Lecture 3. Multi-Armed Bandits

Peng Zhao

School of AI

Nanjing University

July 15, 2026

Outline

- Bandits Problem
- Adversarial Multi-Armed Bandit
  - Exp3 and IW Estimator
  - Regret Analysis
- Stochastic Multi-Armed Bandit
  - Explore-then-Commit
  - ε-Greedy

Part 1. Bandits

- Bandit Problems

- Multi-Armed Bandits

Online Learning

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a feasible set $\mathcal{X} \subseteq \mathbb{R}^d$;

(2) and environments pick an online function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

on the feedback information:

- ***full information***: observe entire f_t (or at least gradient $\nabla f_t(\mathbf{x}_t)$)

- ***partial information (bandits)***: observe function value $f_t(\mathbf{x}_t)$ only

less information

full information

horse racing

partial information

multi-armed bandits

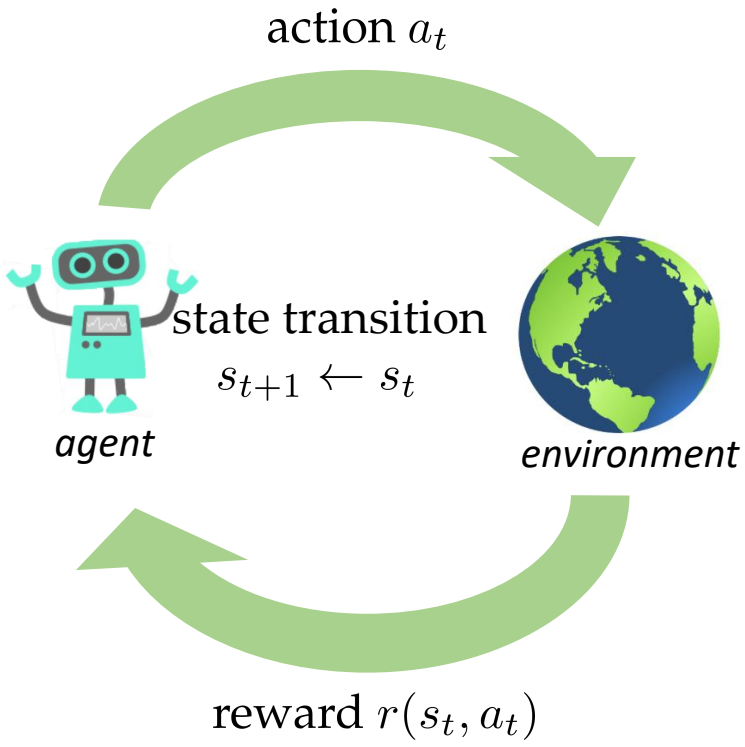
Bandits

- Bandit problems
  - named after a ***one-armed bandit***
  - ***arm***: a colloquial term for a slot machine that is pulled to try to win
  - ***bandit***: comes from the idea that the machine is a "thief" that takes your money without offering a guaranteed return

- Multi-armed bandits
  - Context: there are multiple slot machines, each with its own probability of payout
  - Goal: the player (gambler) places her bets on a slot machine to maximize the total reward
  - **Exploration-Exploitation tradeoff**

Bandits as Interactive Learning

- Bandit is ***"single-step"*** decision version of Reinforcement Learning



Reinforcement learning:

- Sequential decision making
- With state transition

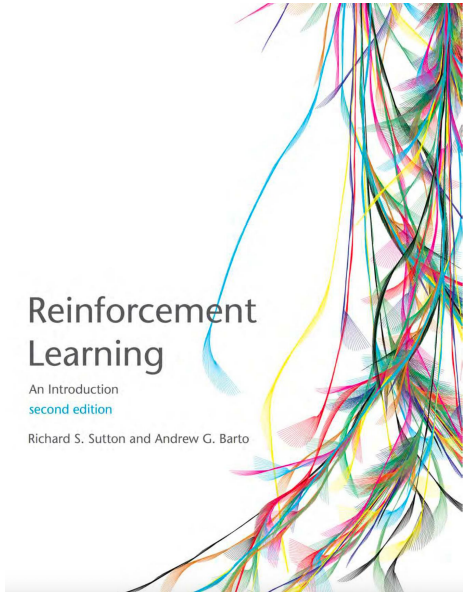
Bandits:

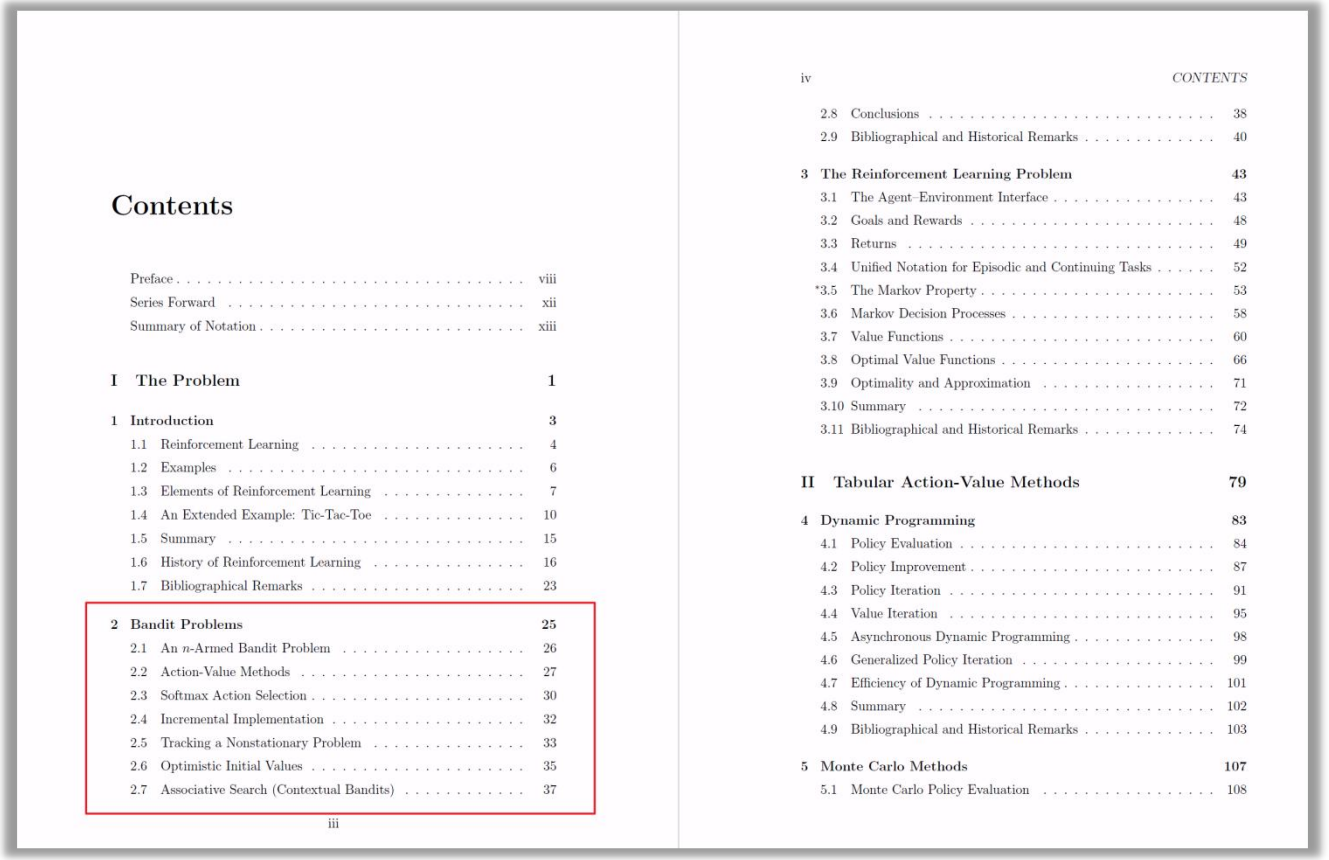
- Single-step decision making
- No state transition

Bandits as Interactive Learning

Sutton & Barto. Reinforcement Learning, second edition: An Introduction. MIT Press, 2018.

