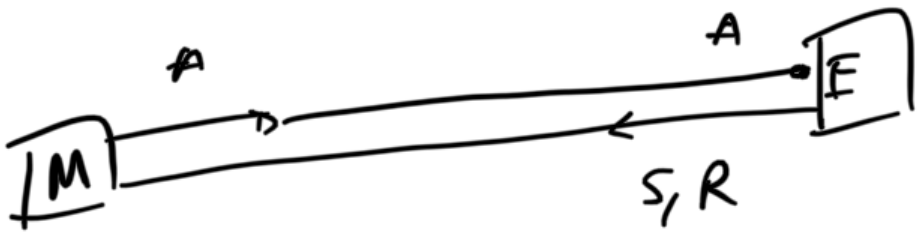
APS 1080 LECTURE 2

REINFORCEMENT LEARNING

If observability is missing and if what you see could be in different states

Markov Decision Process require observability.



$A = \{A^{(1)}, \ldots, A^{(N)}\}$ ↑ The goal

Goal

Reward signal does not come intrinsically from the environment.

- -1 → for the steps missed
- $-\infty$ → Loss
- 100 → win

Agent's mechanism to generate an action A based on S is called a policy.

↓

$\pi(A|S)$

Policy of an action given a state.

Policy is to design a distribution for the policy

Our goal is

so that an agent interacts with the environment appropriately

$$A = \{ A^{(1)}, \dots, A^{(A)} \}$$

($A^{(i)}$: i → index of action)

