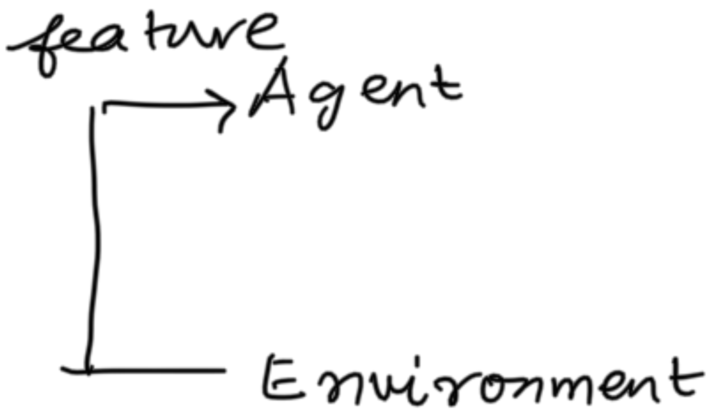
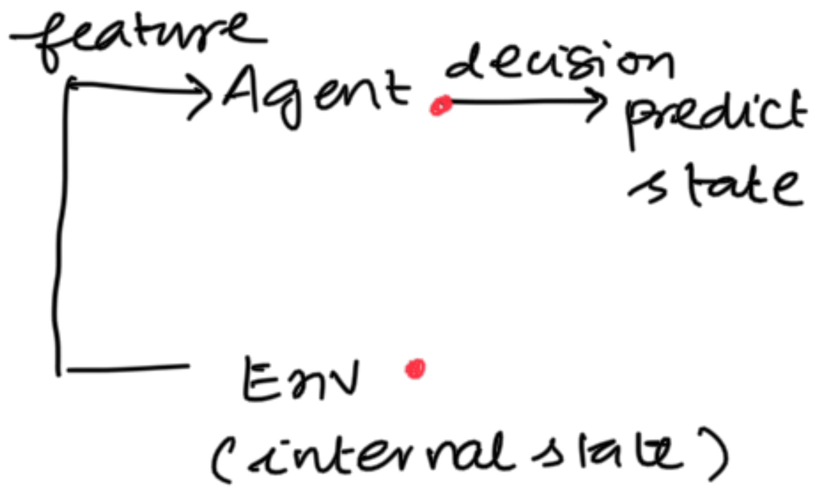
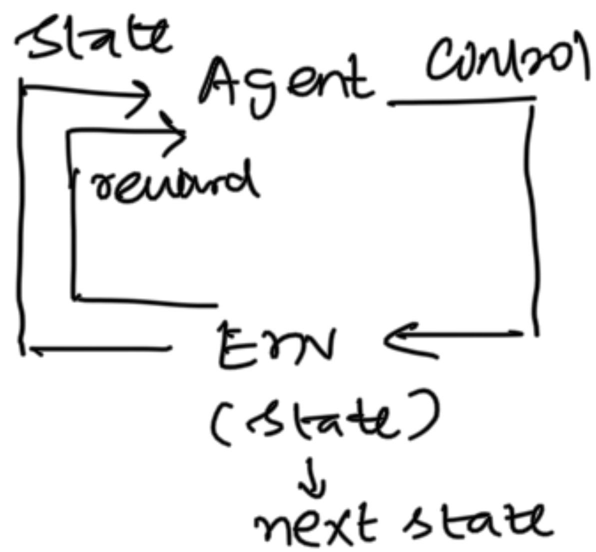






Descriptive

unsupervised

No decision

Predictive

Supervised

Predictive decision

No loop

Control

Reinforced

control decision

+ loop

Multi Armed Bandits

Goal: To look at the challenges in the control or reinforcement learning setting and slice out the multi-armed bandits

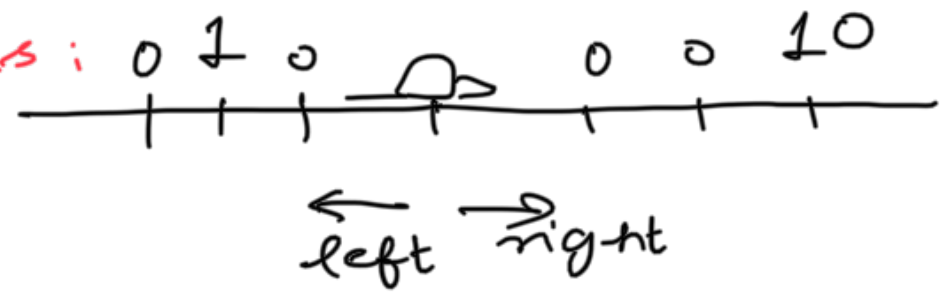
Application of control: (learning to play) playing chess/Go, autonomous driving, robotics (Mujoco), play video games (Atari)

inventory control, traffic signal control, optimal investment, online ad placement