2024 Turing award winners





Contents

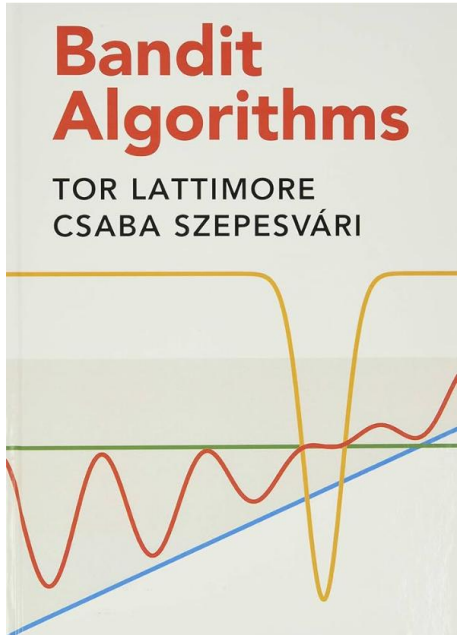
iii

iv CONTENTS

Bandit Problems

- Also called ***partial-information*** online learning.

- There are a variety of bandit problems:
  - ❑ multi-armed bandits (MAB)
  - ❑ linear bandits/convex bandits
  - ❑ generalized linear bandits/graph bandits
  - ❑ contextual bandits
  - ❑ partial monitoring
  - ❑ … (even MDP for RL)



Bandit Algorithms
Tor Lattimore, Csaba Szepesvári
Cambridge University Press, 2021

Online Learning

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a feasible set $\mathcal{X} \subseteq \mathbb{R}^d$;

(2) and environments pick an online function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

on the difficulty of environments:

- stochastic setting

- adversarial setting
  - oblivious
  - adaptive (non-oblivious)

less restricted but harder

oblivious adversary

examination

adaptive adversary

interview

Bandit Problems

- Also called ***partial-information*** online learning.

- According to the environments, it can be roughly classified as
  - ❑ ***Stochastic bandits***: environment is generated by a stochastic model
  - ❑ ***Adversarial bandits***: environment can be chosen against the learner
    - by "adversarial", one can think it as "non-stochastic"
    - *oblivious* adversary: thinking of the final exam
    - *non-oblivious* adversary: thinking of the online games

Adversarial Bandits

- Continuing the OCO protocol:

At each round $t = 1, 2, \cdots$

(1) the player first picks a model $\mathbf{x}_t$ from a convex set $\mathcal{X} \subseteq \mathbb{R}^d$;

(2) and environments pick an online convex function $f_t : \mathcal{X} \to \mathbb{R}$;

(3) the player suffers loss $f_t(\mathbf{x}_t)$, observes some information about f_t and updates the model.

⇨ at round t, the learner can only observe the function value $f_t(\mathbf{x}_t)$ information

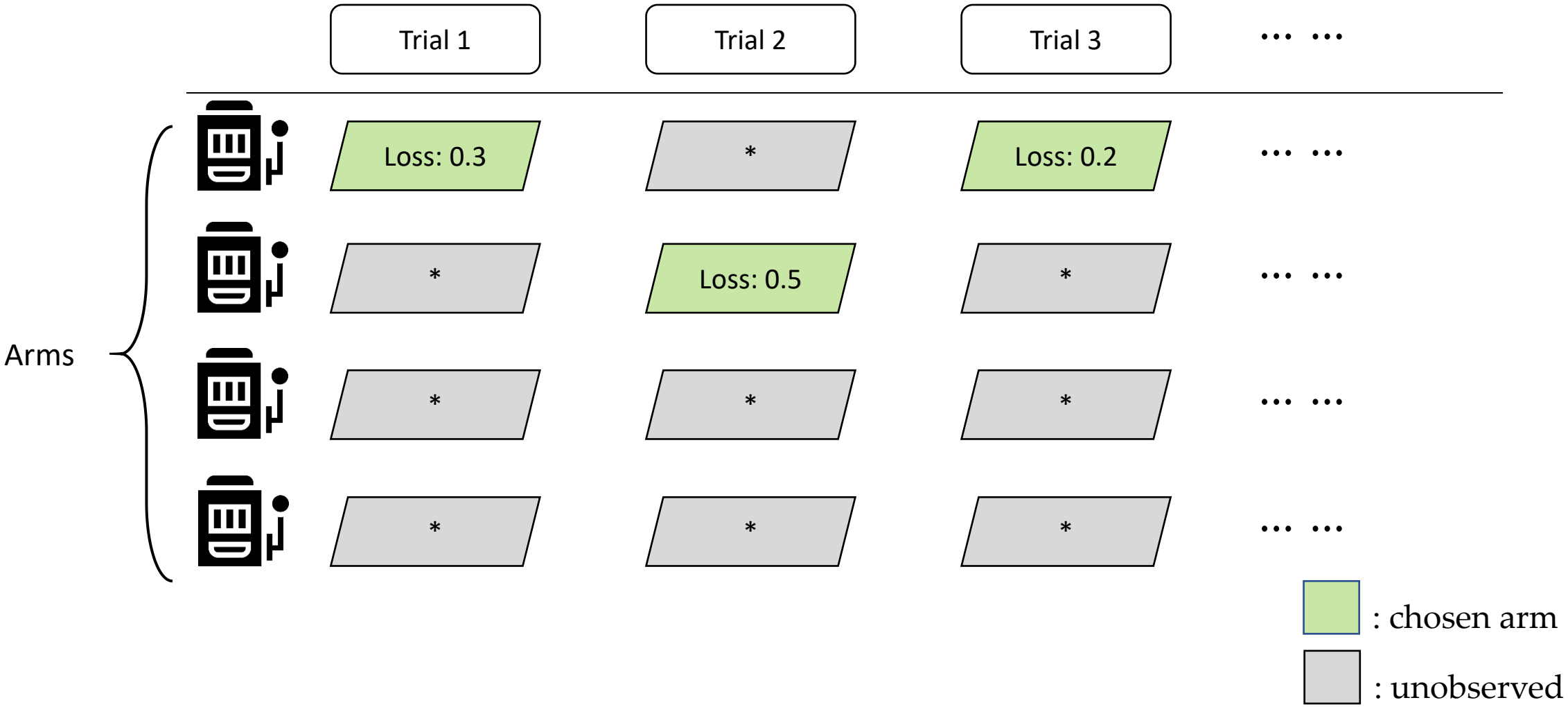
We focus on the ***oblivious*** setting (non-oblivious bandits are usually challenging)

i.e., environments decide online functions of all the rounds before the online game starts.

Part 2. Adversarial MAB

- Formulation
- Loss Estimator
- Exp3 and Regret Analysis

Multi-Armed Bandit



Formulation

At each round $t = 1, 2, \cdots$

(1) the player first picks an arm $a_t \in [K]$ from K candidate arms;

(2) and simultaneously environments pick a loss vector $\boldsymbol{\ell}_t \in [0,1]^K$;

(3) the player suffers and only observes loss $\ell_{t,a_t} \in [0,1]$, then updates the model.

on the difficulty of environments:

- ***adversarial*** setting
  - ***oblivious***: $\{\boldsymbol{\ell}_t\}_{t=1}^T$ are chosen before the game starts.
  - ***non-oblivious***: $\boldsymbol{\ell}_t(a_1, \ell_{1,a_1}, \ldots, a_{t-1}, \ell_{t-1,a_{t-1}})$ can depend on the past history.
- ***stochastic*** setting: $\boldsymbol{\ell}_t \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, where $\mathcal{D}$ is a fixed unknown distribution.

Formulation

At each round $t = 1, 2, \cdots$

(1) the player first picks an arm $a_t \in [K]$ from K candidate arms;

(2) and simultaneously environments pick a loss vector $\boldsymbol{\ell}_t \in [0, 1]^K$;

(3) the player suffers and only observes loss $\ell_{t,a_t} \in [0, 1]$, then updates the model.

Goal: to minimize ***expected regret***

$$\mathbb{E}[\text{REG}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_{t,a},$$

where the expectation is taken over the ***randomness of algorithms***.

deterministic algorithms will suffer an $\Omega(T)$ regret in the worst case under the bandit setting!

Comparison

Full-Information Problem	Domain	Loss Functions	Feedback
Prediction with Experts' Advice	Δ_d	$f_t(\boldsymbol{p}_t) = \langle \boldsymbol{\ell}_t, \boldsymbol{p}_t \rangle$	$f_t(\boldsymbol{p}_t),\ \boldsymbol{\ell}_t$
Online Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t),\ \nabla f_t(\mathbf{x}_t), \dots$

Bandit Problem	Domain	Loss Functions	Feedback
Multi-Armed Bandits	$\{\mathbf{e}_1, \dots, \mathbf{e}_K\}$	$f_t(\mathbf{e}_{a_t}) = \langle \boldsymbol{\ell}_t, \mathbf{e}_{a_t} \rangle$	$f_t(\mathbf{e}_{a_t}) = \ell_{t,a_t}$
Bandit Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t)$

Notation: $\mathbf{e}_i \in \mathbb{R}^K$ is the one-hot vector, with i-th entry being 1.

