Computer Engineering Department

Multi-Armed Bandits

Mohammad Hossein Rohban, Ph.D.

Hosein Hasani

Spring 2023

Outline

- Exploration/exploitation tradeoff
- Multi-armed bandit problem
- Regret
- Methods:
  - ϵ-greedy strategies
  - Upper confidence bounds
  - Thompson sampling

Exploration/Exploitation Tradeoff

- **Exploration**: Choosing an arbitrary action with *unknown* outcome
  - Can lead to higher reward in the long run while sacrificing short term rewards.

- **Exploitation**: Taking the best action w.r.t. to the *current knowledge*
  - Ensures an immediate reward in the short-term, with uncertainty in the long run.

Exploration/Exploitation Tradeoff

In the presence of incomplete information:

- Neither exploitation nor exploration can be pursued exclusively!
- We need to have a balance.
- The best long-term strategy may involve taking short-term sacrifices.

Exploration/Exploitation Tradeoff

- How to study exploration/exploitation strategies?
- Can we analyze optimality of methods theoretically?

multi-armed bandits (1-step stateless RL problems)	contextual bandits (1-step RL problems)	small, finite MDPs (e.g., tractable planning, model-based RL setting)	large, infinite MDPs, continuous spaces

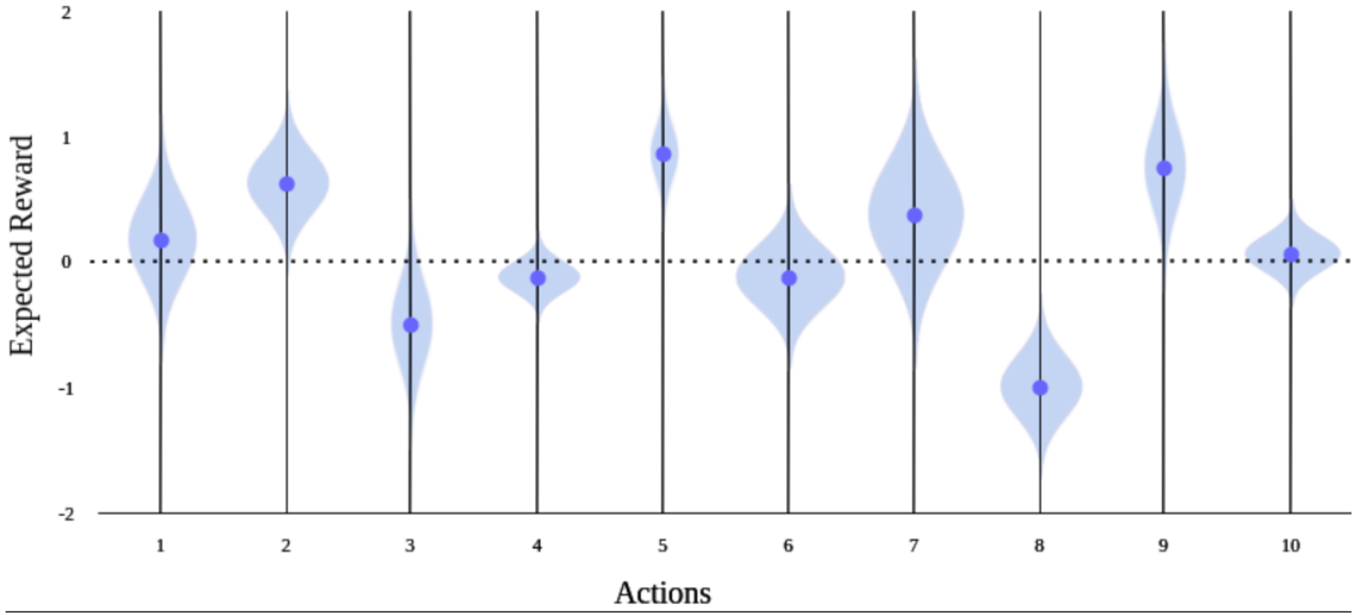
theoretically tractable ⟷ theoretically intractable

Multi-Armed Bandit

Multiple bandits with unknown average rewards

Multi-Armed Bandit: Goal

Finding the best arm (in the sense of expected reward) with minimum trial and error.



