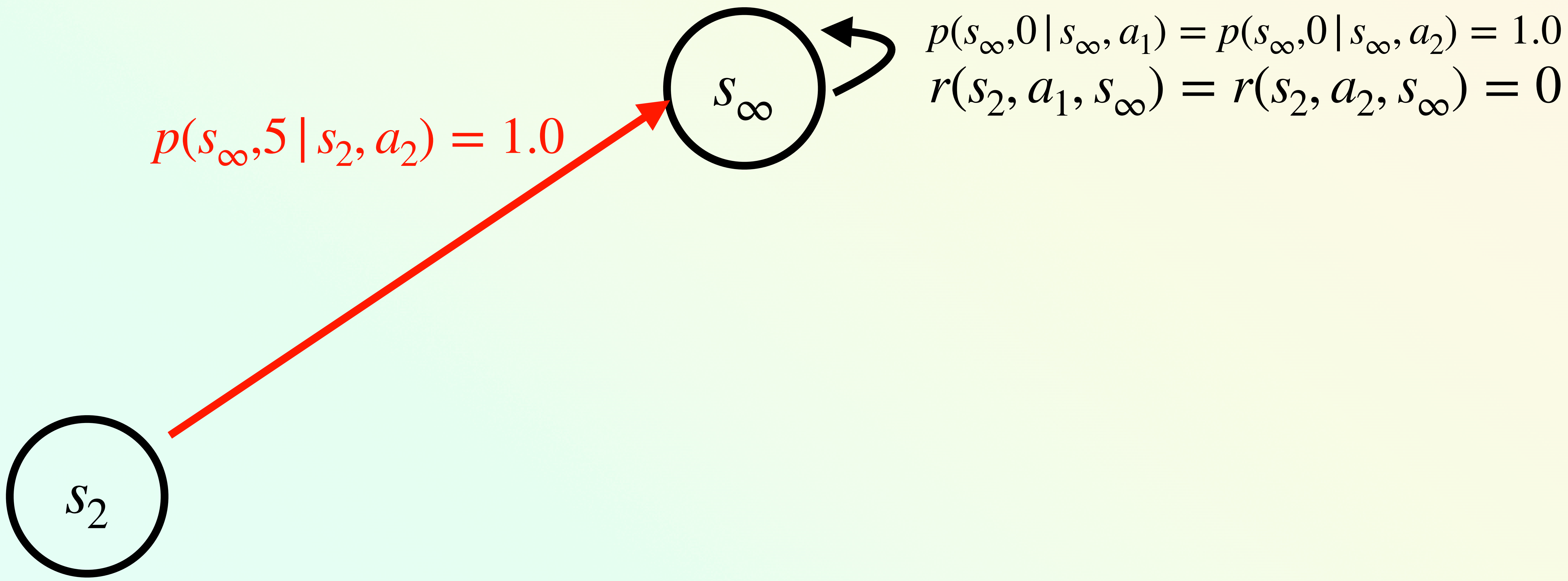
$v_\pi(s_1) = 0$ for all π

All policies are optimal in s_1

COMPUTING VALUES FUNCTIONS

EXAMPLE

$$q_\pi(s_2, a_2) = r(s_2, a_2, s_\infty) = 5$$

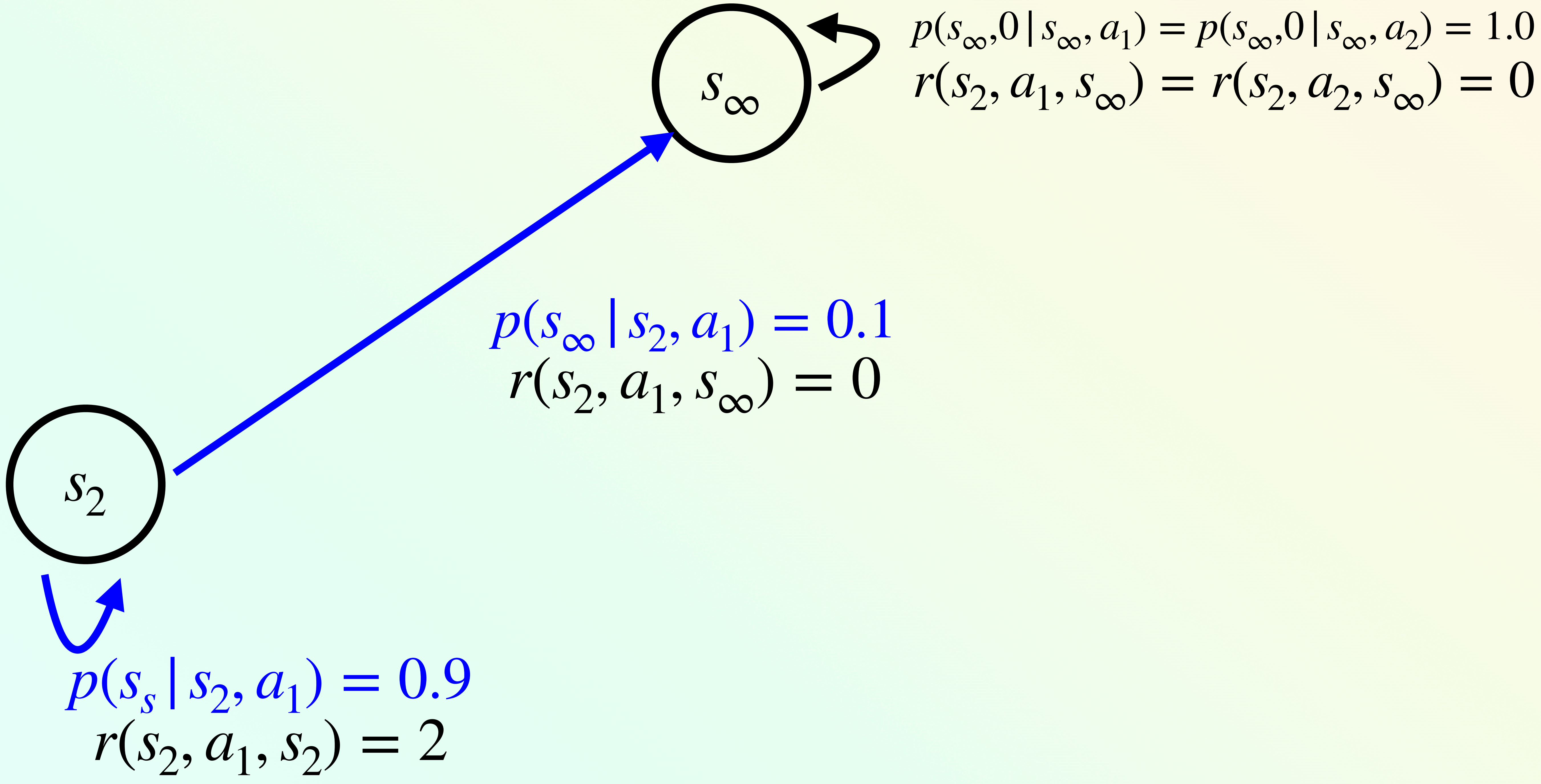


COMPUTING VALUES FUNCTIONS

EXAMPLE

$$q_\pi(s_2, a_1) = ?$$