Caveat: the feasible domain of MAB is actually ***not*** convex.

(simplex is the convex hull of $\{\mathbf{e}_1, \dots, \mathbf{e}_K\}$)

Comparison

Full-Information Problem	Domain	Loss Functions	Feedback
Prediction with Experts' Advice	Δ_d	$f_t(\boldsymbol{p}_t) = \langle \boldsymbol{\ell}_t, \boldsymbol{p}_t \rangle$	$f_t(\boldsymbol{p}_t),\ \boldsymbol{\ell}_t$
Online Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t),\ \nabla f_t(\mathbf{x}_t), \dots$

Bandit Problem	Domain	Loss Functions	Feedback
Multi-Armed Bandits	$\{\mathbf{e}_1, \dots, \mathbf{e}_K\}$	$f_t(\mathbf{e}_{a_t}) = \langle \boldsymbol{\ell}_t, \mathbf{e}_{a_t} \rangle$	$f_t(\mathbf{e}_{a_t}) = \ell_{t,a_t}$
Bandit Convex Optimization	$\mathcal{X}$	$f_t(\cdot)$	$f_t(\mathbf{x}_t)$

Notation: $\mathbf{e}_i \in \mathbb{R}^K$ is the one-hot vector, with i-th entry being 1.

Caveat: the feasible domain of MAB is actually ***not*** convex.

(simplex is the convex hull of $\{\mathbf{e}_1, \dots, \mathbf{e}_K\}$)

A Natural Solution for MAB

- MAB bears much similarity with the PEA problem (except for the amount feedback information).

⇨ Deploying Hedge to MAB problem.

Hedge for PEA

At each round $t = 1, 2, \cdots$

(1) compute $\boldsymbol{p}_t \in \Delta_K$ such that $p_{t,i} \propto \exp\left(-\eta L_{t-1,i}\right)$ for $i \in [K]$

(2) the player submits $\boldsymbol{p}_t$, suffers loss $\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, and observes loss $\boldsymbol{\ell}_t \in \mathbb{R}^K$

(3) update $\boldsymbol{L}_t = \boldsymbol{L}_{t-1} + \boldsymbol{\ell}_t$

A Natural Solution for MAB

- However, Hedge does not fit for MAB setting due to ***limited feedback***.

Hedge requires $\ell_{t,i}$ for all $i \in [K]$, but only ℓ_{t,a_t} is available in MAB.

Hedge for PEA

At each round $t = 1, 2, \cdots$

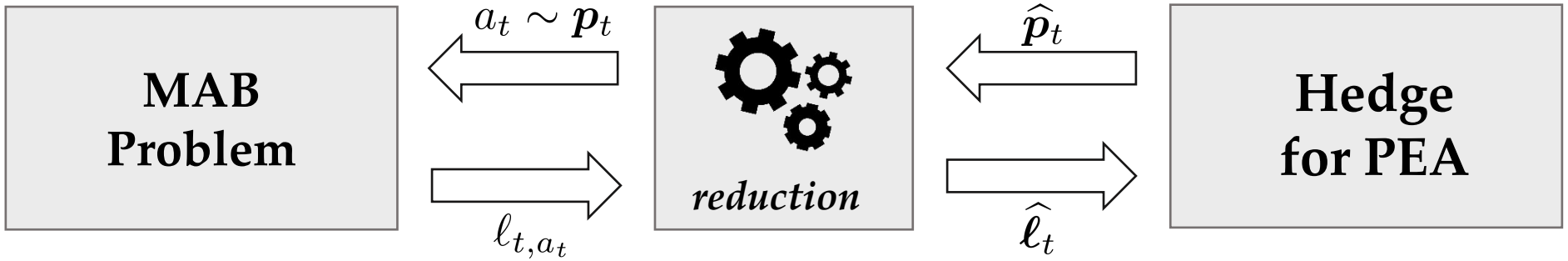
(1) compute $\boldsymbol{p}_t \in \Delta_K$ such that $p_{t,i} \propto \exp\left(-\eta L_{t-1,i}\right)$ for $i \in [K]$

(2) the player submits $\boldsymbol{p}_t$, suffers loss $\langle \boldsymbol{p}_t, \boldsymbol{\ell}_t \rangle$, and observes loss $\boldsymbol{\ell}_t \in \mathbb{R}^K$

(3) update $\boldsymbol{L}_t = \boldsymbol{L}_{t-1} + \boldsymbol{\ell}_t$

Reduction of MAB to PEA

- Given the similarity of MAB and PEA, can we realize the reduction?



- sample an arm $a_t \sim \boldsymbol{p}_t$, and we simply set $\boldsymbol{p}_t = \widehat{p}_t$ (henceforth using notation $\boldsymbol{p}_t$)
- $\widehat{\boldsymbol{\ell}}_t \in \mathbb{R}_+^K$ is the estimated loss fed to Hedge for PEA

$$\Rightarrow \quad \boxed{\mathrm{REG}_T^{\mathrm{MAB}} \overset{\text{by reduction}}{\sim} \mathrm{REG}_T^{\mathrm{PEA}} = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t \rangle - \sum_{t=1}^{T} \widehat{\ell}_{t,i} \leq \mathcal{O}(\sqrt{T})}$$

Reduction of MAB to PEA

- **Importance-Weighted (IW)** Loss Estimator

Define $\widehat{\boldsymbol{\ell}}_t \in \mathbb{R}^K$, for all $a \in [K]$,

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

Loss Estimator

IW Loss Estimator

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

- Property 1. $\ell_{t,a_t} = \langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t \rangle$

$$\mathrm{REG}_T^{\mathrm{MAB}} = \sum_{t=1}^{T} \ell_{t,a_t} - \sum_{t=1}^{T} \ell_{t,i}$$

relate loss of MAB algorithm to that of PEA algorithm

$$\mathrm{REG}_T^{\mathrm{PEA}} = \sum_{t=1}^{T} \langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t \rangle - \sum_{t=1}^{T} \widehat{\ell}_{t,i} \leq \mathcal{O}(\sqrt{T})$$

Loss Estimator

IW Loss Estimator

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

- Property 1. $\ell_{t,a_t} = \langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t \rangle$

- Property 2. $\mathbb{E}_{a_t \sim \boldsymbol{p}_t}[\widehat{\ell}_{t,a}] = \ell_{t,a}, \ \forall a \in [K]$ ***unbiasedness***

Proof.

$$\mathbb{E}_{a_t \sim \boldsymbol{p}_t}[\widehat{\ell}_{t,a}] = \mathbb{E}_{a_t \sim \boldsymbol{p}_t}\left[\frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\}\right] = \mathbb{E}_{a_t \sim \boldsymbol{p}_t}\left[\frac{\ell_{t,a}}{p_{t,a}} \mathbb{1}\{a = a_t\}\right]$$

$$= \frac{\ell_{t,a}}{p_{t,a}} \mathbb{E}_{a_t \sim \boldsymbol{p}_t}[\mathbb{1}\{a = a_t\}] = \frac{\ell_{t,a}}{p_{t,a}} p_{t,a} = \ell_{t,a}. \qquad \square$$

Exp3: Algorithm

Exp3 (Exponential-weight for Exploration and Exploitation)

At each round $t = 1, 2, \cdots$

(1) compute $\boldsymbol{p}_t \in \Delta_K$ such that $p_{t,i} \propto \exp\left(-\eta \widehat{L}_{t-1,a}\right)$ for $a \in [K]$

(2) chooses $a_t \sim \boldsymbol{p}_t$, suffers and observe loss ℓ_{t,a_t}, and construct loss estimator $\widehat{\boldsymbol{\ell}}_t \in \mathbb{R}^K$ as

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

(3) update $\widehat{\boldsymbol{L}}_t = \widehat{\boldsymbol{L}}_{t-1} + \widehat{\boldsymbol{\ell}}_t$

Exp3: Regret Bound

Theorem 1. *Suppose that* $\forall t \in [T]$ *and* $a \in [K], 0 \leq \ell_{t,a} \leq 1$, *then Exp3 with learning rate* $\eta = \sqrt{(\ln K)/(TK)}$ *guarantees*

$$\mathbb{E}[\mathrm{REG}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_{t,a} \leq \mathcal{O}\left(\sqrt{TK \log K}\right),$$

where the expectation is taken over the randomness of the algorithm.

