



Reinforcement Learning

Lecture 14: Exploration; Approximation Intro

23 Sep 2025

Instructor: Raunak Bhattacharyya



ScAI

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Exploration vs Exploitation

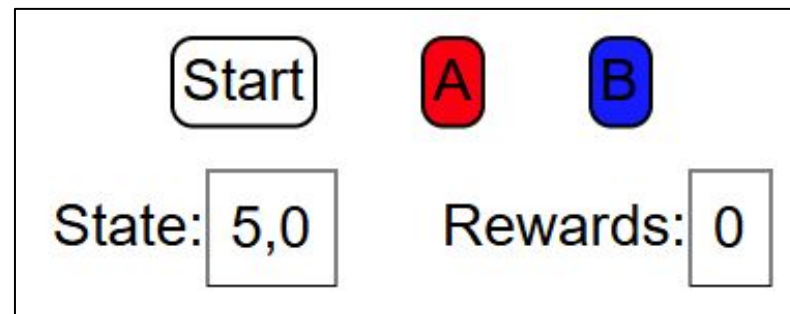
How can an agent decide whether to attempt new behaviors (to discover ones with higher reward) or continue to do the best thing it knows so far?

Exploitation: Do what you think will yield the highest reward

Exploration: Do the things that you haven't done before hoping that it will yield even higher reward



Seaquest, [Source](#)



[Mystery Game, Liang](#)



Montezuma's Revenge, [Source](#)



Pitfall, [Source](#)

Single-armed Bandit



One-arm bandit, [Source](#)

Multi-armed Bandit



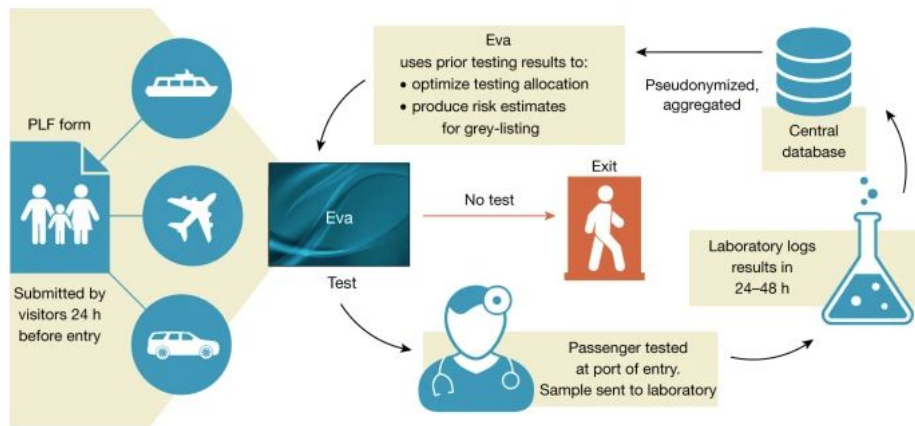
Multi-armed Bandit, Source: [Wikipedia](https://en.wikipedia.org/wiki/Multi-armed_bandit)

- Faced repeatedly with a choice among k diff options
- Make a choice
- Receive a numerical reward chosen from a stationary prob dist depending on the action you selected
- Goal is to maximise expected total reward over some time period: eg. 1000 action selections

Applications of Bandits

- Advert placement
- Recommender systems
- Tree search
- A/B testing

Fig. 1: A reinforcement learning system for COVID-19 testing (Eva).



Source: [Bastani, Nature 2021](#)

Formulation

$$\mathcal{A} = \{\text{pull}_1, \text{pull}_2, \dots, \text{pull}_n\}$$

$$p(r_i = 1) = \theta_i, \quad p(r_i = 0) = 1 - \theta_i$$

estimate $\hat{Q}_t(a) \approx Q(a) = \mathbb{E}[R(a)]$

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^{t-1} r_i \mathbb{1}(a_i = a)$$

Example: Treating a Broken Toe

- Surgery
 - Buddy taping
 - Do nothing
- Surgery: $Q(a^1) = \theta_1 = 0.95$
 - Buddy taping: $Q(a^2) = \theta_2 = 0.9$
 - Do nothing: $Q(a^3) = \theta_3 = 0.1$
- After 6 weeks: do an X-ray to check whether healed (1) or not (0)