$$\pi(s_2) = a_1$$

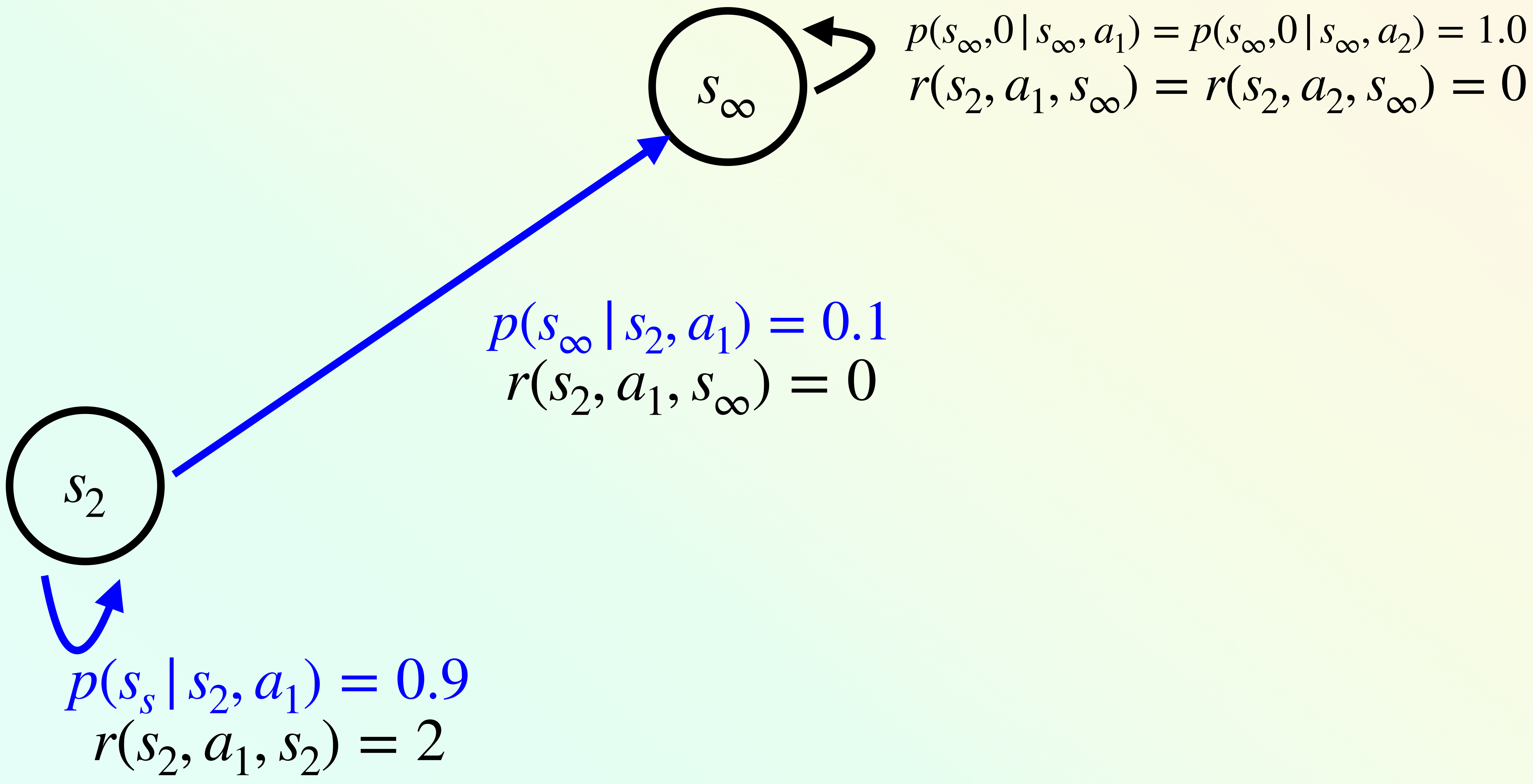


COMPUTING VALUES FUNCTIONS

EXAMPLE

$$q_\pi(s_2, a_1) = ?$$

$$\pi(s_2) = a_1$$



COMPUTING VALUES FUNCTIONS

EXAMPLE

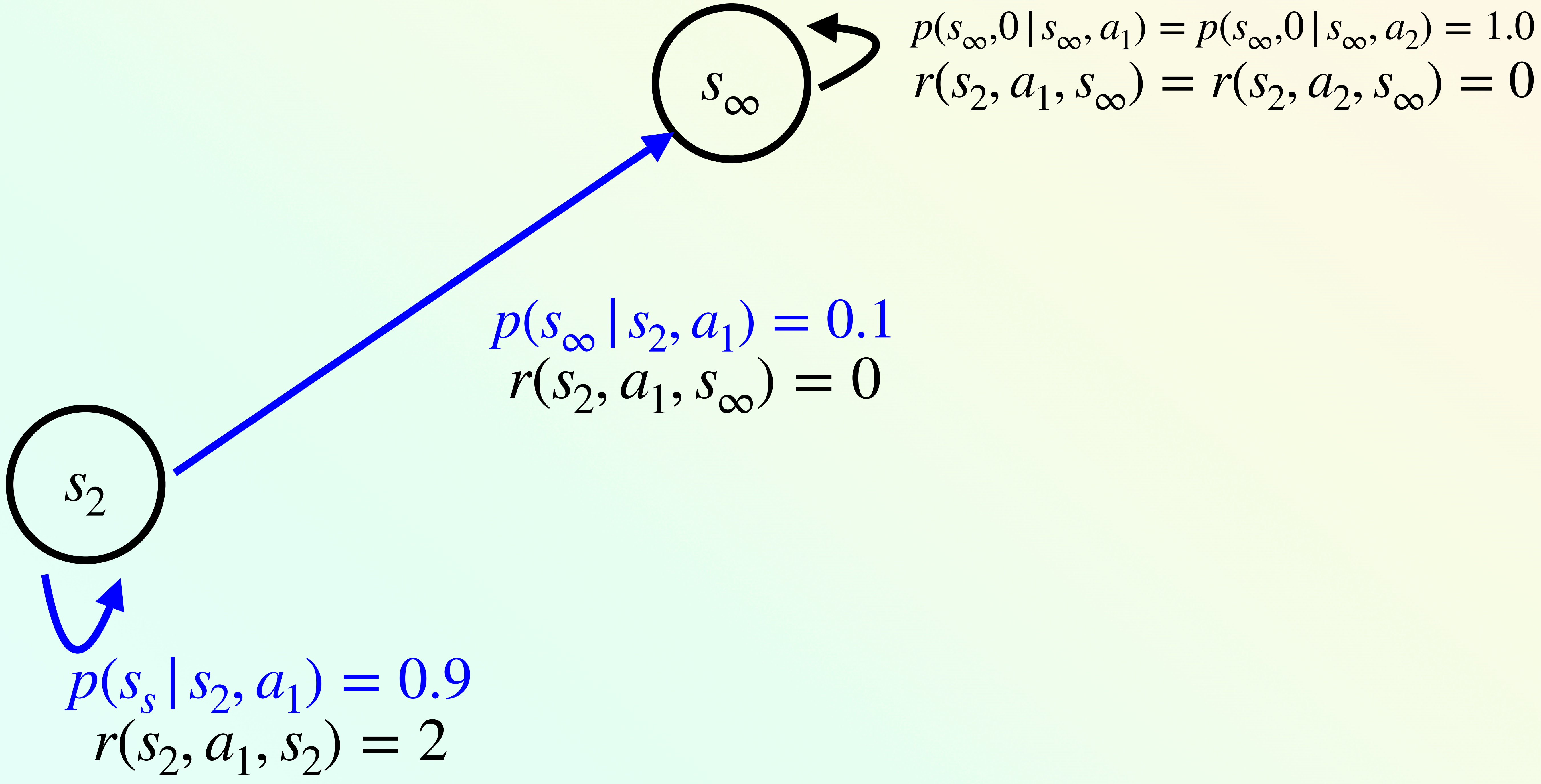
$$\pi(s_2) = a_1$$

Agent gets a reward of 2 every step until the episode ends

