Johns Hopkins University

Multi-armed Bandit II

Algorithms

Thanh Nguyen-Tang

The slides are partially credited to: Patrick Rebeschini

Formal set up of Multi-armed Bandits

- Arm set: $\mathcal{A} = \{1,2,\ldots,\mathrm{k}\}$ and $|\mathcal{A}| = \mathrm{k}$
- At every round $\mathrm{t} = 1,2,\ldots,\mathrm{n}$:
  - Learner chooses an arm $\mathrm{a_t} \in \mathcal{A} = \{1,\ldots,\mathrm{k}\}$
  - A data point $z_t = (z_{t,1}, z_{t,2}, \ldots, z_{t,k}) \in [0,1]^k$ is sampled independently from an unknown distribution with unknown means $(\mu_1, \ldots, \mu_k) \in [0,1]^k$
  - Learner observes reward z_{t,a_t} (but not other rewards $z_{t,a}$ for any $a \neq a_t$) (**bandit feedback**)
- **Goal**: Minimize the **pseudo-regret** R_n defined as

$$\boxed{\mathrm{R_n} := \mathrm{n}\,\mu_{\mathrm{a}^*} - \sum_{\mathrm{t}=1}^{\mathrm{n}} \mu_{\mathrm{a_t}}}$$

$$\boxed{\mathrm{a}_* \in \operatorname*{argmax}_{\mathrm{a}\in\mathcal{A}} \mu_{\mathrm{a}}}$$

- $\mathrm{a_t}$ is a function of $\mathrm{a_1}, \ldots, \mathrm{a_{t-1}}$ and $\mathrm{z_{1,a_1}}, \mathrm{z_{2,a_2}}, \ldots, \mathrm{z_{t-1,a_{t-1}}}$
- For simplicity, assume the optimal arm a_* is unique
- Note: Learning occurs when algorithm achieves sub-linear growth in n, i.e. $\frac{\mathbb{E}R_n}{n} \to 0$

Multi-armed Bandit problem

- Number of times arm a is pulled up to time t: $\mathrm{N}_{t,a} \coloneqq \sum_{i=1}^{t} 1\{a_i = a\}$
- Sub-optimality of arm a: $\Delta_a \coloneqq \mu_{a^*} - \mu_a$

Lemma 1: $R_n = \sum_{a=1}^{k} \Delta_a N_{n,a}$

Proof: $n = \sum_{a=1}^{k} N_{n,a}$ and $\sum_{t=1}^{n} \mu_{a_t} = \sum_{a=1}^{k} \mu_a N_{n,a}$

Multi-armed Bandit problem

Q: How to construct an algorithm?

A: Use sample mean

$$\hat{\mu}_{t,a} := \frac{1}{N_{t,a}} \sum_{i=1}^{t} z_{t,a_t} 1\{a_t = a\}$$

Attempt #1: Explore-then-Commit

- **Idea**: Explore all arms for m times and then commit to the arm with the highest sample mean
- Exploration-exploitation trade-off controlled by m

Algorithm 1 Explore-Then-Commit(m)

1: **for** $t = 1, \ldots, mk$ **do**
2: Set $a_t = t \pmod{k} + 1$ % Explore
3: **end for**
4: **for** $t = mk + 1, \ldots, n$ **do**
5: Set $a_t \in \arg\max_{a \in [k]} \widehat{\mu}_{mk,a}$ % Commit
6: **end for**

Pseudo-regret of Explore-then-Commit

Explore-then-commit suffers linear pseudo-regret

Proposition 1: Linear pseudo-regret for Explore-Then-Commit

For any $m \in \mathbb{N}_+$, there exists a stochastic multi-armed bandit problem such that

$$\boxed{\mathbb{E}R_n \geq c_1 n + c_2}$$

for some absolute constants $c_1 \geq 0$ and $c_2 \in \mathbb{R}$ that are independent of n

