Foundations of Reinforcement Learning

Multi-arm bandits: stochastic bandits

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2023

Outline

Introduction and formulation

A/B testing

How do you decide which variation leads to higher traffic/revenue?

Figure credit: internet.

A/B testing: explore each variation equally first, then deploy the statistically better one.

From A/B testing to multi-arm bandits

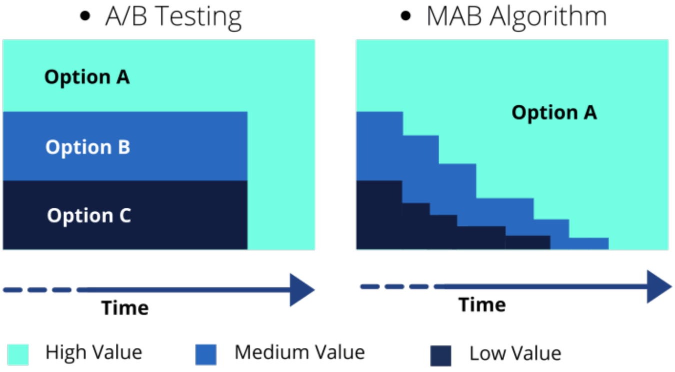
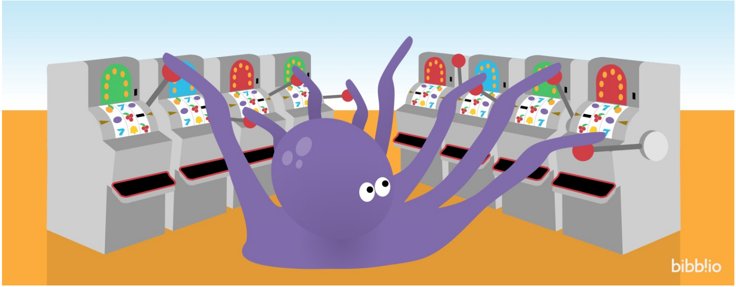


Figure credit: internet.

Multi-arm bandits: simultaneous exploration and exploitation, dynamic allocation.

Multi-arm bandit

Which slot machine will give me the most money?



First proposed in [Thompson, 1933], popularized by [Robbins, 1952]

Learning the best arm

Can we learn which slot machine gives the most money?

$1
$0
$0

$1
$4
$0
$2
$1
$3
$5

$1
$0
$1
$2

Formulation

We can play multiple rounds $t = 1, 2, \ldots, T$.

In each round, we select an arm i_t from a fixed set $i = 1, 2, \ldots, n$; and observe the reward r_t that the arm gives.

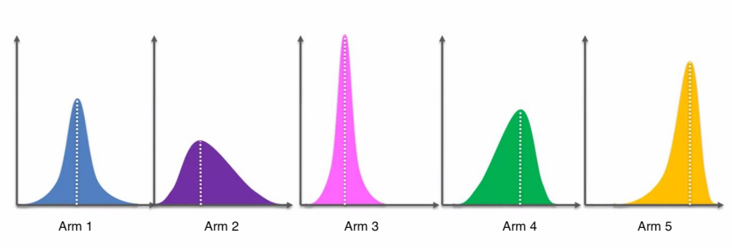
Arm 1

Arm 2

Arm 3

Objective: Maximize the total reward over time.

Stochastic bandit



- The reward at each arm is stochastic (e.g., 1 with probability p_i and otherwise 0).
- Suppose the rewards are independent over time. The best arm is then the arm with highest expected reward.

Example of online ads: arm $=$ ad, reward $=1$ if the user clicks on the ad and 0 otherwise

Stochastic bandit with i.i.d. rewards

We consider a simple setting with i.i.d. bounded rewards.

- Each arm distributes rewards according to some (unknown) distribution over $[0, 1]$, with

$$\mathbb{E}[r_{i,t}] = \mu_i, \quad \forall i \in [n],\ t = 1, 2 \ldots$$

- Suppose we play arm i_t at round t, and receive the reward

$$r_{i_t,t}$$

  drawn i.i.d. from the arm i_t's distribution.