Greedy Approach

Take action a^1 ($r \sim \text{Bernoulli}(0.95)$), get 0, $\hat{Q}(a^1) = 0$

Take action a^2 ($r \sim \text{Bernoulli}(0.90)$), get +1, $\hat{Q}(a^2) = 1$

Take action a^3 ($r \sim \text{Bernoulli}(0.1)$), get 0, $\hat{Q}(a^3) = 0$

$$a_t^* = \arg \max_{a \in \mathcal{A}} \hat{Q}_t(a)$$

Regret

- T denote the total number of rounds or time steps.
- K denote the number of possible actions (arms).
- $a_t \in \{1, \dots, K\}$ denote the action (arm) chosen by the agent at time t .
- $r_{a_t}(t)$ denote the reward obtained by choosing action a_t at time t .
- a^* denote the optimal action, which is the action with the highest expected reward:

$$a^* = \arg \max_{a \in \{1, \dots, K\}} \mathbb{E}[r_a]$$

Regret

The *cumulative regret* $R(T)$ after T rounds

$$R(T) = \sum_{t=1}^T (\mathbb{E}[r_{a^*}] - \mathbb{E}[r_{a_t}])$$

- Maximise cumulative reward
 - Equivalent to minimise total regret

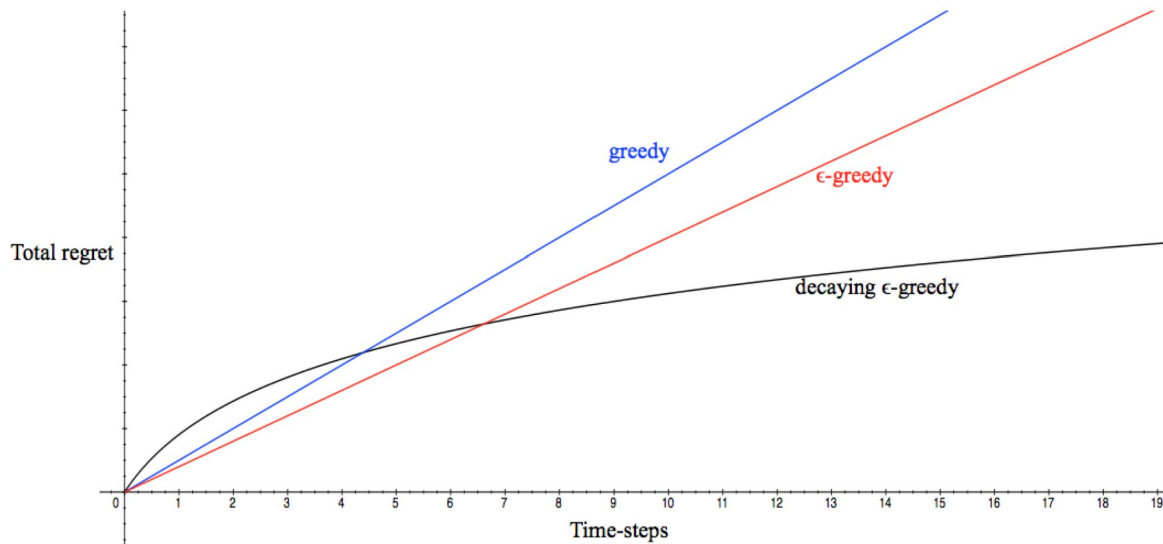
Regret: Greedy Approach

$$R(T) = \sum_{t=1}^T (\mathbb{E}[r_{a^*}] - \mathbb{E}[r_{a_t}]) \qquad R(T) = T \cdot \mathbb{E}[r_{a^*}] - \sum_{t=1}^T \mathbb{E}[r_{a_t}]$$

Action	Optimal Action	Observed Reward	Regret
a^1	a^1	0	0
a^2	a^1	1	0.05
a^3	a^1	0	0.85
a^2	a^1	1	0.05
a^2	a^1	0	0.05

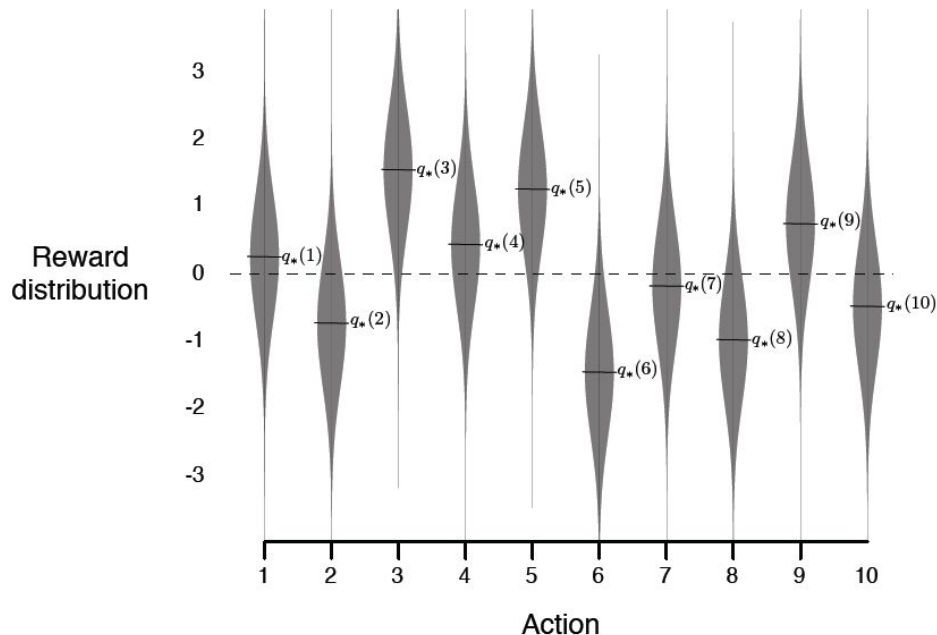
Cannot evaluate regret in real settings because requires knowledge of the true best action