$S_0 \; A_0$

↑

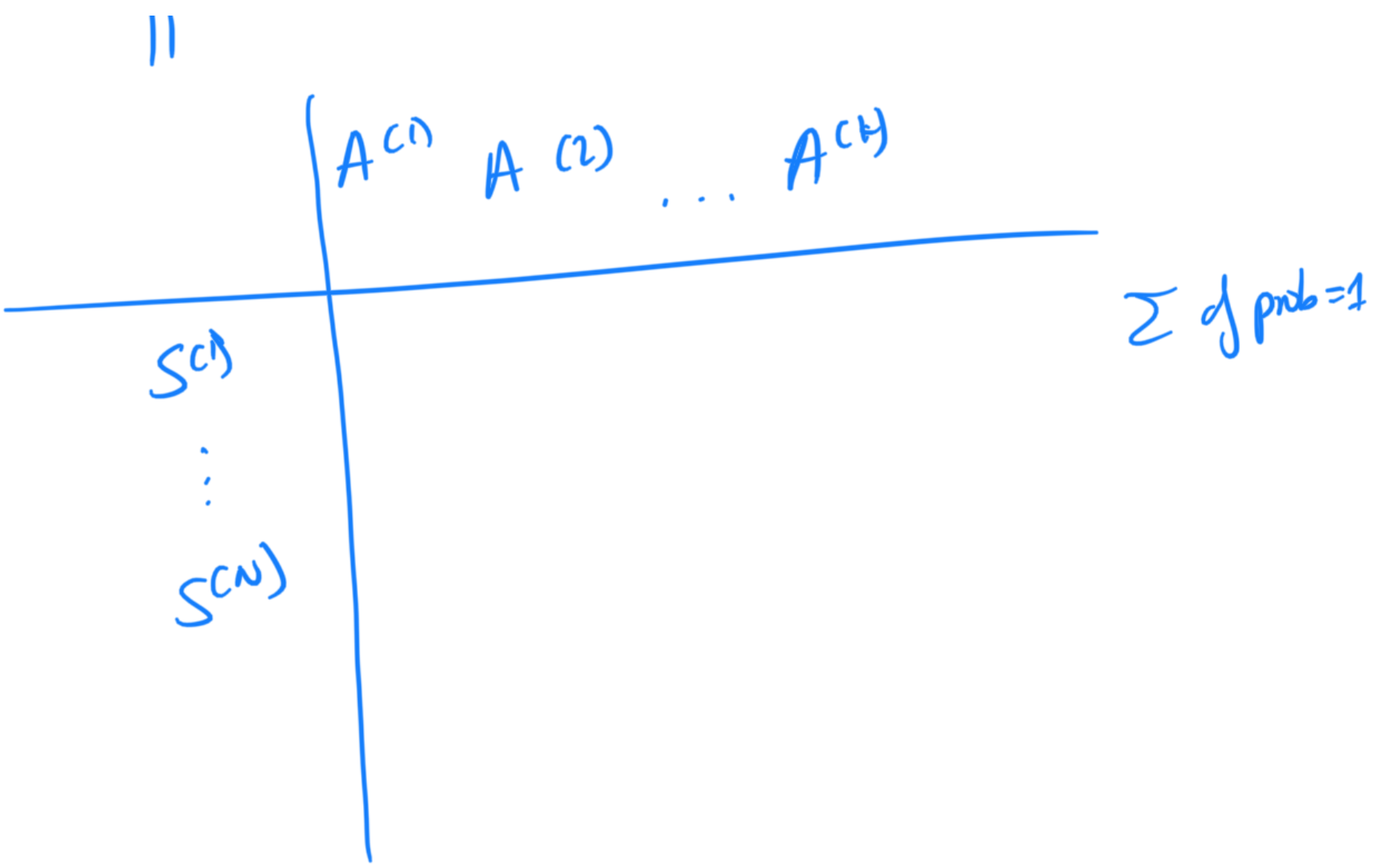
time index

Action selection Problem

$$S_0 A_0 \; R_1 \; S_1 A_1 \; \dots \; S_i A_i \; R_{i+1} \; S_{i+1} A_{i+1} \dots R_S S_T$$

This sequence is called an episode or a trace

what action do I select given a state?



The sum of the Pars must equal 1.

We may initialize the policy to random values.

What matters is not the initial conditions but rather how those values evolve.

How is the agent selecting the best action?

Subsequent sequence of rewards is maximized.

From time step i to the end of the episode

at time t.

We want to maximize the sum of future rewards.

Definition

Return:

$$G_T = \sum_{K=t+1}^{T} \gamma^{K-(t+1)} R_K$$

Cumulative sum of all future rewards

γ → discount factor

↳ A number between 0 and 1 if $T \to \infty$

$\gamma \in [0,1)$ $T \to \infty$

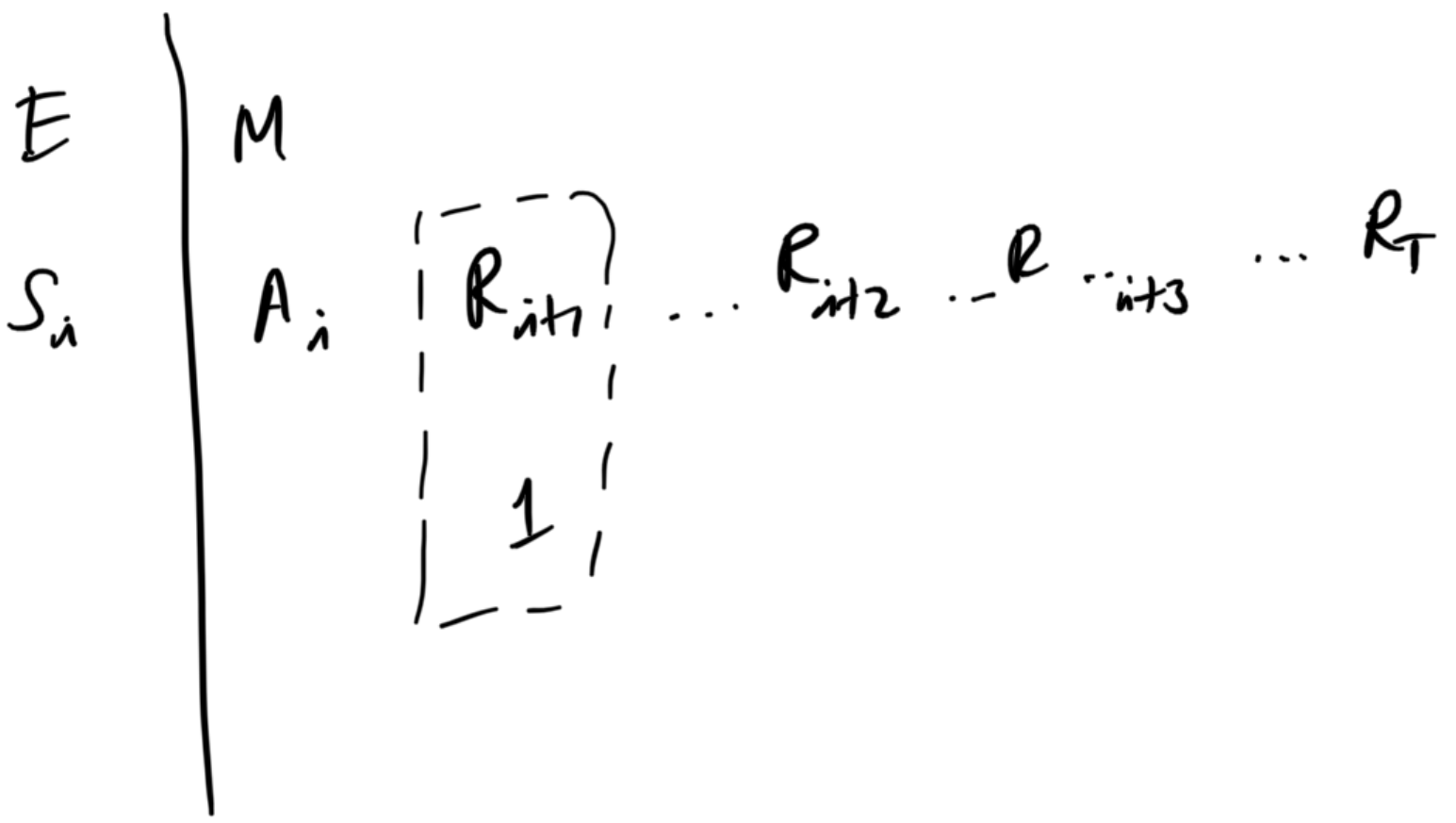
$\gamma \in [0,1]$ $T \to$ Finite.