Partial information: Every round we cannot observe the reward of all arms: we just know the reward of the arm that we played.

Regret: performance metric

We design algorithms that determine the sequence $\{i_t\}$, i.e. *policies*.

How to evaluate the performance?

Definition 1 (Expected regret)

The expected regret over T rounds is defined as

$$R_T = \max_{1\le i\le n} \mathbb{E}\left[\sum_{t=1}^{T}(r_{i,t} - r_{i_t,t})\right] = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_{i_t,t}\right],$$

where $\mu^\star = \max_{1\le i\le n}\mu_i$ is the highest expected reward over all arms.

- 1st term captures the highest cumulative reward in *hindsight*.
- 2nd term captures the *actual* accumulated reward.

Regret decomposition lemma

Since $\mathbb{E}[r_{i_t,t}] = \mathbb{E}\left[\sum_{i=1}^n \mu_i \mathbb{I}_{i_t=i}\right] = \sum_{i=1}^n \mu_i \left(\mathbb{E}\mathbb{I}_{i_t=i}\right)$, then

$$\begin{aligned}
R_T &= \sum_{t=1}^T \left[\sum_{i=1}^n \mu^\star \left(\mathbb{E}\mathbb{I}_{i_t=i}\right) - \sum_{i=1}^n \mu_i \left(\mathbb{E}\mathbb{I}_{i_t=i}\right)\right] \\
&= \sum_{i=1}^n \Delta_i \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}_{i_t=i}\right] \\
&=: \sum_{i=1}^n \Delta_i \mathbb{E}\left[T_{i,T}\right]
\end{aligned}$$

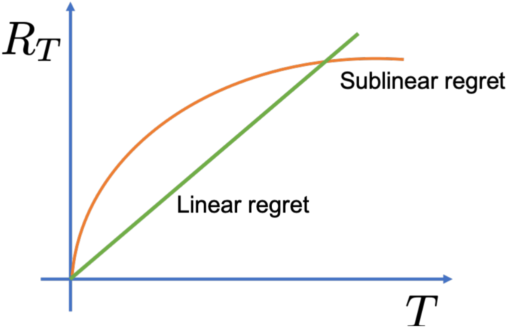
where

- $\Delta_i = \mu^\star - \mu_i$ is the sub-optimality gap of arm i;
- $T_{i,T} = \sum_{t=1}^T \mathbb{I}_{i_t=i}$ is the number of times arm i is played in T rounds.

Sublinear regret

Sublinear regret: most MAB algorithms aim to achieve sublinear regret, so that the average regret goes to 0 as $T \to \infty$:

$$\lim_{T\to\infty} \frac{R_T}{T} = 0$$



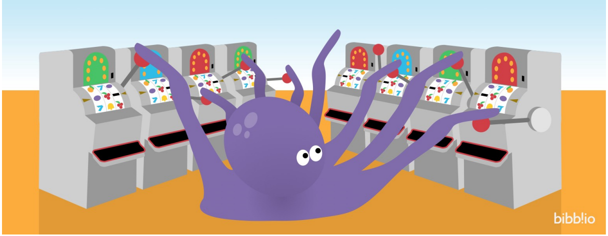
From ϵ-greedy to UCB algorithm

Learning the best arm via trial-and-error

Which arm do I pick next, so that I maximize my reward over time?

\$1	\$1	\$1
\$0	\$4	\$0
\$0	\$0	\$1
	\$2	\$2
	\$1	\$12
	\$3	\$11
	\$5	

Exploration-exploitation trade-off



Which arm should I play?

- Best arm observed so far? (exploitation)
- Or should I look around to try and find a better arm? (exploration)

We need both in order to maximize the total reward.

An ϵ-greedy approach

Exploit, but explore a random arm ϵ fraction of the time.

