

# CMPT 419/983: Theoretical Foundations of Reinforcement Learning

## Lecture 1

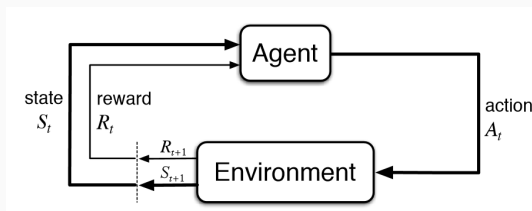
---

Sharan Vaswani

September 8, 2023

- Supervised machine learning involves learning from a fixed, static dataset.
- Once a dataset is collected, supervised learning does not typically reason about how the data was acquired nor does it involve further interactions with the world.
- Applications in computational advertising, robotics, clinical trials involve collecting data in an online fashion, and reasoning about the decisions used to gather it.
- **Sequential decision-making** under uncertainty focuses on problems that involve interacting with the world, collecting data and reasoning about it, all with incomplete information about the world.

# Introduction



- A typical problem in sequential decision-making involves an *agent* (e.g: marketer, robot, investor) sequentially interacting with the *environment* (e.g: online advertising platform, Mars terrain, stock market).
- An interaction involves the agent choosing an *action* and receiving feedback.
- For example, the feedback can be in the form of a *reward*) designed to measure the agent's performance in achieving its goal.
- One possible objective: Find a sequence of actions (referred to as a *policy*) that maximizes the *cumulative reward* across the sequence of interactions.

# Motivating Applications

- Games. E.g: Go and Atari by DeepMind.
- Conversational agents. Eg: ChatGPT by OpenAI.
- Chip design by Google AI
- Cooling the interior of large commercial buildings by DeepMind
- Recommendation system by Microsoft
- Healthcare and Clinical Trials.
- Autonomous Navigation of Stratospheric Balloons by Google AI.
- For more applications, refer to Glen Berseth's and Csaba Szepesvari's lists.

## Motivation

- Typical algorithms used in practice are often (a) brittle (their performance is sensitive to hyper-parameters) (b) inefficient (require a large number of interactions to learn to make good decisions) and (c) do not have theoretical guarantees on their performance and can fail on simple problems.
- Numerous fundamental theoretical questions remain unanswered and there is a large discrepancy between the theory and practice.

## Objective:

- Understand the foundational concepts in bandits and reinforcement learning (RL) from a theoretical perspective.
- Use this knowledge to inform the design of theoretically-principled, statistically and computationally efficient algorithms.

## Topics:

- **Bandits:** Multi-armed/Contextual Bandit framework, Algorithms for regret minimization
- **Markov Decision Processes:** Structural properties, (Approximate) Value/Policy Iteration, Linear Programming, Temporal Difference Learning, Policy Gradients
- **Online & Batch RL:** Q Learning, LSVI-UCB, Learning with access to a simulator

**What we won't cover:** Continuous state-action spaces, Constrained MDPs, Multi-objective RL

- **Instructor:** Sharan Vaswani. [sharan\_vaswani@sfu.ca]
- **Teaching Assistant:** Michael Lu. [michael\_lu\_3@sfu.ca]
- **Course Webpage:** [https://vaswanis.github.io/419\\_983-F23.html](https://vaswanis.github.io/419_983-F23.html)
- **Piazza:** <https://piazza.com/sfu.ca/fall2023/cmpt419983/home>
- **Prerequisites:** Probability, Linear Algebra, Calculus, Undergraduate Machine Learning

# Course Logistics – Grading

## Assignments $[4 \times 12\% = 48\%]$

- Assignments to be submitted online (via Coursys), typed up in Latex with accompanying code submitted as a zip file.
- Each assignment will be due in 3 weeks (at 11.59 pm PST).

## Final Project [50%]

- Aim is to give you a taste of research in RL Theory.
- Projects to be done in groups of 3-4. Will maintain a list of possible topics. Can choose from the list or propose your own topic. (more details will be on Piazza)
- Project Proposal [10%] – Discussion (before 20 October) + Report (due 20 October)
- Project Milestone [5%] – Update (before 20 November)
- Project Presentation [10%] – (*tentatively* 1, 4 December)
- Project Report [25%] (15 December)

## Participation [2%] In class (during lectures, project presentations), on Piazza

# Stochastic Multi-armed Bandits



# Motivating Application: Clinical Trials

- Do not have complete information about the effectiveness or side-effects of the drugs.
- **Aim:** Maximize the number of patients healed.
- Each drug choice is mapped to an *arm* and the drug's effectiveness is mapped to the arm's *mean reward*.
- Administering a drug is an *action* that is equivalent to *pulling* the corresponding arm.
- Each time an arm is pulled, we get a *noisy* reward that models a patient's reaction to the drug.
- The trial goes on for  $T$  rounds.
- Other motivating applications: Recommendation systems, computational advertising.



# Problem Formulation

**Input:**  $K$  arms (possible actions) and their corresponding unknown reward distributions  $\{\nu_a\}_{a=1}^K$ . Define  $\mu_a := \mathbb{E}_{r \sim \nu_a}[r]$  as the expected reward obtained by choosing action  $a$ .