$$G_t = \sum_{k=0}^{K-1} \gamma^k 2 = 2\frac{1-\gamma^K}{1-\gamma}$$

K is a random variable for the number of time steps until termination

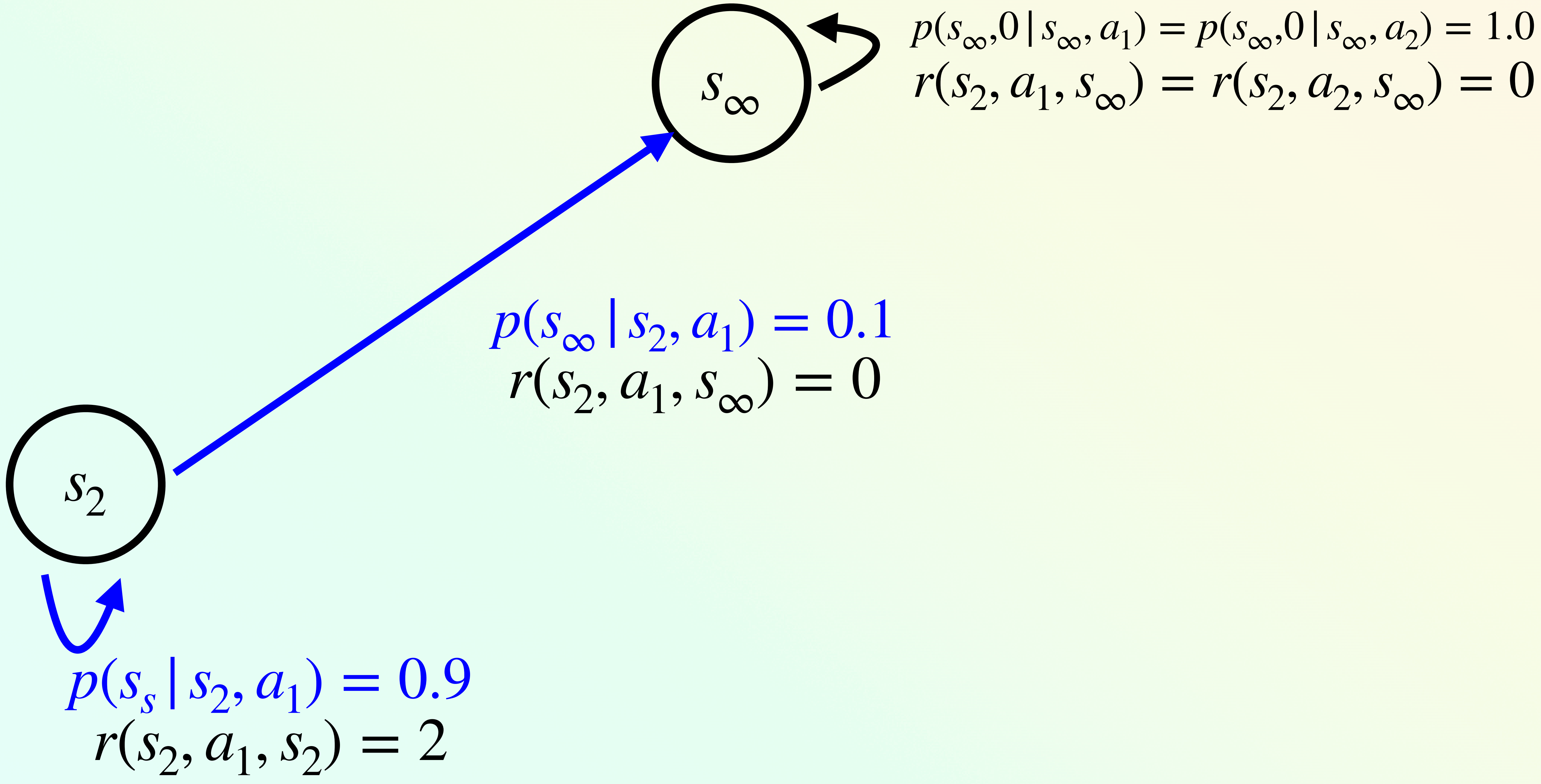
$$\Pr(K = k) = (0.9)^{(k-1)}(0.1)$$



COMPUTING VALUES FUNCTIONS

EXAMPLE

$$
\begin{aligned}
\mathbb{E}[G_t \mid S_t = s_2, A_t = a_1] &= \mathbb{E}\left[2\frac{1-\gamma^K}{1-\gamma} \mid S_t = s_2, A_t = a_1\right] \\
&= \sum_{k=1}^{\infty} \Pr(K=k) 2\frac{1-\gamma^k}{1-\gamma} \\
&= \sum_{k=1}^{\infty} (0.9)^{(k-1)}(0.1) 2\frac{1-\gamma^k}{1-\gamma} \\
&= \frac{2}{1-0.9\gamma}
\end{aligned}
$$