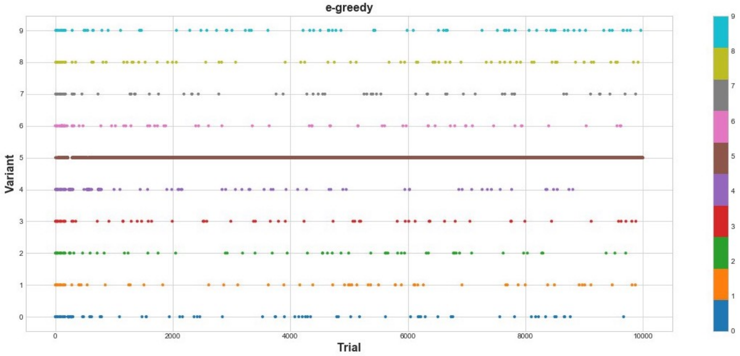
1. **Initial phase:** Try each arm and observe the reward.
2. For each round $t = n + 1, \ldots, T$:
   - Calculate the empirical average reward for each arm i:

     $$\overline{\mu}_{i,t} = \frac{\text{total reward from pulling this arm in the past}}{\text{number of times I pulled this arm}} = \frac{\sum_{t:i_t=i} r_t}{\sum_{t:i_t=i} 1},$$

     where i_t is the index of the arm played at time t, r_t is the reward.

   - With probability $1 - \epsilon$, play the arm with highest $\overline{\mu}_{i,t}$ and observe the reward. Otherwise, choose an arm at random and observe the reward.

Understanding ϵ-greedy



- In the first thousand iterations, all arms are chosen fairly frequently.
- Eventually the algorithm realizes that arm 5 has the highest expected reward.

Regrets of greedy policies

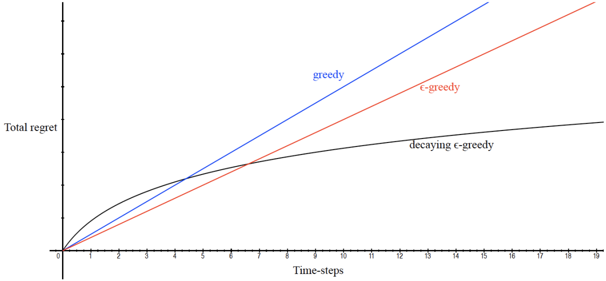


Figure credit: David Silver's lecture.

- Greedy policy incurs linear regret since it can lock on a sub-optimal policy.
- ϵ-greedy always explores by ϵ fraction and therefore its regret is still linear (recall the regret decomposition lemma).
- Decaying ϵ helps, however it is hard to design the schedule.

The UCB algorithm

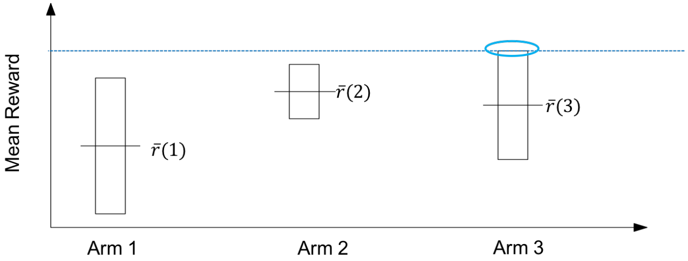
[Auer et al., 2002]: the idea is to always try the best arm, where "best" includes exploration and exploitation.

1. **Initial phase:** try each arm and observe the reward.
2. For each round $t = n + 1, \ldots, T$:
   - Calculate the UCB (upper confidence bound) index for each arm i:

     $$\mathrm{UCB}_{i,t} = \overline{\mu}_{i,t} + \sqrt{\frac{\log t}{T_{i,t}}},$$

     where $\overline{\mu}_{i,t}$ is the empirical average reward for arm i and $T_{i,t}$ is the number of times arm i has been played up to round t.
   - Play the arm with the highest UCB index and observe the reward.

