Bandits: Upper Confidence Bound Algorithm

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2022

Today

- Feedback from last lecture
- Recap
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm
- UCB regret analysis

Feedback from feedback forms

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2. Main feedback: **Too fast!**

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2. Main feedback: **Too fast!**
    - I will slow down! (please let me know in today's feedback how the pace is)

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2. Main feedback: **Too fast!**
    - I will slow down! (please let me know in today's feedback how the pace is)
    - I will review the last 15 minutes from last lecture (which was always supposed to be the first part of today's lecture)

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2. Main feedback: **Too fast!**
   - I will slow down! (please let me know in today's feedback how the pace is)
   - I will review the last 15 minutes from last lecture (which was always supposed to be the first part of today's lecture)
3. Common points of confusion:

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2. Main feedback: **Too fast!**
    - I will slow down! (please let me know in today's feedback how the pace is)
    - I will review the last 15 minutes from last lecture (which was always supposed to be the first part of today's lecture)
3. Common points of confusion:
    i. Why we bound the regret with high probability instead of in expectation (I will also talk about this a bit)

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2. Main feedback: **Too fast!**
   - I will slow down! (please let me know in today's feedback how the pace is)
   - I will review the last 15 minutes from last lecture (which was always supposed to be the first part of today's lecture)
3. Common points of confusion:

   i. Why we bound the regret with high probability instead of in expectation (I will also talk about this a bit)

   ii. Union bound trick (I will review this today in more detail)

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2. Main feedback: **Too fast!**
   - I will slow down! (please let me know in today's feedback how the pace is)
   - I will review the last 15 minutes from last lecture (which was always supposed to be the first part of today's lecture)
3. Common points of confusion:
   i. Why we bound the regret with high probability instead of in expectation (I will also talk about this a bit)
   ii. Union bound trick (I will review this today in more detail)
4. Use bit.ly and QR code for feedback form: done!

Today

- Feedback from last lecture
- Recap
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm
- UCB regret analysis

Recap

Recap

- Pure greedy and pure exploration achieve linear regret

Recap

- Pure greedy and pure exploration achieve linear regret
- Explore-then-commit (ETC) and ε-greedy:

Recap

- Pure greedy and pure exploration achieve linear regret
- Explore-then-commit (ETC) and ε-greedy:
  - balance exploration with exploitation

Recap

- Pure greedy and pure exploration achieve linear regret
- Explore-then-commit (ETC) and ε-greedy:
  - balance exploration with exploitation
  - Achieve sublinear regret of $\tilde{O}(T^{2/3})$

Recap

- Pure greedy and pure exploration achieve linear regret
- Explore-then-commit (ETC) and ε-greedy:
    - balance exploration with exploitation
    - Achieve sublinear regret of $\tilde{O}(T^{2/3})$
    - Exploration is non-adaptive

Recap

- Pure greedy and pure exploration achieve linear regret
- Explore-then-commit (ETC) and ε-greedy:
  - balance exploration with exploitation
  - Achieve sublinear regret of $\tilde{O}(T^{2/3})$
  - Exploration is non-adaptive
- Today: can we do better than a rate of $T^{2/3}$?

Recap

- Pure greedy and pure exploration achieve linear regret
- Explore-then-commit (ETC) and ε-greedy:
    - balance exploration with exploitation
    - Achieve sublinear regret of $\tilde{O}(T^{2/3})$
    - Exploration is non-adaptive
- Today: can we do better than a rate of $T^{2/3}$?
- First, review a couple points of common confusion from last lecture

Some points of confusion (cont’d)

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret_T

Some points of confusion (cont'd)

Should we bound Regret$_T$ with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret$_T$
    i. Bounds *typical* behavior of Regret$_T$

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret_T
  1. Bounds *typical* behavior of Regret_T
  2. Leaves possibility of some rare but really bad Regret_T values

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret_T
    i. Bounds *typical* behavior of Regret_T
    ii. Leaves possibility of some rare but really bad Regret_T values
- Bound on $\mathbb{E}[\text{Regret}_T]$

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret_T
  i. Bounds *typical* behavior of Regret_T
  ii. Leaves possibility of some rare but really bad Regret_T values
- Bound on $\mathbb{E}[\text{Regret}_T]$
  iii. Bounds *average* behavior of Regret_T

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret_T
  - i. Bounds *typical* behavior of Regret_T
  - ii. Leaves possibility of some rare but really bad Regret_T values
- Bound on $\mathbb{E}[\text{Regret}_T]$
  - iii. Bounds *average* behavior of Regret_T
  - iv. Leaves possibility that the behavior of $\mathbb{E}[\text{Regret}_T]$ is *atypical*

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret_T
  1. Bounds *typical* behavior of Regret_T
  2. Leaves possibility of some rare but really bad Regret_T values
- Bound on $\mathbb{E}[\text{Regret}_T]$
  3. Bounds *average* behavior of Regret_T
  4. Leaves possibility that the behavior of $\mathbb{E}[\text{Regret}_T]$ is *atypical*
- E.g., $\text{Regret}_T \sim T^{100} \times \text{Bernoulli}(10^{-10})$:

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret_T
  i. Bounds *typical* behavior of Regret_T
  ii. Leaves possibility of some rare but really bad Regret_T values
- Bound on $\mathbb{E}[\text{Regret}_T]$
  iii. Bounds *average* behavior of Regret_T
  iv. Leaves possibility that the behavior of $\mathbb{E}[\text{Regret}_T]$ is *atypical*
- E.g., $\text{Regret}_T \sim T^{100} \times \text{Bernoulli}(10^{-10})$:
  - $\text{Regret}_T = 0$ w/p 0.9999999999

Some points of confusion (cont'd)

Should we bound Regret_T with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret_T
  i. Bounds *typical* behavior of Regret_T
  ii. Leaves possibility of some rare but really bad Regret_T values
- Bound on $\mathbb{E}[\text{Regret}_T]$
  iii. Bounds *average* behavior of Regret_T
  iv. Leaves possibility that the behavior of $\mathbb{E}[\text{Regret}_T]$ is *atypical*
- E.g., $\text{Regret}_T \sim T^{100} \times \text{Bernoulli}(10^{-10})$:
  - $\text{Regret}_T = 0$ w/p 0.9999999999
  - $\mathbb{E}[\text{Regret}_T] = 10^{-10}\, T^{100}$ ($\approx 10^{20}$ when $T = 2$)

Some points of confusion (cont'd)

Should we bound Regret$_T$ with high probability (i.e., probability $\geq 1 - \delta$), or should we bound $\mathbb{E}[\text{Regret}_T]$?