COMPUTING VALUES FUNCTIONS

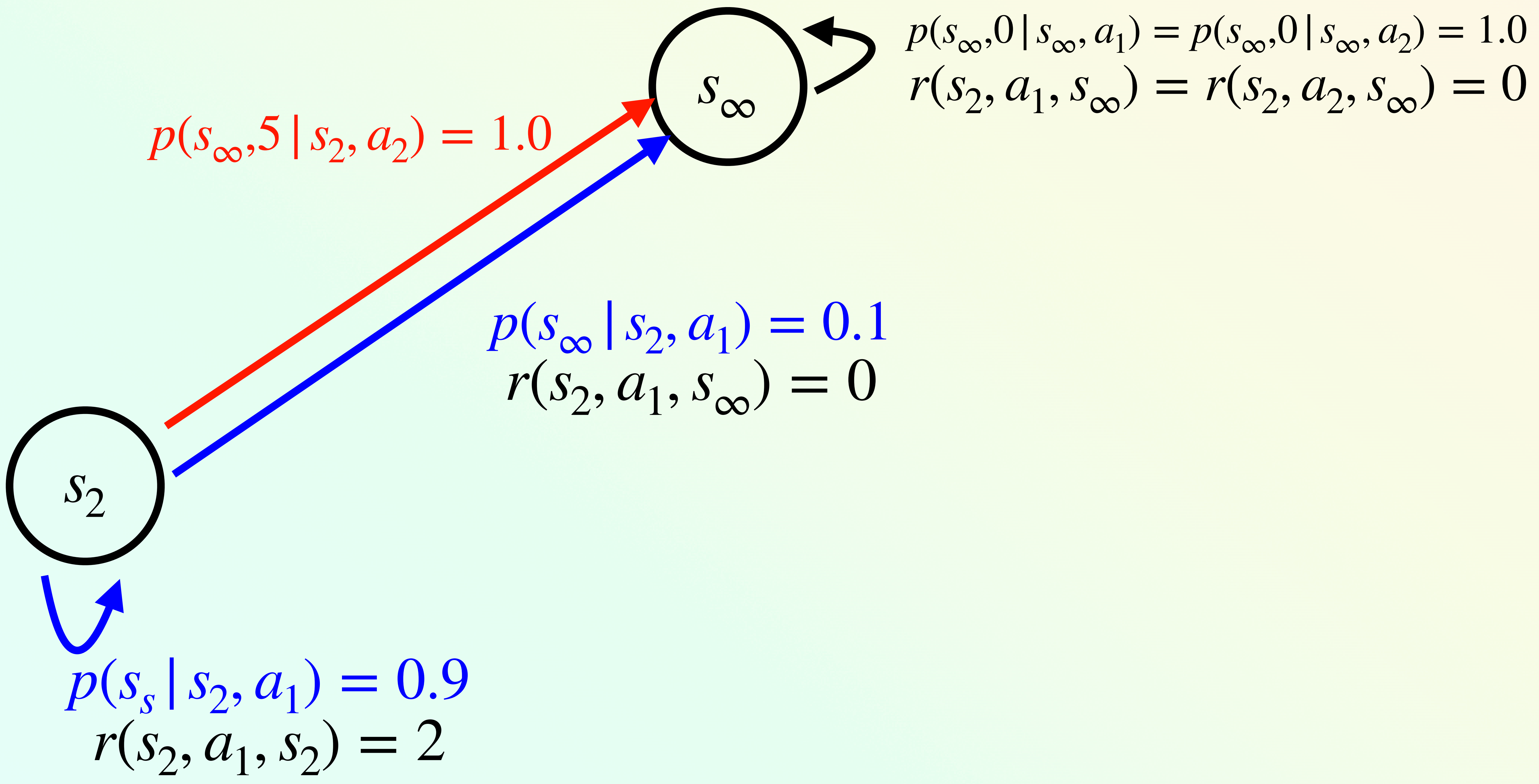
EXAMPLE

For $\pi(s_2) = a_1$

$$q_\pi(s_2, a_1) = \frac{2}{1 - 0.9\gamma}$$

$$q_\pi(s_2, a_2) = 5$$

$$v_\pi(s_2) = q_\pi(s, a_1)$$



COMPUTING VALUES FUNCTIONS