Understanding UCB



$$\mathrm{UCB}_{i,t} = \overline{\mu}_{i,t} + \sqrt{\frac{\log t}{T_{i,t}}},$$

- Exploitation: $\overline{\mu}_{i,t}$ is the average observed reward. High observed rewards of an arm leads to high UCB index.
- Exploration: $\sqrt{\frac{\log t}{T_{i,t}}}$ decreases as we make more observations ($T_{i,t}$ grows). Few observations of an arm leads to high UCB index.

Theory of UCB algorithm

Theorem 2 (Instance-dependent regret bound of UCB)

For $T \geq n$, the expected regret of UCB algorithm is upper bounded as

$$R_T \leq \sum_{i:\Delta_i>0} \left(\frac{4 \log T}{\Delta_i} + 8\Delta_i \right) \leq \sum_{i:\Delta_i>0} \frac{4 \log T}{\Delta_i} + 8n,$$

where $\Delta_i = \mu^\star - \mu_i$ is the sub-optimality gap of arm i.

- The regret bound scales with the *harmonic mean* of the gaps,

$$R_T \lesssim \frac{n \log T}{\text{harmonic mean}(\{\Delta_i\})}.$$

  - E.g. $\Delta_2 = \frac{1}{2}$, $\Delta_3 = \frac{1}{2}$, harmonic mean $= \frac{1}{2}$.
  - E.g. $\Delta_2 = \frac{1}{10}$, $\Delta_3 = \frac{1}{2}$, harmonic mean $= \frac{1}{6}$.
- When Δ_i's are constants, the regret scales as (ignoring n)

$$R_T = O\big(\log T\big),$$

which is nearly the best we can hope for! (We'll see why later.)

Gap-free bound of UCB algorithm

The gap-dependent bound may become too loose when Δ_i is, say, asymptotically small, $\Delta_i \sim \log T/T$.

Fortunately, this can be fixed by studying the following instance-independent (aka worst-case) bound.

Theorem 3 (Instance-independent regret bound of UCB)

For $T \geq n$, the expected regret of UCB algorithm is upper bounded as

$$R_T \leq 4\sqrt{nT\log T} + 8n.$$

- When $n = O(1)$, the regret scales as

$$R_T = O\big(\sqrt{T\log T}\big) = \widetilde{O}\big(\sqrt{T}\big)$$

- The logarithmic factor can be shaved away [Audibert and Bubeck, 2009].

Analysis

Toolkit: Hoeffding's inequality

Theorem 4 (Hoeffding's inequality)

Let $X_1, X_2, \ldots, X_n$ be independent random variables satisfying $a_i \leq X_i \leq b_i$. Then for all $\delta \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Setting $a_i = 0$, $b_i = 1$, and $\mathbb{E}[X_i] = \mu$, we obtain

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right)$$

$$\implies \quad \left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \qquad \text{with prob. } 1-\delta.$$

This will allow us to talk about how the mean reward concentrates around the true mean.

Implications of Hoeffding's inequality

For each arm i at time t, with probability at least $1 - 2/t^2$,

$$\left|\overline{\mu}_{i,t} - \mu_i\right| < \sqrt{\frac{\log t}{T_{i,t}}}$$

$$\Longrightarrow \qquad \mathsf{UCB}_{i,t} = \overline{\mu}_{i,t} + \sqrt{\frac{\log t}{T_{i,t}}} \geq \mu_i .$$

Optimism in the face of uncertainty:
acting according to the UCB index, which is an upper bound of the true mean μ_i.

Bound the number of sub-optimal pulls

Recall that

$$R_T = \sum_{i=1}^{n} \Delta_i \underbrace{\mathbb{E}\left[T_{i,T}\right]}_{\text{control target}}.$$

Key observation: at each t, the UCB index of the sub-optimal arms $i \neq i^\star$ will be sufficiently apart from the optimal one and arm i will not get pulled (i.e. $i_{t+1} \neq i$), as long as $T_{i,t}$ is sufficiently large:

$$\begin{aligned} \mathsf{UCB}_{i,t} = \overline{\mu}_{i,t} + \sqrt{\frac{\log t}{T_{i,t}}} &\leq \mu_i + 2\sqrt{\frac{\log t}{T_{i,t}}} \quad \text{(Hoeffding)} \\ &\leq \mu_{i^\star} \leq \mathsf{UCB}_{i^\star,t} \quad \text{(optimism/Hoeffding)} \end{aligned}$$

as long as

$$T_{i,t} \geq \frac{4\log t}{\Delta_i^2}$$

with probability at least $1 - 4/t^2$ (we applied Hoeffding twice).