Challenge 1:

env: deterministic,

rewards:



$$\sum_{t=1}^{T} r_t$$

$$\sum_{t=1}^{T} (0.9)^t r_t$$

Say T = 2, go left ⟆ switched
T = 3, go right
T = 4, go right

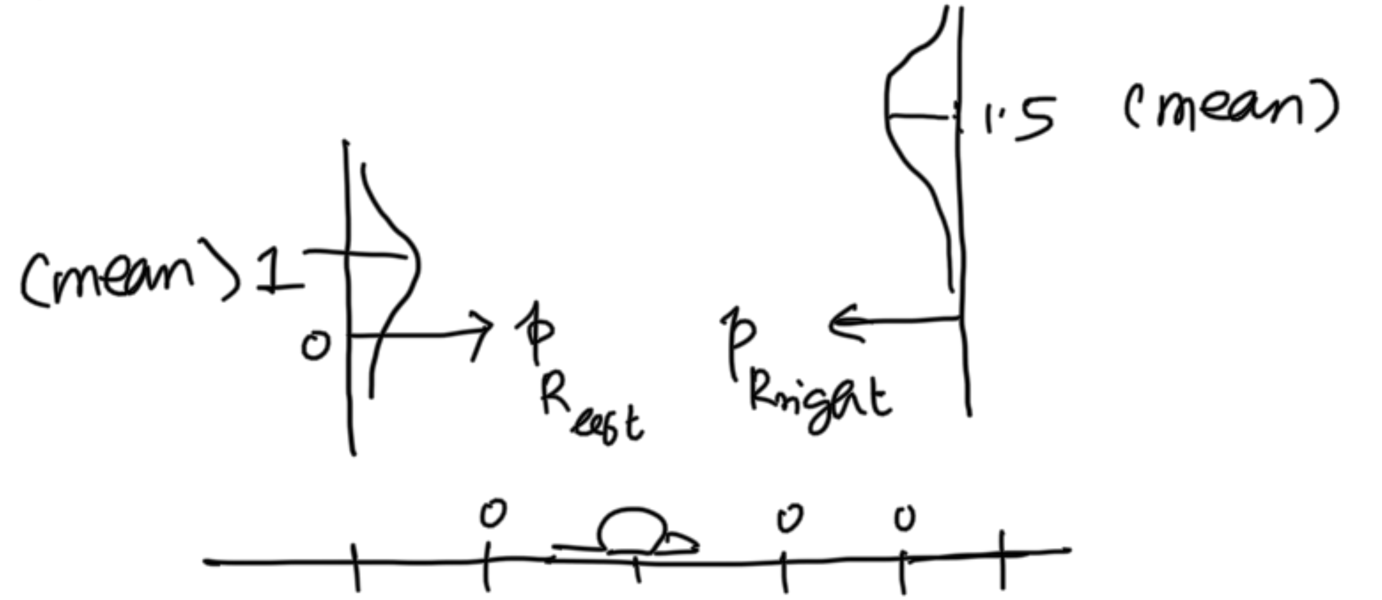
T = 5, go right
:
:
T = 8, go right
T = 9, go right for 3 steps then left

Moral: Immediate vs Future

Challenge 2: Say I play a game of chess, and I win, is it possible to pin point the winning move?

Moral: Temporal credit assignment

Challange 3! Environment is not deterministic



$$\max \mathbb{E}\left[\sum_{t=1}^{T} r_t \right]$$

* you have to visit each of these places multiple times

* say I visit each location 10 times, collect samples, look at sample mean and decide

(few good samples at left location and few bad samples at right location can screw us)

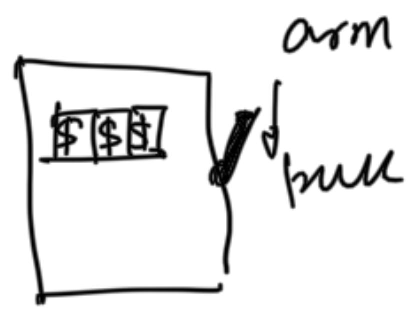
Moral: Cannot explore and commit

* in order to gain information we actually need ∞ samples (keep exploring)

* while we explore, we should not over explore bad choices, (exploit what we know)

Exploration vs Exploitation

Notation for Multi-Armed Bandits



- $A = \{1, \ldots, k\}$ arms

- arm a is associated with distribution P_a

- at time t, we pick A_t (pulling arm)