Proof idea for explore-then-commit

- Consider a bandit instance with two arms (i.e., $k = 2$)
- The optimal arm has deterministic reward $\mu_1 \in (0, 1)$
- The sub-optimal arm has reward distribution $\text{Bernoulli}(\mu_2)$ where $0 < \mu_2 < \mu_1$
- The probability that the explore-then-commit algorithm chooses the sub-optimal arm after its exploration phase is

$$p \coloneqq \Pr\left(\hat{\mu}_{2m,1} < \hat{\mu}_{2m,2}\right) = \Pr\left(m\mu_1 < \text{Binomial}(m, \mu_2)\right) > 0$$

- Thus, we have

$$\begin{aligned}
\mathbb{E}R_n &= \mathbb{E}[R_n 1\{\hat{\mu}_{2m,1} < \hat{\mu}_{2m,2}\}] + \mathbb{E}[R_n 1\{\hat{\mu}_{2m,1} \geq \hat{\mu}_{2m,2}\}] \\
&\geq \mathbb{E}[R_n 1\{\hat{\mu}_{2m,1} < \hat{\mu}_{2m,2}\}] \\
&= \underbrace{m\Delta_2}_{\text{exploration}} + \underbrace{(n - 2m)\Delta_2 p}_{\text{commit after 2m rounds}} \\
&= n\ \underbrace{p\,\Delta_2}_{c_1} + \underbrace{(1 - 2p)\Delta_2 p}_{c_2}
\end{aligned}$$

Attempt #2: ϵ-Greedy

- **Idea**: keep exploration on
- Exploration-exploitation trade-off controlled by ϵ

Algorithm 2 ϵ-Greedy

```
for t = 1, ..., k do
    Set a_t = t % Init explore
end for
for t = k + 1, ..., n do
    Set
```

$$a_t \begin{cases} \in \arg\max_{a \in [k]} \widehat{\mu}_{t-1,a} & \text{with probability } 1-\epsilon \\ \sim \mathrm{Uniform}(\{1,\ldots,k\}) & \text{with probability } \epsilon \end{cases}$$

```
end for
```

Pseudo-regret for ϵ-Greedy

ϵ-Greedy suffers linear pseudo-regret!

Proposition 2: Linear pseudo-regret for ϵ-greedy

For any $\epsilon > 0$ in ϵ-Greedy, there exists a stochastic multi-armed bandit problem such that

$$\boxed{\mathbb{E}R_n \geq c_1 n + c_2}$$

for some absolute constants $c_1 \geq 0$ and $c_2 \in \mathbb{R}$ that are independent of n