Regret: Rate of Growth



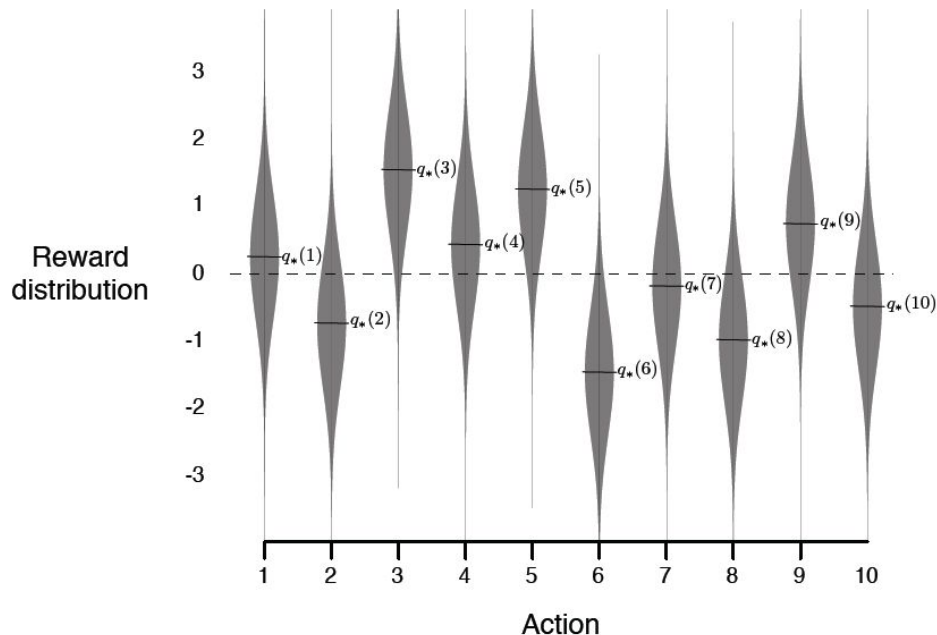
10-armed Testbed

- 2000 10-armed bandit problems
- For each bandit problem, set the true action values by sampling from a standard Normal distribution
- Then, when we test our proposed learning algo, the obtained reward at each timestep is sampled from a $\text{Normal}(Q(a_i), 1)$

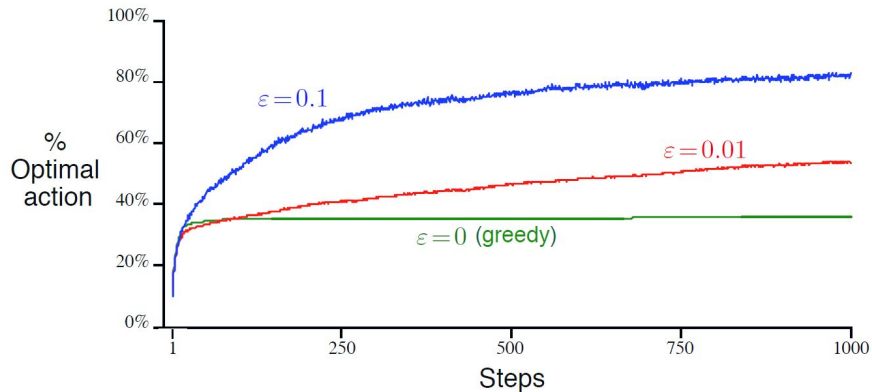
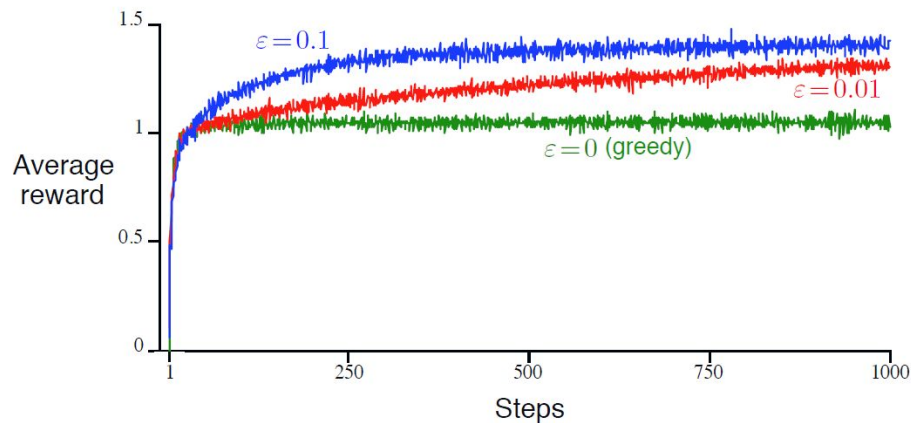