Comparison:

Hedge for PEA	***Exp3 for MAB***
full-information feedback	bandit feedback
$\mathrm{REG}_T \leq \mathcal{O}(\sqrt{T \log K})$	$\mathrm{REG}_T \leq \mathcal{O}(\sqrt{TK \log K})$

suffer a larger arm dependence

Upper and Lower Bounds for MAB

Theorem 1 (Upper Bound for Exp3). *Suppose that* $\forall t \in [T]$ *and* $a \in [K]$, $0 \le \ell_{t,a} \le 1$, *then Exp3 with learning rate* $\eta = \sqrt{(\ln K)/(TK)}$ *guarantees*

$$\mathbb{E}[\text{REG}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_{t,a} \le \mathcal{O}\left(\sqrt{TK \log K}\right),$$

where the expectation is taken over the randomness of the algorithm.

Theorem 2 (Lower Bound for MAB). *For any algorithm* $\mathcal{A}$, *there exists a sequence of loss vectors* $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \cdots, \boldsymbol{\ell}_T$ *constituing an MAB problem such that*

$$\inf_{\mathcal{A}} \sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \mathbb{E}\left[\text{REG}_T\right] = \Omega(\sqrt{TK}).$$

Proof of Exp3 Regret Bound

Proof.

Recall that (Lecture 6, OMD), Hedge under PEA setting guarantees,

$$\sum_{t=1}^{T}\langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t\rangle - \sum_{t=1}^{T}\widehat{\ell}_{t,a} \le \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T}\sum_{a=1}^{K} p_{t,a}\left(\widehat{\ell}_{t,a}\right)^2,\ \forall a \in [K]$$

(potential-based analysis allows $\widehat{\ell}_t \in \mathbb{R}_+^K$.)

Note that our previous reduction ensures $\ell_{t,a_t} = \langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t\rangle$,

$$\Longrightarrow \quad \sum_{t=1}^{T}\ell_{t,a_t} - \sum_{t=1}^{T}\widehat{\ell}_{t,a} \le \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T}\sum_{a=1}^{K} p_{t,a}\left(\widehat{\ell}_{t,a}\right)^2$$

Proof of Exp3 Regret Bound

Proof. $$\sum_{t=1}^{T} \ell_{t,a_t} - \sum_{t=1}^{T} \widehat{\ell}_{t,a} \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} p_{t,a} \left(\widehat{\ell}_{t,a}\right)^2$$

We have the following upper bound for the variance,

$$\mathbb{E}_{a_t \sim \boldsymbol{p}_t} \left[\left(\widehat{\ell}_{t,a}\right)^2 \right] = \mathbb{E}_{a_t \sim \boldsymbol{p}_t} \left[\left(\frac{\ell_{t,a_t}}{p_{t,a}}\right)^2 \mathbb{1}\left\{a = a_t\right\} \right] = \mathbb{E}_{a_t \sim \boldsymbol{p}_t} \left[\left(\frac{\ell_{t,a}}{p_{t,a}}\right)^2 \mathbb{1}\left\{a = a_t\right\} \right]$$

$$= \left(\frac{\ell_{t,a}}{p_{t,a}}\right)^2 \mathbb{E}_{a_t \sim \boldsymbol{p}_t} \left[\mathbb{1}\left\{a = a_t\right\}\right] = \frac{\left(\ell_{t,a}\right)^2}{p_{t,a}}.$$

Proof of Exp3 Regret Bound

Proof. $$\sum_{t=1}^{T} \ell_{t,a_t} - \sum_{t=1}^{T} \widehat{\ell}_{t,a} \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} p_{t,a} \left(\widehat{\ell}_{t,a}\right)^2$$

By the Law of total expectation and the above inequality,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t} - \sum_{t=1}^{T} \widehat{\ell}_{t,a}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_t\left[\ell_{t,a_t} - \widehat{\ell}_{t,a}\right]\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_t\left[\ell_{t,a_t}\right] - \ell_{t,a}\right] \qquad (\mathbb{E}_t[\widehat{\ell}_{t,a}] = \ell_{t,a})$$

regret bound

$$= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a_t}\right] - \sum_{t=1}^{T} \ell_{t,a}$$

$$\leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} \mathbb{E}\left[\mathbb{E}_t\left[p_{t,a} \cdot \left(\widehat{\ell}_{t,a}\right)^2\right]\right]$$

Proof of Exp3 Regret Bound

Proof.

$$\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,a_t}\right]-\sum_{t=1}^{T}\ell_{t,a}$$

$$\leq \frac{\ln K}{\eta}+\eta\sum_{t=1}^{T}\sum_{a=1}^{K}\mathbb{E}\left[\mathbb{E}_t\left[p_{t,a}\cdot\left(\widehat{\ell}_{t,a}\right)^2\right]\right]$$

$$=\frac{\ln K}{\eta}+\eta\sum_{t=1}^{T}\sum_{a=1}^{K}p_{t,a}\cdot\frac{\ell_{t,a}^2}{p_{t,a}} \qquad \left(\mathbb{E}_t\left[p_{t,a}\cdot\widehat{\ell}_{t,a}^2\right]=p_{t,a}\cdot\mathbb{E}_t\left[\widehat{\ell}_{t,a}^2\right]=p_{t,a}\cdot\frac{\ell_{t,a}^2}{p_{t,a}}\right)$$

$$=\frac{\ln K}{\eta}+\eta\sum_{t=1}^{T}\sum_{a=1}^{K}\ell_{t,a}^2$$

$$\leq \frac{\ln K}{\eta}+\eta TK\leq\mathcal{O}(\sqrt{TK\log K}) \qquad \left(\ell_{t,a}^2\leq 1,\eta=\sqrt{(\ln K)/TK}\right)$$

□

Key: Loss Estimator in Bandits

- Importance weighting estimator,

$$\widehat{\boldsymbol{\ell}}_t = [0, \ldots, 0, \underbrace{\frac{\ell_{t,a_t}}{p_{t,a_t}}}_{a_t\text{-th entry}}, 0, \ldots, 0]^\top$$

$\Longrightarrow$ balancing exploitation ℓ_{t,a_t} and exploration p_{t,a_t}

Other Choice

- Other estimators coming to mind,

$$\widehat{\boldsymbol{\ell}}_t = [0, \ldots, 0, \underbrace{\ell_{t,a_t}}_{a_t\text{-th entry}}, 0, \ldots, 0]^\top$$

pure exploitation!

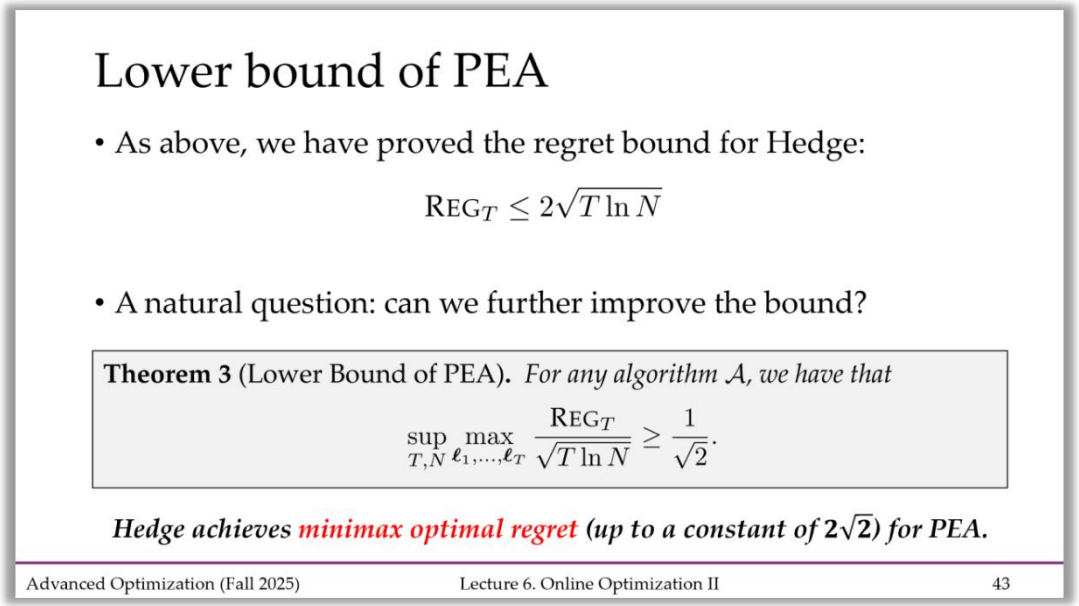
$$\Longrightarrow \quad \underbrace{\langle \boldsymbol{p}_t, \widehat{\boldsymbol{\ell}}_t \rangle}_{\text{loss for Hedge}} = p_{t,a_t} \ell_{t,a_t} \neq \underbrace{\ell_{t,a_t}}_{\text{loss for MAB}}$$

and it ***cannot*** apply Hedge

Lower Bound for MAB

Theorem 2 (Lower Bound for MAB). *For any algorithm $\mathcal{A}$, there exists a sequence of loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \cdots, \boldsymbol{\ell}_T$ constituing an MAB problem such that*

$$\inf_{\mathcal{A}} \sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \mathbb{E}\left[\mathrm{REG}_T\right] = \Omega(\sqrt{TK})$$



Lower bound of PEA

- As above, we have proved the regret bound for Hedge:

$$\mathrm{REG}_T \leq 2\sqrt{T \ln N}$$

- A natural question: can we further improve the bound?

