Lecture 11. Stochastic Bandits

Advanced Optimization (Fall 2025)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

Outline

- Multi-Armed Bandits
- Algorithm for Stochastic MAB
- Comparison and Extension

Part 1. Multi-Armed Bandits

- Problem Formulation
- Exploration-Exploitation Dilemma
- Lower Bound

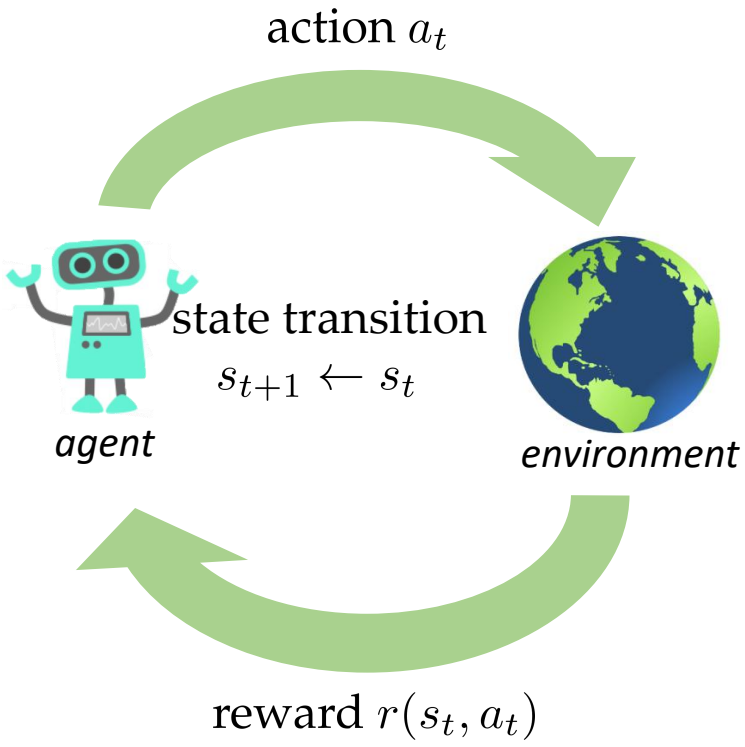
Bandits

- 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

Bandits as Interactive Learning

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



Reinforcement learning:

- Sequential decision making
- With state transition

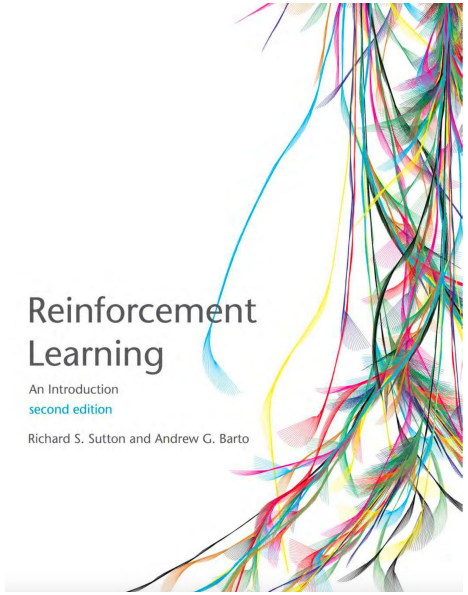
Bandits:

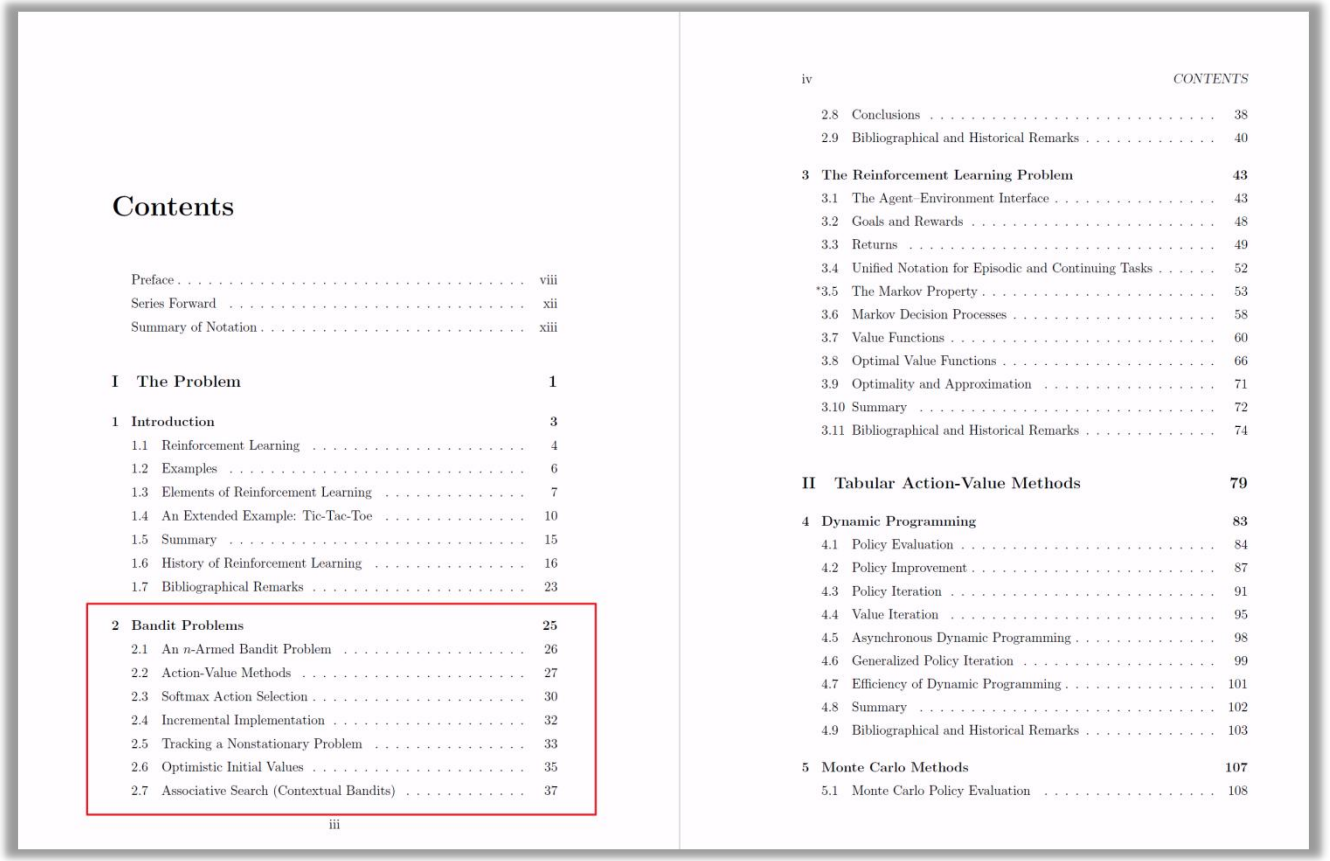
- Single-step decision making
- No state transition

Bandits as Interactive Learning

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

2024 Turing award winner





Contents

iii

iv CONTENTS

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.

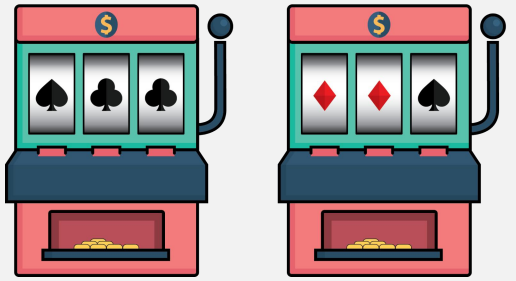
Formulation

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*.