Long term view

shorter term view

As $\gamma \to 0$ → Distant Rewards get attenuated

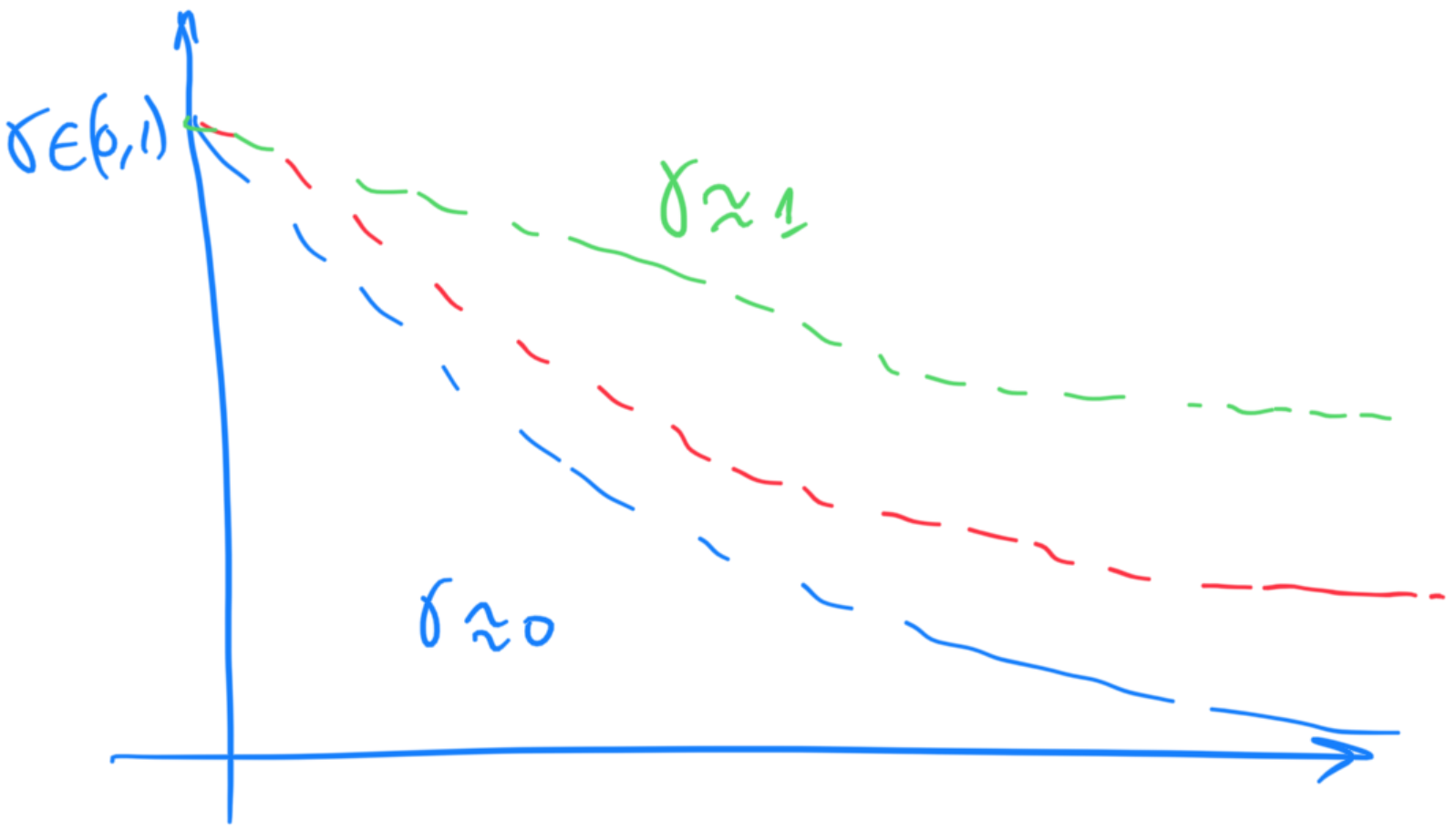
$\gamma \to 1$ → Short term rewards get attenuated



Maximize feature reward

we want to maximize all the things related to my work

We always account for next reward with full strength



long term view → $\gamma > 0$

Short term view → $\gamma = 0$.

The ACTION SELECTION PROBLEM

Given S_i, chose A_i such that we maximize the cumulative future rewards (G_i) which is called the return.

Our goal is to create a policy that maximizes the Cumulative Return G_i.

→ How will the rewards look like for every combination of actions

→ Suppose we can learn by historical episodes how being in state S_i relates to the return G_i for all states.

→ How being in S_1

Combinations of states and Actions and how they relate to the returns that they generate

VALUE FUNCTIONS

Keep track of what results in the best result historically

State value function

$$V_\pi(s) = \mathbb{E}\left[G_t \mid S_t = s\right] \quad \forall s$$

G_t → Return

$\forall s$ → for all states

s	Expected value of Future returns. $\mathbb{E}(G_t \mid S_t = s^{(i)})$
$s^{(1)}$	#
$s^{(2)}$	#
⋮	⋮
$s^{(n)}$	#

"

We populate this table by processing the traces or sequences that we pick up by having the agent interacting with the environment.

ACTION VALUE FUNCTION ($q_\pi(s,a)$)

Expected future return given a current state and a current action.

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

$S \times A$		
$S^{(1)}$	$A^{(1)}$	# $\mathbb{E}\left[G_i \middle| S_i = S^{(1)}, A_i = A^{(1)} \right]$
	$A^{(2)}$	#
	⋮	
	$A^{(k)}$	
$S^{(2)}$	$A^{(1)}$	
	$A^{(2)}$	
	⋮	
	$A^{(k)}$	
⋮		
$S^{(N)}$	$A^{(1)}$	
	$A^{(2)}$	
	⋮	
	$A^{(k)}$	

Given a state, choose the action that maximizes the reward.

$Q_\pi(s,a)$ → helps to solve the ACTION SELECTION PROBLEM as the maximum value can be computed directly.