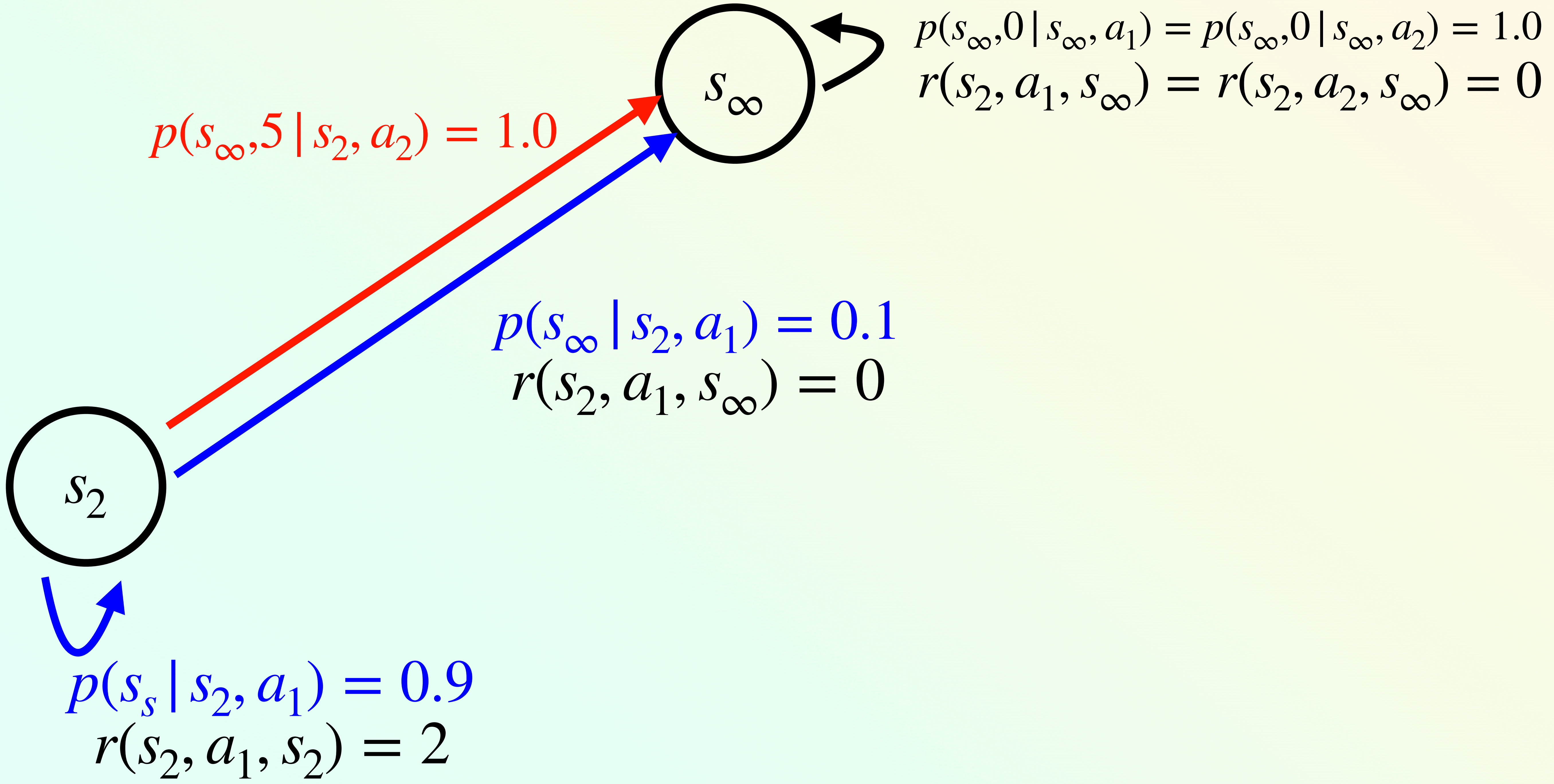
EXAMPLE

For $\pi(s_2) = a_2$

$$\begin{aligned} q_\pi(s_2, a_1) = &\ 0.9\left(r(s_2, a_1, s_2) + \gamma r(s_2, a_2, s_\infty)\right) + 0.1(0) \\ = &\ 0.9(2 + \gamma 5) \end{aligned}$$