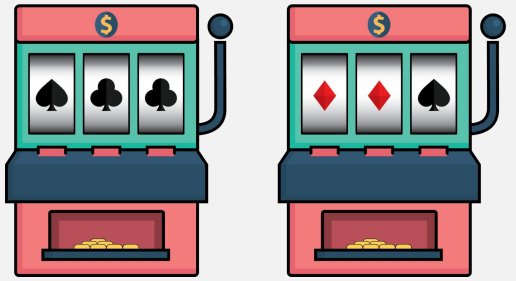
Formulation

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***:

$$\bar{R}_T = T\mu(a^*) - \sum_{t=1}^{T} \mu(a_t)$$

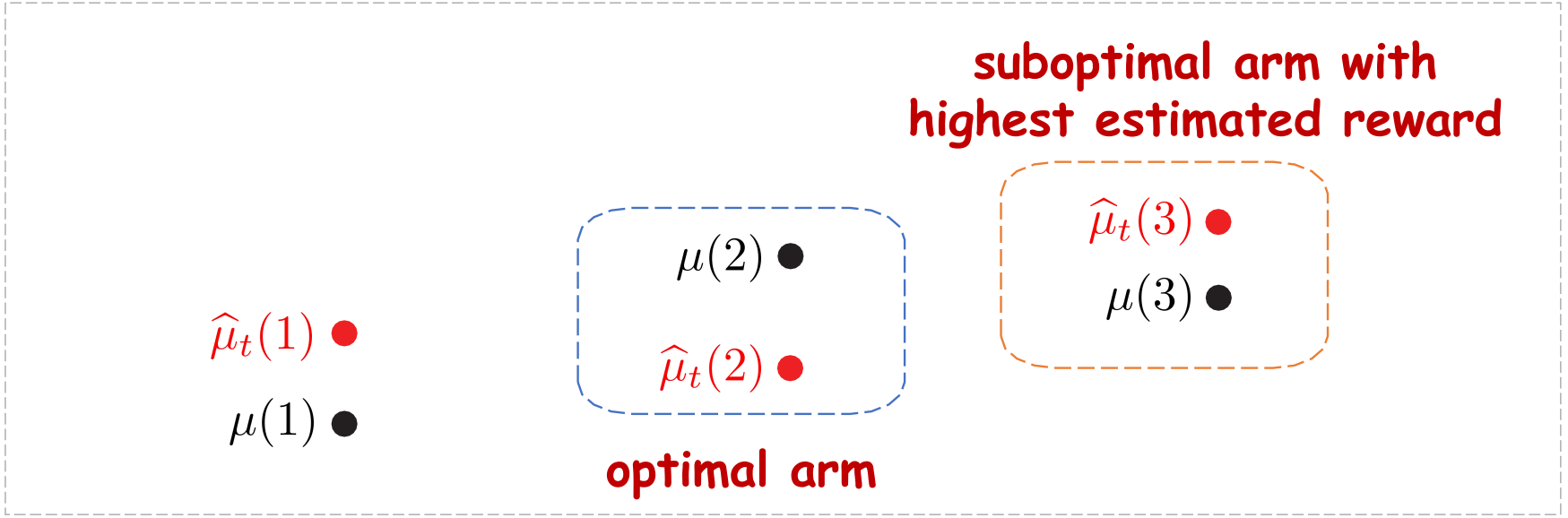
i.e., difference between the cumulative reward of the best arm and that obtained by the bandit algorithm

Exploration vs Exploitation

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

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, recommendation systems, and related areas.

Solving Stochastic MAB: Deploying Exp3

- Stochastic MAB is a special case of Adversarial MAB

⇨ Deploying Exp3 achieves the expected regret (though having gap to pseudo regret).

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*

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

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

⇨ Not yet to exploit benign ***stochastic*** modeling….

instance-dependent analysis

Regret Decomposition

- 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: $n_t(a) = \sum_{s=1}^{t} \mathbf{1}\{a_s = a\}$.

- Regret Decomposition Lemma:

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

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

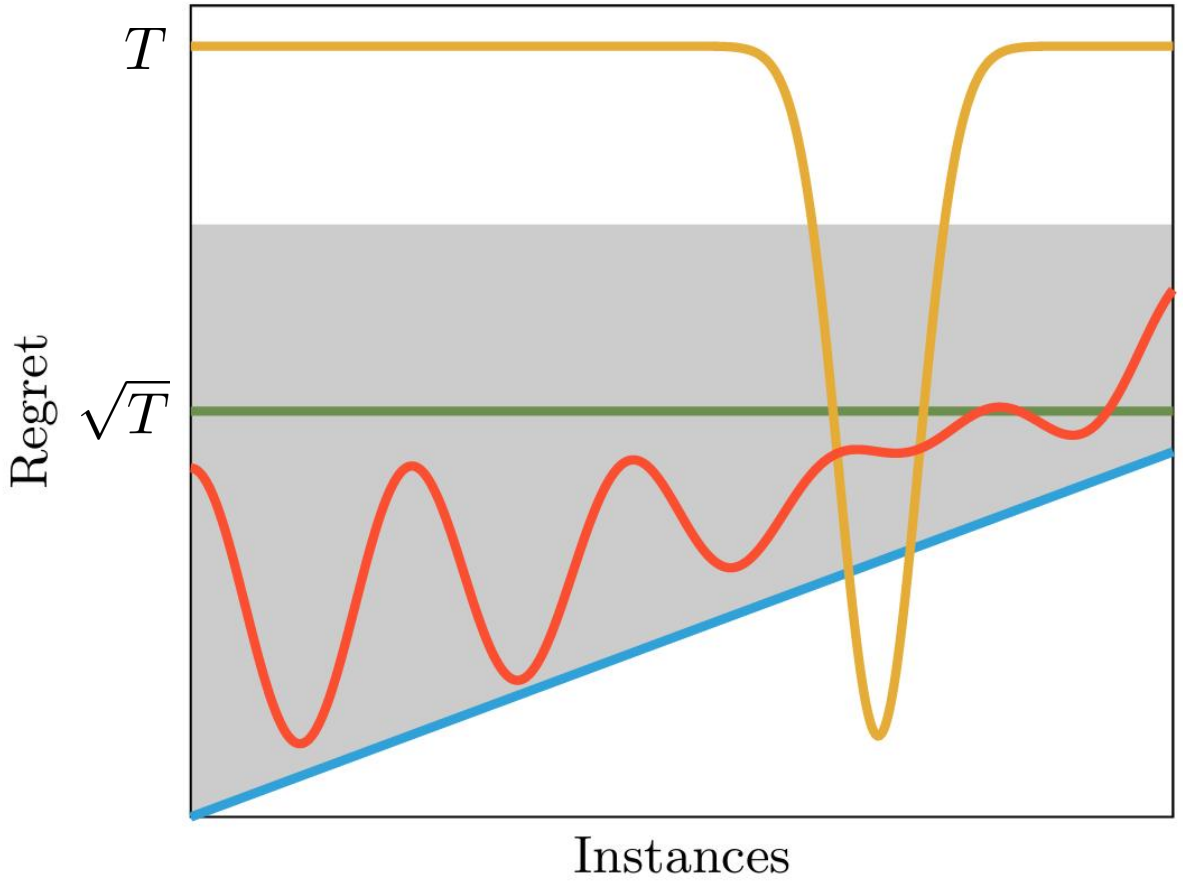
Lower Bound

- How hard is the stochastic MAB problem?
- Two types:
  - **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*

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

Lower Bound



over-specialized

reasonable, not instance optimal

minimax optimality limit

instance optimality limit

Slide credit: Chapter 16,
Bandit Algorithm book

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.