↳ value function

If you have access to the V value function

$V_\pi(s)$ → What is the expected Return being in each state? which state is the best one to be in?

$s^{(1)}$	#
$s^{(2)}$	#
⋮	⋮
$s^{(4)}$	#

Some states are more desirable than others, $V_\pi(s)$ tells you which are the most desirable states to be in.

Know what the tables look like for the TEST

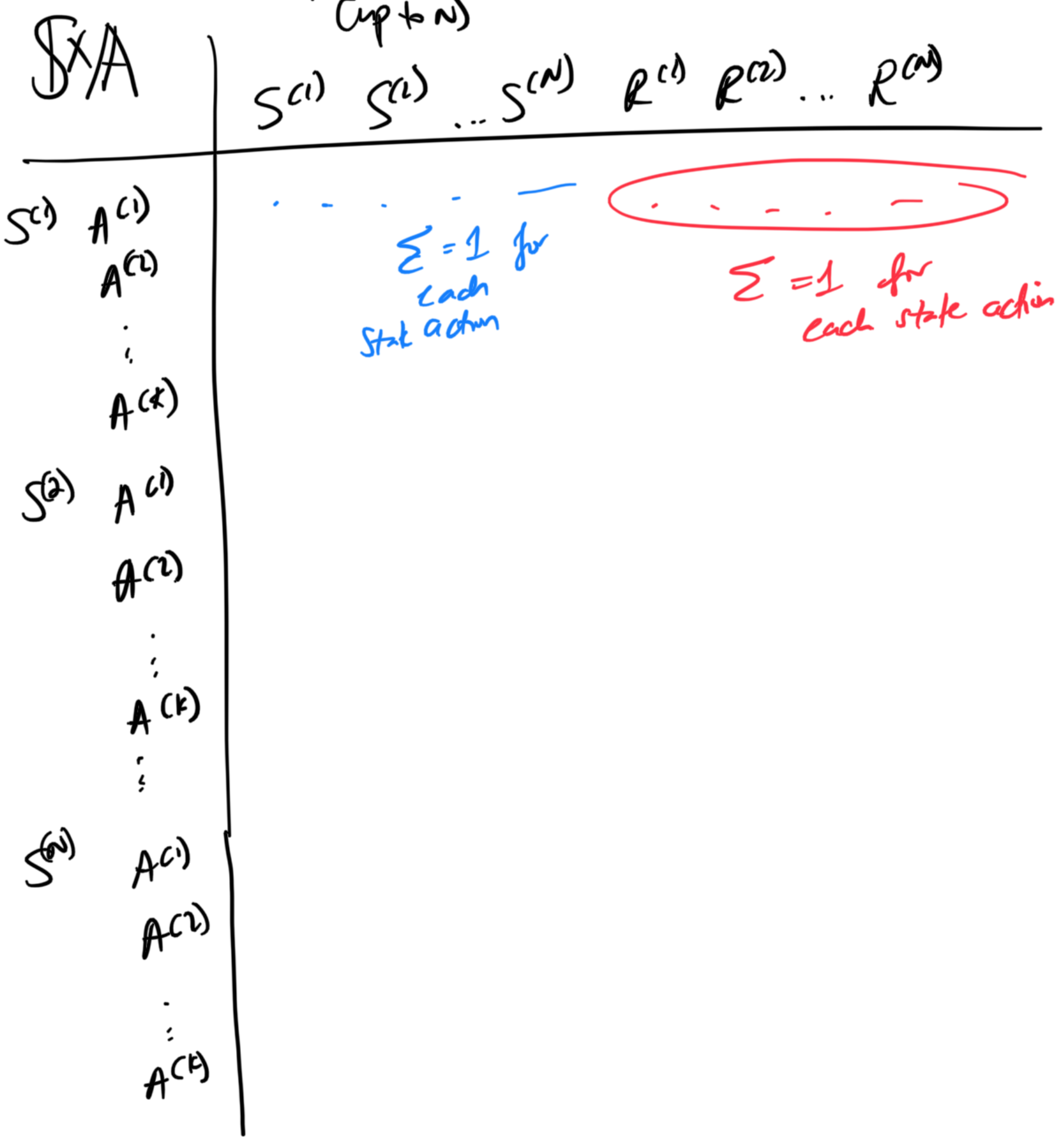
You can select a value with Q only

with V, you don't have a mean to select an

action.