- High probability bound on Regret$_T$
  i. Bounds *typical* behavior of Regret$_T$
  ii. Leaves possibility of some rare but really bad Regret$_T$ values
- Bound on $\mathbb{E}[\text{Regret}_T]$
  iii. Bounds *average* behavior of Regret$_T$
  iv. Leaves possibility that the behavior of $\mathbb{E}[\text{Regret}_T]$ is *atypical*
- E.g., Regret$_T \sim T^{100} \times$ Bernoulli(10^{-10}):
  - Regret$_T = 0$ w/p 0.9999999999
  - $\mathbb{E}[\text{Regret}_T] = 10^{-10}\, T^{100}$ ($\approx 10^{20}$ when $T = 2$)

Some points of confusion

Some points of confusion

Union bound trick: want to translate a bound that holds for *each* random variable X_i in a collection $X_1, \ldots, X_n$,

$$X_i \leq B(\delta) \quad \text{w/p } 1 - \delta, \quad \forall i$$

Some points of confusion

Union bound trick: want to translate a bound that holds for *each* random variable X_i in a collection $X_1, \ldots, X_n$,

$$X_i \leq B(\delta) \quad \text{w/p } 1 - \delta, \quad \forall i$$

to one that holds *uniformly* over a collection of random variables $X_1, \ldots, X_n$, i.e.,

$$X_i \leq B_n(\delta) \quad \forall i, \quad \text{w/p } 1 - \delta$$

Some points of confusion

Union bound trick: want to translate a bound that holds for *each* random variable X_i in a collection $X_1, \ldots, X_n$,

$$X_i \leq B(\delta) \quad \text{w/p } 1 - \delta, \quad \forall i$$

to one that holds *uniformly* over a collection of random variables $X_1, \ldots, X_n$, i.e.,

$$X_i \leq B_n(\delta) \quad \forall i, \quad \text{w/p } 1 - \delta$$

What is B_n?

Some points of confusion

Union bound trick: want to translate a bound that holds for *each* random variable X_i in a collection $X_1, \ldots, X_n$,

$$X_i \leq B(\delta) \quad \text{w/p } 1-\delta, \quad \forall i$$

to one that holds *uniformly* over a collection of random variables $X_1, \ldots, X_n$, i.e.,

$$X_i \leq B_n(\delta) \quad \forall i, \quad \text{w/p } 1-\delta$$

What is B_n?

i.e., $\mathbb{P}(X_i \leq B(\delta)) \geq 1-\delta, \ \forall i \quad \longrightarrow \quad \mathbb{P}(\forall i, X_i \leq B_n(\delta)) \geq 1-\delta$

$$\mathbb{P}(\forall i, X_i \leq B(\delta)) = 1 - \mathbb{P}(\exists i, X_i > B(\delta))$$

$$\text{U.B.} \rightarrow \geq 1 - \sum_{i=1}^{n} \mathbb{P}(X_i > B(\delta)) = 1 - \sum_{i=1}^{n} \underbrace{(1 - \overbrace{\mathbb{P}(X_i \leq B(\delta))}^{\geq 1-\delta})}_{\leq \delta} \geq 1 - n\delta$$

$$\delta \rightarrow \frac{\delta}{n} \Rightarrow \mathbb{P}(\forall i, X_i \leq B(\delta/n)) \geq 1 - n\frac{\delta}{n} = 1 - \delta$$

$$\Rightarrow B_n(\delta) = B(\delta/n) \ \checkmark$$

Today

- ✓ Feedback from last lecture
- ✓ Recap
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm
- UCB regret analysis

Upper Confidence Bound (UCB)

Upper Confidence Bound (UCB)

Intuition: maintain confidence intervals for mean of each arm and use them to focus exploration on most promising arms

Upper Confidence Bound (UCB)

Intuition: maintain confidence intervals for mean of each arm and use them to focus exploration on most promising arms

First: how to construct confidence intervals?

Upper Confidence Bound (UCB)

Intuition: maintain confidence intervals for mean of each arm and use them to focus exploration on most promising arms

First: how to construct confidence intervals?

Recall Hoeffding inequality:

Sample mean of N i.i.d. samples on $[0,1]$ satisfies

$$\left|\hat{\mu} - \mu\right| \leq \sqrt{\frac{\ln(2/\delta)}{2N}} \quad \text{w/p } 1 - \delta$$

Upper Confidence Bound (UCB)

Intuition: maintain confidence intervals for mean of each arm and use them to focus exploration on most promising arms

First: how to construct confidence intervals?