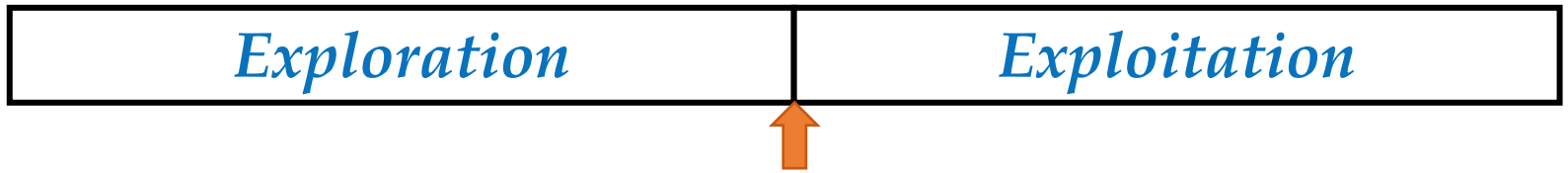
Part 2. Algorithms for Stochastic MAB

- Explore-then-Commit (ETC)
- ε-Greedy
- Upper Confidence Bound (UCB)
- Thompson Sampling

A Natural Solution: 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

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

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

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

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

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

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

Proof of ETC Regret Bound

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

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

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

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

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

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

Issue of ETC

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

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

- 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* $a \in [K], 0 \le r_t(a) \le 1$*, then ETC with exploration period* T_0 *guarantees*

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

- Need to tune T_0 :

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

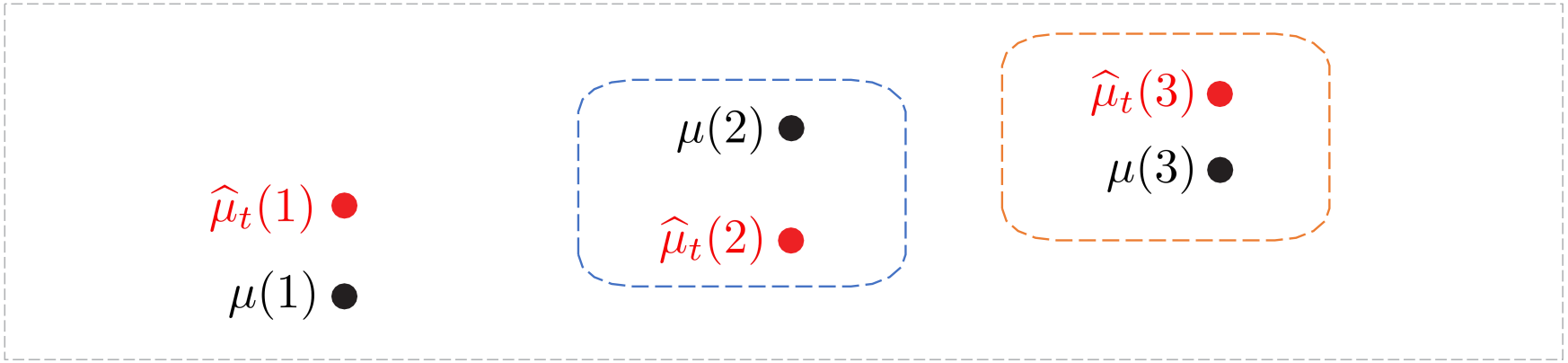
  Tune T_0 without prior of suboptimality gap $\Delta_{\min}$: $\mathbb{E}[\bar{R}_T] = \widetilde{\mathcal{O}}(T^{2/3})$

- ETC is ***not*** a minimax optimal algorithm.

  ⇨ Solution: need ***strategic exploration***.

Strategic Exploration

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



- Strategic exploration methods:
  - ***ε-Greedy:*** explore with certain randomness
  - ***Upper Confidence Bound (UCB):*** explore optimistically
  - ***Thompson Sampling:*** explore by randomness in posterior sampling