---

**Algorithm** Generic Bandit Framework ( $K$  arms,  $T$  rounds)

---

- 1: **for**  $t = 1 \rightarrow T$  **do**
  - 2:   **SELECT:** Use a bandit algorithm to decide which arm(s) to pull.
  - 3:   **OBSERVE:** Pull the selected arm  $a_t \in [K]$  and observe reward  $R_t \sim \nu_{a_t}$ .
  - 4:   **UPDATE:** Update the estimated reward for arm  $a_t$ .
  - 5: **end for**
- 

**Bandit Feedback:** Can only observe the noisy reward  $R_t$  from the pulled arm  $a_t$ .

**Objective:** Maximize  $\mathbb{E}[\sum_{t=1}^T R_t]$  where the expectation is over both the randomness of the algorithm (if any) and the distribution of rewards.

Bandit problems are a special case of RL, and capture a lot of the underlying intricacy.

# Problem Formulation

- Define  $a^* := \arg \max_{a \in [K]} \mu_a$  as the *best or optimal arm* in hindsight, and  $\mu_* := \max_a \mu_a$ .
- Maximizing cumulative rewards  $\implies$  Select  $a^*$  as much as possible  $\implies$  Minimize the *cumulative regret*.
- **Cumulative Regret:**  $\text{Regret}(T) := \sum_{t=1}^T [\mu^* - \mathbb{E}[R_t]] = T\mu^* - \sum_{t=1}^T \mathbb{E}[R_t]$ .
- Since the optimal arm is unknown, the algorithm needs to *explore* to narrow down on the best arm. If we can identify the best arm, the algorithm should *exploit* and always choose it.
- Need to find a *policy* that trades off exploration and exploitation to minimize  $\text{Regret}(T)$ .
- Ideally, want  $\text{Regret}(T) = o(T)$  i.e. the regret grows sub-linearly with  $T$ , meaning that  $\lim_{T \rightarrow \infty} \frac{\text{Regret}(T)}{T} = 0$ .

# Regret Decomposition

**Claim:** If  $\Delta_a := \mu^* - \mu_a$  and  $N_a(T)$  is the number of times arm  $a$  was chosen *until* round  $T$ , then,

$$\text{Regret}(T) = \sum_{a=1}^K \Delta_a \mathbb{E}[N_a(T)].$$

*Proof:*

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

(Taking the expectation w.r.t to the reward distribution)

$$= \sum_{a=1}^K [\mu^* - \mu_a] \mathbb{E} \left[ \sum_{t=1}^T \mathcal{I} \{a_t = a\} \right] = \sum_{a=1}^K [\mu^* - \mu_a] \mathbb{E}[N_a(T)]$$

$$\implies \text{Regret}(T) = \sum_{a=1}^K \Delta_a \mathbb{E}[N_a(T)].$$

- Hence, to minimize the regret, an algorithm should (i) not pull arms with  $\Delta_a > 0$  too often (**exploit**) which requires (ii) estimating the values of  $\Delta_a$  to sufficient accuracy (**explore**).

# Naive Strategy

---

**Algorithm** Naive Strategy

---

- 1: **for**  $t = 1 \rightarrow K$  **do**
  - 2:   Select arm  $a_t = t$  and observe reward  $R_t$
  - 3: **end for**
  - 4: Calculate empirical mean reward for arm  $a \in [K]$  as  $\hat{\mu}_a(K) := \frac{\sum_{t=1}^K R_t \mathcal{I}\{a_t=a\}}{N_a(K)}$
  - 5: **for**  $t = K + 1 \rightarrow T$  **do**
  - 6:   Pull arm  $\hat{a} := \arg \max_{a \in [K]} \hat{\mu}_a(K)$  (*choose lower-indexed arm if there is a tie*).
  - 7: **end for**
- 

Q: Will this naive strategy result in sublinear regret?

# Explore-Then-Commit (ETC)

---

**Algorithm** Explore-Then-Commit

---

```
1: Input:  $m \in \{1, \dots, \lfloor \frac{T}{K} \rfloor\}$ .  
2: for  $t = 1 \rightarrow mK$  do  
3:   Select arm  $a_t = t \bmod K + 1$  and observe reward  $R_t$    (Explore)  
4: end for  
5: Calculate empirical mean reward for arm  $a \in [K]$  as  $\hat{\mu}_a(mK) := \frac{\sum_{t=1}^{mK} R_t \mathbb{I}\{a_t=a\}}{N_a(mK)}$   
6: for  $t = mK + 1 \rightarrow T$  do  
7:   Pull arm  $\hat{a} := \arg \max_{a \in [K]} \hat{\mu}_a(mK)$    (Commit)  
8: end for
```

---

**Q:** Will ETC result in sublinear regret?

Yes! under suitable distributional assumptions on the rewards.

In particular, if  $r \sim \nu_a$ , we will assume that  $r - \mu_a$  are sub-Gaussian random variables, then we will prove that ETC results in sub-linear regret. For this, we need to first recap some concentration (tail) inequalities from undergraduate probability.