Proof idea for ϵ-Greedy

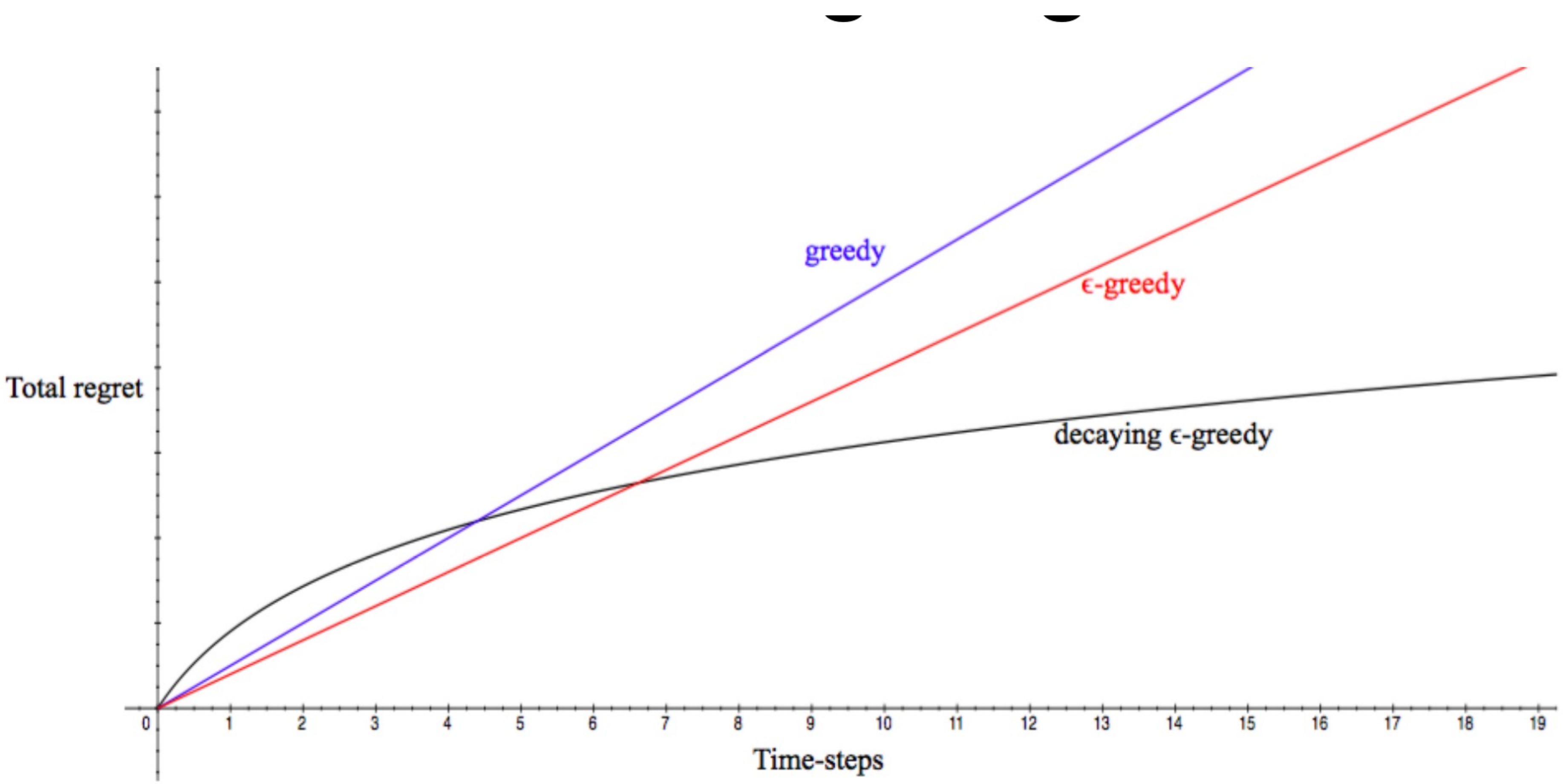
- The probability that each arm is played at any round after the initial phase is $\frac{\epsilon}{k}$
- The expected number of times arm a is pulled up to round n:

$$\mathbb{E}N_{n,a} \geq 1 + \frac{\epsilon}{\mathrm{k}}(\mathrm{n} - \mathrm{k})$$