ε-Greedy

- Simple idea on balancing 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

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

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

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

Proof of ε-Greedy Regret Bound

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

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

No exploitation before t_0+1

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

union bound similar to splitting analysis in ETC

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

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

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

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

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

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

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

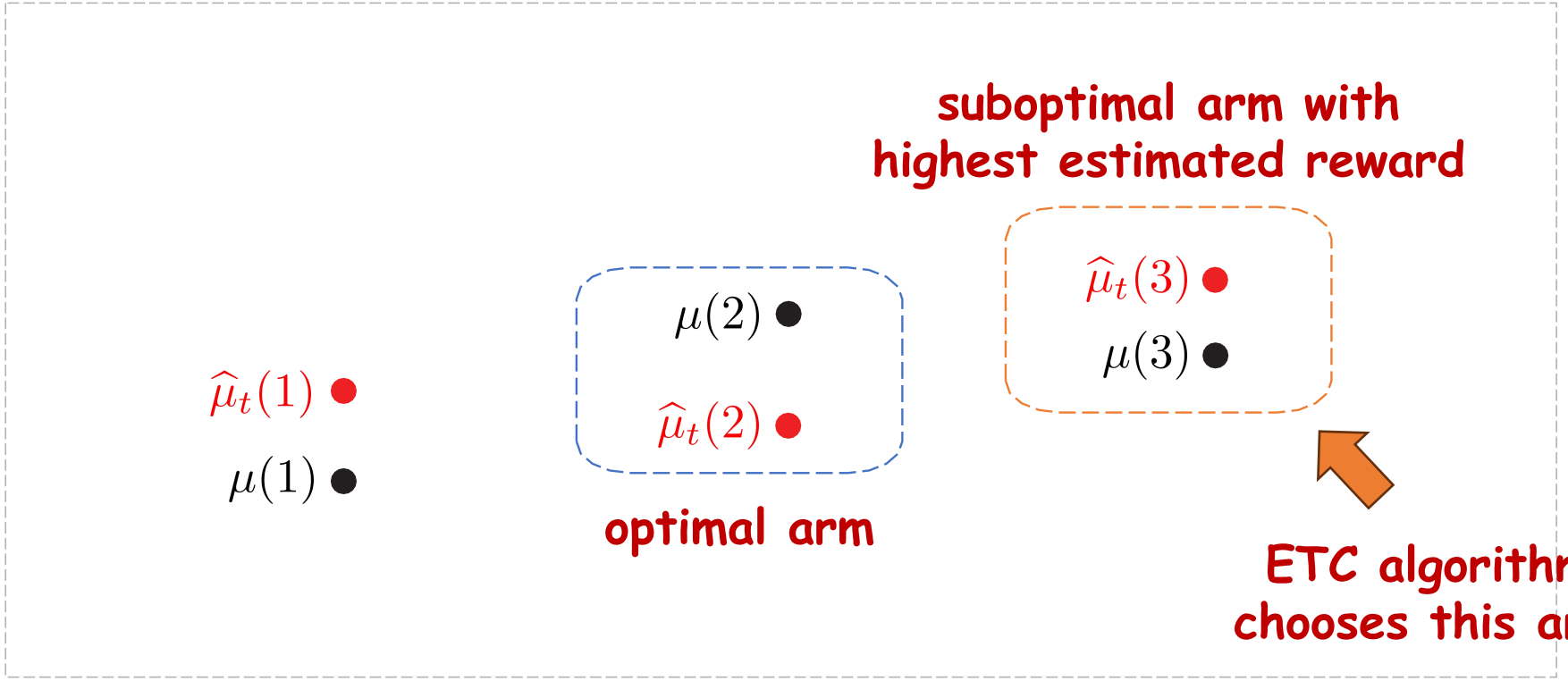
Choosing $C = 16$

Finally, we have

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

Upper Confidence Bound

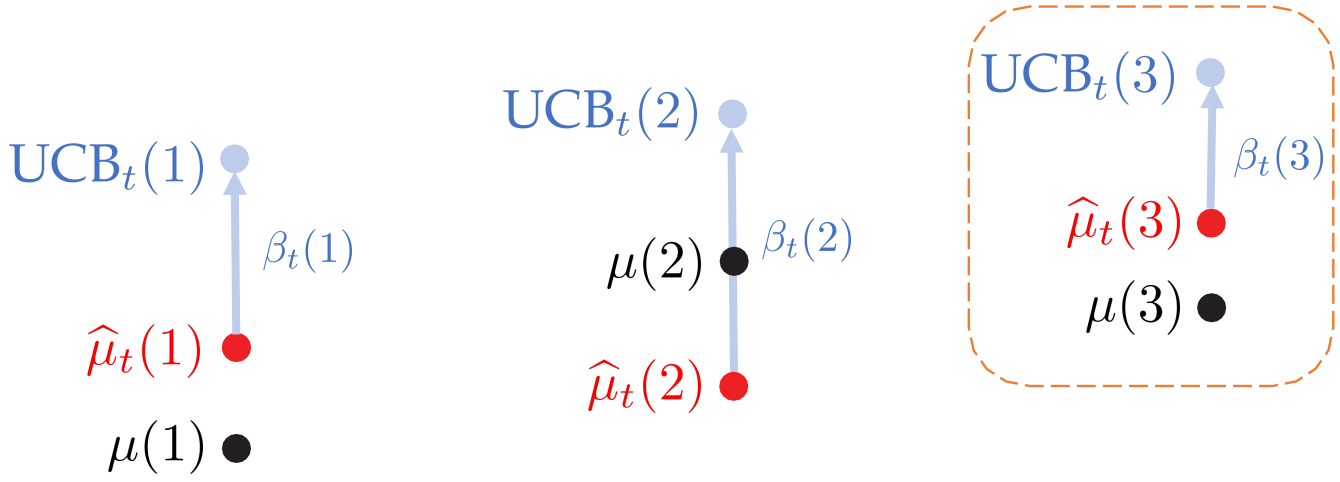
- UCB



Upper Confidence Bound

- UCB

With high probability $\mu(a) \le \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

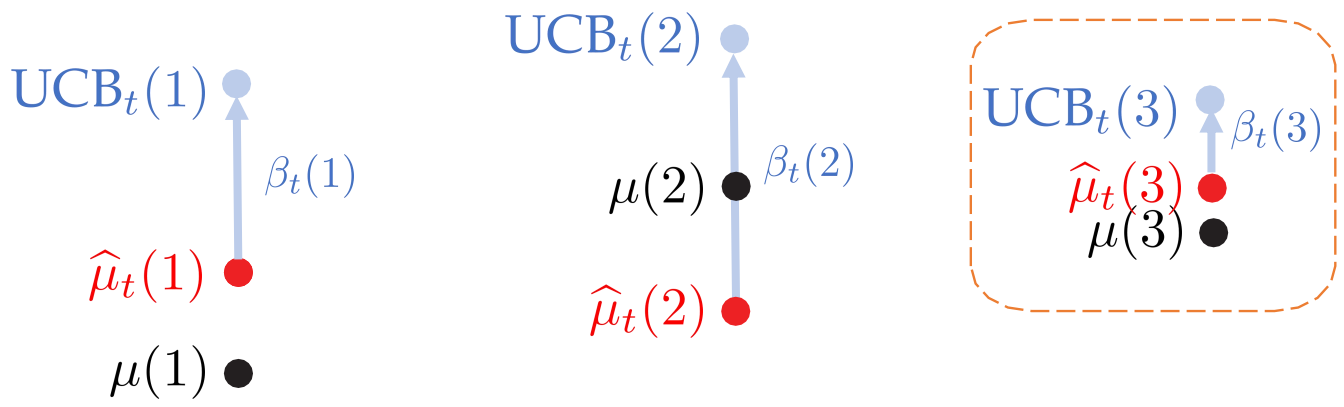


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \le \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

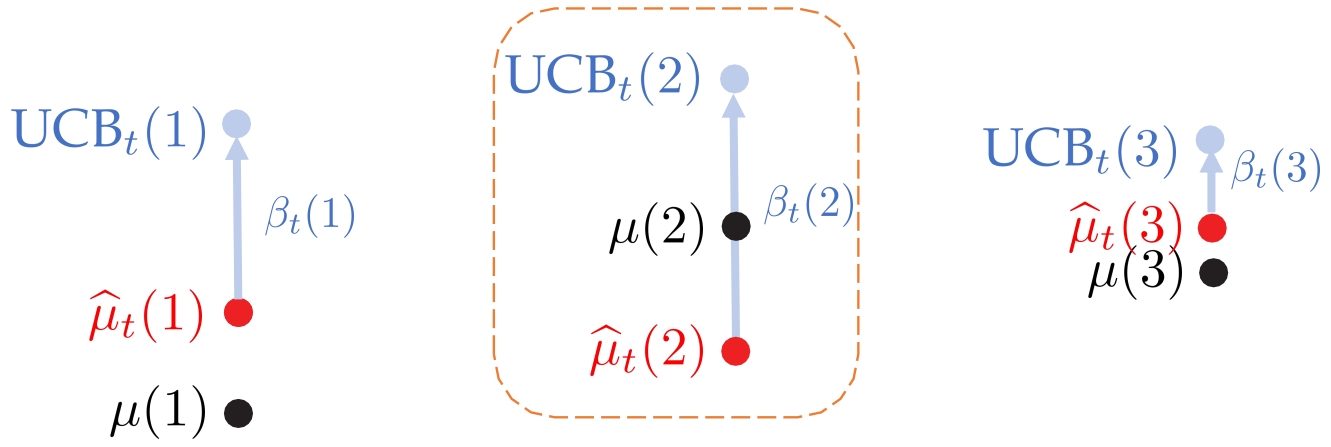


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \le \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

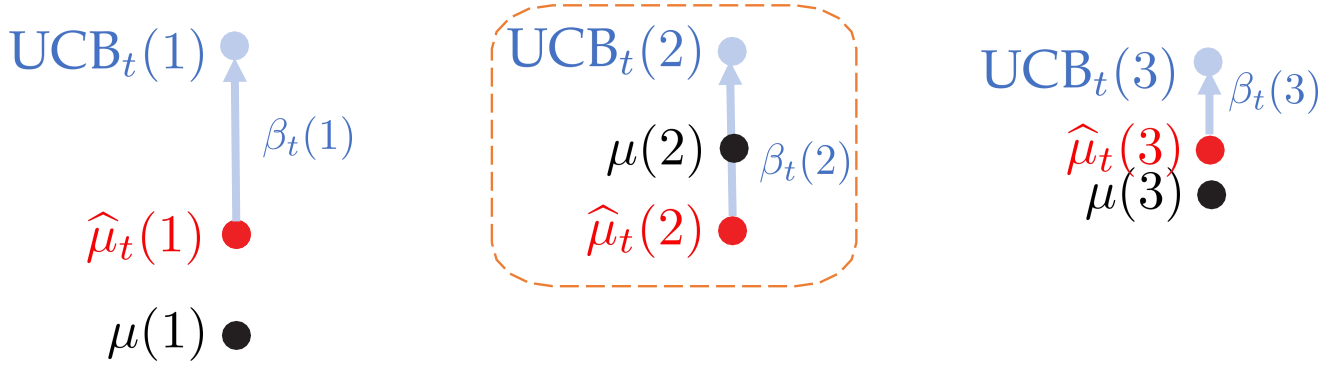


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \leq \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

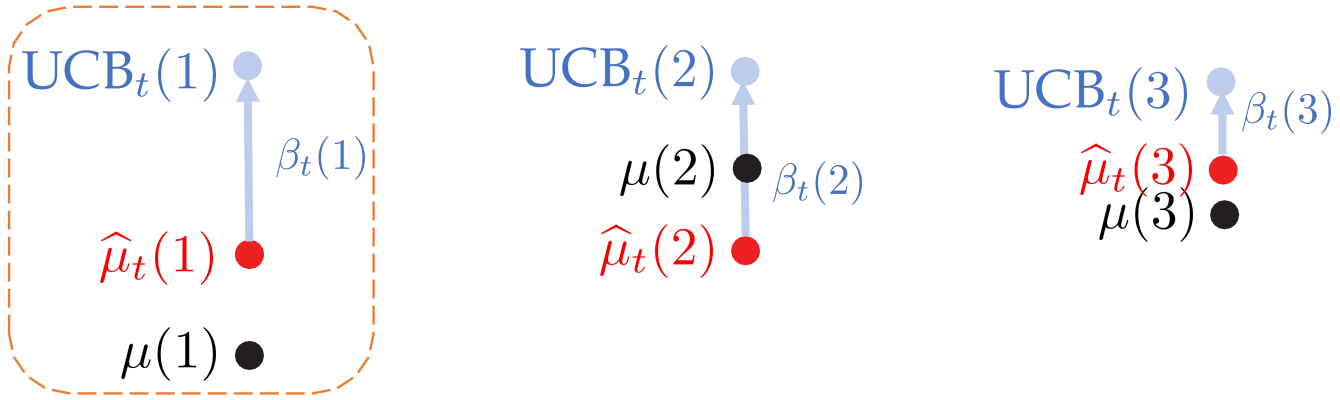


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \le \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

