

# Reinforcement Learning

---

## Exercise 9: Multi-armed Bandits

08.07.2025

Alexander Mattick

# Exploration vs. Exploitation



# 3 Views on RL

## Why and How

---

### Dynamic Programming

- Table of State-Actions
- Recursion
- Try to find fixed point

Guarantees:  
Tabular: Easy  
Function Approx:  
None/Good luck

### Probabilistic Inference

- Try to find a distribution that represents optimal choices
- KL constraints to ensure stable information gain
- Naturally a variational inference problem

Guarantees:  
Sufficiently powerful models will eventually converge

### Decision Theory

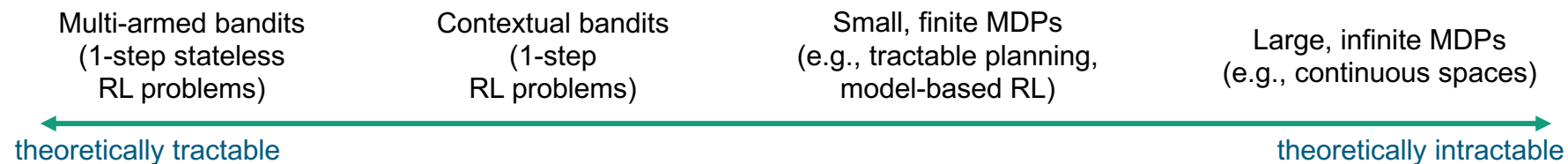
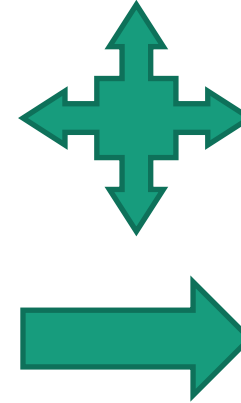
- Find optimal action amongst set
- Maximize utility/minimize Regret

Guarantees:  
Convergence is well understood + exact guarantees are possible!

# Exploration vs. Exploitation

## Why and How

- Both definitions stem from the same problem:
  - Exploration:** do things you haven't done before (in the hopes of getting even higher reward)  
→ increase knowledge
  - Exploitation:** do what you know to yield highest reward  
→ maximize performance based on knowledge

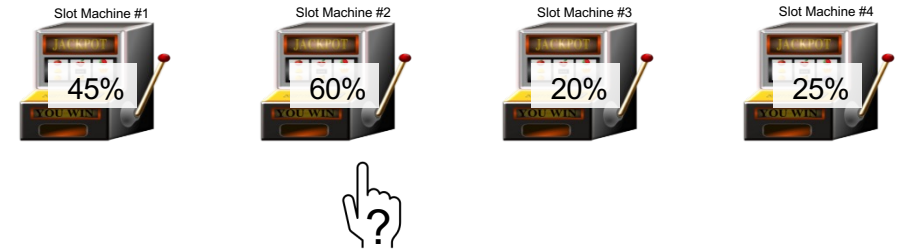


*(illustration adapted from Sergey Levine's CS285 class from UC Berkeley)*

# Exploration vs. Exploitation

## Multi-Armed Bandits and Regret

- The multi-armed-bandit problem is a classic problem used to study the exploration vs. exploitation dilemma
- Imagine you are in a casino with multiple slot machines, each configured with an unknown reward probability:



- Under the assumption of an infinite number of trials:  
→ **What is the best strategy to achieve highest long-term rewards?**

- Our loss function is the total regret we might have by not select the optimal action up to the time step  $T$ :

$$\mathcal{L}_T = \mathbb{E} \left[ \sum_{t=1}^T (\underbrace{\theta^*}_{\text{what we should have been doing}} - \underbrace{Q(a_t)}_{\text{what we did}}) \right] = \sum_{a \in \mathcal{A}} N_T(a) \Delta_a$$