$$q_\pi(s_2, a_2) = 5$$

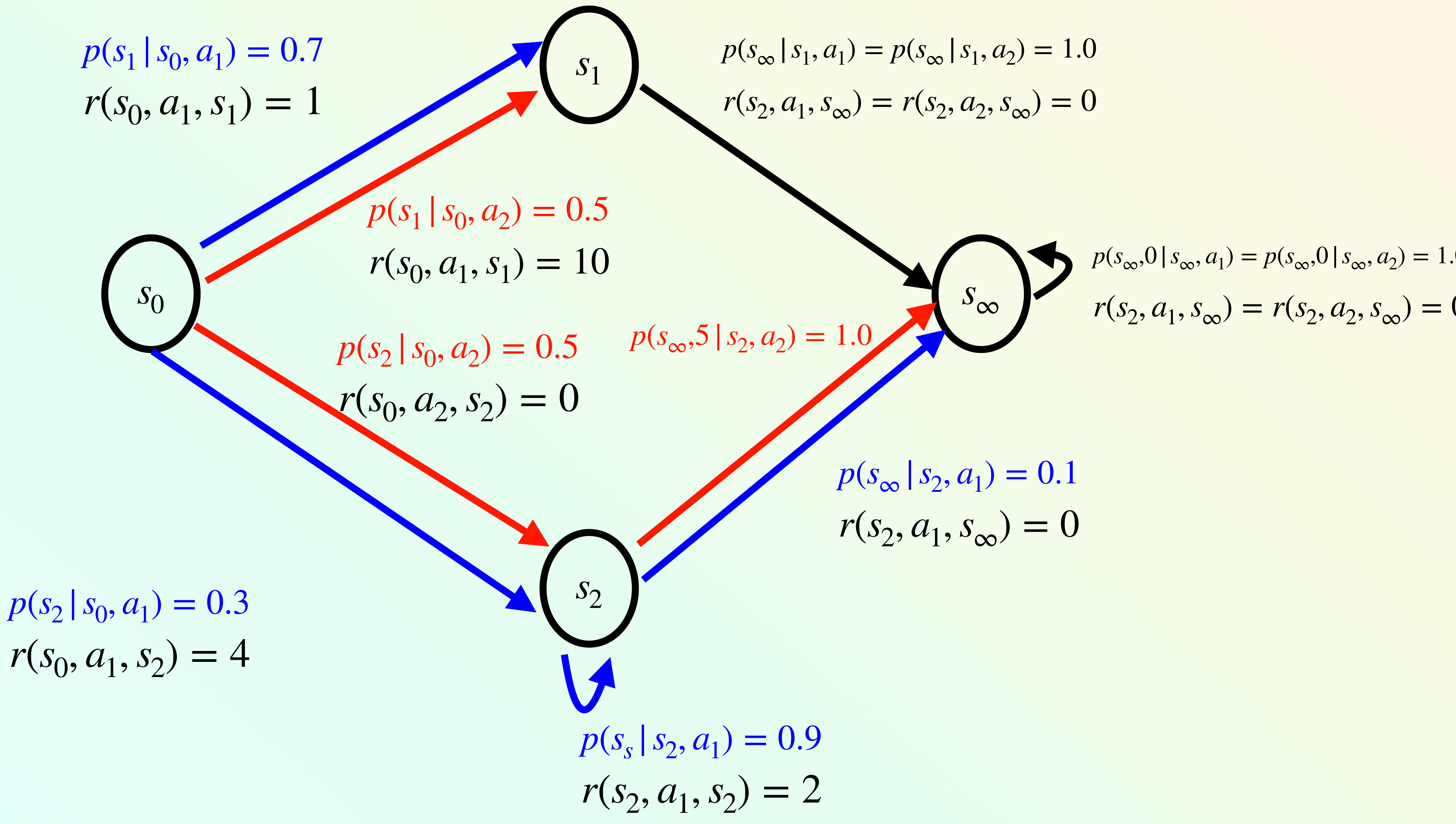
$$v_\pi(s_2) = q_\pi(s_2, a_2) = 5$$



COMPUTING VALUES FUNCTIONS

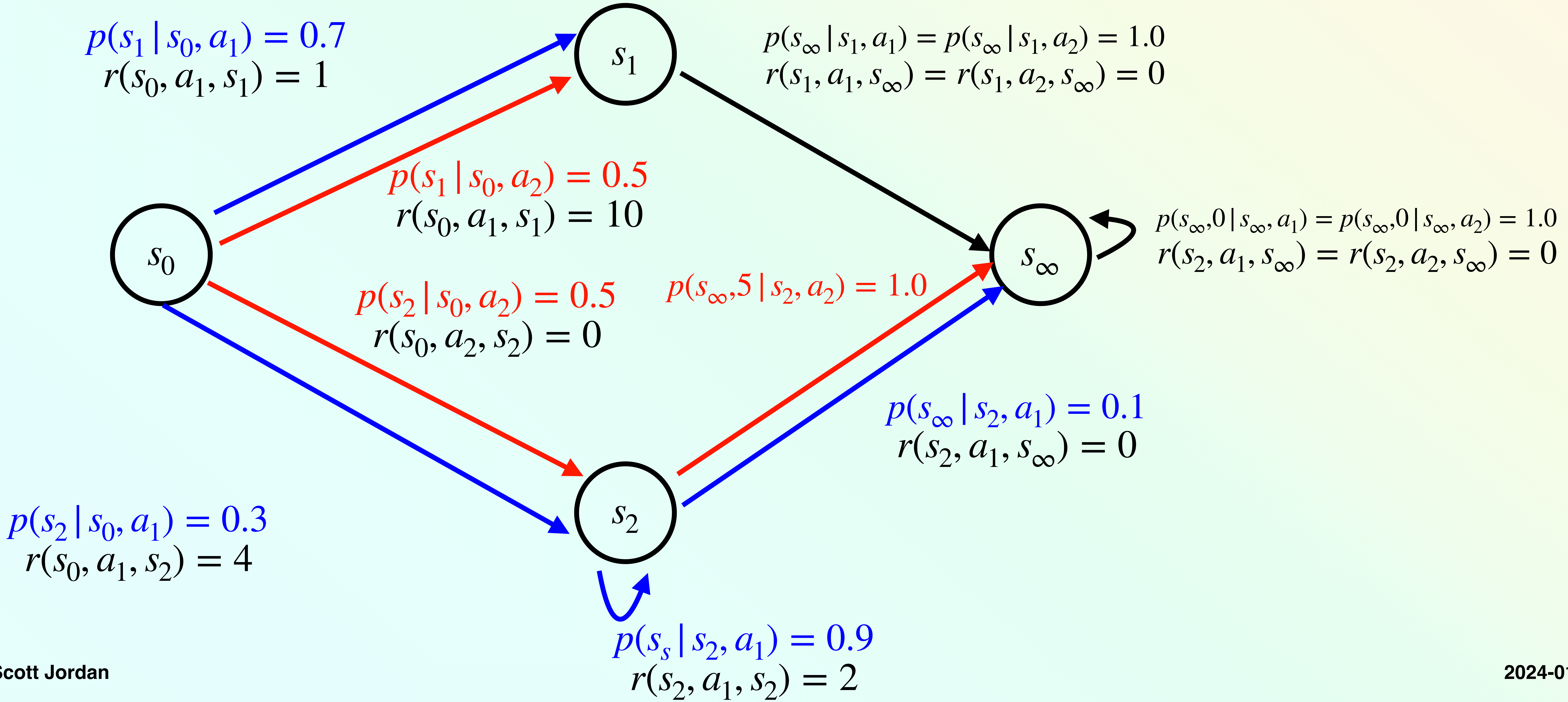
EXAMPLE

$$q_\pi(s_0, a) = ?$$



COMPUTING VALUES FUNCTIONS

EXAMPLE



COMPUTING VALUE FUNCTIONWS

$q_\pi(s_0, a)$

$$
\begin{aligned}
q_\pi(s_0, a_1) &= \mathbb{E}[G_t \,|\, S_t = s_0, A_t = a_1] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} \,|\, S_t = s_0, A_t = a_1] \\
&= \underbrace{\mathbb{E}[R_{t+1} \,|\, S_t = s_0, A_t = a_1]}_{=r(s_0, a_1)} + \gamma \mathbb{E}[G_{t+1} \,|\, S_t = s_0, A_t = a_1]
\end{aligned}
$$

COMPUTING VALUE FUNCTIONWS

$q_\pi(s_0, a)$

$$
\begin{aligned}
r(s_0, a_1) &= \mathbb{E}[R_{t+1} \mid S_t = s_0, A_t = a_1] \\
&= \sum_{s'} p(s_0, a_1, s') r(s_0, a_1, s') \\
&= p(s_0, a_1, s_1) r(s_0, a_1, s_1) + p(s_0, a_1, s_2) r(s_0, a_1, s_2) \\
&= 0.7(1) + 0.3(4) = 1.9
\end{aligned}
$$

COMPUTING VALUE FUNCTIONWS

$q_\pi(s_0, a)$

$$
\begin{aligned}