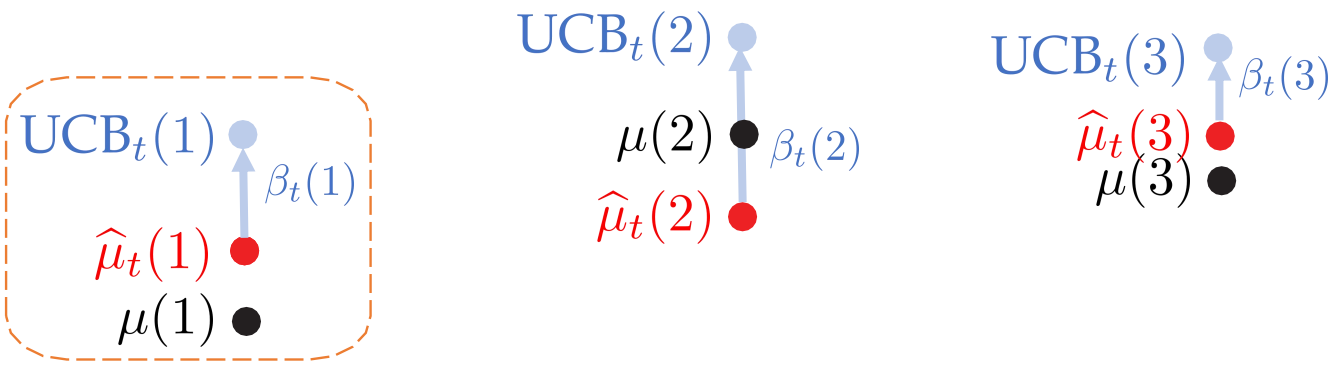


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \le \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$

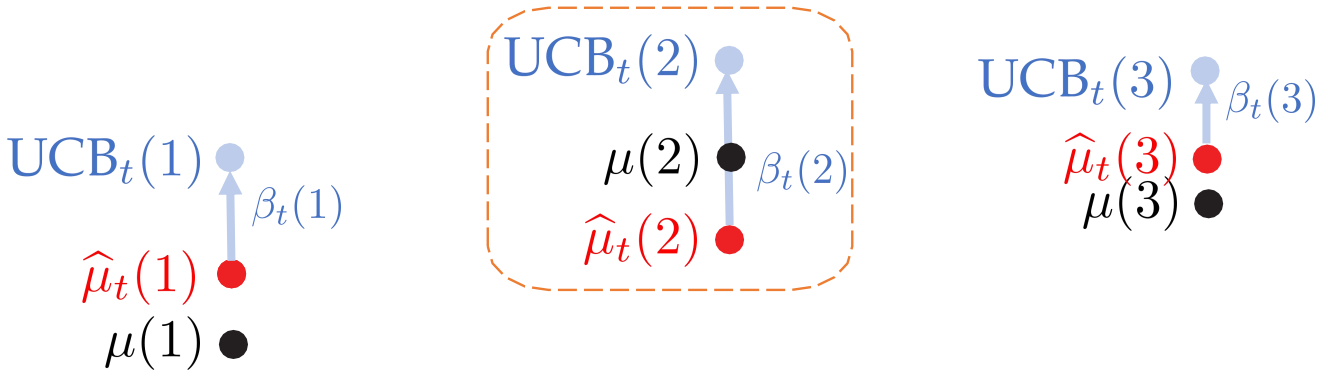


Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

Upper Confidence Bound

- UCB

With high probability $\mu(a) \le \mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$



Optimism in Face of Uncertainty: $a_t = \arg\max_{a \in [K]} \mathrm{UCB}_t(a)$

A large UCB means uncertainty or good arm.

Choosing the largest UCB means either exploring or exploiting.

Optimism in the Face of Uncertainty

- A general principle for dealing with uncertainty, or a strategy for balancing exploration and exploitation

$$\mathrm{UCB}_t(a) = \widehat{\mu}_t(a) + \beta_t(a)$$

Decision-Making Under Uncertainty: ***optimism drives exploration***, encouraging to try new things or take controllable risks, which can lead to better long-term outcomes

UCB Algorithm: Formulation

UCB Algorithm (known as UCB1)

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

(1) Choose arm $a_t = \arg\max_{a \in [K]} \mathbf{UCB}_{t-1}(a)$

(2) Observe reward r_t and update the estimation $\widehat{\mu}_t$

(3) Update upper confidence bounds $\mathbf{UCB}_t(a)$ by new estimation

- Estimation: empirical average

$$\widehat{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{s=1}^{t} \mathbf{1}\{a_s = a\} r_s(a), \quad \text{where } n_t(a) \text{ is the pulled times of arm } a$$

- UCB construction: Hoeffding's inequality

Construct UCB

Lemma 2 (Estimation error). *With probability at least* $1 - 2K/T$*, we have*

$$\forall a \in [K], t \in [T], |\mu(a) - \widehat{\mu}_t(a)| \leq \sqrt{\frac{\log T}{n_t(a)}}.$$

Therefore, it suggests $\mathbf{UCB}_t(a) \triangleq \widehat{\mu}_t(a) + \sqrt{\frac{\log T}{n_t(a)}}$*, ensuring* $\mu(a) \leq \mathbf{UCB}_t(a)$.

Proof. For each arm a, by Hoeffding inequality, we have

$$\Pr\left\{|\mu(a) - \widehat{\mu}_t(a)| \leq \sqrt{\frac{\log(1/\delta)}{2n_t(a)}}\right\} \geq 1 - 2\delta$$

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

Furthermore, by the union bound over all arms and all rounds and letting $\delta = 1/T^2$,

$$\Pr\left\{\forall a \in [K], t \in [T], |\mu(a) - \widehat{\mu}_t(a)| \leq \sqrt{\frac{\log T}{n_t(a)}}\right\} \geq 1 - 2\frac{K}{T} \qquad \square$$

UCB: Gap-Dependent Bound

Theorem 6 (Gap-dependent). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \le r_t(a) \le 1$, then with probability at least $1 - 2K/T$, UCB satisfies*

$$\bar{R}_T \le \sum_{a:\Delta_a>0} \frac{4\log T}{\Delta_a} + \Delta_a = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right).$$

Proof. With probability at least $1 - 2K/T$

$$\begin{aligned}
\Delta_{a_t} = \mu(a^*) - \mu(a_t) &\le \mathbf{UCB}_{t-1}(a^*) - \mu(a_t) && \forall a \in [K], \mu(a) \le \mathbf{UCB}_t(a) \\
&\le \mathbf{UCB}_{t-1}(a_t) - \mu(a_t) && a_t = \arg\max_{a\in[K]} \mathbf{UCB}_{t-1}(a) \\
&\le 2\sqrt{\frac{\log T}{n_{t-1}(a_t)}} && |\mu(a) - \widehat{\mu}_t(a)| \le \sqrt{\tfrac{\log(1/\delta)}{n_t(a)}} \\
& && \mathbf{UCB}_t(a) \triangleq \widehat{\mu}_t(a) + \sqrt{\tfrac{\log T}{n_t(a)}}
\end{aligned}$$