per-action regret

action-selection counter

# Exploration vs. Exploitation

Straightforward but usually bad: Greedy or  $\epsilon$ -greedy

- Greedy may select a suboptimal action forever  
→ Greedy has hence linear expected total regret
- $\epsilon$ -greedy continues to explore forever
  - with probability  $1 - \epsilon$  it selects  $a = \arg \max_{a \in \mathcal{A}} Q_T(a)$
  - with probability  $\epsilon$  it selects a random action
- Will hence continue to select all suboptimal actions with (at least) a probability of  $\frac{\epsilon}{|\mathcal{A}|}$   
→  $\epsilon$ -greedy, with a constant  $\epsilon$  has a linear expected total regret
- **Option #1: decrease  $\epsilon$  over course of training might work**
  - It is not easy to tune the parameters
- **Option #2: be optimistic with options of high uncertainty**
  - Prefer actions for which you do not have a confident value estimation yet  
→ Those have a great potential to be high-rewarding!
  - This idea is called **Upper Confidence Bounds**

# Exploration vs. Exploitation

## Upper Confidence Bounds (UCB1)

- Idea: estimate an upper confidence  $U_t(a)$  for each action value, such that with a high probability we satisfy

$$Q(a) \leq \hat{Q}_t(a) + U_t(a)$$

- Next, we select the action that maximizes the upper confidence bound:

$$a_t^{UCB} = \arg \max_{a \in \mathcal{A}} [Q_t(a) + U_t(a)]$$

Large  $N_t(a) \rightarrow$  small bound  $U_t(a)$  (estimated value is *certain/accurate*)

Small  $N_t(a) \rightarrow$  large bound  $U_t(a)$  (estimated value is *uncertain*)

- The vanilla **UCB1** algorithm uses  $p = t^{-4}$ :

$$U_t(a) = \sqrt{\frac{2 \log t}{N_t(a)}} \quad \text{and} \quad a_t^{UCB} = \arg \max_{a \in \mathcal{A}} Q(a) + \sqrt{\frac{2 \log t}{N_t(a)}}$$

Derived from Hoeffding's Inequality:

$$P(\mathbb{E}[X] \geq \bar{X}_t + u) \leq e^{-2tu^2}$$

- This ensures that we always keep exploring
- But we select the optimal action much more often as  $t \rightarrow \infty$

# Exploration vs. Exploitation

## Probability Matching via Thompson Sampling

We can also try the idea of directly sampling the action

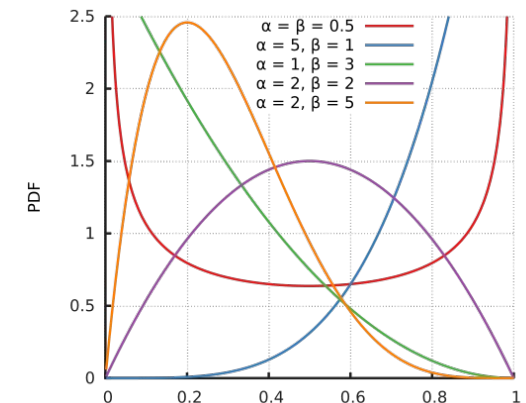
- Select action  $a$  according to probability that  $a$  is the optimal action (given the history of everything we observed so far):

$$\begin{aligned}\pi_t(a|h_t) &= P[Q(a) > Q(a'), \forall a' \neq a | h_t] \\ &= \mathbb{E}_{r|h_t} \left[ \mathbb{I} \left( a = \arg \max_{a \in \mathcal{A}} Q(a) \right) \right]\end{aligned}$$

Probability matching via Thompson Sampling:

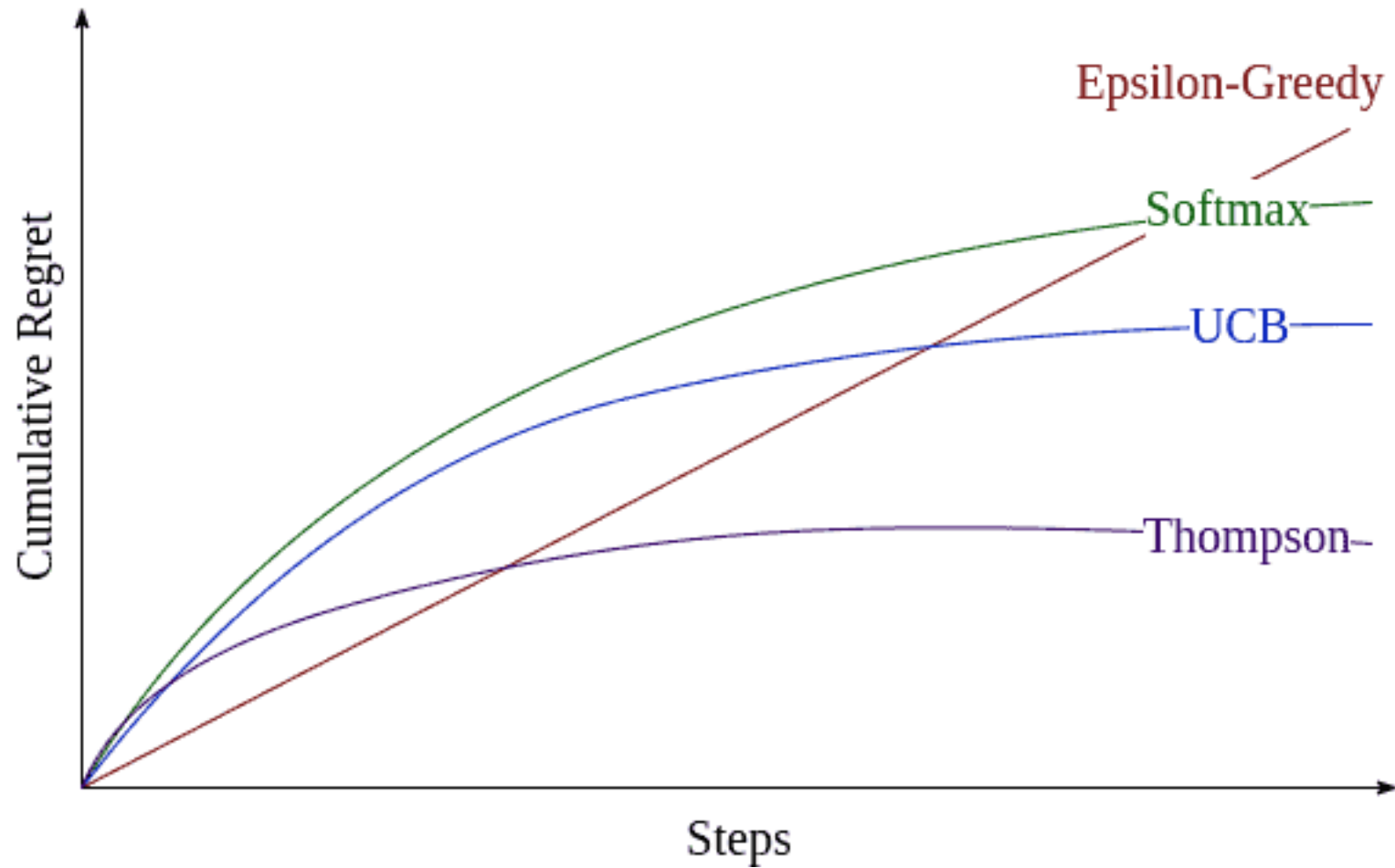
1. Assume  $Q(a)$  follows a Beta distribution for the Bernoulli bandit
  - As  $Q(a)$  is the success probability of  $\theta$
  - Beta( $\alpha, \beta$ ) is within  $[0,1]$ , and  $\alpha$  and  $\beta$  relate to the counts of success/failure
2. Initialize prior (e.g.,  $\alpha = \beta = 1$  or something different/what we think it is)
3. At each time step  $t$  we sample an expected reward  $\hat{Q}(a)$  from the prior Beta( $\alpha_i, \beta_i$ ) for every action
  - We select and execute the best action among the samples:  $a_t^{TS} = \arg \max_{a \in \mathcal{A}} \hat{Q}(a)$
4. With the newly observed experience we update the Beta distribution:

$$\begin{aligned}\alpha_i &\leftarrow \alpha_i + r_i \mathbb{I}[a_t^{TS} = a_i] \\ \beta_i &\leftarrow \beta_i + (1 - r_i) \mathbb{I}[a_t^{TS} = a_i]\end{aligned}$$



# Exploration vs. Exploitation

Characteristic curves



# Exercise Sheet 10

## Bandits



**Thank you for your attention!**