Multi-Armed Bandit: Examples

- Online advertisement
- Recommender systems
- Clinical trials
- Mining
- Network (packet routing)

Stochastic K-Armed Bandits

Formal definition:

A tuple $< \mathcal{A}, \mathcal{Y}, P, r >$ where

- $\mathcal{A}$ is the set of actions, and $|\mathcal{A}| = K$
- $\mathcal{Y}$ is the set of possible outcomes
- $P(y|a)$ is the probability of outcome $y \in \mathcal{Y}$ conditioned on action $a \in \mathcal{A}$
- $r(y) \in \mathbb{R}$ the reward associated with the outcome $y \in \mathcal{Y}$

We can simplify this definition by considering $r = y$

Regret

- Expected reward: $q(a) = \mathbb{E}_{y\sim p(.|a)}[r(y)|a]$ or simply $q(a) = \mathbb{E}[r|a]$
- Expected best reward: $q^* = \max_a q(a)$
- Best action: $a^* = \operatorname{argmax}_a q(a)$

- Difference between the expected best reward and the actual reward:

$$Regret(T) = \sum_{t=1}^{T} q^* - r(a_t)$$

- Expected cumulative regret:

$$\mathbb{E}[Regret(T)] = \sum_{t=1}^{T} q^* - q(a_t)$$

Methods: Simple Heuristics

- **Greedy strategy**:

select the arm with the highest average so far

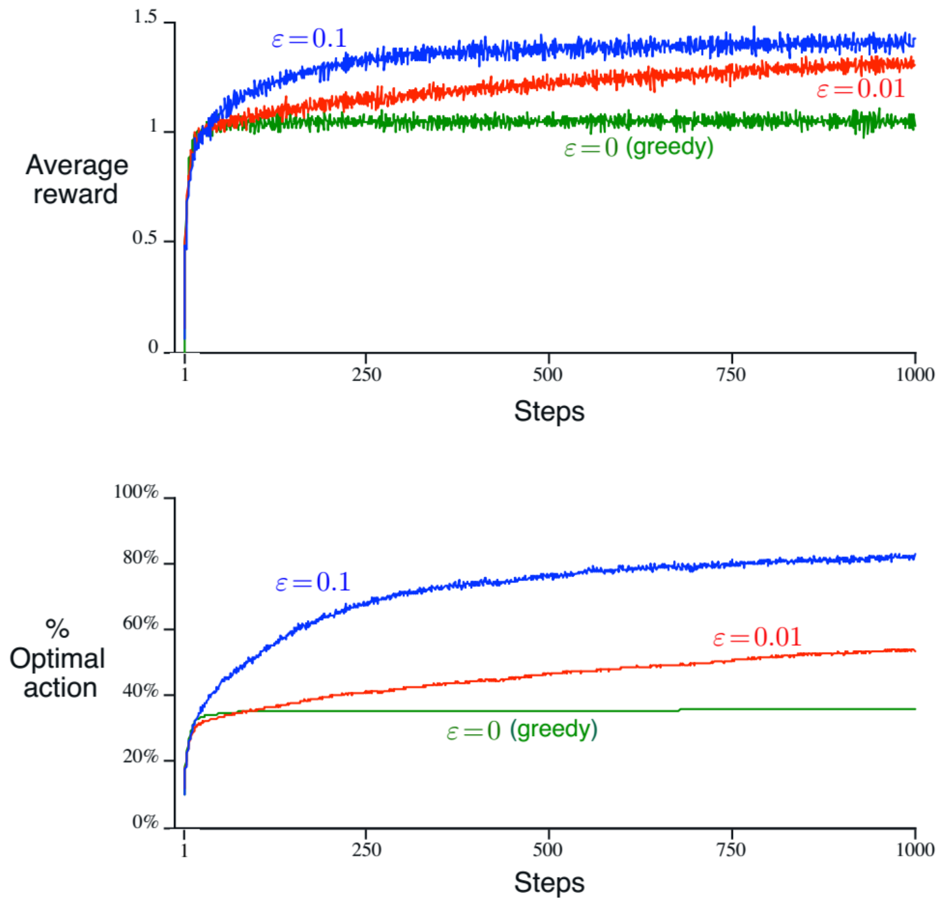
$$a_t = \underset{a}{argmax}\, \hat{q}\,(a)$$