## Digression – Concentration inequalities

**Concentration inequalities** bound the probability that the r.v. takes a value much different from its mean.

*Example:* Consider a r.v.  $X$  that can take on only non-negative values and  $\mathbb{E}[X] = 99.99$ . Show that  $\Pr[X \geq 300] \leq \frac{1}{3}$ .

$$\begin{aligned} \text{Proof: } \mathbb{E}[X] &= \sum_{x \in \text{Range}(X)} x \Pr[X = x] = \sum_{x|x \geq 300} x \Pr[X = x] + \sum_{x|0 \leq x < 300} x \Pr[X = x] \\ &\geq \sum_{x|x \geq 300} (300) \Pr[X = x] + \sum_{x|0 \leq x < 300} x \Pr[X = x] \\ &= (300) \Pr[X \geq 300] + \sum_{x|0 \leq x < 300} x \Pr[X = x] \end{aligned}$$

If  $\Pr[X \geq 300] > \frac{1}{3}$ , then,  $\mathbb{E}[X] > (300) \frac{1}{3} + \sum_{x|0 \leq x < 300} x \Pr[X = x] > 100$  (since the second term is always non-negative). Hence, if  $\Pr[X \geq 300] > \frac{1}{3}$ ,  $\mathbb{E}[X] > 100$  which is a contradiction since  $\mathbb{E}[X] = 99.99$ .

## Digression – Markov's Theorem

Markov's theorem formalizes the intuition on the previous slide, and can be stated as follows.

**Markov's Theorem:** If  $X$  is a non-negative random variable, then for all  $x > 0$ ,

$$\Pr[X \geq x] \leq \frac{\mathbb{E}[X]}{x}.$$

*Proof:* Define  $\mathcal{I}\{X \geq x\}$  to be the indicator r.v. for the event  $[X \geq x]$ . Then for all values of  $X$ ,  $x\mathcal{I}\{X \geq x\} \leq X$ .

$$\begin{aligned}\mathbb{E}[x\mathcal{I}\{X \geq x\}] &\leq \mathbb{E}[X] \implies x\mathbb{E}[\mathcal{I}\{X \geq x\}] \leq \mathbb{E}[X] \implies x\Pr[X \geq x] \leq \mathbb{E}[X] \\ &\implies \Pr[X \geq x] \leq \frac{\mathbb{E}[X]}{x}. \quad \square\end{aligned}$$

Since the above theorem holds for all  $x > 0$ , we can set  $x = c\mathbb{E}[X]$  for  $c \geq 1$ . In this case,  $\Pr[X \geq c\mathbb{E}[X]] \leq \frac{1}{c}$ . Hence, the probability that  $X$  is “far” from the mean in terms of the multiplicative factor  $c$  is upper-bounded by  $\frac{1}{c}$ .



## Digression – Sub-Gaussian random variables

If a centered r.v.  $X$  (meaning that  $\mathbb{E}[X] = 0$ ) is  $\sigma$  sub-Gaussian, then for all  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

*Example 1:* If  $X \sim N(0, 1)$ , then its moment generating function  $\mathbb{E}[\exp(\lambda X)] = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$ , meaning that Gaussian r.v. are sub-Gaussian.

*Example 2:* If  $X \in [a, b]$  and  $\mathbb{E}[X] = 0$ , then  $X$  is  $(b - a)$  sub-Gaussian.

**Properties:** If  $X$  is centered and  $\sigma$  sub-Gaussian, then,

- (a)  $\mathbb{E}[X] = 0$ ,  $\text{Var}[X] \leq \sigma^2$
- (b) For a constant  $c \in \mathbb{R}$ ,  $cX$  is  $|c| \sigma$  sub-Gaussian.
- (c) If  $\{X_i\}_{i=1}^n$  are independent and  $\sigma_i$  sub-Gaussian respectively, then,  $\sum_{i=1}^n X_i$  is  $\sqrt{\sum_{i=1}^n \sigma_i^2}$  sub-Gaussian.

Need to prove some of these properties in Assignment 1!

## Digression – Concentration inequalities for sub-Gaussian r.v's

**Claim:** If  $X$  is centered and  $\sigma$  sub-Gaussian, then for any  $\epsilon \geq 0$ ,  $\Pr[X \geq \epsilon] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$ .

*Proof:* For some constant  $c > 0$  to be tuned later,

$$\begin{aligned}\Pr[X \geq \epsilon] &= \Pr[cX \geq c\epsilon] = \Pr[\exp(cX) \geq \exp(c\epsilon)] \\ &\leq \mathbb{E}[\exp(cX)] \exp(-c\epsilon) && \text{(Markov's inequality)} \\ &\leq \exp\left(\frac{c^2\sigma^2}{2} - c\epsilon\right) && \text{(Def. of sub-Gaussian r.v's)} \\ &= \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \quad \square && \text{(Setting } c = \epsilon/\sigma^2\text{)}\end{aligned}$$

Similarly,  $\Pr[X \leq -\epsilon] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$ . By the union bound,  $\Pr[|X| \geq \epsilon] \leq 2 \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$ .