10-armed Testbed

- 2000 10-armed bandit problems
- One run: 1000 timesteps on a bandit problem
- Do 2000 runs, each run for a different bandit problem. This establishes the learning algo performance



Performance on 10-armed Testbed: Eps-Greedy



Value Function Approximation

'Real-World' RL

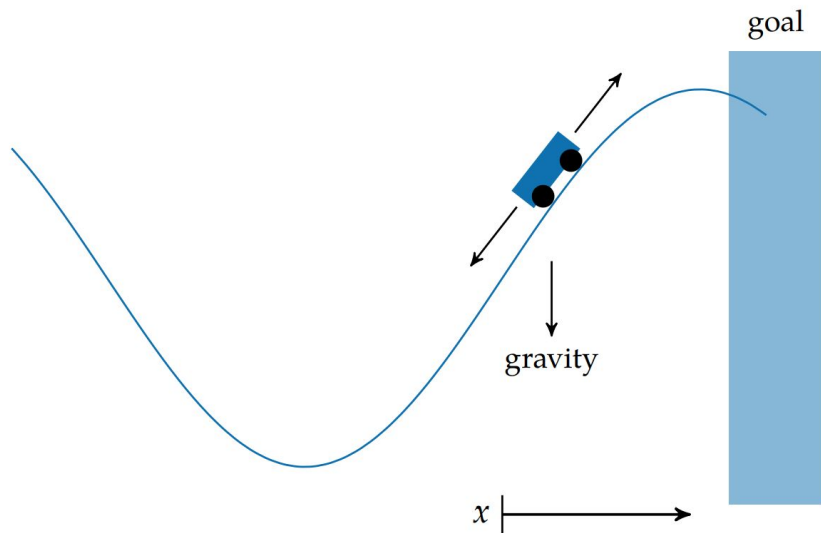


[Source: Deepmind, DQN](#)

128^{33600}



Mountain Car



[Source](#)

$$x \in [-1.2, 0.6]$$

$$v \in [-0.07, 0.07]$$

$$(a = -1)$$

$$(a = 1)$$

$$(a = 0)$$

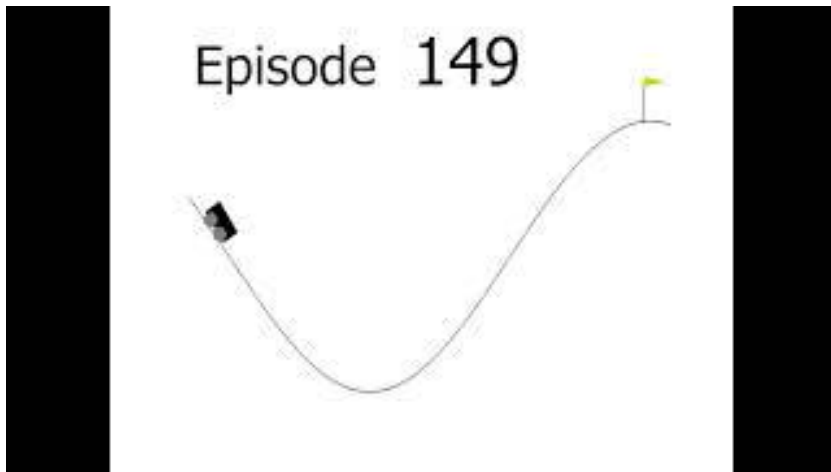
-1 reward every turn

past $x = 0.6$

Mountain Car

$$v' \leftarrow v + 0.001a - 0.0025 \cos(3x)$$

$$x' \leftarrow x + v'$$



[Source](#)