- The expected pseudo-regret

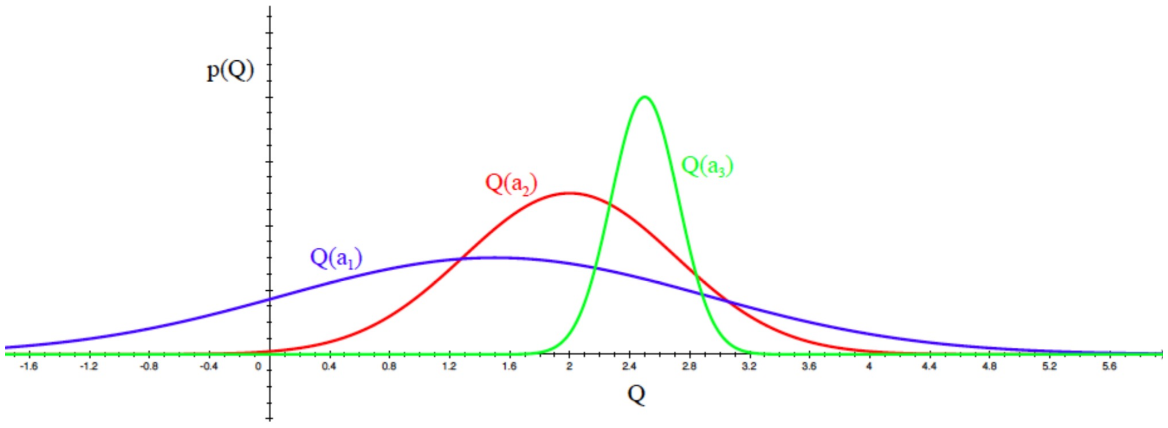
$$\begin{aligned} \mathbb{E}[R_n] &= \sum_{a=1}^{k} \mathbb{E}[N_{n,a}]\Delta_a \geq \sum_{a=1}^{k}(1 + \frac{\epsilon}{k}(n-k))\Delta_a \\ &= n\underbrace{\frac{\epsilon}{k}\sum_{a=1}^{k}\Delta_a}_{c_1} + \underbrace{(1-\epsilon)\sum_{a=1}^{k}\Delta_a}_{c_2} \end{aligned}$$

Practical performance of ϵ-Greedy



Upper confidence bound (UCB)

- **Idea**: Let exploration depend on the confidence of mean estimates
- Exploration-exploitation trade-offs controlled by $\{\beta_t\}_{t\in\{1,\ldots,n\}}$
- **Optimism in the face of uncertainty principle**: we explore arms that are highly uncertain and high sample estimates



Algorithm 3 UCB($\{\beta_t\}_{t=1}^{n}$)

for $t = 1, \ldots, k$ **do**
  Set $a_t = t$ % Init explore
end for
for $t = k+1, \ldots, n$ **do**
  Set $a_t = \arg\max_{a\in\{1,\ldots,k\}} \widehat{\mu}_{t,a} + \sqrt{\frac{\beta_t}{N_{t,a}}}$
end for

Gap-dependent bounds for UCB

Proposition 3: Gap-dependent bounds for UCB

In UCB($\{\beta_t\}_{t=1}^n$), set $\beta_t = 0.5 \log\left(4(n-k)/\delta\right)$ for any $\delta > 0$. The expected pseudo-regret of UCB($\{\beta_t\}_{t=1}^n$) is:

$$\mathbb{E}R_n \leq 2\log(4(n-k)/\delta) \sum_{a \neq a_*} \frac{1}{\Delta_a} + n\delta \sum_{a=1}^{k} \Delta_a.$$

- Set $\delta = \frac{1}{n}$, we have $\mathbb{E}\mathrm{R_n} = \mathcal{O}\left((\log \mathrm{n}) \sum_{\mathrm{a} \neq \mathrm{a}_*} \frac{1}{\Delta_\mathrm{a}} + \sum_{\mathrm{a}=1}^{\mathrm{k}} \Delta_\mathrm{a}\right)$

Gap-dependent bounds for UCB: Proof idea

Step 1: Construct a confidence region around the sample mean

- Hoeffding's inequality: Let $X_1, \ldots, X_n$ be i.i.d. samples from $[0, 1]$ with mean μ

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}\right) \geq 1 - \delta$$

- Application: For any $t \in [k + 1, n]$ and $a \in \{1, \ldots, k\}$, we have

$$\Pr\left(\left|\hat{\mu}_{t,a} - \mu_a\right| \leq \sqrt{\frac{\log(2/\delta)}{2N_{t,a}}}\right) \geq 1 - \delta$$

Proof idea for UCB (con't)

- By the **union bound**,

$$\Pr\left(\left|\hat{\mu}_{t,a} - \mu_a\right| \le \sqrt{\frac{\log(4(n-k)/\delta)}{2N_{t,a}}}, \forall t \in [k+1, n]\right) \ge 1 - \delta/2$$