\mathbb{E}[G_{t+1} \,|\, S_t = s_0, A_t = a_1] &= \sum_{s'} p(s_0, a_1, s')\mathbb{E}[G_{t+1} \,|\, S_t = s_0, A_t = a_1, S_{t+1} = s'] \\
&= \sum_{s'} p(s_0, a_1, s')\underbrace{\mathbb{E}[G_{t+1} \,|\, S_{t+1} = s']}_{t'=t+1} \\
&= \sum_{s'} p(s_0, a_1, s')\mathbb{E}[G_{t'} \,|\, S_{'} = s'] \\
&= \sum_{s'} p(s_0, a_1, s')v_\pi(s') \\
&= p(s_0, a_1, s_1)v_\pi(s_1) + p(s_0, a_1, s_2)v_\pi(s_2)
\end{aligned}
$$

COMPUTING VALUE FUNCTIONWS

$q_\pi(s_0, a)$

$$\begin{aligned}
q_\pi(s_0, a_1) &= \mathbb{E}[G_t \,|\, S_t = s_0, A_t = a_1] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} \,|\, S_t = s_0, A_t = a_1] \\
&= r(s_0, a_1) + \gamma \sum_{s'} p(s_0, a_1, s') v_\pi(s')
\end{aligned}$$

This is one version of the Bellman equation. We will finish this example and cover the bellman equation next class.

NEXT CLASS

WHAT YOU SHOULD DO

1. Quiz due Friday night: Value Functions and Bellman Equations 1

Monday: Continuation and Bellman Equation