Setting  $\delta = 2 \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \implies \epsilon = \sqrt{2\sigma^2 \log(2/\delta)}$ . Hence, w.p.  $1 - \delta$ ,  $X$  will take on values in the range  $\left[-\sqrt{2\sigma^2 \log(2/\delta)}, +\sqrt{2\sigma^2 \log(2/\delta)}\right]$ .

## Digression – Concentration inequalities for sub-Gaussian r.v's

**Claim:** Consider  $n$  i.i.d r.v's  $X_i$  such that  $\mathbb{E}[X_i] = \mu$ . If  $X_i - \mu$  are  $\sigma$  sub-Gaussian and  $\hat{\mu} := \frac{1}{n} \sum_{i=1}^n X_i$  is the empirical mean, then,  $\Pr[|\hat{\mu} - \mu| \geq \epsilon] \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$ .

*Proof:* Using property (c) of  $\sigma$  sub-Gaussian r.v's,  $\sum_{i=1}^n [X_i - \mu]$  is  $\sqrt{n\sigma^2}$  sub-Gaussian. Using property (b) of  $\sigma$  sub-Gaussian r.v's,  $\frac{\sum_{i=1}^n [X_i - \mu]}{n}$  is  $\frac{\sigma}{\sqrt{n}}$  sub-Gaussian.

$\implies \hat{\mu} - \mu$  is  $\frac{\sigma}{\sqrt{n}}$  sub-Gaussian. Using the concentration result from the previous slide,  $\Pr[|\hat{\mu} - \mu| \geq \epsilon] \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$ . □

Hence, as we collect more data, the empirical mean concentrates around the true mean at an exponential rate.

# Back to Explore-Then-Commit (ETC)

---

**Algorithm** Explore-Then-Commit

---

- 1: **Input:**  $m \in \{1, \dots, \lfloor \frac{T}{K} \rfloor\}$ .
  - 2: **for**  $t = 1 \rightarrow mK$  **do**
  - 3:   Select arm  $a_t = (t \bmod K) + 1$  and observe reward  $R_t$    **(Explore)**
  - 4: **end for**
  - 5: Calculate empirical mean reward for arm  $a \in [K]$  as  $\hat{\mu}_a(mK) := \frac{\sum_{t=1}^{mK} R_t \mathbb{I}\{a_t=a\}}{N_a(mK)}$
  - 6: **for**  $t = mK + 1 \rightarrow T$  **do**
  - 7:   Pull arm  $\hat{a} := \arg \max_{a \in [K]} \hat{\mu}_a(mK)$    **(Commit)**
  - 8: **end for**
- 

**Distributional Assumption:** The noise  $\eta_t := R_t - \mu_{a_t}$  is 1 sub-Gaussian.  $\implies$  after pulling each arm  $m$  times in the **exploration** phase, for all  $a \in [K]$ ,  $|\hat{\mu}_a - \mu_a|$  is  $\frac{\sigma}{\sqrt{m}}$  sub-Gaussian and hence,  $\Pr[|\hat{\mu}_a - \mu_a| \geq \epsilon] \leq 2 \exp\left(-\frac{m\epsilon^2}{2\sigma^2}\right)$ .

Intuitively, the **exploration** phase estimates the gap  $\Delta_a$  for each arm upto a certain error. After this initial estimation, the algorithm **commits** to the *best empirical arm*.

## Explore-Then-Commit – Regret Analysis

**Claim:** For any  $m \in \{1, \dots, \lfloor T/K \rfloor\}$ ,

$$\text{Regret}(\text{ETC}, T) \leq m \sum_{a=1}^K \Delta_a + (T - mK) \sum_{a=1}^K \Delta_a \exp\left(-\frac{m \Delta_a^2}{4}\right)$$

*Proof:* Without loss of generality, assume that arm 1 is the best arm. Using the regret decomposition, we know that  $\text{Regret}(\text{ETC}, T) = \sum_a \Delta_a \mathbb{E}[N_a(T)]$ . For each arm  $a \in [K]$ ,  $\mathbb{E}[N_a(T)] = m + (T - mK) \Pr[\text{algorithm commits to arm } a]$ .

$$\Pr[\text{algorithm commits to arm } a] = \Pr[\hat{\mu}_a > \max_{j \neq a} \hat{\mu}_j] \leq \Pr[\hat{\mu}_a > \hat{\mu}_1]$$

(Since  $\{\hat{\mu}_a > \max_{j \neq a} \hat{\mu}_j\}$  is a subset of  $\{\hat{\mu}_a > \hat{\mu}_1\}$ )

$$= \Pr[\hat{\mu}_a - \mu_a > \hat{\mu}_1 - \mu_1 + [\mu_1 - \mu_a]] = \Pr[\underbrace{[\hat{\mu}_a - \mu_a]}_{X_a} - \underbrace{[\hat{\mu}_1 - \mu_1]}_{X_1} \geq \Delta_a]$$