Proof of UCB Regret Bound

Proof. $\Delta_{a_t} \le 2\sqrt{\dfrac{\log T}{n_{t-1}(a_t)}}$

Let t be the last time a is selected, then with probability at least $1 - 2K/T$,

$$\Delta_a \le 2\sqrt{\frac{\log T}{n_{t-1}(a)}} = 2\sqrt{\frac{\log T}{n_T(a) - 1}}$$

$$\Rightarrow\ n_T(a) \le 4\frac{\log T}{\Delta_a^2} + 1$$

$$\Rightarrow\ \bar{R}_T = \sum_{a\in[K]} \Delta_a n_T(a) \le \sum_{a:\Delta_a>0} \Delta_a \left(4\frac{\log T}{\Delta_a^2} + 1\right) = \sum_{a:\Delta_a>0} 4\frac{\log T}{\Delta_a} + \Delta_a.$$

□

UCB: Gap-Dependent Bound

Theorem 6 (Gap-dependent). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \le r_t(a) \le 1$, then with probability at least $1 - 2K/T$, UCB satisfies*

$$\bar{R}_T \le \sum_{a:\Delta_a>0} \frac{4\log T}{\Delta_a} + \Delta_a = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right).$$

- Smaller the Δ_a, larger the regret. Its harder to distinguish the optimal arm from the suboptimal one.
- However, tiny Δ_a should not lead to larger regret. Always pick arm a should just lead to $\bar{R}_T = \Delta_a T$.

$$\Rightarrow \quad \bar{R}_T \le \min\left\{\max_{a\in[K]} \Delta_a T, \sum_{a:\Delta_a>0} \frac{4\log T}{\Delta_a} + \Delta_a\right\}$$

distribution-dependent
also called gap/instance-dependent

Gap-dependent Upper and Lower Bounds

Theorem 6 (Gap-dependent). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \le r_t(a) \le 1$, then with probability at least $1 - 2K/T$, UCB satisfies*

$$\bar{R}_T \le \sum_{a:\Delta_a>0} \frac{4\log T}{\Delta_a} + \Delta_a = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right).$$

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*

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

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

UCB: Gap-Independent Bound

Theorem 7 (Gap-independent). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \le r_t(a) \le 1$, then UCB satisfies with probability at least $1 - 2K/T$,*

$$\bar{R}_T \le 2\sqrt{TK \log T} + \sum_{a \in [K]} \Delta_a = \mathcal{O}\left(\sqrt{TK \log T}\right).$$

Proof.

$$\bar{R}_T = \sum_{a \in [K]} \Delta_a n_T(a) = \sum_{a : \Delta_a < \Delta} \Delta_a n_T(a) + \sum_{a : \Delta_a \ge \Delta} \Delta_a n_T(a)$$

$$\le T\Delta + \sum_{a : \Delta_a \ge \Delta} \Delta_a \left(4 \frac{\log T}{\Delta_a^2} + 1\right) \le T\Delta + 4\frac{K \log T}{\Delta} + \sum_{a \in [K]} \Delta_a \qquad n_T(a) \le 4\tfrac{\log T}{\Delta_a^2} + 1$$

$$\le 2\sqrt{TK \log T} + \sum_{a \in [K]} \Delta_a \qquad \text{Choosing } \Delta = 2\sqrt{K(\log T)/T}$$

□

Gap-Independent Upper and Lower Bounds

Theorem 7 (Gap-independent). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \le r_t(a) \le 1$, then UCB satisfies with probability at least $1 - 2K/T$,*

$$\bar{R}_T \le 2\sqrt{TK \log T} + \sum_{a \in [K]} \Delta_a = \mathcal{O}\left(\sqrt{TK \log T}\right).$$

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*

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

Thompson Sampling

- Suppose for each arm $a \in [K]$, $r_t(a) \in \{0, 1\}$ and $r_t(a) \sim \mathrm{Ber}(\mu_a)$ (μ_a is unknown).

Thompson Sampling

Initialization: Choose fake prior $\mathrm{Beta}(\alpha_{a,1}, \beta_{a,1})$ for $a \in [K]$ following some strategy.

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

(1) For each arm a, sample $\tilde{\mu}_t(a) \sim \mathrm{Beta}(\alpha_{a,t}, \beta_{a,t})$

(2) Choose $a_t = \arg\max_{a \in [K]} \tilde{\mu}_t(a)$ and observe reward $r_t \in \{0, 1\}$

(3) Update the posterior of arm a_t by

$$(\alpha_{a_t,t+1}, \beta_{a_t,t+1}) = \begin{cases} (\alpha_{a_t,t} + 1,\ \beta_{a_t,t}), & \text{if } r_t = 1, \\ (\alpha_{a_t,t},\ \beta_{a_t,t} + 1), & \text{if } r_t = 0. \end{cases}$$

Thompson Sampling

Thompson Sampling

Initialization: Choose fake prior $\mathrm{Beta}(\alpha_{a,1}, \beta_{a,1})$ for $a \in [K]$ following some strategy.

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

(1) For each arm a, sample $\tilde{\mu}_t(a) \sim \mathrm{Beta}(\alpha_{a,t}, \beta_{a,t})$

(2) Choose $a_t = \arg\max_{a \in [K]} \tilde{\mu}_t(a)$ and observe reward $r_t \in \{0, 1\}$

(3) Update the posterior of arm a_t by

$$(\alpha_{a_t,t+1}, \beta_{a_t,t+1}) = \begin{cases} (\alpha_{a_t,t} + 1,\ \beta_{a_t,t}), & \text{if } r_t = 1, \\ (\alpha_{a_t,t},\ \beta_{a_t,t} + 1), & \text{if } r_t = 0. \end{cases}$$

$\mathrm{Beta}(\alpha, \beta)$

- Mean: $\mu = \frac{\alpha}{\alpha+\beta}$ Exploitation
- Var: $\sigma \approx \frac{\mu(1-\mu)}{\alpha+\beta}$ Exploration

A large $\tilde{\mu}_t(a_t)$ means large mean (good arm) or large variance (uncertainty).

Choosing the largest $\tilde{\mu}_t(a_t)$ means either exploring or exploiting.

TS Regret Bound

- Gap-dependent bound

Theorem 8. *For every $\epsilon > 0$ there exists a problem-dependent constant $C(\epsilon, \mu_1, \ldots, \mu_K)$ such that the regret of Thompson Sampling satisfies:*

$$\mathbb{E}\left[\bar{R}_T\right] \leq (1+\epsilon) \sum_{a \in A: \mu_a \neq \mu^*} \frac{\Delta_a (\log(T) + \log\log(T))}{KL(\mu_a, \mu^*)} + C(\epsilon, \mu_1, \ldots, \mu_K).$$

By the Pinsker's inequality $2\mathrm{KL}(\mu_a, \mu^*) > \Delta_a^2$, we have the asymptotically optimal bound

$$\mathbb{E}\left[\bar{R}_T\right] \leq 2(1+\epsilon) \sum_{a \in A: \mu_a \neq \mu^*} \frac{\log(T) + \log\log(T)}{\Delta_a} + C(\epsilon, \mu_1, \ldots, \mu_K).$$

Summary of All those methods

- ETC $$\mathbb{E}[\bar{R}_T^{\text{ETC}}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$$
- ε-greedy $$\mathbb{E}[\bar{R}_T^{\varepsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$$
- UCB1 $$\bar{R}_T^{\text{UCB}} = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right), \text{with probability } \geq 1 - \frac{2K}{T}$$
- TS $$\mathbb{E}[\bar{R}_T^{\text{TS}}] = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a} + C\right)$$

(C is a problem-dependent constant)

Part 3. Comparison

- ETC vs ε-greedy
- ε-greedy vs UCB
- UCB vs Thompson Sampling

Summary of All those methods

- ETC $\mathbb{E}[\bar{R}_T^{\text{ETC}}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$
- ε-greedy $\mathbb{E}[\bar{R}_T^{\varepsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$
- UCB1 $\bar{R}_T^{\text{UCB}} = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right)$, with probability $\geq 1 - \frac{2K}{T}$
- TS $\mathbb{E}[\bar{R}_T^{\text{TS}}] = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a} + C\right)$

(C is a problem-dependent constant)