- Define lower confidence bound (LCB) and upper confidence bound (UCB):
  - $\mathrm{L}_{\mathrm{t,a}} := \hat{\mu}_{\mathrm{t,a}} - \sqrt{\frac{\log(4(\mathrm{n}-\mathrm{k})/\delta)}{2\mathrm{N}_{\mathrm{t,a}}}}$ and $\mathrm{U}_{\mathrm{t,a}} := \hat{\mu}_{\mathrm{t,a}} + \sqrt{\frac{\log(4(\mathrm{n}-\mathrm{k})/\delta)}{2\mathrm{N}_{\mathrm{t,a}}}}$
- For any $a \in \{1, \dots, k\}$, we have

$$\Pr\left(\mu_a \in \left[\mathrm{L}_{\mathrm{t,a}}, \mathrm{U}_{\mathrm{t,a}}\right], \forall \mathrm{t} \in [\mathrm{k}+1, \mathrm{n}]\right) \ge 1 - \delta/2$$

Proof idea for UCB (con't)

Step 2: Show that any non-optimal action a cannot be pulled too frequently

- Fix any non-optimal arm a
- Consider the event

$$\mathrm{E_a} \coloneqq \{\mu_\mathrm{a} \in [\mathrm{L_{t,a}}, \mathrm{U_{t,a}}], \forall \mathrm{t} \in [\mathrm{k}+1, \mathrm{n}]\} \cap \{\mu_{\mathrm{a}_*} \in [\mathrm{L}_{\mathrm{t},\mathrm{a}_*}, \mathrm{U}_{\mathrm{t},\mathrm{a}_*}], \forall \mathrm{t} \in [\mathrm{k}+1, \mathrm{n}]\}$$

- $\Pr(\mathrm{E_a}) \geq 1 - \delta$
- Let n_a be the largest round $t \in \{1, \dots, n\}$ in which arm a is played. We must have

$$\mathrm{U}_{\mathrm{n_a},\mathrm{a}} \geq \mathrm{U}_{\mathrm{n_a},\mathrm{a}_*}$$

Proof idea for UCB (con't)

- This implies that, under event E_a, we have

$$\mu_a \geq L_{n_a,a} = U_{n_a,a} - 2\sqrt{\frac{\log 4(n-k)/\delta}{2N_{n_a,a}}} \geq U_{n_a,\mathbf{a}_*} - 2\sqrt{\frac{\log 4(n-k)/\delta}{2N_{n_a,a}}} \geq \mu_{\mathbf{a}_*} - 2\sqrt{\frac{\log 4(n-k)/\delta}{2N_{n_a,a}}}$$

- It implies that $N_{n,a} = N_{n_a,a} \leq \frac{2\log(4(n-k)/\delta)}{\Delta_a^2}$
- Thus, we have $\Delta_a \mathbb{E} N_{n,a} = \Delta_a \mathbb{E}[N_{n,a} 1\{E_a\}] + \Delta_a \mathbb{E}[N_{n,a} 1\{E_a^c\}] \leq \frac{2\log\left(\frac{4(n-k)}{\delta}\right)}{\Delta_a} + n\Delta_a\delta$
- Combing via $\mathbb{E}R_n = \sum_{a=1}^{k} \Delta_a \mathbb{E}N_{n,a}$

Gap-independent bounds for UCB

Proposition 4: Gap-independent bounds for UCB

Set $\beta_t = 0.5 \log\left(4(n-k)/\delta\right)$ for any $\delta > 0$

$$\mathbb{E}[R_n] \leq \sqrt{2nk \log(4n(n-k))} + k = \mathcal{O}(\sqrt{nk \log n})$$