$$= \Pr[X_a - X_1 \geq \Delta_a]$$

## Explore-Then-Commit – Regret Analysis

Recall that  $\text{Regret}(\text{ETC}, T) = \sum_a \Delta_a [m + (T - mK) \Pr[\text{algorithm commits to arm } a]]$  and  $\Pr[\text{algorithm commits to arm } a] \leq \Pr[X_a - X_1 \geq \Delta_a]$  where  $X_a = \hat{\mu}_a - \mu_a$ . Because of our assumption, both  $X_a$  and  $X_1$  are  $\frac{1}{\sqrt{m}}$  sub-Gaussian. Using property (c) of sub-Gaussian r.v's,  $X_a - X_1$  is  $\frac{\sqrt{2}}{\sqrt{m}}$  sub-Gaussian. Using the concentration result for sub-Gaussian r.v's,

$$\Pr[X_a - X_1 \geq \Delta_a] \leq \exp\left(-\frac{m \Delta_a^2}{4}\right)$$

Putting everything together,

$$\begin{aligned} \text{Regret}(\text{ETC}, T) &\leq \sum_a \Delta_a \left[ m + (T - mK) \exp\left(-\frac{m \Delta_a^2}{4}\right) \right] \\ \Rightarrow \text{Regret}(\text{ETC}, T) &\leq m \sum_{a=1}^K \Delta_a + (T - mK) \sum_{a=1}^K \Delta_a \exp\left(-\frac{m \Delta_a^2}{4}\right) \quad \square \end{aligned}$$

## Explore-Then-Commit – Regret Analysis

Recall that  $\text{Regret}(\text{ETC}, T) \leq m \sum_{a=1}^K \Delta_a + (T - mK) \sum_{a=1}^K \Delta_a \exp\left(-\frac{m\Delta_a^2}{4}\right)$ .

In order to gain some intuition about how to set  $m$ , consider  $K = 2$  with  $\Delta := \mu_1 - \mu_2$ .

$$\text{Regret}(\text{ETC}, T) \leq m\Delta + (T - 2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right) < m\Delta + T\Delta \exp\left(-\frac{m\Delta^2}{4}\right)$$

Optimizing the RHS w.r.t  $m$ , we get  $m = \frac{4}{\Delta^2} \log\left(\frac{\Delta^2 T}{4}\right)$ . Since  $m$  is an integer  $\geq 1$ , we should set  $m = \max\left\{1, \lceil \frac{4}{\Delta^2} \log\left(\frac{\Delta^2 T}{4}\right) \rceil\right\}$ . Plugging this value back,

$$\implies \text{Regret}(\text{ETC}, T) \leq \Delta + \frac{4}{\Delta} \left[1 + \log_+\left(\frac{\Delta^2 T}{4}\right)\right] \quad (\log_+(x) := \max\{0, \log(x)\})$$

Hence, ETC with  $m = O(1/\Delta^2)$  achieves  $O\left(\frac{\log(T)}{\Delta}\right)$  *instance or gap-dependent* regret.

**Q:** What is the problem with this bound?

## Explore-Then-Commit – Regret Analysis

To overcome the previous problem, one can bound the *worst-case problem-independent regret*.

**Claim:** For  $\Delta \leq 1$ , ETC results in an  $O(1 + \sqrt{T})$  worst-case bound on the regret.

*Proof:* In the worst-case, we pull the sub-optimal arm in every round. Hence, the regret for any algorithm is upper-bounded by  $T\Delta$ . Putting this together with the bound on the previous slide,

$$\text{Regret}(\text{ETC}, T) \leq \min \left\{ T\Delta, \Delta + \frac{4}{\Delta} \left[ 1 + \log_+ \left( \frac{\Delta^2 T}{4} \right) \right] \right\}$$

If  $\Delta < \frac{1}{\sqrt{T}}$ ,  $\text{Regret}(\text{ETC}, T) \leq \sqrt{T}$ . On the other hand, if  $\Delta \geq \frac{1}{\sqrt{T}}$ ,

$$\text{Regret}(\text{ETC}, T) \leq \Delta + 4\sqrt{T} + \left[ \frac{4}{\Delta} \log_+ \left( \frac{\Delta^2 T}{4} \right) \right] = \Delta + 4\sqrt{T} + 4 \max_{z>0} \frac{\log_+(Tz^2/4)}{z}$$

$$\text{Regret}(\text{ETC}, T) \leq \Delta + 4\sqrt{T} + \frac{4\sqrt{T}}{e} \leq 1 + \sqrt{T} \left( 4 + \frac{4}{e} \right) \quad (\text{Since } \Delta \leq 1)$$

- In general, for  $K$  arms, it can be shown that ETC results in  $O(\sqrt{KT})$  worst-case regret.



## Explore-Then-Commit – Regret Analysis

We have seen that ETC with  $m = O(1/\Delta^2)$  achieves an  $O(\Delta + \sqrt{T})$  regret for any instance.

**Q:** What is the problem with the ETC algorithm?

**Claim:** For  $\Delta \leq 1$ , there exists  $C > 0$  s.t. ETC with  $m = T^{2/3}$  results in  $(1 + C) T^{2/3}$  regret.

*Proof:* Need to prove this in Assignment 1!