Recall Hoeffding inequality:

Sample mean of N i.i.d. samples on $[0,1]$ satisfies

$$\left|\hat{\mu} - \mu\right| \leq \sqrt{\frac{\ln(2/\delta)}{2N}} \quad \text{w/p } 1 - \delta$$

Worked for ETC b/c exploration phase was i.i.d., but in general the rewards from a given arm are *not* i.i.d. due to adaptivity of action selections

Constructing confidence intervals

Constructing confidence intervals

Notation:

Constructing confidence intervals

Notation:

Let $N_t^{(k)} = \sum_{\tau=0}^{t-1} 1_{\{a_\tau = k\}}$ be the number of times arm k is pulled before time t

Constructing confidence intervals

Notation:

Let $N_t^{(k)} = \sum_{\tau=0}^{t-1} 1_{\{a_\tau = k\}}$ be the number of times arm k is pulled before time t

Let $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{\tau=0}^{t-1} 1_{\{a_\tau = k\}} r_\tau$ be the sample mean reward of arm k up to time t

Constructing confidence intervals

Notation:

Let $N_t^{(k)} = \sum_{\tau=0}^{t-1} 1_{\{a_\tau = k\}}$ be the number of times arm k is pulled before time t

Let $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{\tau=0}^{t-1} 1_{\{a_\tau = k\}} r_\tau$ be the sample mean reward of arm k up to time t

So want Hoeffding to give us something like

$$\left| \hat{\mu}_t^{(k)} - \mu \right| \leq \sqrt{\frac{\ln(2/\delta)}{2N_t^{(k)}}} \text{ w/p } 1 - \delta$$

Constructing confidence intervals

Notation:

Let $N_t^{(k)} = \sum_{\tau=0}^{t-1} 1_{\{a_\tau = k\}}$ be the number of times arm k is pulled before time t

Let $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{\tau=0}^{t-1} 1_{\{a_\tau = k\}} r_\tau$ be the sample mean reward of arm k up to time t

So want Hoeffding to give us something like

$$\cancel{\left| \hat{\mu}_t^{(k)} - \mu \right| \le \sqrt{\frac{\ln(2/\delta)}{2N_t^{(k)}}} \text{ w/p } 1 - \delta}$$

But this is generally FALSE

(unless a_t chosen very simply, like exploration phase of ETC)

Constructing confidence intervals (cont'd)

The problem: Although $r_\tau \mid a_\tau = k$ is an i.i.d. draw from $\nu^{(k)}$,

Constructing confidence intervals (cont'd)

The problem: Although $r_\tau \mid a_\tau = k$ is an i.i.d. draw from $\nu^{(k)}$, (all *arm* indexing (k) now in superscripts; subscripts reserved for time index *t*)

Constructing confidence intervals (cont'd)

The problem: Although $r_\tau \mid a_\tau = k$ is an i.i.d. draw from $\nu^{(k)}$, (all *arm* indexing (k) now in <u>super</u>scripts; <u>sub</u>scripts reserved for time index t)

$\hat{\mu}_t^{(k)}$ is the sample mean of a random number $N_t^{(k)}$ of returns

Constructing confidence intervals (cont'd)

The problem: Although $r_\tau \mid a_\tau = k$ is an i.i.d. draw from $\nu^{(k)}$, (all *arm* indexing (k) now in superscripts; subscripts reserved for time index *t*)

$\hat{\mu}_t^{(k)}$ is the sample mean of a random number $N_t^{(k)}$ of returns

in general $N_t^{(k)}$ will depend on those returns themselves

Constructing confidence intervals (cont'd)

The problem: Although $r_\tau \mid a_\tau = k$ is an i.i.d. draw from $\nu^{(k)}$, (all *arm* indexing (k) now in superscripts; subscripts reserved for time index *t*)

$\hat{\mu}_t^{(k)}$ is the sample mean of a random number $N_t^{(k)}$ of returns

in general $N_t^{(k)}$ will depend on those returns themselves

(i.e., how often we select arm k depends on the historical returns of arm k)

Constructing confidence intervals (cont'd)

The problem: Although $r_\tau \mid a_\tau = k$ is an i.i.d. draw from $\nu^{(k)}$, (all *arm* indexing (k) now in superscripts; subscripts reserved for time index *t*)

$\hat{\mu}_t^{(k)}$ is the sample mean of a random number $N_t^{(k)}$ of returns

in general $N_t^{(k)}$ will depend on those returns themselves

(i.e., how often we select arm k depends on the historical returns of arm k)

Solution: First, imagine an infinite sequence of *hypothetical* i.i.d. draws from $\nu^{(k)}$:

$$\tilde{r}_0^{(k)}, \tilde{r}_1^{(k)}, \tilde{r}_2^{(k)}, \tilde{r}_3^{(k)}, \ldots$$

Constructing confidence intervals (cont'd)

The problem: Although $r_\tau \mid a_\tau = k$ is an i.i.d. draw from $\nu^{(k)}$, (all *arm* indexing (k) now in superscripts; subscripts reserved for time index *t*)

$\hat{\mu}_t^{(k)}$ is the sample mean of a random number $N_t^{(k)}$ of returns

in general $N_t^{(k)}$ will depend on those returns themselves

(i.e., how often we select arm k depends on the historical returns of arm k)

Solution: First, imagine an infinite sequence of *hypothetical* i.i.d. draws from $\nu^{(k)}$:

$$\tilde{r}_0^{(k)}, \tilde{r}_1^{(k)}, \tilde{r}_2^{(k)}, \tilde{r}_3^{(k)}, \ldots$$

Then we can think of every time we pull arm k, just pulling the next $\tilde{r}_i^{(k)}$ off this list,

Constructing confidence intervals (cont'd)

The problem: Although $r_\tau \mid a_\tau = k$ is an i.i.d. draw from $\nu^{(k)}$, (all *arm* indexing (k) now in superscripts; subscripts reserved for time index t)

$\hat{\mu}_t^{(k)}$ is the sample mean of a random number $N_t^{(k)}$ of returns

in general $N_t^{(k)}$ will depend on those returns themselves

(i.e., how often we select arm k depends on the historical returns of arm k)

Solution: First, imagine an infinite sequence of *hypothetical* i.i.d. draws from $\nu^{(k)}$:

$$\tilde{r}_0^{(k)}, \tilde{r}_1^{(k)}, \tilde{r}_2^{(k)}, \tilde{r}_3^{(k)}, \ldots$$