- Independent of gap Δ_a

Proof idea for gap-independent bound of UCB

Proof of Lemma 1.6 For any $\delta > 0$, $\epsilon > 0$, we have

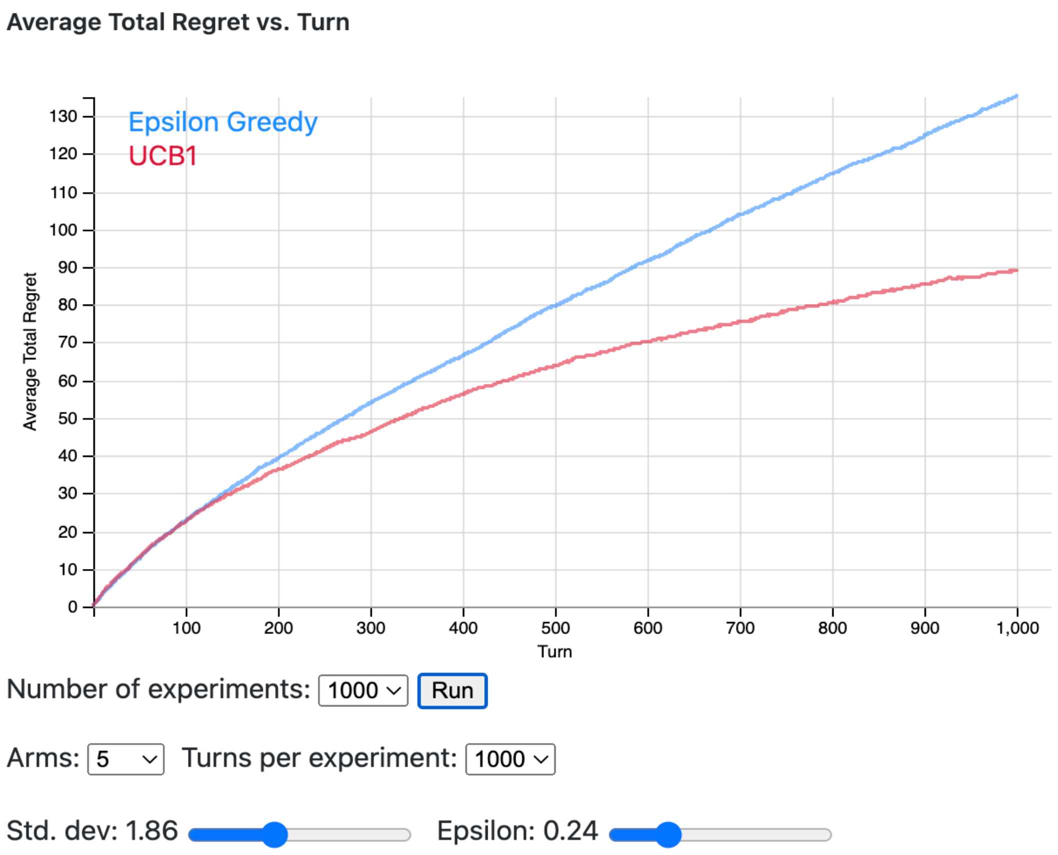
$$\begin{aligned}
\mathbb{E}R_n &= \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}](\text{Lemma 1.1}) \\
&= \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}]\mathbb{1}\{\Delta_a < \epsilon\} + \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}]\mathbb{1}\{\Delta_a \geq \epsilon\} \\
&\leq \epsilon n + \Delta_a \mathbb{E}[N_{n,a}]\mathbb{1}\{\Delta_a \geq \epsilon\} \\
&= \epsilon n + \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}\mathbb{1}\{E_a\}]\mathbb{1}\{\Delta_a \geq \epsilon\} + \sum_{a\in[k]} \Delta_a \mathbb{E}[N_{n,a}\mathbb{1}\{E_a^c\}]\mathbb{1}\{\Delta_a \geq \epsilon\} \quad (E_a \text{ defined in Eq. (1.5)}) \\
&\leq \epsilon n + \sum_{a\in[k]} \frac{2\log(4(n-k)/\delta)}{\Delta_a}\mathbb{1}\{\Delta_a \geq \epsilon\} + \sum_{a\in[k]} n\Delta_a\delta\mathbb{1}\{\Delta_a \geq \epsilon\}(\text{Eq. (1.6)}) \\
&\leq \epsilon n + \sum_{a\in[k]} \frac{2\log(4(n-k)/\delta)}{\epsilon}\mathbb{1}\{\Delta_a \geq \epsilon\} + \sum_{a\in[k]} n\delta(\Delta_a \leq 1) \\
&\leq \epsilon n + \frac{2k\log(4(n-k)/\delta)}{\epsilon} + nk\delta.
\end{aligned}$$