- Reward $X_t \sim P_{A_t}$

- Mean rewards $\mu(a) = \mathbb{E}_{P_a}[X_t]$

- Best reward $\mu_* = \max_a \mu(a)$

- Best arm $a_* = \arg\max_a \mu(a)$

- Suboptimality Gaps $\Delta_a = \mu_* - \mu(a)$

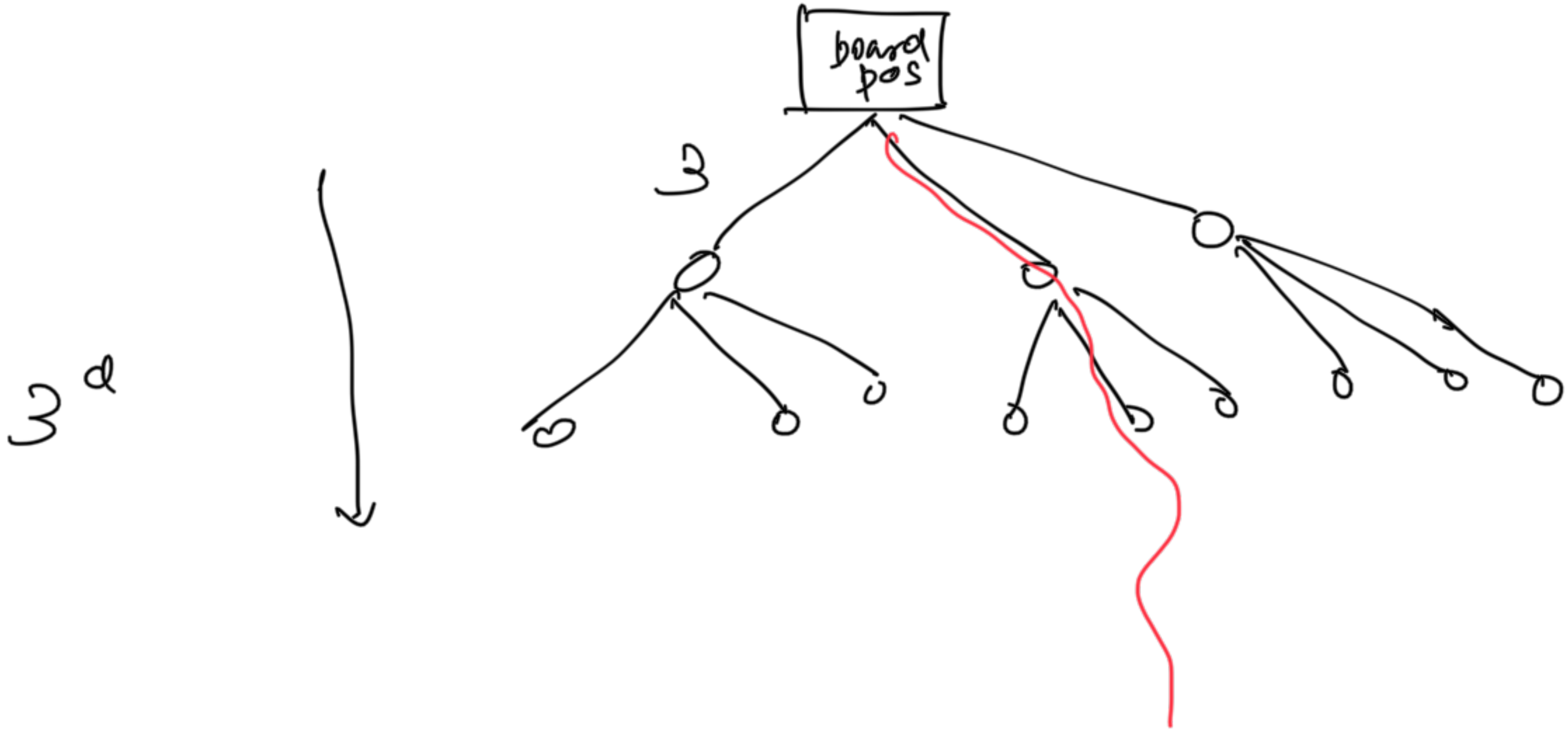
- Goal ; Minimize

$$R_n = \text{Regret}_n = \mathbb{E}\left[\sum_{t=1}^{n} \underbrace{(\mu_* - X_t)}_{R.V}\right]$$

$n = 10$

Why / where do we see the impact of explore vs exploit

Example: Game



How to solve the MAB problem

- Sample mean → True Mean

Sequence / Convergence of random variables → $\{X_i\}_{i \geq 0}$ iid $\quad \dfrac{\sum_{i=1}^{n} X_i}{n} \to \mathbb{E}[X_1]$

- Probability concentrates around true mean

Concentration of probability