Then we can think of every time we pull arm k, just pulling the next $\tilde{r}_i^{(k)}$ off this list,

i.e., $r_\tau \mid a_\tau = k$ to simply equal to $\tilde{r}_{N_\tau^k}^{(k)}$, and hence $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{i=0}^{N_t^{(k)}-1} \tilde{r}_i^{(k)}$

Constructing confidence intervals (cont'd)

Recall: $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{i=0}^{N_t^{(k)}-1} \tilde{r}_i^{(k)}$

Constructing confidence intervals (cont'd)

Recall: $\hat{\mu}_t^{(k)} = \dfrac{1}{N_t^{(k)}} \displaystyle\sum_{i=0}^{N_t^{(k)}-1} \tilde{r}_i^{(k)}$ Now define: $\tilde{\mu}_n^{(k)} = \dfrac{1}{n} \displaystyle\sum_{i=0}^{n-1} \tilde{r}_i^{(k)}$ $\left(\Rightarrow \hat{\mu}_t^{(k)} = \tilde{\mu}_{N_t^{(k)}}^{(k)} \right)$

Constructing confidence intervals (cont'd)

Recall: $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{i=0}^{N_t^{(k)}-1} \tilde{r}_i^{(k)}$ Now define: $\tilde{\mu}_n^{(k)} = \frac{1}{n} \sum_{i=0}^{n-1} \tilde{r}_i^{(k)}$ $\left(\Rightarrow \hat{\mu}_t^{(k)} = \tilde{\mu}_{N_t^{(k)}}^{(k)} \right)$

Now Hoeffding applies to $\tilde{\mu}_n^{(k)}$ because n fixed/nonrandom

Constructing confidence intervals (cont'd)

Recall: $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{i=0}^{N_t^{(k)}-1} \tilde{r}_i^{(k)}$ Now define: $\tilde{\mu}_n^{(k)} = \frac{1}{n} \sum_{i=0}^{n-1} \tilde{r}_i^{(k)}$ $\left(\Rightarrow \hat{\mu}_t^{(k)} = \tilde{\mu}_{N_t^{(k)}}^{(k)} \right)$

Now Hoeffding applies to $\tilde{\mu}_n^{(k)}$ because n fixed/nonrandom

and we know $\hat{\mu}_t^{(k)} = \tilde{\mu}_n^{(k)}$ for some $n \leq t$ (but which one is *random*)

Constructing confidence intervals (cont'd)

Recall: $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{i=0}^{N_t^{(k)}-1} \tilde{r}_i^{(k)}$ Now define: $\tilde{\mu}_n^{(k)} = \frac{1}{n} \sum_{i=0}^{n-1} \tilde{r}_i^{(k)}$ $\left(\Rightarrow \hat{\mu}_t^{(k)} = \tilde{\mu}_{N_t^{(k)}}^{(k)} \right)$

Now Hoeffding applies to $\tilde{\mu}_n^{(k)}$ because n fixed/nonrandom

and we know $\hat{\mu}_t^{(k)} = \tilde{\mu}_n^{(k)}$ for some $n \leq t$ (but which one is *random*)

Recall union bound in ETC analysis made Hoeffding hold simultaneously over $k \leq K$

Constructing confidence intervals (cont'd)

Recall: $\hat{\mu}_t^{(k)} = \frac{1}{N_t^{(k)}} \sum_{i=0}^{N_t^{(k)}-1} \tilde{r}_i^{(k)}$ Now define: $\tilde{\mu}_n^{(k)} = \frac{1}{n} \sum_{i=0}^{n-1} \tilde{r}_i^{(k)}$ $(\Rightarrow \hat{\mu}_t^{(k)} = \tilde{\mu}_{N_t^{(k)}}^{(k)})$

Now Hoeffding applies to $\tilde{\mu}_n^{(k)}$ because n fixed/nonrandom

and we know $\hat{\mu}_t^{(k)} = \tilde{\mu}_n^{(k)}$ for some $n \leq t$ (but which one is *random*)

Recall union bound in ETC analysis made Hoeffding hold simultaneously over $k \leq K$

Hoeffding + union bound over $n \leq t$:

$$\Rightarrow \mathbb{P}\left(\forall n \leq t, \, |\tilde{\mu}_n^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2n}\right) \geq 1 - \delta$$

Constructing confidence intervals (cont'd)

Hoeffding + union bound over $n \leq t$:

$$\Rightarrow \mathbb{P}\left(\forall n \leq t,\ |\tilde{\mu}_n^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2n}\right) \geq 1 - \delta$$

Constructing confidence intervals (cont'd)

Hoeffding + union bound over $n \leq t$:

$$\Rightarrow \mathbb{P}\left(\forall n \leq t, \, |\tilde{\mu}_n^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2n} \right) \geq 1 - \delta$$

But since in particular $N_t^{(k)} \leq t$, this immediately implies

$$\mathbb{P}\left(|\tilde{\mu}_{N_t^{(k)}}^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}} \right) \geq 1 - \delta$$

Constructing confidence intervals (cont'd)

Hoeffding + union bound over $n \leq t$:

$$\Rightarrow \mathbb{P}\left(\forall n \leq t, \, |\tilde{\mu}_n^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2n}\right) \geq 1 - \delta$$

But since in particular $N_t^{(k)} \leq t$, this immediately implies

$$\mathbb{P}\left(|\tilde{\mu}_{N_t^{(k)}}^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right) \geq 1 - \delta$$

And then since $\tilde{\mu}_{N_t^{(k)}}^{(k)} = \hat{\mu}_t^{(k)}$, we immediately get the kind of result we want:

$$\mathbb{P}\left(|\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right) \geq 1 - \delta$$

Constructing confidence intervals (cont'd)