Theorem 3 (Lower Bound of PEA). *For any algorithm $\mathcal{A}$, we have that*

$$\sup_{T,N} \max_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T} \frac{\mathrm{REG}_T}{\sqrt{T \ln N}} \geq \frac{1}{\sqrt{2}}.$$

Hedge achieves minimax optimal regret (up to a constant of $2\sqrt{2}$) for PEA.

Advanced Optimization (Fall 2025) Lecture 6. Online Optimization II 43

MAB Problem $\Omega(\sqrt{TK})$

PEA Problem $\Omega(\sqrt{T \log K})$

Proof Sketch

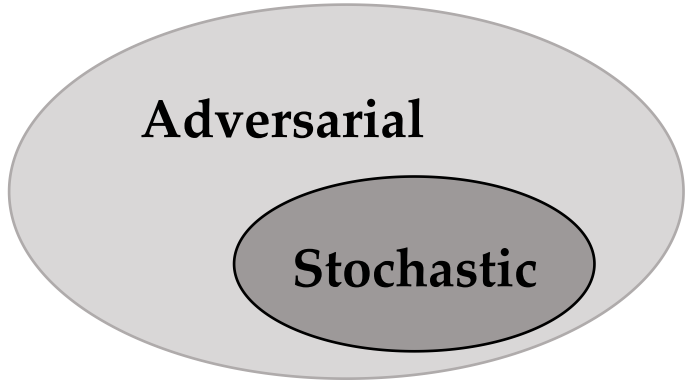
Proof (Sketch).

We prove the theorem under stochastic MAB setting, since the stochastic setting is strictly easier than the adversarial one.

We construct two hard distributions over arms $\mathcal{D}_1, \mathcal{D}_2$ and show that,

$\forall A \in \mathcal{A}$, the following holds

$$\max\left\{\mathbb{E}[\text{REG}_T(A); \mathcal{D}_1], \mathbb{E}[\text{REG}_T(A); \mathcal{D}_2]\right\} = \Omega(\sqrt{TK})$$



Advanced Topics

- How to shave off the extra $\sqrt{\log K}$ factor?

⇨ Using OMD with ***Tsallis entropy*** regularizer, also using the IW estimator

$$\psi(\boldsymbol{p}) = \frac{1 - \sum_{a=1}^{K} p_a^{\beta}}{1 - \beta}$$

which is actually a *generalization* of negative-entropy used in Hedge, as we have the following fact due to the L'Hôpital'srule

$$\lim_{\beta \to 1} \frac{1 - \sum_a p_a^{\beta}}{1 - \beta} = \sum_a p_a \ln(p_a).$$

Reference: Jean-Yves Audibert and Sebastien Bubeck. Regret bounds and minimax policies under partial monitoring. Journal of Machine Learning Research, 11(Oct):2785–2836, 2010.

Advanced Topics

- How to boost from expected guarantee to a ***high-probability*** one?

⇨ Using an improved estimator: **Implicit eXploration (IX) Loss Estimator**

IW Loss Estimator

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a}} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t}} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

IX Loss Estimator

$$\widehat{\ell}_{t,a} = \frac{\ell_{t,a_t}}{p_{t,a} + \gamma} \mathbb{1}\{a = a_t\} = \begin{cases} \frac{\ell_{t,a_t}}{p_{t,a_t} + \gamma} & \text{if } a = a_t; \\ 0 & \text{else.} \end{cases}$$

Reference: Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. NIPS 2015.

Part 3. Stochastic MAB

- Exploration-Exploitation Dilemma
- Lower Bound
- Algorithms
  - Explore-then-Commit (ETC)
  - ε-Greedy

Stochastic Multi-Armed Bandit (MAB)

- **MAB**: A player is facing K arms. At each time t, the player pulls one arm $a \in [K]$ and then receives a reward $r_t(a) \in [0, 1]$:

Arm 1	$r_1(1)$	$r_2(1)$	0.6	$r_4(1)$	$r_5(1)$
Arm 2	1	$r_2(2)$	$r_3(2)$	0.2	$r_5(2)$
Arm 3	$r_1(3)$	0.7	$r_3(3)$	$r_4(3)$	0.3

- **Stochastic**:

  Each arm $a \in [K]$ has an unknown distribution $\mathcal{D}_a$ with mean $\mu(a)$,

  such that rewards $r_1(a), r_2(a), ..., r_T(a)$ are i.i.d samples from $\mathcal{D}_a$.

For conventional issue, we will use the "***reward language***" in stochastic bandits.

Formulation

At each round $t = 1, 2, \cdots$

(1) player first chooses an arm $a_t \in [K]$;

(2) environment reveals a reward $r_t(a_t) \sim$ distribution $\mathcal{D}_{a_t}$;

(3) player updates the model by the pair $(a_t, r_t(a_t))$.

- The goal is to minimize the ***pseudo regret***:

$$\bar{R}_T \triangleq \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t)\right] = T\mu(a^*) - \sum_{t=1}^{T} \mu(a_t)$$

where $a^* \in \arg\max_{a \in [K]} \mu(a)$ is the best arm in the sense of expectation.

- Caveat: note the difference between *pseudo regret* and the *(expected) regret*.

Lower Bound

- Two types of lower bound
  - **Instance-dependent lower bound**: characterize the difficulty of a specific bandit instance.
  - **Instance-independent (minimax) lower bound**: hold for all algorithms and all stochastic bandit environments.

Theorem 2 (Minimax Lower Bound for MAB)**.** *For any bandit algorithm $\mathcal{A}$, there exists an instance ν with a stochastic loss sequence such that*

$$\inf_{\mathcal{A}} \sup_{\nu} \mathbb{E}\left[\bar{R}_T(\mathcal{A}, \nu)\right] = \Omega(\sqrt{TK})$$

Instance-Dependent Lower Bound

Theorem 3 (Lai-Robbins Lower Bound for Stochastic MAB). *For any algorithm $\mathcal{A}$ and any stochastic MAB instance ν, with arm a's reward distribution denoted by ν_a and optimal arm a^*, we have*

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[\bar{R}_T(\mathcal{A},\nu)\right]}{\log T} \geq \sum_{a:\Delta_a>0} \frac{\Delta_a}{\mathrm{KL}(\nu_a\|\nu_{a^*})}.$$

- In typical reward models (e.g., Bernoulli or sub-Gaussian), we have that $\mathrm{KL}(\nu_a\|\nu_{a^*}) = \Theta(\Delta_a^2)$. This indicates that

$$\mathbb{E}\left[\bar{R}_T\right] = \Omega\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right).$$

- This instance-dependent guarantee is (usually) called ***gap-dependent*** in MAB.

Gap-dependent vs Gap-independent

- Consider this gap-dependent lower bound: $\mathbb{E}\left[\bar{R}_T\right] = \Omega\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right)$.

- This does *not* contradict with the (gap-independent) minimax lower bound, since we can construct *hard* instances with vanishing gap $\Delta_a = \Theta(\sqrt{K/T})$.

  - Suppose there is an arm a with a small gap Δ_a, then always picking arm a should just lead to $\bar{R}_T = \Delta_a T$.

  - So if $\Delta_a \leq \sqrt{K/T}$, this will not contradict with $\bar{R}_T = \sqrt{KT}$ minimax rate.

  - Otherwise ($\Delta_a \geq \sqrt{K/T}$), the gap-dependent lower bound implies $\sum_a \log T/\Delta_a \geq \log T\sqrt{KT}$, also collaborates with minimax lower bound.