*Hint:* Starting from the expression,  $\text{Regret}(\text{ETC}, T) \leq m\Delta + T\Delta \exp\left(-\frac{m\Delta^2}{4}\right)$ , upper-bound the second term independent of  $\Delta$  and then choose  $m$ .

# $\epsilon$ -greedy Algorithm

---

**Algorithm**  $\epsilon$ -greedy

---

- 1: **Input:**  $\{\epsilon_t\}_{t=1}^T$
  - 2: **for**  $t = 1 \rightarrow K$  **do**
  - 3:   Select arm  $a_t = t$  and observe  $R_t$
  - 4: **end for**
  - 5: Calculate empirical mean reward for arm  $a \in [K]$  as  $\hat{\mu}_a(K) := \frac{\sum_{t=1}^K R_t \mathcal{I}\{a_t=a\}}{N_a(K)}$
  - 6: **for**  $t = K + 1 \rightarrow T$  **do**
  - 7:   Select arm  $\begin{cases} a_t = \arg \max_{a \in [K]} \hat{\mu}_a(t-1) \text{ w.p. } 1 - \epsilon_t \\ a_t \sim \mathcal{U}\{1, 2, \dots, K\} \text{ w.p. } \epsilon_t \end{cases}$
  - 8:   Observe reward  $R_t$  and update for  $a \in [K]$ :
$$N_a(t) = N_a(t-1) + \mathcal{I}\{a_t = a\} \quad ; \quad \hat{\mu}_a(t) = \frac{N_a(t-1) \hat{\mu}_a(t-1) + R_t \mathcal{I}\{a_t = a\}}{N_a(t)}$$
  - 9: **end for**
- 

- $\epsilon$ -greedy with a fixed  $\epsilon_t = \epsilon$  can result in linear regret.
- For  $K = 2$ ,  $\epsilon$ -greedy with  $\epsilon_t = O\left(\frac{1}{\Delta^2 T}\right)$  incurs  $O\left(\frac{\log(T)}{\Delta^2}\right)$  regret.

# Upper Confidence Bound (UCB) Algorithm

- Based on the principle of *optimism in the face of uncertainty*.

---

**Algorithm** Upper Confidence Bound

---

- 1: **Input:**  $\delta$
- 2: For each arm  $a \in [K]$ , initialize  $U_a(0, \delta) := \infty$ .
- 3: **for**  $t = 1 \rightarrow T$  **do**
- 4:   Select arm  $a_t = \arg \max_{a \in [K]} U_a(t-1, \delta)$  (*Choose the lower-indexed arm in case of a tie*)
- 5:   Observe reward  $R_t$  and update for  $a \in [K]$ :

$$N_a(t) = N_a(t-1) + \mathcal{I}\{a_t = a\} \quad ; \quad \hat{\mu}_a(t) = \frac{N_a(t-1) \hat{\mu}_a(t-1) + R_t \mathcal{I}\{a_t = a\}}{N_a(t)}$$

$$U_a(t, \delta) = \hat{\mu}_a(t) + \sqrt{\frac{2 \log(1/\delta)}{N_a(t)}}$$

- 6: **end for**
- 

- Intuitively, UCB pulls a “promising” arm (with higher empirical mean  $\hat{\mu}_a$ ) or one that has not been explored enough (with lower  $N_a(t)$ ).

# UCB – Regret Analysis

**Claim:** UCB with  $\delta = \frac{1}{T^2}$  achieves the following problem-dependent bound on the regret,

$$\text{Regret}(\text{UCB}, T) \leq 2 \sum_{a=1}^K \Delta_a + \sum_{a \in [K] | \Delta_a > 0} \frac{16 \log(T)}{\Delta_a}$$

*Proof:* Without loss of generality, assume that arm 1 is the best arm. Using the regret decomposition, we know that  $\text{Regret}(\text{UCB}, T) = \sum_a \Delta_a \mathbb{E}[N_a(T)]$ . Define a threshold  $\tau_a$  and  $\hat{\mu}_{a, \tau_a}$  as the mean for arm  $a$  after pulling it for the first  $\tau_a$  times. Define a “good” event  $G_a$  for each  $a \neq 1$ .

$$G_a = \left\{ \mu_1 < \min_{t \in [T]} U_1(t, \delta) \right\} \cap \left\{ \hat{\mu}_{a, \tau_a} + \sqrt{\frac{2 \log(1/\delta)}{\tau_a}} < \mu_1 \right\}$$

Consider two cases when bounding  $\mathbb{E}[N_a(T)]$ . Using the law of total expectation,

$$\begin{aligned} \mathbb{E}[N_a(T)] &= \mathbb{E}[N_a(T) | G_a] \Pr[G_a] + \mathbb{E}[N_a(T) | G_a^c] \Pr[G_a^c] \\ &\leq \underbrace{\mathbb{E}[N_a(T) | G_a]}_{\text{Term (i)}} + T \underbrace{\Pr[G_a^c]}_{\text{Term (ii)}} \quad (N_a(T) \leq T \text{ for all } a, \Pr[G_a] \leq 1) \end{aligned}$$

## UCB – Regret Analysis

Recall that  $G_a = \{\mu_1 < \min_{t \in [T]} U_1(t, \delta)\} \cap \left\{ \hat{\mu}_{a, \tau_a} + \sqrt{\frac{2 \log(1/\delta)}{\tau_a}} < \mu_1 \right\}$ . We will show (by contradiction) that Term (i)  $= \mathbb{E}[N_a(T) | G_a] \leq \tau_a$ .