Hoeffding + union bound over $n \leq t$:

$$\Rightarrow \mathbb{P}\left(\forall n \leq t, \, |\tilde{\mu}_n^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2n}\right) \geq 1 - \delta$$

But since in particular $N_t^{(k)} \leq t$, this immediately implies

$$\mathbb{P}\left(|\tilde{\mu}_{N_t^{(k)}}^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right) \geq 1 - \delta$$

And then since $\tilde{\mu}_{N_t^{(k)}}^{(k)} = \hat{\mu}_t^{(k)}$, we immediately get the kind of result we want:

$$\mathbb{P}\left(|\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right) \geq 1 - \delta$$

<u>Summary</u>: to deal with problem of non-i.i.d. rewards that enter into $\hat{\mu}_t^{(k)}$, we used rewards' *conditional* i.i.d. property along with a union bound to get Hoeffding bound that is wider by just a factor of t in the log term

Uniform confidence intervals

Uniform confidence intervals

So we have a valid $(1-\delta)$ confidence interval (CI) for $\mu^{(k)}$ at time t from last equation:

$$\mathbb{P}\left(|\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right) \geq 1-\delta,$$

i.e., $\left[\hat{\mu}_t^{(k)} - \sqrt{\ln(2t/\delta)/2N_t^{(k)}},\ \hat{\mu}_t^{(k)} + \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right]$

Uniform confidence intervals

So we have a valid $(1-\delta)$ confidence interval (CI) for $\mu^{(k)}$ at time t from last equation:

$$\mathbb{P}\left(|\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}} \right) \geq 1-\delta,$$

i.e., $\left[\hat{\mu}_t^{(k)} - \sqrt{\ln(2t/\delta)/2N_t^{(k)}},\ \hat{\mu}_t^{(k)} + \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right]$

Valid for any bandit algorithm!
Of independent statistical interest
for interpreting results

Uniform confidence intervals

So we have a valid $(1-\delta)$ confidence interval (CI) for $\mu^{(k)}$ at time t from last equation:

$$\mathbb{P}\left(|\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}} \right) \geq 1-\delta,$$

i.e., $\left[\hat{\mu}_t^{(k)} - \sqrt{\ln(2t/\delta)/2N_t^{(k)}},\ \hat{\mu}_t^{(k)} + \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right]$

Valid for any bandit algorithm!
Of independent statistical interest
for interpreting results

But analysis easier if CIs are *uniformly valid* over time t and arm k

Uniform confidence intervals

So we have a valid $(1-\delta)$ confidence interval (CI) for $\mu^{(k)}$ at time t from last equation:

$$\mathbb{P}\left(|\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right) \geq 1-\delta,$$

i.e., $\left[\hat{\mu}_t^{(k)} - \sqrt{\ln(2t/\delta)/2N_t^{(k)}},\ \hat{\mu}_t^{(k)} + \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right]$

Valid for any bandit algorithm!
Of independent statistical interest
for interpreting results

But analysis easier if CIs are *uniformly valid* over time t and arm k

By same argument as last two slides using a union bound over Hoeffding applied to all $\tilde{\mu}_n^{(k)}$ for $n \leq T$, and noting that $N_t^{(k)} \leq T$ for all $t < T$, we get:

Uniform confidence intervals

So we have a valid $(1-\delta)$ confidence interval (CI) for $\mu^{(k)}$ at time t from last equation:

$$\mathbb{P}\left(|\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}} \right) \geq 1-\delta,$$

i.e., $\left[\hat{\mu}_t^{(k)} - \sqrt{\ln(2t/\delta)/2N_t^{(k)}},\ \hat{\mu}_t^{(k)} + \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right]$

Valid for any bandit algorithm!
Of independent statistical interest
for interpreting results

But analysis easier if CIs are *uniformly valid* over time t and arm k

By same argument as last two slides using a union bound over Hoeffding applied to all $\tilde{\mu}_n^{(k)}$ for $n \leq T$, and noting that $N_t^{(k)} \leq T$ for all $t < T$, we get:

$$\mathbb{P}\left(\forall t < T,\ |\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2T/\delta)/2N_t^{(k)}} \right) \geq 1-\delta$$

Uniform confidence intervals

So we have a valid $(1-\delta)$ confidence interval (CI) for $\mu^{(k)}$ at time t from last equation:

$$\mathbb{P}\left(|\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right) \geq 1-\delta,$$

i.e., $\left[\hat{\mu}_t^{(k)} - \sqrt{\ln(2t/\delta)/2N_t^{(k)}},\ \hat{\mu}_t^{(k)} + \sqrt{\ln(2t/\delta)/2N_t^{(k)}}\right]$

Valid for any bandit algorithm!
Of independent statistical interest
for interpreting results

But analysis easier if CIs are *uniformly valid* over time t and arm k

By same argument as last two slides using a union bound over Hoeffding applied to all $\tilde{\mu}_n^{(k)}$ for $n \leq T$, and noting that $N_t^{(k)} \leq T$ for all $t < T$, we get:

$$\mathbb{P}\left(\forall t < T,\ |\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2T/\delta)/2N_t^{(k)}}\right) \geq 1-\delta$$

By same argument made in ETC analysis, union bound over K makes coverage uniform over k:

$$\mathbb{P}\left(\forall k \leq K, t < T,\ |\hat{\mu}_t^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(2TK/\delta)/2N_t^{(k)}}\right) \geq 1-\delta$$

Today

- Feedback from last lecture
- Recap
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm
- UCB regret analysis

Upper Confidence Bound (UCB) algorithm

Upper Confidence Bound (UCB) algorithm

For $t = 0,\ldots,T-1$:

Choose the arm with the highest upper confidence bound, i.e.,

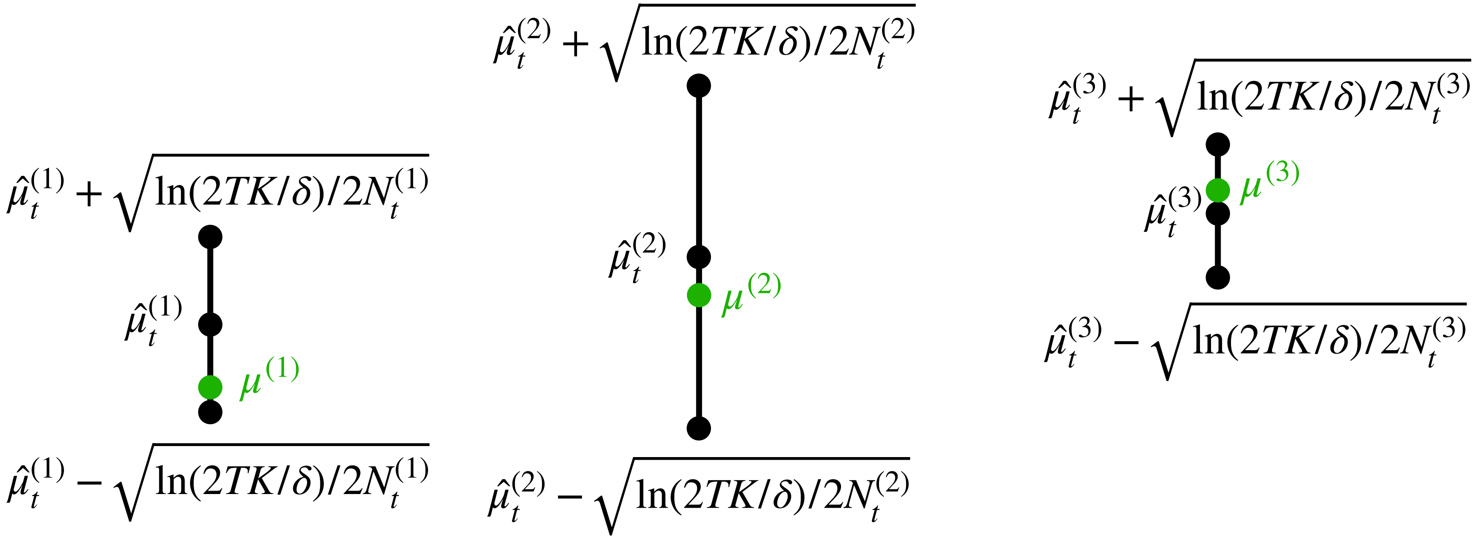
$$a_t = \arg\max_{k\in\{1,\ldots,K\}} \hat{\mu}_t^{(k)} + \sqrt{\ln(2TK/\delta)/2N_t^{(k)}}$$

Upper Confidence Bound (UCB) algorithm

For $t = 0, \ldots, T-1$:

Choose the arm with the highest upper confidence bound, i.e.,

$$a_t = \arg\max_{k \in \{1,\ldots,K\}} \hat{\mu}_t^{(k)} + \sqrt{\ln(2TK/\delta)/2N_t^{(k)}}$$

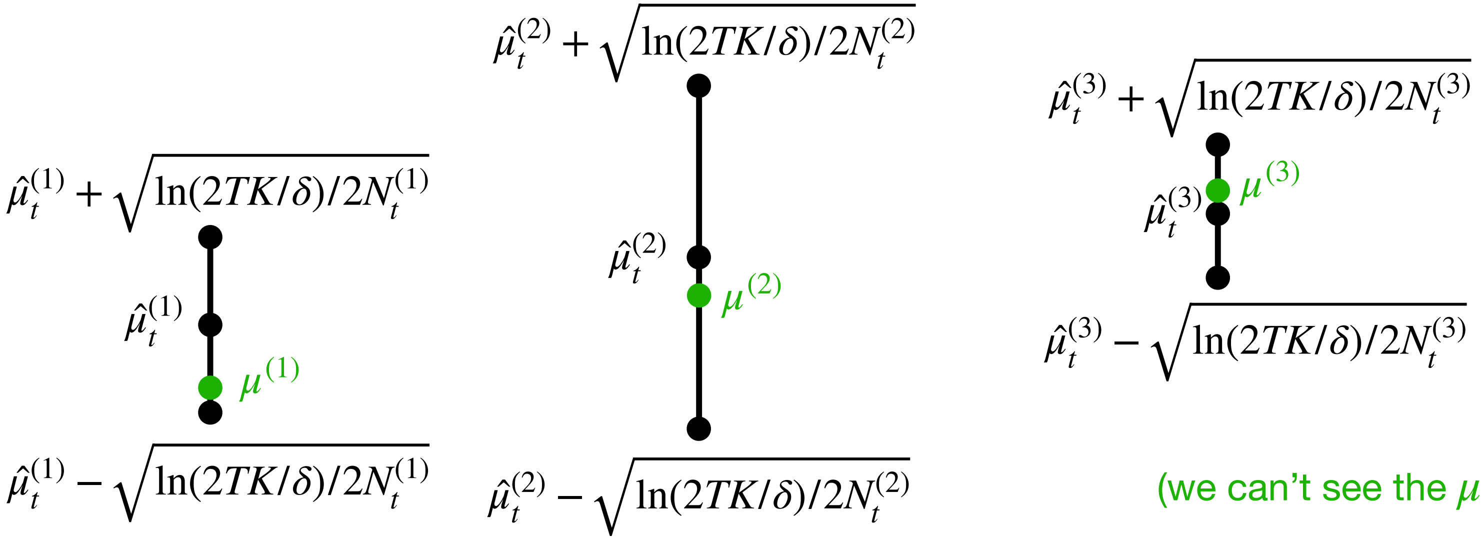


Upper Confidence Bound (UCB) algorithm

For $t = 0,\ldots,T-1$:

Choose the arm with the highest upper confidence bound, i.e.,

$$a_t = \arg\max_{k\in\{1,\ldots,K\}} \hat{\mu}_t^{(k)} + \sqrt{\ln(2TK/\delta)/2N_t^{(k)}}$$



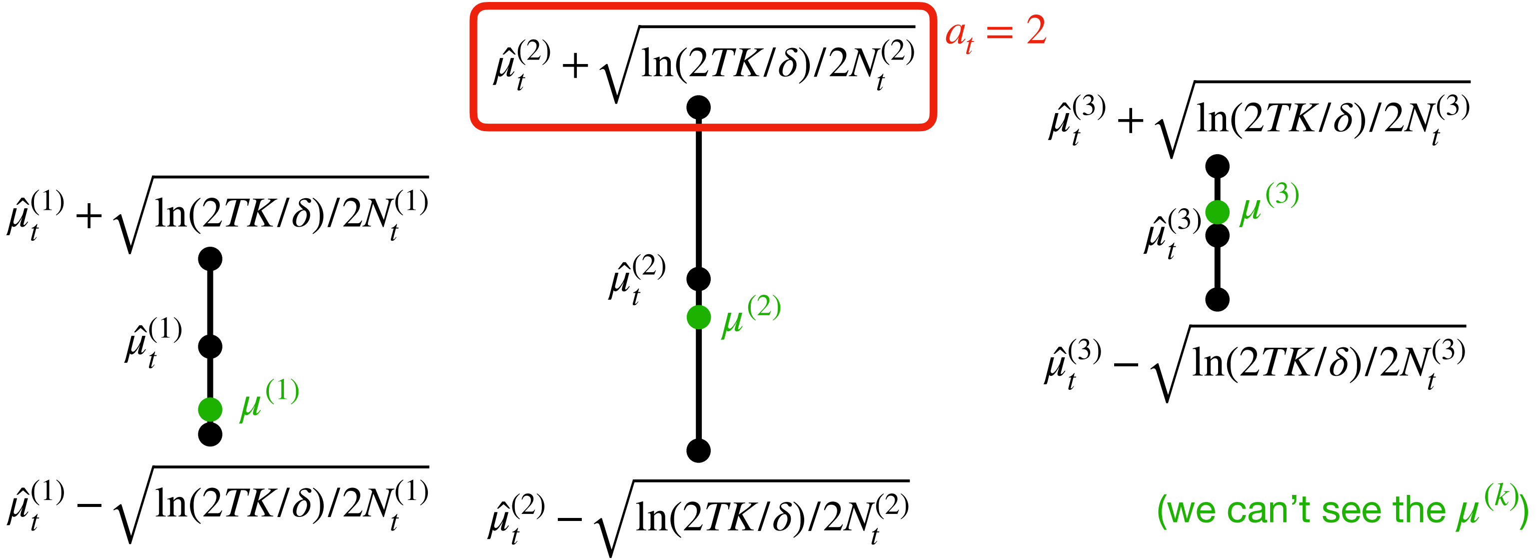
(we can't see the $\mu^{(k)}$)

Upper Confidence Bound (UCB) algorithm

For $t = 0,\ldots,T-1$:

Choose the arm with the highest upper confidence bound, i.e.,

$$a_t = \arg\max_{k \in \{1,\ldots,K\}} \hat{\mu}_t^{(k)} + \sqrt{\ln(2TK/\delta)/2N_t^{(k)}}$$



UCB Intuition: *optimism in the face of uncertainty*

Optimism in the face of uncertainty is an important principle in RL
It basically says to give each arm the benefit of the doubt, and basically act as if that arm is as good as it could plausibly be in choosing an action

UCB Intuition: *optimism in the face of uncertainty*

Optimism in the face of uncertainty is an important principle in RL
It basically says to give each arm the benefit of the doubt, and basically act as if that arm is as good as it could plausibly be in choosing an action

In UCB, this means constructing a CI (i.e., set of plausible values) for each $\mu^{(k)}$, and being greedy with respect to the upper bound of the CIs

UCB Intuition: *optimism in the face of uncertainty*

Optimism in the face of uncertainty is an important principle in RL

It basically says to give each arm the benefit of the doubt, and basically act as if that arm is as good as it could plausibly be in choosing an action

In UCB, this means constructing a CI (i.e., set of plausible values) for each $\mu^{(k)}$, and being greedy with respect to the upper bound of the CIs

Since each upper bound is $\hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$, this means when we select $a_t = k$, at least one of the two terms is large, i.e., either

UCB Intuition: *optimism in the face of uncertainty*

Optimism in the face of uncertainty is an important principle in RL
It basically says to give each arm the benefit of the doubt, and basically act as if that arm is as good as it could plausibly be in choosing an action

In UCB, this means constructing a CI (i.e., set of plausible values) for each $\mu^{(k)}$, and being greedy with respect to the upper bound of the CIs

Since each upper bound is $\hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$, this means when we select $a_t = k$, at least one of the two terms is large, i.e., either

1. $\sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$ large, i.e., we haven't explored arm k much (exploration)

UCB Intuition: *optimism in the face of uncertainty*

Optimism in the face of uncertainty is an important principle in RL

It basically says to give each arm the benefit of the doubt, and basically act as if that arm is as good as it could plausibly be in choosing an action

In UCB, this means constructing a CI (i.e., set of plausible values) for each $\mu^{(k)}$, and being greedy with respect to the upper bound of the CIs

Since each upper bound is $\hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$, this means when we select $a_t = k$, at least one of the two terms is large, i.e., either

1. $\sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$ large, i.e., we haven't explored arm k much (exploration)
2. $\hat{\mu}_t^{(k)}$ large, i.e., based on what we've seen so far, arm k is the best (exploitation)