Solving Stochastic MAB

- Regret Decomposition:

$$\bar{R}_T = \max_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t)\right] = T\mu(a^*) - \sum_{t=1}^{T} \mu(a_t)$$

$$= \sum_{a\in[K]} (\mu(a^*) - \mu(a_t)) \cdot n_T(a) = \sum_{a\in[K]} \Delta_a \cdot n_T(a)$$

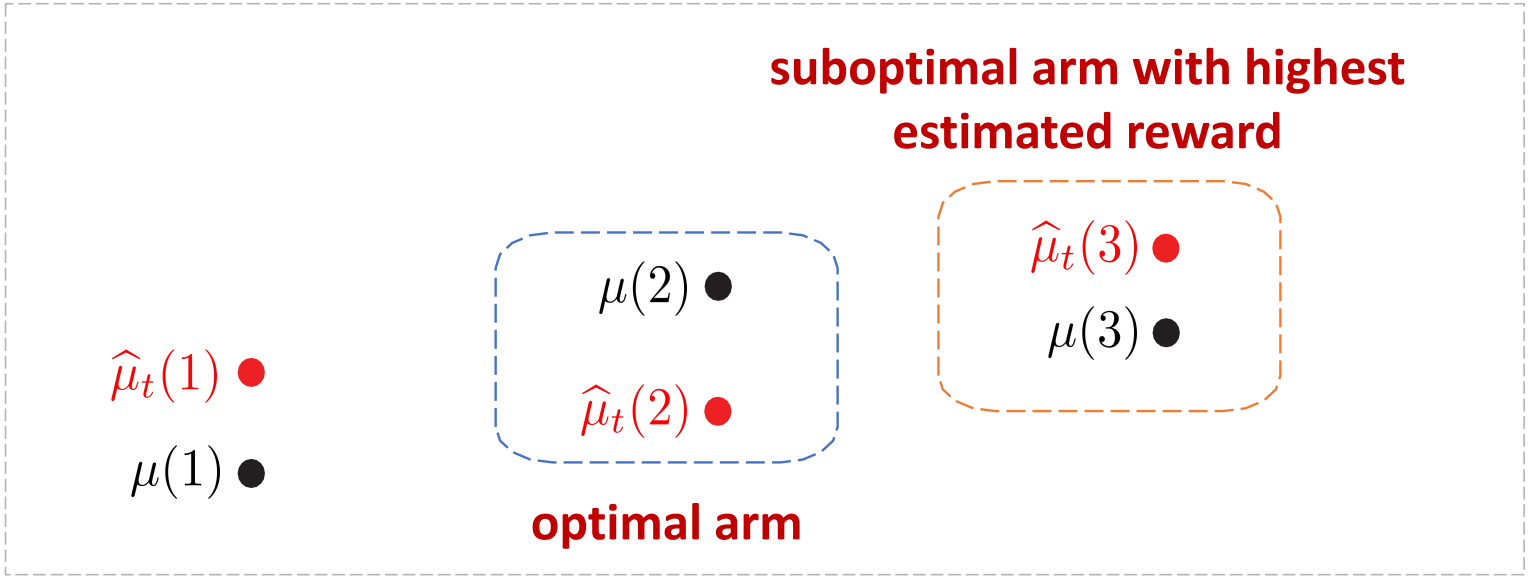
- For stochastic MAB, a natural characterization of the arms:

(i) Suboptimality gap: $\Delta_a = \mu(a^*) - \mu(a)$;

(ii) Number of times arm a is pulled in t rounds: $n_t(a) = \sum_{s=1}^{t} \mathbf{1}\{a_s = a\}$.

Exploration-Exploitation Dilemma

- How to balance exploration and exploitation?



Exploration vs Exploitation

- **Exploitation**: pull the best arm so far
- **Exploration**: try other arms that may be better

- This is a fundamental problem in bandits, reinforcement learning, control, decision-making problems, and related areas.

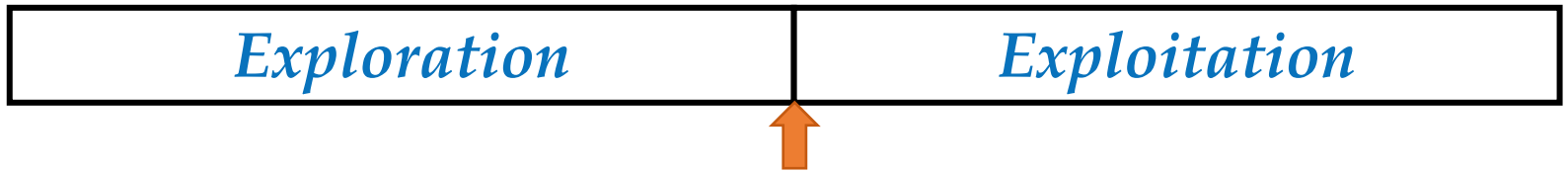
MAB Algorithms

- Explore-then-Commit (ETC)
- ε-Greedy
- Comparisons

Explore-then-Commit

(1) Do ***explore*** for the first T_0 round by pulling each arm for T_0/K times;

(2) Do ***exploit*** for the rest $T - T_0$ round by always pulling $\widehat{a} = \arg\max_{a \in [K]} \widehat{\mu}_{T_0}(a)$.



Theorem 4. *Suppose that* $\forall t \in [T]$ *and* $a \in [K], 0 \le r_t(a) \le 1$, *then ETC with exploration period* T_0 *guarantees*

$$\mathbb{E}[\bar{R}_T] \le \sum_{a \in [K]} \left(\frac{T_0}{K} + 2T \exp\left(-\frac{T_0 \Delta_a^2}{2K} \right) \right) \Delta_a.$$

Proof of ETC Regret Bound

$\Delta_a = \mu(a^*) - \mu(a)$

Proof. By regret decomposition lemma: $\mathbb{E}[\bar{R}_T] = \sum_{a \in [K]} \Delta_a \cdot \mathbb{E}[n_T(a)]$

Below we estimate $\mathbb{E}[n_T(a)]$, the expected number of pulls of arm a:

Exploration Exploitation

$$\begin{aligned}\mathbb{E}[n_T(a)] &= T_0/K + (T - T_0)\Pr\{\widehat{a} = a\} \\ &\leq T_0/K + (T - T_0)\Pr\{\widehat{\mu}_{T_0}(a) \geq \widehat{\mu}_{T_0}(a^*)\}\end{aligned}$$

Optimal arm $a^* = \arg\max_{a \in [K]} \mu(a)$

Pulling strategy $\hat{a} = \arg\max_{a \in [K]} \hat{\mu}_{T_0}(a)$

Note that when $\widehat{\mu}_{T_0}(a) \geq \widehat{\mu}_{T_0}(a^*)$ happens, it implies one of the following two rare events must happen:

$$\widehat{\mu}_{T_0}(a) \geq (\mu(a) + \mu(a^*))/2, \text{ and } \widehat{\mu}_{T_0}(a^*) \leq (\mu(a) + \mu(a^*))/2.$$

Otherwise, $\widehat{\mu}_{T_0}(a) < (\mu(a) + \mu(a^*))/2 < \widehat{\mu}_{T_0}(a^*)$.

Proof of ETC Regret Bound

$\Delta_a = \mu(a^*) - \mu(a)$

Proof. By regret decomposition lemma: $\mathbb{E}[\bar{R}_T] = \sum_{a \in [K]} \Delta_a \cdot \mathbb{E}[n_T(a)]$

Below we estimate $\mathbb{E}[n_T(a)]$, the expected number of pulls of arm a:

Exploration Exploitation

$$\begin{aligned}
\mathbb{E}[n_T(a)] &= T_0/K + (T - T_0)\Pr\{\widehat{a} = a\} \\
&\leq T_0/K + (T - T_0)\Pr\{\widehat{\mu}_{T_0}(a) \geq \widehat{\mu}_{T_0}(a^*)\} \\
&\leq T_0/K + (T - T_0)\Pr\left\{\widehat{\mu}_{T_0}(a) \geq \frac{\mu(a) + \mu(a^*)}{2} \cup \widehat{\mu}_{T_0}(a^*) \leq \frac{\mu(a) + \mu(a^*)}{2}\right\} \\
&\leq T_0/K + (T - T_0)\left(\Pr\left\{\widehat{\mu}_{T_0}(a) \geq \frac{\mu(a) + \mu(a^*)}{2}\right\} + \Pr\left\{\widehat{\mu}_{T_0}(a^*) \leq \frac{\mu(a) + \mu(a^*)}{2}\right\}\right)
\end{aligned}$$

Union bound $\Pr\{X \cup Y\} \leq \Pr\{X\} + \Pr\{Y\}$

Proof of ETC Regret Bound

Proof. $\mathbb{E}[n_T(a)] \leq T_0/K + T\left(\Pr\left\{\widehat{\mu}_{T_0}(a) \geq \frac{\mu(a)+\mu(a^*)}{2}\right\} + \Pr\left\{\widehat{\mu}_{T_0}(a^*) \leq \frac{\mu(a)+\mu(a^*)}{2}\right\}\right)$