Suppose  $\mathbb{E}[N_a(T) | G_a] > \tau_a$ , then there is a round  $t$  s.t.  $N_a(t-1) = \tau_a$ ,  $a_t = a$ . Since  $a_t = \arg \max_a U_a(t-1, \delta)$ , it follows that  $U_a(t-1, \delta) > U_1(t-1, \delta)$ . However, we know that,

$$\begin{aligned} U_a(t-1, \delta) &= \hat{\mu}_a(t-1) + \sqrt{\frac{2 \log(1/\delta)}{N_a(t-1)}} = \hat{\mu}_a(t-1) + \sqrt{\frac{2 \log(1/\delta)}{\tau_a}} \\ &\hspace{15em} \text{(By assumption, } N_a(t-1) = \tau_a \text{)} \\ &= \hat{\mu}_{a, \tau_a} + \sqrt{\frac{2 \log(1/\delta)}{\tau_a}} \hspace{10em} \text{(Since arm } a \text{ has been pulled } \tau_a \text{ times)} \\ &\leq \mu_1 < U_1(t-1, \delta), \hspace{10em} \text{(Since we are conditioning on } G_a \text{)} \end{aligned}$$

which is a contradiction. Hence,  $\mathbb{E}[N_a(T) | G_a] \leq \tau_a$ .

## UCB – Regret Analysis

$$\text{Bounding Term (ii)} = \Pr[G_a^c] \leq \Pr[\mu_1 \geq \min_{t \in [T]} U_1(t, \delta)] + \Pr\left[\hat{\mu}_{a, \tau_a} + \sqrt{\frac{2 \log(1/\delta)}{\tau_a}} \geq \mu_1\right].$$

$$\begin{aligned}\left\{\mu_1 \geq \min_{t \in [T]} U_1(t, \delta)\right\} &= \left\{\mu_1 \geq \min_{t \in [T]} \left\{\hat{\mu}_1(t) + \sqrt{\frac{2 \log(1/\delta)}{N_1(t)}}\right\}\right\} \\ &= \left\{\mu_1 \geq \min_{s \in [T]} \left\{\hat{\mu}_{1,s} + \sqrt{\frac{2 \log(1/\delta)}{s}}\right\}\right\} \\ &= \bigcup_{s=1}^T \left\{\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2 \log(1/\delta)}{s}}\right\}\end{aligned}$$

$$\begin{aligned}\Rightarrow \Pr\left[\mu_1 \geq \min_{t \in [T]} U_1(t, \delta)\right] &\leq \sum_{s=1}^T \Pr\left[\mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2 \log(1/\delta)}{s}}\right] && \text{(Union Bound)} \\ &\leq \sum_{s=1}^T \delta = \delta T && \text{(Using concentration for sub-Gaussian r.v's)}\end{aligned}$$

# UCB – Regret Analysis

Recall that Term (ii) =  $\Pr[G_a^c] \leq \delta T + \Pr\left[\hat{\mu}_{a,\tau_a} + \sqrt{\frac{2\log(1/\delta)}{\tau_a}} \geq \mu_1\right]$ . Assume that  $\tau_a$  is chosen such that  $\Delta_a - \frac{2\log(1/\delta)}{\tau_a} \geq \frac{\Delta_a}{2}$ .

$$\begin{aligned}\Pr\left[\hat{\mu}_{a,\tau_a} + \sqrt{\frac{2\log(1/\delta)}{\tau_a}} \geq \mu_1\right] &= \Pr\left[\hat{\mu}_{a,\tau_a} - \mu_a + \sqrt{\frac{2\log(1/\delta)}{\tau_a}} \geq \Delta_a\right] \leq \Pr\left[\hat{\mu}_{a,\tau_a} - \mu_a \geq \frac{\Delta_a}{2}\right] \\ &\leq \exp\left(-\frac{\tau_a \Delta_a^2}{8}\right) \\ &\quad \text{(Using concentration for sub-Gaussian r.v's)}\end{aligned}$$

Putting everything together,

$$\begin{aligned}\implies \Pr[G_a^c] &\leq \delta T + \exp\left(-\frac{\tau_a \Delta_a^2}{8}\right) \\ \implies \mathbb{E}[N_a(T)] &\leq \tau_a + T \left[ \delta T + \exp\left(-\frac{\tau_a \Delta_a^2}{8}\right) \right]\end{aligned}$$

# UCB – Regret Analysis

Recall that  $\mathbb{E}[N_a(T)] \leq \tau_a + T \left[ \delta T + \exp\left(-\frac{\tau_a \Delta_a^2}{8}\right) \right]$ .