ETC vs ε-greedy

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

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

• 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 of All those methods

- ETC $\mathbb{E}[\bar{R}_T^{\text{ETC}}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$
- ε-greedy $\mathbb{E}[\bar{R}_T^{\varepsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$
- UCB1 $\bar{R}_T^{\text{UCB}} = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right)$, with probability $\geq 1 - \frac{2K}{T}$
- TS $\mathbb{E}[\bar{R}_T^{\text{TS}}] = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a} + C\right)$

(C is a problem-dependent constant)

ε-greedy vs UCB

- ε-greedy $\quad \mathbb{E}[\bar{R}_T^{\epsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$
- UCB1 $\quad \bar{R}_T^{\mathrm{UCB}} = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right)$, with probability $\geq 1 - \frac{2K}{T}$

Example 1: All gaps similar

When $\Delta_a = \Delta$ for all $a \in [K]$, we have

ε-greedy	$\mathbb{E}[\bar{R}_T^{\varepsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta^2} \cdot K\Delta\right) = \mathcal{O}\left(K\frac{\log T}{\Delta}\right),$
UCB	$\bar{R}_T^{\mathrm{UCB}} = \mathcal{O}\left(\log T \cdot K\frac{1}{\Delta}\right) = \mathcal{O}\left(K\frac{\log T}{\Delta}\right).$

the same order

ε-greedy vs UCB

- ε-greedy $\quad \mathbb{E}[\bar{R}_T^{\epsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$
- UCB1 $\quad \bar{R}_T^{\mathrm{UCB}} = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right)$, with probability $\geq 1 - \frac{2K}{T}$

Example 2: One nearly-optimal arm, many clearly bad arms

When $\mu_1 = 0.99$, $\mu_2 = 0.98$, and $\mu_a = 0$ for $a = 3, \ldots, K$, we have $\Delta_2 = 0.01$ and $\Delta_a \approx 1$ for $a = 3, \ldots, K$.

ε-greedy	$\mathbb{E}[\bar{R}_T^{\varepsilon}] \approx \frac{\log T}{(0.01)^2} \cdot K = 10^4 K \log T,$
UCB1	$\bar{R}_T^{\mathrm{UCB}} \approx (100 + K) \log T.$

UCB incurs significantly lower regret

ε-greedy vs UCB

- ε-greedy $\quad \mathbb{E}[\bar{R}_T^{\epsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$
- UCB1 $\quad \bar{R}_T^{\text{UCB}} = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right)$, with probability $\geq 1 - \frac{2K}{T}$

❑ Prior information dependence

- ε-greedy requires the knowledge of $\Delta_{\min}$ to achieve the desired regret
- UCB1 doesn't need any prior knowledge of gaps.

❑ Exploration mechanism

- ε-greedy explores all the arms uniformly
- UCB1 drives exploration through a confidence bonus, allocating more trials to arms with greater uncertainty.

Summary of All those methods

- ETC $$\mathbb{E}[\bar{R}_T^{\text{ETC}}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$$
- ε-greedy $$\mathbb{E}[\bar{R}_T^{\varepsilon}] = \mathcal{O}\left(\frac{\log T}{\Delta_{\min}^2} \sum_{a:\Delta_a>0} \Delta_a\right)$$
- UCB1 $$\bar{R}_T^{\text{UCB}} = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right), \text{with probability } \geq 1 - \frac{2K}{T}$$
- TS $$\mathbb{E}[\bar{R}_T^{\text{TS}}] = \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a} + C\right)$$

(C is a problem-dependent constant)

UCB vs Thompson Sampling

	UCB	**Thompson Sampling**
Decision Style	**Deterministic**	**Probabilistic**
Core Principle	**Frequentist**	**Bayesian**
Prior Knowledge	Subgaussian parameter σ	Prior distribution type
Exploration	Empirical mean $\hat{\mu}_t$	Mean of distribution μ
Exploitation	Uncertainty $\beta_t(a)$	Var of distribution σ
Guarantee	$\mathcal{O}\left(\log T \sum_a \frac{1}{\Delta_a}\right)$ with high probability	$\mathcal{O}\left(\log T \sum_a \frac{1}{\Delta_a} + C\right)$ asymptotically

Table 1: Comparison between UCB and Thompson Sampling

(C is a problem-dependent constant)

Part 4. Extension

- Best of Both Words
- Extensions of UCB1
- Best Arm Identification (BAI)
- UCB in Online RL

Advanced Topic: Best of Both Worlds

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

- Best of stochastic MAB: $\bar{R}_T = \max_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t)\right] \le \mathcal{O}\left(\sum_{a:\Delta_a>0} \frac{\log T}{\Delta_a}\right)$

Can one algorithm achieve the ***best of both worlds***, without knowing whether the world is stochastic or adversarial?

- UCB: can get almost linear regret under the adversarial setting.
- Exp3: can't have adaptive regret bound in the stochastic case.

⇨ Surprisingly, using OMD with ***Tsallis entropy*** regularizer.

Reference: Julian Zimmert, Yevgeny Seldin. An Optimal Algorithm for Stochastic and Adversarial Bandits. AISTATS 2019.

Advanced Topic: Extension of UCB1

- Recall that UCB1 algorithm is ***nearly*** minimax optimal (up to some logarithmic factor).

Theorem 7 (Gap-independent). *Suppose that for all $t \in [T]$ and $a \in [K]$, $0 \leq r_t(a) \leq 1$, then UCB1 with $\delta = 1/T^2$ satisfies with high probability,*

$$\bar{R}_T \leq 2\sqrt{TK \log T} + \sum_{a \in [K]} \Delta_a = \mathcal{O}\left(\sqrt{TK \log T}\right).$$

- How to achieve ***minimax optimality***?
  - Carefully tuning the (adaptive) confidence level $\delta_t = 1/f(t)$ with $f(t) = 1 + t \log^2 t$ achieves ***asymptotic*** optimality (see Chapter 8, *Bandit Algorithm* book).

Advanced Topic: Extension of UCB1

- Recall that UCB1 algorithm is ***nearly*** minimax optimal (up to some logarithmic factor).
- How to achieve ***minimax optimality***?
  - Carefully tuning the confidence level achieves ***asymptotic*** optimality.
  - MOSS algorithm [Audibert and Bubeck, 2009] uses bonus term chosen based on T and K, as well as the number of plays of the individual arms. This achieves ***minimax optimality*** (See Chapter 9, *Bandit Algorithm* book).

Arm selection in UCB1: $a_t = \arg\max_a \hat{\mu}_{t-1}(a) + \sqrt{\frac{2\log(1/\delta)}{n_{t-1}(a)}}$

Arm selection in MOSS: $a_t = \arg\max_a \hat{\mu}_{t-1}(a) + \sqrt{\frac{4}{n_{t-1}(a)} \log^+ \frac{T}{K n_{t-1}(a)}}$

Advanced Topic: Extension of UCB1

- Recall that UCB1 algorithm is ***nearly*** minimax optimal (up to some logarithmic factor).
- How to achieve ***minimax optimality***?
  - Carefully tuning the confidence level achieves ***asymptotic*** optimality.
  - MOSS achieves ***minimax optimality*** by bonus term chosen based on T and K, as well as the number of plays of the individual arms.
- UCB1 only suites for sub-Gaussian noise (to use Hoeffding's inequality). How to deal with ***Bernoulli noise***?
  - Needs different concentration (for sums of Bernoulli r.v.).
  - KL-UCB algorithm [Garivier and Cappe, 2011; Maillard et al., 2011] solves that for ***Bernoulli bandits*** (see Chapter 10, *Bandit Algorithm* book).

Advanced Topic: Best Arm Identification

- Previously we mainly focus on *regret minimization*, which seeks for exploration-exploitation trade-off.
- Another topic in stochastic bandit: **best arm identification** (or an ε-optimal arm)

Essentially, a *pure exploration* problem.

