Fraunhofer-Institut für Integrierte
Schaltungen IIS

Reinforcement Learning

—

Exercise 9: Multi-armed Bandits

25.06.2024

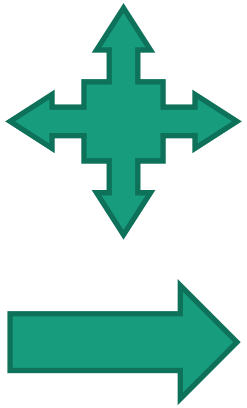
Alexander Mattick

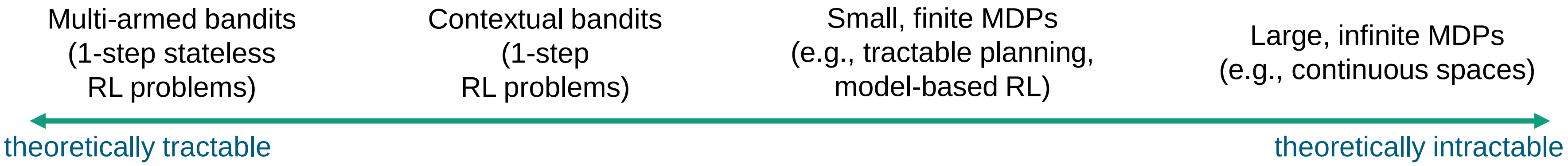
Exploration vs. Exploitation

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





Multi-armed bandits (1-step stateless RL problems)

Contextual bandits (1-step RL problems)

Small, finite MDPs (e.g., tractable planning, model-based RL)

Large, infinite MDPs (e.g., continuous spaces)

theoretically tractable

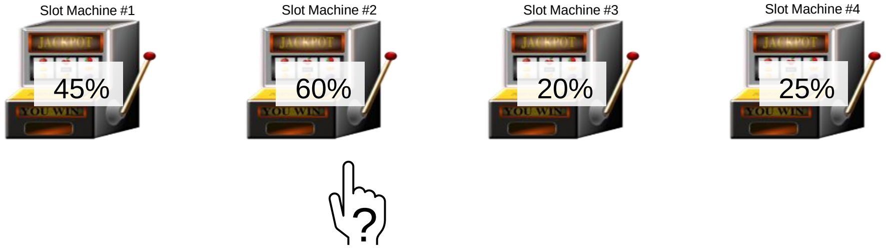
theoretically intractable

(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:

$$\mathcal{L}_T = \mathbb{E}\left[\sum_{t=1}^{T}\left(\underset{\substack{\uparrow\\ \text{what we should have been doing}}}{\theta^*} - \overset{\substack{\text{what we did}\\ \downarrow}}{Q(a_t)}\right)\right] = \sum_{a\in\mathcal{A}} \underset{\substack{\nwarrow\\ \text{action-selection counter}}}{N_T(a)}\,\overset{\substack{\text{per-action regret}\\ \swarrow}}{\Delta_a}$$

Exploration vs. Exploitation

Straightforward but usually bad: Greedy or ε-greedy

- Greedy may select a suboptimal action forever
  → Greedy has hence linear expected total regret

- ϵ-greedy continues to explore forever
  - with probability $1 - \epsilon$ it selects $a = \arg\max_{a \in \mathcal{A}} Q_T(a)$
  - with probability ϵ it selects a random action

- Will hence continue to select all suboptimal actions with (at least) a probability of $\frac{\epsilon}{|\mathcal{A}|}$
  → ϵ-greedy, with a constant ϵ has a linear expected total regret

- **Option #1: decrease ϵ 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 $N_t(a)$ → small bound $U_t(a)$ *(estimated value is certain/accurate)*

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

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

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

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

Derived from Hoeffding's Inequality:

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

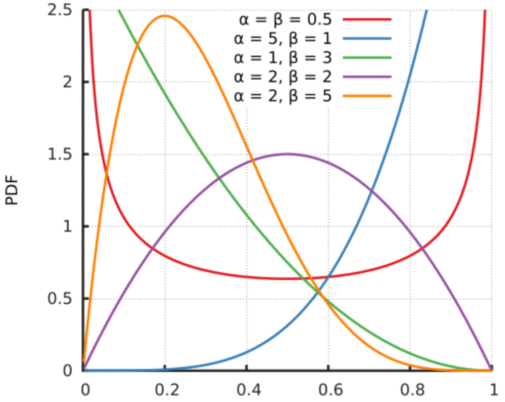
Exploration vs. Exploitation

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):

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



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 θ
   - Beta(α, β) is within $[0,1]$,and α and β 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(α_i, β_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:

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

Exercise Sheet 10

Bandits

Fraunhofer
IIS

Fraunhofer-Institut für Integrierte
Schaltungen IIS

Thank you for your attention!