Bound the number of sub-optimal pulls

$$
\begin{aligned}
\mathbb{E}\left[T_{i,T}\right] &\le \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{I}(i_{t+1}=i)\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{I}\left(i_{t+1}=i, T_{i,t} < \frac{4\log t}{\Delta_i^2}\right)\right] + \mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{I}\left(i_{t+1}=i, T_{i,t} \ge \frac{4\log t}{\Delta_i^2}\right)\right] \\
&\le \frac{4\log T}{\Delta_i^2} + \sum_{t=n}^{T-1} \mathbb{P}\left(i_{t+1}=i, T_{i,t} \ge \frac{4\log t}{\Delta_i^2}\right) \\
&\le \frac{4\log T}{\Delta_i^2} + \sum_{t=n}^{T-1} \mathbb{P}\left(i_{t+1}=i \middle| T_{i,t} \ge \frac{4\log t}{\Delta_i^2}\right) \mathbb{P}\left(T_{i,t} \ge \frac{4\log t}{\Delta_i^2}\right) \\
&\le \frac{4\log T}{\Delta_i^2} + \sum_{t=n}^{T-1} \frac{4}{t^2} \\
&\le \frac{4\log T}{\Delta_i^2} + 8.
\end{aligned}
$$

A key lemma

Lemma 5 (bounding the number of pulls of sub-optimal arms)

For any arm with $\Delta_i > 0$*, it holds that*

$$\mathbb{E}\left[T_{i,T}\right] \leq \frac{4\log T}{\Delta_i^2} + 8.$$

Proof of Theorem 2:

$$\begin{aligned} R_T = \sum_{i=1}^{n} \Delta_i \mathbb{E}\left[T_{i,T}\right] &\leq \sum_{\Delta_i > 0} \Delta_i \left(\frac{4\log T}{\Delta_i^2} + 8 \right) \\ &= \sum_{\Delta_i > 0} \left(\frac{4\log T}{\Delta_i} + 8\Delta_i \right). \end{aligned}$$

From gap-dependent to gap-independent bounds

Intuition: for some Δ to be determined later,

- For arms $\{i : \Delta_i \geq \Delta\}$ with large gaps: use the gap-dependent bound
$$\mathbb{E}[T_{i,T}] \leq \frac{4 \log T}{\Delta_i^2} + 8;$$
- For arms $\{i : \Delta_i < \Delta\}$ with small gaps: use the naive bound
$$\sum_i \mathbb{E}[T_{i,T}] \leq T.$$

Hence,

$$\begin{aligned} R_T = \sum_{i=1}^n \Delta_i \mathbb{E}[T_{i,T}] &\leq \sum_{i:\ \Delta_i \geq \Delta} \Delta_i \left(\frac{4 \log T}{\Delta_i^2} + 8 \right) + \sum_{i:\ \Delta_i < \Delta} \Delta_i \mathbb{E}[T_{i,T}] \\ &\leq \frac{4n \log T}{\Delta} + 8n + \Delta T. \end{aligned}$$

Choosing $\Delta = \sqrt{\frac{4n \log T}{T}}$, we obtain $R_T \leq 4\sqrt{nT \log T} + 8n$.

References I

Audibert, J.-Y. and Bubeck, S. (2009).
Minimax policies for adversarial and stochastic bandits.
In *COLT*, pages 217–226.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002).
Finite-time analysis of the multiarmed bandit problem.
Machine learning, 47(2):235–256.

Robbins, H. (1952).
Some aspects of the sequential design of experiments.
Bulletin of the American Mathematical Society, 58(5):527–535.

Thompson, W. R. (1933).
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25(3-4):285–294.