- May get stuck due to lack of exploration

- ***ϵ*-greedy**:

select an arm randomly with probability ϵ and otherwise do a greedy selection

- Convergence rate depends on choice of ϵ



[image credit: Sutton's RL book]

$\epsilon - Greedy$: Theoretical Guarantees

- Constant ϵ:
  - For large enough t: $p(a_t \neq a^*) \approx \epsilon$
  - $\mathbb{E}[Regret(T)] \approx \sum_{t=1}^{T} \epsilon \Rightarrow \mathcal{O}(T)$

linear regret

- Decaying ϵ:
  - $\epsilon_t \propto \frac{1}{t}$
  - For large enough t: $p(a_t \neq a^*) \approx \epsilon_t$
  - $\mathbb{E}[Regret(T)] \approx \sum_{t=1}^{T} \epsilon_t = \sum_{t=1}^{T} \frac{1}{t} \Rightarrow \mathcal{O}(logT)$

logarithmic regret

Optimistic Algorithm

Positivism in the Face of uncertainty:

- If the difference between empirical and true mean is bounded:

$$|q(a) - \hat{q}(a)| \leq bound \Longrightarrow q(a) \leq \hat{q}(a) + bound$$

- We could select arms based on best $\hat{q}(a) + bound$
- Overtime, additional data will allow us to refine $\hat{q}(a)$ and compute a tighter bound

Probabilistic Upper Bound

- **Problem**:
  We can't compute an upper bound with certainty since we are sampling

- **Solution**:
  Consider an upper bound in probability with Hoeffding's inequality:

$$p(q(a) - \hat{q}(a) > \mathcal{B}) < e^{-2N_a\mathcal{B}^2}$$

for $0 \leq r \leq 1$

Upper Confidence Bound (UCB)

$$p(q(a) - \hat{q}(a) > \mathcal{B}) < e^{-2N_a\mathcal{B}^2} = \delta \implies \mathcal{B} = \sqrt{\frac{\log(\frac{1}{\delta})}{2N_a}}$$

UCB algorithm:

- Set $\delta \propto \left(\frac{1}{t}\right)^c$
- Choose a based on highest Hoeefding bound

$$a = \underset{a}{argmax}\, \hat{q}(a) + c' \sqrt{\frac{\log(t)}{N_a}}$$

Upper Confidence Bound (UCB)

UCB Algorithm

$N_a \leftarrow 0 \; \forall a$

For $t = 1$ to T:

$$a_t \leftarrow \underset{a}{argmax}\, \hat{q}(a) + c' \sqrt{\frac{\log(t)}{N_a}}$$

Execute a_t and receive r_t

$$\hat{q}(a) \leftarrow \frac{N_a \hat{q}(a) + r_t}{N_a + 1}$$

$$N_a \leftarrow N_a + 1$$

intuition: try each arm until you are sure it's not good!

Bayesian Bandit

Intuition:

- Consider a probability distribution $p(r^a|\theta^a)\ \forall a$
- Express uncertainty about θ^a by a prior $p(\theta^a)$
- Observe samples $r_1^a, r_2^a, r_3^a, \ldots, r_n^a$
- Belief update:

$$p(\theta^a|r_1^a, r_2^a, r_3^a, \ldots, r_n^a) \propto p(r_1^a, r_2^a, r_3^a, \ldots, r_n^a|\theta^a)p(\theta^a)$$

Bayesian Bandit

Posterior over θ^a allows us to estimate

- Distribution over the next reward

$$p(r^a|r_1^a, r_2^a, \dots, r_n^a) = \int p(r^a|\theta^a)p(\theta^a|r_1^a, r_2^a, \dots, r_n^a)\mathrm{d}\theta^a$$

- Distribution over $q(a)$ when θ^a includes the mean

$$p(q(a)|r_1^a, r_2^a, \dots, r_n^a) = p(\theta^a|r_1^a, r_2^a, \dots, r_n^a) \text{ if } \theta^a = q(a)$$