With V alone you need a model of the environment to make it useful.

$$p(s', r, s, a) = \Pr\{S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a\}$$



We want to chose an action that has the highest likelihood of getting me to the state that gives me the highest value. (S^*)

Given a stochastic model of the environment and a value function, how can you solve the Action selection Problem

What all you need

If you have $q \rightarrow$ then ... π

If you have $V \rightarrow$ You need V_π. and p

$$V_\pi(s) = \sum_{\forall a \in \mathbb{A}} \pi(a|s) \sum_{s,R} p(s', R, s, a) \left[R + \gamma V_\pi(s') \right]$$

Next Lecture:

How to obtain V and q given p

Next Next Lecture

How to obtain V and q if we do not have p.

S: State space → Surface or abstract state.

A: Action Space → Click, Impression, Repin...

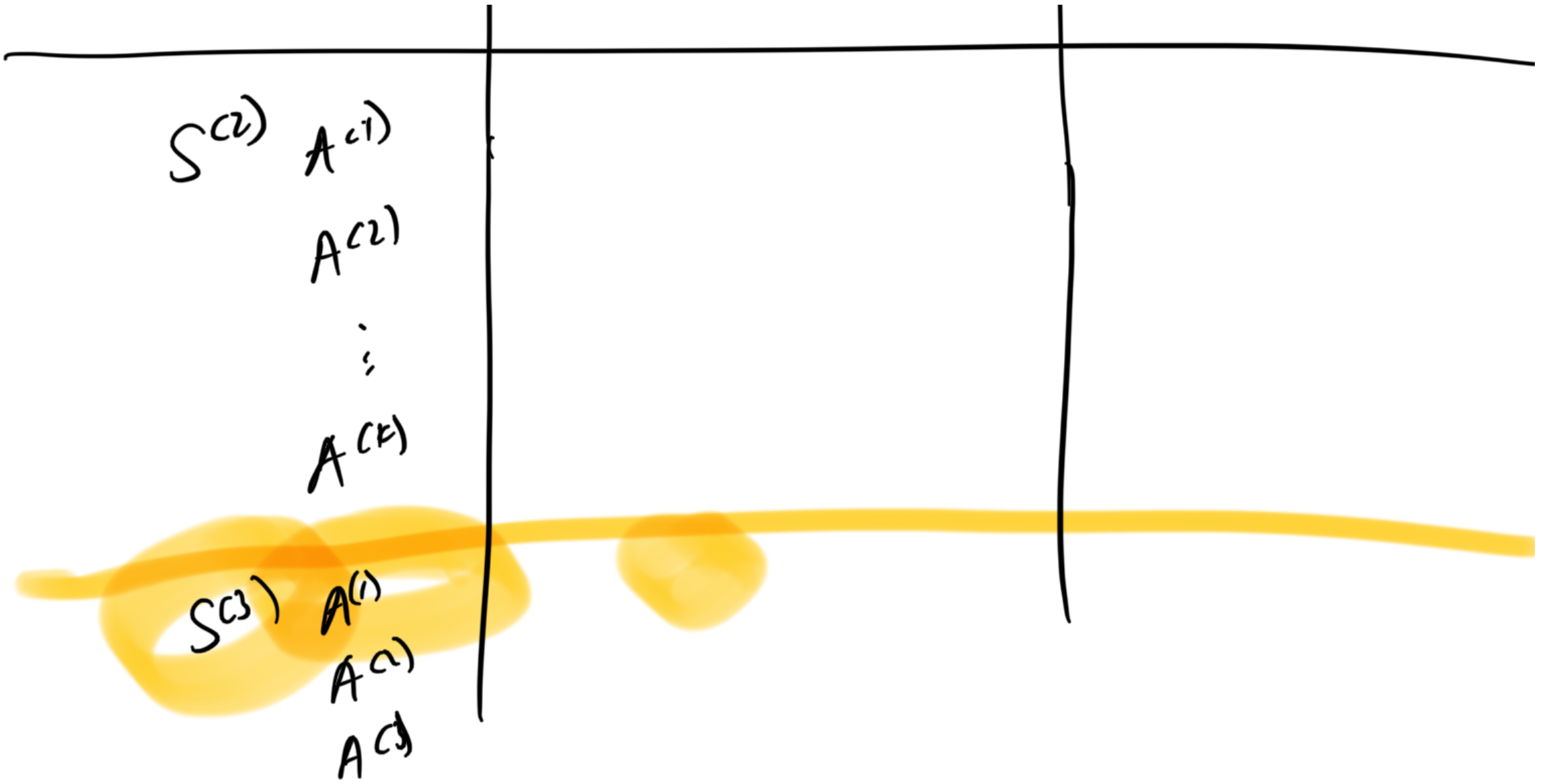
q: Transition rule function → Likelihood of going from one state to the other.