$$\mathbb{E}[N_a(T)] \leq \frac{8 \log(1/\delta)}{\Delta_a^2} + T [\delta T + \delta] \quad (\text{Setting } \tau_a = \frac{8 \log(1/\delta)}{\Delta_a^2})$$

$$\leq \frac{8 \log(1/\delta)}{\Delta_a^2} + 2\delta T^2$$

$$= \frac{16 \log(T)}{\Delta_a^2} + 2 \quad (\text{Setting } \delta = 1/T^2)$$

$$\implies \text{Regret}(\text{UCB}, T) = \sum_a \Delta_a \mathbb{E}[N_a(T)] = 2 \sum_{a=1}^K \Delta_a + \sum_{a=2}^K \frac{16 \log(T)}{\Delta_a} \quad \square$$



# UCB – Regret Analysis

**Claim:** For  $\Delta \leq 1$ , UCB with  $\delta = \frac{1}{T^2}$  achieves the following worst-case regret,

$$\text{Regret}(\text{UCB}, T) \leq 2K + 8\sqrt{K T \log(T)}$$

*Proof:* Define  $C > 0$  to be a constant to be tuned later. From the regret decomposition result,

$$\begin{aligned} \text{Regret}(\text{UCB}, T) &= \sum_{a=1}^K \Delta_a \mathbb{E}[N_a(T)] = \sum_{a|\Delta_a < C} \Delta_a \mathbb{E}[N_a(T)] + \sum_{a|\Delta_a \geq C} \Delta_a \mathbb{E}[N_a(T)] \\ &\leq CT + \sum_{a|\Delta_a \geq C} \Delta_a \mathbb{E}[N_a(T)] && \text{(Since } \sum_{a=1}^K N_a(T) = T \text{)} \\ &\leq CT + \sum_{a|\Delta_a \geq C} \left[ \frac{16 \log(T)}{\Delta_a} + 2\Delta_a \right] && \text{(From the previous slide)} \\ &\leq CT + \left[ \frac{16K \log(T)}{C} + \sum_{a|\Delta_a \geq C} 2\Delta_a \right] && \text{(Setting } C = \sqrt{\frac{16K \log(T)}{T}} \text{)} \end{aligned}$$

$$\implies \text{Regret}(\text{UCB}, T) \leq 8\sqrt{K T \log(T)} + 2K\Delta_a \leq 2K + 8\sqrt{K T \log(T)}$$

# UCB vs ETC

- Similar to best-tuned ETC, UCB results in an  $\tilde{O}(\sqrt{KT})$  problem-independent regret.
- Unlike best-tuned ETC, UCB does not need to know the gaps  $\Delta$  to set algorithm parameters, but does require knowledge of the horizon  $T$ .

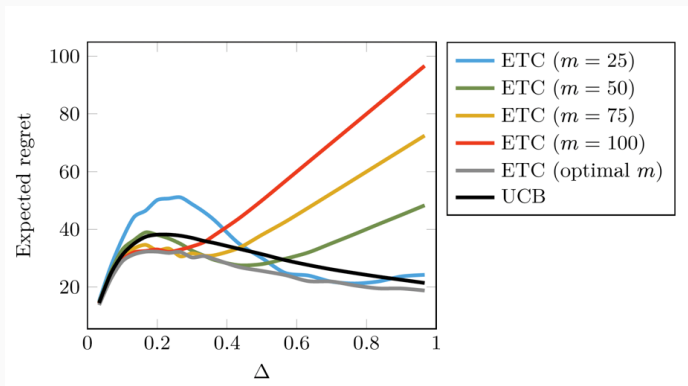





Figure 1: For  $K = 2$ ,  $T = 1000$ , Gaussian rewards, comparing UCB and ETC( $m$ ) as a function of the gap  $\Delta$ .

# Improvements to UCB

- **Problem:** UCB requires knowledge of  $T$  and hence, the number of rounds needs to be fixed.
- *Sol:* Define UCB as  $\hat{\mu}_a(t) + \sqrt{\frac{2 \log(f(t))}{N_a(t)}}$  where  $f(t) := 1 + t \log^2(t)$ . No dependence on  $T$ , but results in the same  $O(\sqrt{KT \log(T)})$  worst-case regret. (see [LS20, Chapter 8])
- **Lower-Bound:** For a fixed  $T$  and for every bandit algorithm, there exists a stochastic bandit problem with rewards in  $[0, 1]$  such that  $\text{Regret}(T) = \Omega(\sqrt{KT})$ . (see [LS20, Chapter 15]).
- **Problem:** UCB is sub-optimal by a  $\sqrt{\log(T)}$  factor compared to the lower-bound. Is it possible to develop an algorithm that does not incur this log factor?
- *Sol:* [Lat18, MG17] propose modifications of UCB that achieve  $O(\sqrt{KT})$  regret.

-  Tor Lattimore, *Refining the confidence level for optimistic bandit strategies*, The Journal of Machine Learning Research **19** (2018), no. 1, 765–796.
-  Tor Lattimore and Csaba Szepesvári, *Bandit algorithms*, Cambridge University Press, 2020.
-  Pierre Ménard and Aurélien Garivier, *A minimax and asymptotically optimal algorithm for stochastic bandits*, International Conference on Algorithmic Learning Theory, PMLR, 2017, pp. 223–237.