- ❑ Setting 1: *fixed-budget* → minimize simple regret

$$\mathrm{REG}_T^{\mathrm{simple}} = \sum_{a \in [K]} \Delta_a \Pr(a_{T+1} = a)$$

- ❑ Setting 2: *fixed confidence* → a δ-PAC sample complexity guarantee

$$\Pr(\text{return suboptimal arm}) \leq \delta$$

Advanced Topic: UCB in Online RL

- Recall that reinforcement learning as multi-step bandits.
- ***Exploration-Exploitation dilemma*** is also a central challenge in reinforcement learning.
  - Typically, we need to estimate the state transition, and use dynamic programming to solve for the ***value function*** (cumulative reward).
  - Bellman equation:

$$Q(s,a) = r(s,a) + \sum_{s'} P(s' \mid s,a) \max_{a'} Q(s',a') \; +b(s,a)$$

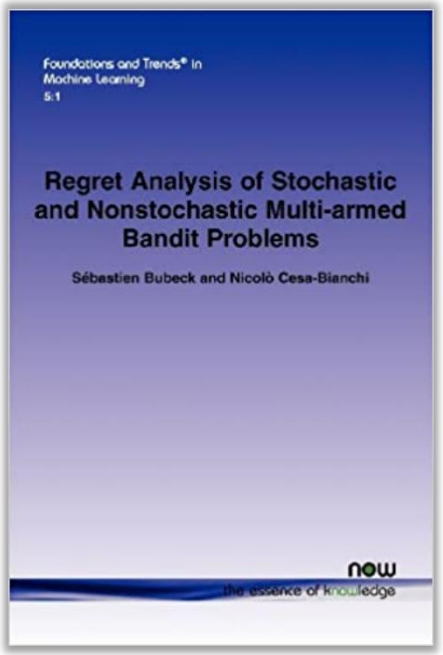
  - UCB-VI algorithm [Azar et al., 2017]

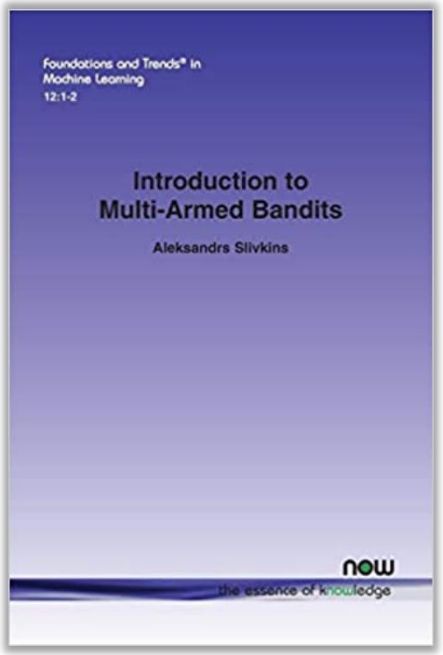
Add bonus term to encourage exploration

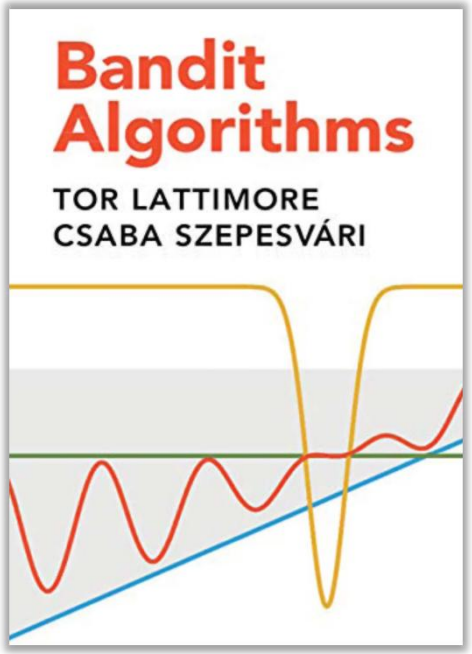
- UCB strategy also used in many RL algorithms, e.g., ***Monte Carlo Tree Search (MCTS)*** in AlphaGo

Many more results

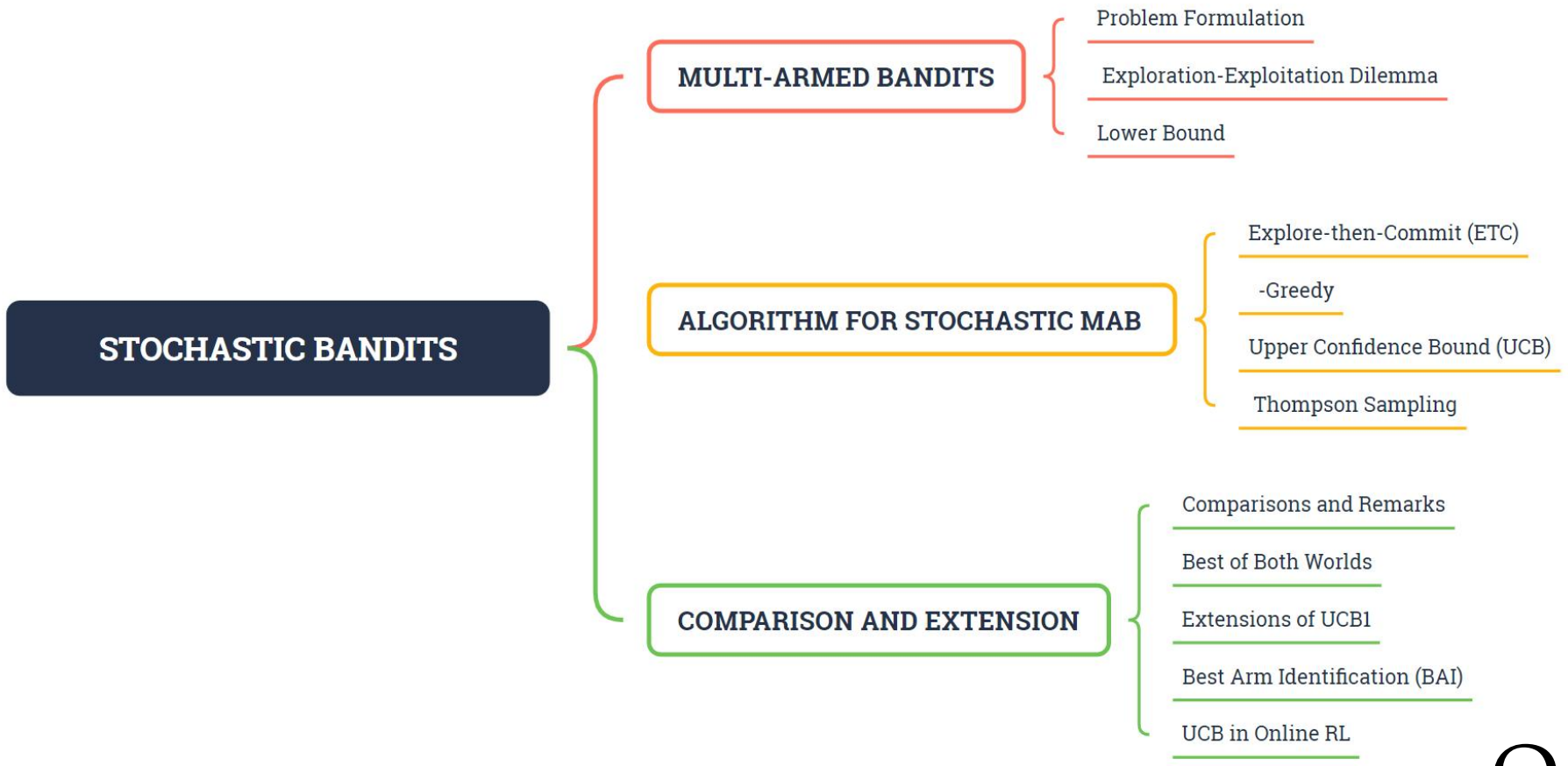
- Techniques developed in bandit problems have been applied in many areas, including machine learning, reinforcement learning, statistics, operational research, and information theory [Bubeck and Cesa-Bianchi, 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020].







Summary



Q & A

Thanks!