UCB Intuition: *optimism in the face of uncertainty*

Optimism in the face of uncertainty is an important principle in RL

It basically says to give each arm the benefit of the doubt, and basically act as if that arm is as good as it could plausibly be in choosing an action

In UCB, this means constructing a CI (i.e., set of plausible values) for each $\mu^{(k)}$, and being greedy with respect to the upper bound of the CIs

Since each upper bound is $\hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$, this means when we select $a_t = k$, at least one of the two terms is large, i.e., either

1. $\sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$ large, i.e., we haven't explored arm k much (exploration)
2. $\hat{\mu}_t^{(k)}$ large, i.e., based on what we've seen so far, arm k is the best (exploitation)

Note that the exploration here is *adaptive*, i.e., focused on most promising arms

Today

- Feedback from last lecture
- Recap
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm
- UCB regret analysis

UCB Regret Analysis Strategy

UCB Regret Analysis Strategy

1. Bound regret at each time step

UCB Regret Analysis Strategy

1. Bound regret at each time step
2. Bound the sum of those bounds over time steps

$$B_t^{(k)} = \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$$

UCB regret at each time step

Recall $k^\star$ is optimal arm, so $\mu^{(k^\star)}$ is true best arm mean. Thus time step t regret is:

$$\mu^{(k^\star)} - \mu^{(a_t)} \leq \hat{\mu}_t^{(k^*)} + B_t^{(k^*)} - \mu^{(a_t)}$$

$$\leq \hat{\mu}_t^{(a_t)} + B_t^{(a_t)} - \mu^{(a_t)}$$

$$\leq B_t^{(a_t)} + B_t^{(a_t)} = 2 B_t^{(a_t)}$$

all w/p $\geq 1-\delta$, $\boxed{\text{uniformly over } t \leq T}$

Sum of UCB per-time-step regrets

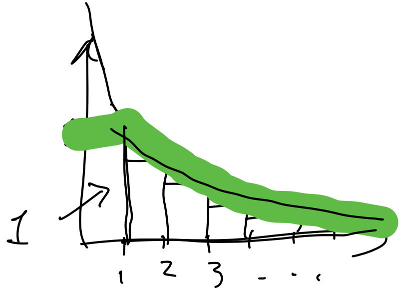
1. per-time-step regret bound $\mu^{(k^\star)} - \mu^{(a_t)} \le \sqrt{2\ln(2KT/\delta)/N_t^{(a_t)}} = 2B_t^{(a_t)}$ w/p $1-\delta$

2. $\text{Regret}_T \le \sum_{t=0}^{T-1}\sqrt{2\ln(2KT/\delta)/N_t^{(a_t)}} = \sqrt{2\ln(2KT/\delta)}\sum_{t=0}^{T-1}\sqrt{\frac{1}{N_t^{(a_t)}}}$

$$\sum_{t=0}^{T-1}\frac{1}{\sqrt{N_t^{(a_t)}}} = \sum_{t=0}^{T-1}\sum_{k=0}^{K}1_{\{a_t=k\}}\sqrt{\frac{1}{N_t^{(k)}}} = \sum_{k=1}^{K}\sum_{n=1}^{N_T^{(k)}}\sqrt{\frac{1}{n}} \le K\sum_{n=1}^{T}\frac{1}{\sqrt{n}}$$

$$\sum_{n=1}^{T}\frac{1}{\sqrt{n}} \le 1 + \int_1^T \frac{1}{\sqrt{x}}\,dx = 1 + 2\sqrt{x}\Big|_{x=1}^{x=T}$$

$$= 2\sqrt{T} - 1 \le 2\sqrt{T}$$



UCB total regret

UCB total regret

Finally, putting it all together, we get:

$$\text{Regret}_T \leq 2K\sqrt{T}\sqrt{2\ln(KT/\delta)} \qquad \text{w/p } 1-\delta$$

UCB total regret

Finally, putting it all together, we get:

$$\text{Regret}_T \leq 2K\sqrt{T}\sqrt{2\ln(KT/\delta)} \quad \text{w/p } 1-\delta$$

$$= \tilde{O}(\sqrt{T}) \quad \text{w/p } 1-\delta$$

UCB total regret

Finally, putting it all together, we get:

$$\text{Regret}_T \le 2K\sqrt{T}\sqrt{2\ln(KT/\delta)} \quad \text{w/p } 1-\delta$$

$$= \tilde{O}(\sqrt{T}) \quad \text{w/p } 1-\delta$$

In fact, a more sophisticated analysis can get: $\text{Regret}_T = \tilde{O}(\sqrt{KT}) \quad \text{w/p } 1-\delta$

Today

- Feedback from last lecture
- Recap
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm
- UCB regret analysis

Today’s summary:

Today's summary:

Upper Confidence Bound (UCB) algorithm:

- Uses uncertainty quantification *inside* algorithm
- Applies principle of *optimism in the face of uncertainty* (OFU)
- Adaptive exploration
- Achieves regret of $\tilde{O}(\sqrt{TK})$

Today's summary:

Upper Confidence Bound (UCB) algorithm:

- Uses uncertainty quantification *inside* algorithm
- Applies principle of *optimism in the face of uncertainty* (OFU)
- Adaptive exploration
- Achieves regret of $\tilde{O}(\sqrt{TK})$

Next time:

- Can't do better than $\tilde{\Omega}(\sqrt{T})$ regret in worst case
- Instance-dependent regret is another way to analyze regret

Today's summary:

Upper Confidence Bound (UCB) algorithm:

- Uses uncertainty quantification *inside* algorithm
- Applies principle of *optimism in the face of uncertainty* (OFU)
- Adaptive exploration
- Achieves regret of $\tilde{O}(\sqrt{TK})$

Next time:

- Can't do better than $\tilde{\Omega}(\sqrt{T})$ regret in worst case
- Instance-dependent regret is another way to analyze regret

1-minute feedback form: https://bit.ly/3RHtIxy