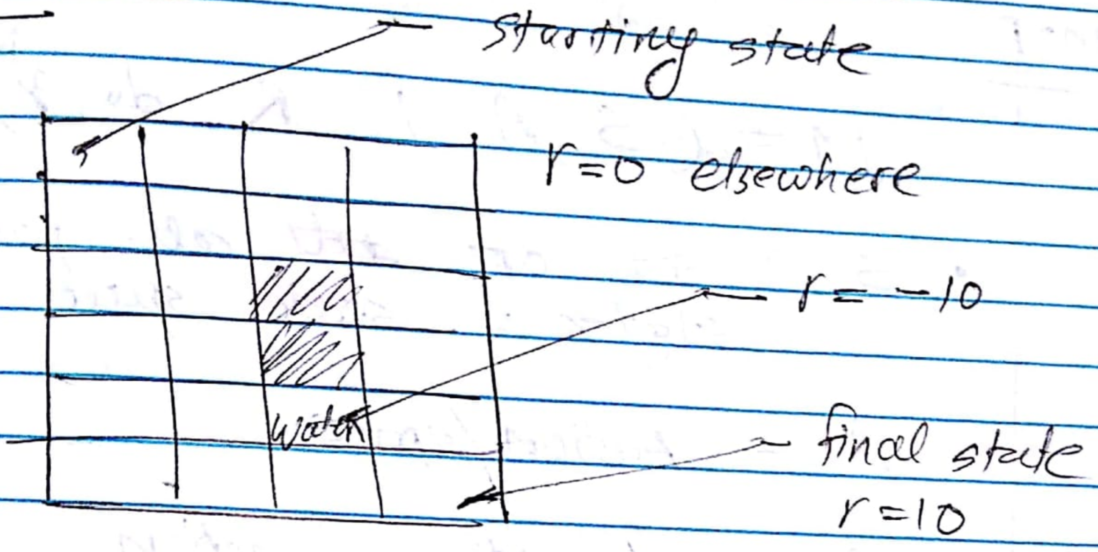
CS - Reinforcement Learning

Example:



$P = 0.8$ actⁿ suceeds

$P = 0.05$ $+90°$

$P = 0.05$ $-90°$

$P = 0.1$ stay

run into wall = stay

MDP :

- Mathematical formulatⁿ of environment, what we want agent to learn

- $t \in \{0, 1, \ldots\}$ time step

- S_t, A_t, R_t

state, action, reward respectively

<u>MDP</u> :

$$M = (S, A, P, R, d_0, \gamma)$$

- S = set of ~~all~~ all possible states = state ~~space~~ Set
- A = Action ~~space~~ set
- P = transition function

$$P : S \times A \times S \Rightarrow [0, 1]$$

$$P(s, a, s') \triangleq P_r(S_{t+1} = s' \mid S_t = s, A_t =$$

$$\ldots \quad \forall \; s \in S, \; a \in A, \; s' \in S$$

<u>NOTE</u>: Deterministic $P \Rightarrow P(s, a, s') \in \{0,$

$$\& \sum_{s'} P(s, a, s') = 1$$

• R = Reward function, determines how rewards are generated!

$R: S \times A \times S \Rightarrow \mathbb{R}$

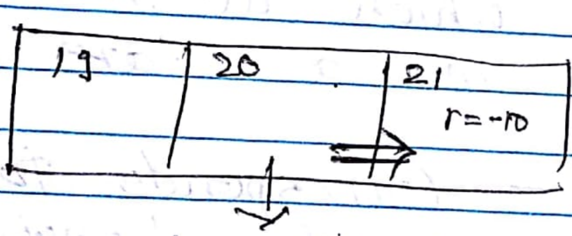
$R(s, a, s')$ ~~=~~

$= E\left[R_t \mid S_t = s, A_t = a, S_{t+1} = s'\right]$

e.g.,

* in robot gridworld problem

$R(20, \text{right}, 21) = -10$



s	a	s'
$= 20$	$=$ right	$= 21$

NOTE: From answering a question — not the notation we will use hereafter.

$R(s,a)$ = ~~P~~ Probability × reward value	$R(s,a,s')$ = deterministic reward value (for deterministic reward not for stochastic reward)

- d_0 = Initial state distribution

$$d_0 : S \rightarrow [0, 1]$$

$$d_0(s) = P(S_0 = s)$$

- γ = reward discount parameter.

Agent Formulation (for MDP)

- Policy – The mechanism within the agent that determines which action to take in a state

- Learning – Corresponds to the agent changing its policy

Policy = $\pi : S \times A \Rightarrow [0, 1]$

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

if $|S|$ and $|A|$ are finite, if R_t is bounded then an optimal policy exists

Applying MDP for gridworld problem :

$S_0 \sim d_0$
$a_0 \sim \pi(S_0, \cdot)$
$S_1 \sim P(S_0, A_0, \cdot)$
R_0 computed with $E[R_0] = R(S_0, A_0, S_1)$

What agent want to do?

Goal : Maximize expected reward

Objective function : $J : \pi \Rightarrow R$

$$J(\pi) \triangleq E\left[\sum_{t=0}^{\infty} R_t \,\middle|\, \pi\right]$$

$$\pi$$

$A_0 \sim \pi(S_0, \cdot)$
$A_1 \sim \pi(S_1, \cdot)$
$A_2 \sim \pi(S_2, \cdot)$
$\vdots$

Optimal policy : $\pi^* \in \arg\max_{\pi \in \Pi} J(\pi)$