Thompson Sampling

Thompson Sampling

Set a prior distribution for $q(a)$

For $t = 1$ to T:

sample $q_1(a), \dots, q_k(a) \sim p\big(q(a)\big) \quad \forall a$

$\hat{q}_t(a) \leftarrow \frac{1}{K} \sum_{i=1}^{K} q_i(a) \quad \forall a$

$a_t \leftarrow \underset{a}{argmax}\, \hat{q}_t(a)$

Execute a_t and receive r_t

update $p(q(a_t))$ based on r_t

Thompson Sampling

- The sample size K and amount of data n, regulate amount of exploration
- As K and n increase, $\hat{q}(a)$ becomes less stochastic, which reduce exploration
  - With larger K, $\hat{q}(a)$ approaches $\mathbb{E}[q(a)|r_1^a, \dots, r_n^a]$
  - With larger n, $p(q(a)|r_1^a, \dots, r_n^a)$ becomes more peaked
- The stochasticity of $\hat{q}(a)$ ensures that all actions are chosen with some probability

Thompson Sampling: Online Advertisement Example

- Formulate as a bandit problem:
  - Arms: the set of possible ads
  - Rewards: 0 (no click) or 1 (click)

- Distribution over $q(a)$:

$$q(a_i) \sim Beta(q; \alpha_i, \beta_i)$$
$$\propto q^{\alpha_i - 1}(1 - q)^{\beta_i - 1}$$

- Conditional distribution of r:

$$p(r|q) = q^r (1 - q)^{1-r}$$

Thompson Sampling: Online Advertisement Example

- Consider, after sampling we have: $a_j = \underset{a}{argmax}\, \hat{q}_j(a)$
  - Posterior after observing reward:
    - If $r = 1$:

$$\begin{aligned} p(q(a_j)|r = 1) &\propto p\left(q(a_j)\right) p\left(r = 1 \middle| q(a_j)\right) \\ &\propto q^{\alpha_j - 1}(1 - q)^{\beta_j - 1} q \\ &= q^{\alpha_j}(1 - q)^{\beta_j - 1} q \\ &\propto Beta(q; \alpha_j + 1, \beta) \end{aligned}$$

    - If $r = 0$:

$$\begin{aligned} p(q(a_j)|r = 0) &\propto p\left(q(a_j)\right) p\left(r = 0 \middle| q(a_j)\right) \\ &\propto q^{\alpha_j - 1}(1 - q)^{\beta_j - 1}(1 - q) \\ &\propto Beta(q; \alpha_j, \beta + 1) \end{aligned}$$

Multi-Armed Bandit Methods: Summary

- Theoretical regret for UCB, Thompson sampling, and $\epsilon - greedy$ (with decaying ϵ): $\boldsymbol{\mathcal{O}(logT)}$
- **UCB** and **Thompson sampling** are harder to analyze theoretically but use exploration more smartly with preference to uncertainty.
- Empirical performance may vary.

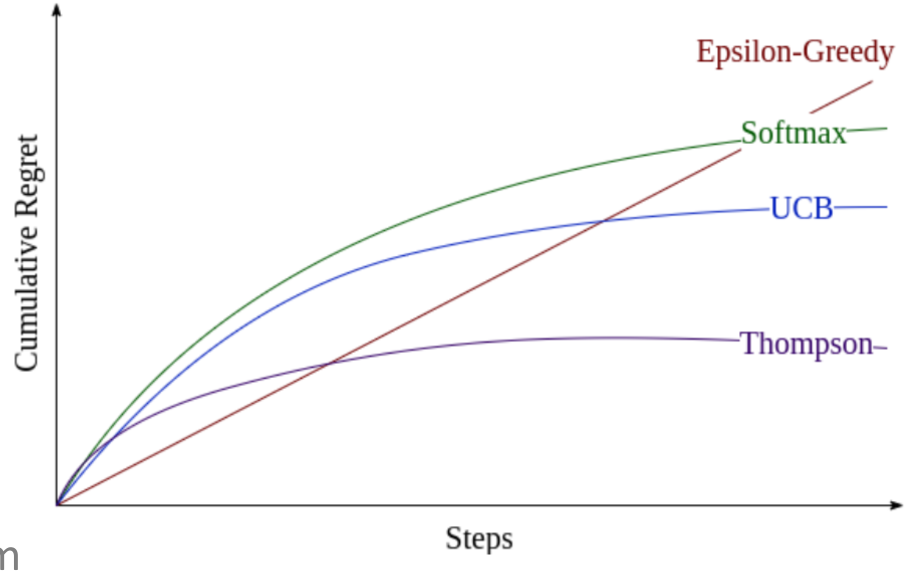


image credit: baeldung.com