Note that the above inequality holds for any $\epsilon > 0$. Picking $\delta = 1/n$ and minimizing the RHS of the above inequality with respect to ϵ yields

$$\mathbb{E}R_n \leq \sqrt{2nk\log(4n(n-k))} + k.$$

Empirical comparison btw ϵ-greedy and UCB

- Simulation: https://cse442-17f.github.io/LinUCB/



Minimax lower bounds

- The UCB algorithms we considered so far yields the regret bound $\mathcal{O}\left(\sqrt{nk\log n}\right)$
- How do we know if this bound is <u>improvable</u>? → Construct minimax lower bounds

Minimax lower bounds

- The minimax lower bound $f(n, k)$ says that: For **any** bandit algorithm $(a_1, \dots, a_n)$, there exists a least a bandit instance M in the bandit family $\mathcal{M}$ such that the regret of $(a_1, \dots, a_n)$ cannot better than $f(n, k)$
- Formally $\sup_{(a_1,\dots,a_n)} \inf_{M \in \mathcal{M}} \mathbb{E}_M R_n \geq f(n, k)$

Minimax lower bound for UCB algorithm

- The theorem says that UCB algorithm is minimax-optimal up to log factors

Theorem 2.2 *Let $\mathcal{M}$ be the set of k-armed bandit problems $\mu = (\mu_1, \ldots, \mu_k)$. For any $n \geq (k-1)/2$, we have*

$$\inf_{(a_1,\ldots,a_n)} \sup_{\mu \in \mathcal{M}} \mathbb{E}_\mu[R_n] \geq c\sqrt{n(k-1)},$$

for some absolute constant $c > 0$.

Proof strategy for minimax lower bounds

- Reduction to hypothesis testing

Lemma 2.1 (Neyman Pearson) *Let $x_1, \ldots, x_n \in \mathcal{X}^n$ the random variable that is distributed according to either P and Q. For any test function $f : \mathcal{X}^n \to \{0, 1\}$, we have*

$$P(f(x_1, \ldots, x_n) = 0) + Q(f(x_1, \ldots, x_n) = 1) \geq 1 - \sqrt{\frac{1}{2}\mathrm{KL}(P, Q)}.$$

- We construct two bandit models μ and ν such that any algorithm $(a_1, \ldots, a_n)$ must commit a high regret in either of these bandits. Specifically,
  - μ and ν have different optimal actions so that they can potentially confuse any algorithm $(a_1, \ldots, a_n)$ (in the sense that the algorithm cannot obtain small regret in both problems simultaneously)
  - μ and ν need to be similar to each other enough so we can have a tight lower bound and can compute $\mathrm{KL}(P_\mu, P_\nu)$ conveniently
- To this end, we consider u and ν be Bernoulli distributions with the means of the following form

$$\mu = (\frac{1}{2} + \Delta, \frac{1}{2}, \ldots, \frac{1}{2}) \in \mathbb{R}^k$$

$$\nu = (\frac{1}{2} + \Delta, \frac{1}{2}, \ldots, \frac{1}{2}, \frac{1}{2} + 2\Delta, \frac{1}{2}, \ldots, \frac{1}{2}) \in \mathbb{R}^k$$