Hoeffding's Inequality. For independent $X_i \in [0,1], i \in [m], \bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$, we have

$$\Pr\left\{\bar{X} - \mathbb{E}[\bar{X}] \geq \epsilon\right\} \leq \exp\left(-2m\epsilon^2\right);$$
$$\Pr\left\{\bar{X} - \mathbb{E}[\bar{X}] \leq -\epsilon\right\} \leq \exp\left(-2m\epsilon^2\right).$$

$$\Rightarrow \Pr\left\{\widehat{\mu}_{T_0}(a) \geq \frac{\mu(a)+\mu(a^*)}{2}\right\} = \Pr\left\{\widehat{\mu}_{T_0}(a) \geq \mu(a) + \frac{\Delta_a}{2}\right\} \leq \exp\left(-\frac{T_0\Delta_a^2}{2K}\right)$$

$$\Rightarrow \Pr\left\{\widehat{\mu}_{T_0}(a^*) \leq \frac{\mu(a)+\mu(a^*)}{2}\right\} = \Pr\left\{\widehat{\mu}_{T_0}(a^*) \leq \mu(a^*) + \frac{\Delta_a}{2}\right\} \leq \exp\left(-\frac{T_0\Delta_a^2}{2K}\right)$$

$$\Rightarrow \mathbb{E}[\bar{R}_T] = \sum_{a\in[K]} \Delta_a \mathbb{E}[n_T(a)] \leq \sum_{a\in[K]} \left(\frac{T_0}{K} + 2T\exp\left(-\frac{T_0\Delta_a^2}{2K}\right)\right)\Delta_a \qquad \square$$

Issue of ETC

Theorem 4. *Suppose that* $\forall t \in [T]$ *and* $a \in [K], 0 \leq r_t(a) \leq 1$, *then ETC with exploration period* T_0 *guarantees*

$$\mathbb{E}[\bar{R}_T] \leq \sum_{a \in [K]} \left(\frac{T_0}{K} + 2T \exp\left(-\frac{T_0 \Delta_a^2}{2K} \right) \right) \Delta_a.$$

- Need to tune T_0 (or more specifically, the number that each arm is pulled in the exploration phase, i.e., $m \triangleq T_0/K$):

  Tune T_0 with prior of suboptimality gap Δ_a:

$$\mathbb{E}[\bar{R}_T] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a > 0} \Delta_a \right) \quad \text{by setting } m = \frac{2}{\Delta_{\min}^2} \log(2T).$$

Issue of ETC

Theorem 4. *Suppose that* $\forall t \in [T]$ *and* $a \in [K], 0 \leq r_t(a) \leq 1$*, then ETC with exploration period* T_0 *guarantees*

$$\mathbb{E}[\bar{R}_T] \leq \sum_{a \in [K]} \left(\frac{T_0}{K} + 2T \exp\left(-\frac{T_0 \Delta_a^2}{2K} \right) \right) \Delta_a .$$

- Need to tune T_0 :

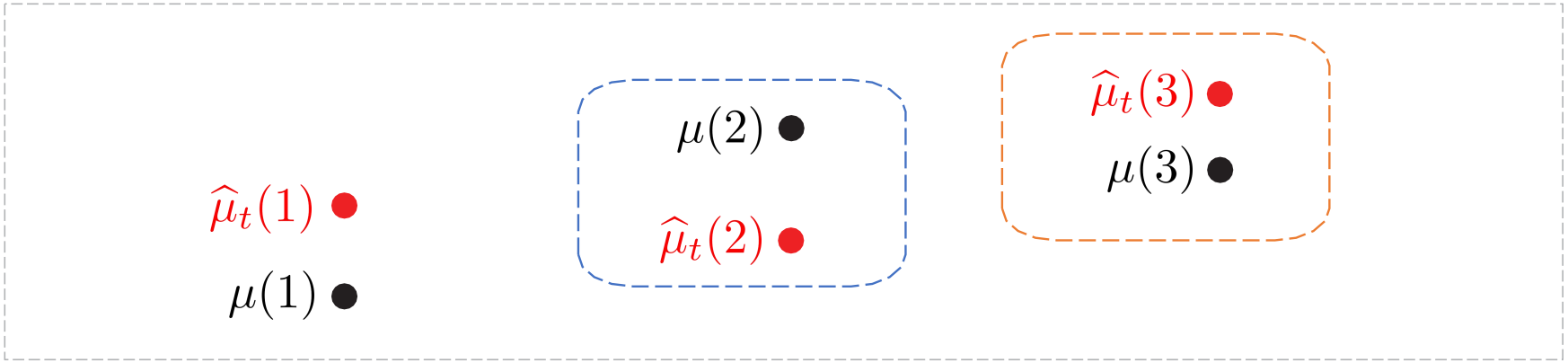
  Tune T_0 with prior of suboptimality gap $\Delta_{\min}$: $\mathbb{E}[\bar{R}_T] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a \in [K]} \Delta_a \right)$

  Tune T_0 without prior of suboptimality gap $\Delta_{\min}$: $\mathbb{E}[\bar{R}_T] = \widetilde{\mathcal{O}}(T^{2/3})$

- In fact, even with a good T_0 with gap information, ETC is ***not optimal***.

Issue of ETC

- ETC algorithm relies on the estimate during the exploration phase. ***There is no way to reverse the estimate!***



- This demonstrates the crucial role of "***adaptive exploration***".

ε-Greedy

- Another simple (yet practical) idea for exploration and exploitation.

ε-Greedy Algorithm

At each round $t = 1, 2, \cdots$

(1) With probability $1 - \varepsilon_t$, choose arm $a_t = \arg\max_a \hat{\mu}_{t-1}(a)$; otherwise choose arm uniformly at random.

(2) Observe reward r_t.

(3) Update the empirical estimate $\hat{\mu}_t(a_t) = \frac{(t-1)\hat{\mu}_{t-1}(a_t)+r_t}{t}$, and $\hat{\mu}_t(a) = \hat{\mu}_{t-1}(a)$ for all $a \neq a_t$.

- In words, do exploration with probability ε_t at round t.

Regret bound of ε-Greedy

- Without adaptive exploration probability: ***asymptotically***,

$$\text{if } \varepsilon_t = \varepsilon > 0, \text{ then } \lim_{T\to\infty} \frac{\mathbb{E}[\bar{R}_T]}{T} = \frac{\varepsilon}{K} \sum_{a\in[K]} \Delta_a.$$

- With adaptive exploration probability:

Theorem 5. *Suppose that* $\forall t \in [T]$ *and* $a \in [K], 0 \le r_t(a) \le 1$, ϵ*-Greedy algorithm with* $\varepsilon_t = \min\{1, CK/(t\Delta_{\min}^2)\}$ *for a sufficiently large universal constant* C*. Then*

$$\mathbb{E}[\bar{R}_T] \le \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a\in[K]} \Delta_a\right).$$

Proof of ε-Greedy Regret Bound

Proof. $\mathbb{E}[\bar{R}_T] = \sum_{a \in [K]} \Delta_a \mathbb{E}[n_T(a)] = \sum_{a \in [K]} \Delta_a \left(\mathbb{E}[n_T^{\text{explore}}(a)] + \mathbb{E}[n_T^{\text{exploit}}(a)] \right)$

Exploration pulls of arm a: $\mathbb{E}[n_T^{\text{explore}}(a)] = \sum_{t=1}^{T} \frac{\varepsilon_t}{K}$

Let $t_0 = \left\lceil \frac{CK}{\Delta_{\min}^2} \right\rceil$ to ensure at least one pull on arm a

$$\leq \frac{1}{K} \left(\sum_{t=1}^{t_0} 1 + \sum_{t=t_0+1}^{T} \frac{CK}{t\Delta_{\min}^2} \right) \quad \text{by the definition of } \varepsilon_t$$

$$\leq \frac{t_0}{K} + \frac{C}{\Delta_{\min}^2} \sum_{t=t_0+1}^{T} \frac{1}{t} \quad \text{by } t_0 \leq \frac{2CK}{\Delta_{\min}^2}$$

$$\leq \frac{2C}{\Delta_{\min}^2} + \frac{C}{\Delta_{\min}^2} \left(1 + \log \frac{T\Delta_{\min}^2}{CK} \right) = \mathcal{O}\left(\frac{1}{\Delta_{\min}^2} \log \frac{T\Delta_{\min}^2}{K} \right)$$