R: Reward function } - Assign points for things you know
Π: Policy. } - Initialize to a random value for things you don't know

γ: Discount factor.
↳ How much you care about the future.

ACTION VALUE FUNCTION (one hop)

	Next states			Next state Reward		
	$S^{(1)}$	$S^{(2)}$	... $S^{(N)}$	$R^{(1)}$	$R^{(2)}$	.. $R^{(N)}$
$S^{(1)}$ $A^{(1)}$			$\Sigma = 1$			$\Sigma = 1$
$A^{(2)}$						
:						
$A^{(K)}$						



$S^{(1)}$ → HF

$S^{(2)}$ → Search

$S^{(3)}$ → Closeup

$A^{(1)}$ → click

$A^{(2)}$ → Impression

RETURNS AND EPISODES

The Agent's goal is to maximize the cumulative reward it receives over time.

Downstream Rewards @ Netflix.

Instant gratification ⟹ Long term objective optimization

↓

Replay.

↓

optimize for Best sequence (long term)

Markov Decision Processes.

$A: \{A_1, A_2 \cdots A_n\}$ → Action space

Click
Repin
Impression ...

$S: \{S_1, S_2, \cdots S_k\}$ → State space.

Abstract state → Exploration, consideration, Fulfillment.

Surfaces.

T: Transition Rate function.

R : Reward Function. → Relative value of different Actions and states.

→ Hide → -100

→ Share → 100

→ π : Policy

γ : Discount factor.

↳ How much you care about the future.

	Next State (up to k)	Reward from the next state.
	$S^{(1)}$ $S^{(2)}$ ... $S^{(k)}$	$R^{(1)}$ $R^{(2)}$ ... $R^{(k)}$
$S^{(1)}$ $A^{(1)}$		$\Sigma = 1$
$A^{(2)}$		
$A^{(N)}$		
$S^{(2)}$ $A^{(1)}$		
$A^{(2)}$		
$A^{(N)}$		
$S^{(K)}$ $A^{(1)}$		

$A^{(1)}$
⋮
$A^{(N)}$

Choosing upto 3-5 Actions

Starting with one surface, i.e. Home Feed.

$S^{(1)}$: Homefeed
$S^{(2)}$: Search
$S^{(3)}$: Related Pins

- User ID
- time stamp begin
- time stamp end
- action (CRUD, repin, click, etc)
- Surface origin
- Surface Destination
- Pin ID.

- Web / Mobile
- Country
- Date joined
- Gender
- Age
- L1 interest
- iOS / Android
- Version

TO normalize population.

$S \times A$	
$S^{(1)}$ $A^{(1)}$	$\# \quad \mathbb{E}\left[G_i \mid S_i = S^{(1)}, A_i = A^{(1)}\right]$
$A^{(2)}$	$\#$
⋮	
$A^{(k)}$	
$S^{(2)}$ $A^{(1)}$	
$A^{(2)}$	
⋮	
$A^{(k)}$	
⋮	
$S^{(N)}$ $A^{(1)}$	
$A^{(2)}$	
⋮	
$A^{(k)}$	