Proof of ε-Greedy Regret Bound

Proof. Exploitation pulls of **suboptimal** arm a:

$$\mathbb{E}[n_T^{\text{exploit}}(a)] \leq \sum_{t=t_0+1}^{T} \Pr\left(\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)\right)$$

No exploitation before t_0+1

$$\leq \sum_{t=t_0+1}^{T} \Pr\left(\hat{\mu}_{t-1}(a) - \mu(a) \geq \tfrac{\Delta_a}{2}\right) + \Pr\left(\mu(a^*) - \hat{\mu}_{t-1}(a^*) \geq \tfrac{\Delta_a}{2}\right)$$

union bound similar to splitting analysis in ETC

Denote $m_t(a) \triangleq \mathbb{E}\left[n_{t-1}^{\text{explore}}(a)\right]$. For $\Pr\left(\hat{\mu}_{t-1}(a) - \mu(a) \geq \tfrac{\Delta_a}{2}\right)$, we have

$$\Pr\left(\widehat{\mu}_{t-1}(a) - \mu(a) \geq \tfrac{\Delta_a}{2}\right)$$

$$= \Pr\left(\widehat{\mu}_{t-1}(a) - \mu(a) \geq \tfrac{\Delta_a}{2},\ n_{t-1}(a) < \frac{m_t(a)}{2}\right) + \sum_{m=\lceil \frac{m_t(a)}{2} \rceil}^{t-1} \Pr\left(\widehat{\mu}_{t-1}(a) - \mu(a) \geq \tfrac{\Delta_a}{2},\ n_{t-1}(a) = m\right)$$

$$\leq \Pr\left(n_{t-1}(a) < \frac{m_t(a)}{2}\right) + \sum_{m=\lceil \frac{m_t(a)}{2} \rceil}^{t-1} \Pr\left(n_{t-1}(a) = m\right) \Pr\left(\widehat{\mu}_{t-1}(a) - \mu(a) \geq \tfrac{\Delta_a}{2} \mid n_{t-1}(a) = m\right)$$

Proof of ε-Greedy Regret Bound

Lemma 1 (Multiplicative Chernoff Bound). *Let N be the sum of independent Bernoulli random variables with mean $\mu = \mathbb{E}[N]$. For any $0 < \delta < 1$, the lower tail bound is given by*

$$\mathbb{P}(N \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right).$$

$$\Pr\left(n_{t-1}(a) < \frac{m_t(a)}{2}\right) \leq \exp\left(-\frac{m_t(a)}{8}\right)$$ By choosing $\delta = \frac{1}{2}$ in Lemma 1

$$\Pr\left(\widehat{\mu}_{t-1}(a) - \mu(a) \geq \tfrac{\Delta_a}{2} \mid n_{t-1}(a) = m\right) \leq \exp\left(-\frac{m}{2\Delta_a^2}\right)$$ Hoeffding's inequality

Plugging them back, we have

$$\Pr\left(\widehat{\mu}_{t-1}(a) - \mu(a) \geq \tfrac{\Delta_a}{2}\right) \leq \exp\left(-\frac{m_t(a)}{8}\right) + \exp\left(-\frac{m_t(a)}{4\Delta_a^2}\right)$$

For $\Pr\left(\mu(a^*) - \hat{\mu}_{t-1}(a^*) \geq \tfrac{\Delta_a}{2}\right)$, we have the same upper bound.

Proof of ε-Greedy Regret Bound

Then, we have

$$\mathbb{E}[n_T^{\text{exploit}}(a)] \leq \sum_{t=t_0+1}^{T} 2\exp\left(-\frac{m_t(a)}{8}\right) + 2\exp\left(-\frac{m_t(a)}{4\Delta_a^2}\right)$$

For $t \geq t_0$, we have

$$m_t(a) \triangleq \mathbb{E}\left[n_{t-1}^{\text{explore}}(a)\right] = \sum_{s=1}^{t-1} \frac{\varepsilon_s}{K} \geq \sum_{s=t_0+1}^{t-1} \frac{C}{s\Delta_{\min}^2} \geq \frac{C}{\Delta_{\min}^2} \log\left(\frac{t}{t_0+1}\right)$$

Plugging back, we have

$$\mathbb{E}[n_T^{\text{exploit}}(a)] \leq \sum_{t=t_0+1}^{T} 2\exp\left(-\frac{m_t(a)}{8}\right) + \sum_{t=t_0+1}^{T} 2\exp\left(-\frac{m_t(a)}{4\Delta_a^2}\right)$$

$$\leq \sum_{t=t_0+1}^{T} \left(\frac{t}{t_0+1}\right)^{-C/8} + \sum_{t=t_0+1}^{T} \left(\frac{t}{t_0+1}\right)^{-C/4} \leq \mathcal{O}(1) \qquad \text{Choosing } C = 16$$

Finally, we have

$$\mathbb{E}[\bar{R}_T] = \sum_{a\in[K]} \Delta_a \left(\mathbb{E}[n_T^{\text{explore}}(a)] + \mathbb{E}[n_T^{\text{exploit}}(a)]\right) \leq \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a\in[K]} \Delta_a\right).$$

Issue of ε-Greedy

- ε-greedy algorithm has certain "adaptivity" on the exploration.

Theorem 5. *Suppose that* $\forall t \in [T]$ *and* $a \in [K], 0 \leq r_t(a) \leq 1$, ϵ*-Greedy algorithm with* $\varepsilon_t = \min\{1, CK/(t\Delta_{\min}^2)\}$ *for a sufficiently large universal constant* C*. Then*

$$\mathbb{E}[\bar{R}_T] \leq \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a \in [K]} \Delta_a\right).$$

- However, the exploration is only over the time-level, and ***ignores the differences over all the arms***.

ETC vs ε-greedy

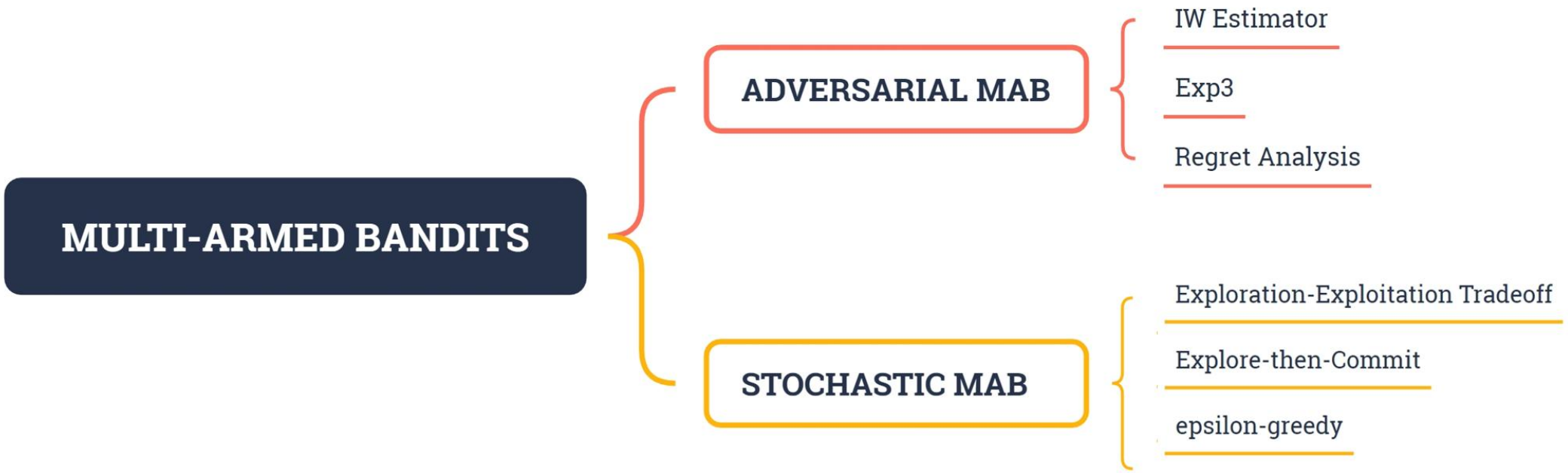
• ETC uses a single exploration length m determined by $\Delta_{\min}$ and explores *every* arm exactly m times.

$$\mathbb{E}[\bar{R}_T^{\text{ETC}}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$$

• ε-greedy performs uniform exploration in each exploration step, but the use of exploitation and the decaying ε_t lead to arm-dependent behaviors.

$$\mathbb{E}[\bar{R}_T^{\varepsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$$

Summary



Q & A

Thanks!