Bandits: Upper Confidence Bound Algorithm

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2023

Today

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

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2.

Today

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

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: UCB does better than a rate of $T^{2/3}$

Today

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

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

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

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

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 $\nu^{(k)}$, (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., $r_\tau \mid a_\tau = k$ 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: $\hat{\mu}_t^{(k)} = \dfrac{1}{N_t^{(k)}} \sum_{i=0}^{N_t^{(k)}-1} \tilde{r}_i^{(k)}$ Now define: $\tilde{\mu}_n^{(k)} = \dfrac{1}{n} \sum_{i=0}^{n-1} \tilde{r}_i^{(k)}$ $\left(\Rightarrow \hat{\mu}_t^{(k)} = \tilde{\mu}_{N_t^{(k)}}^{(k)} \right)$

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*)

Can anyone suggest a strategy for getting a bound for $|\hat{\mu}_t^{(k)} - \mu^{(k)}|$?

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 $n \leq t$:

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

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:

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

Summary: 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

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

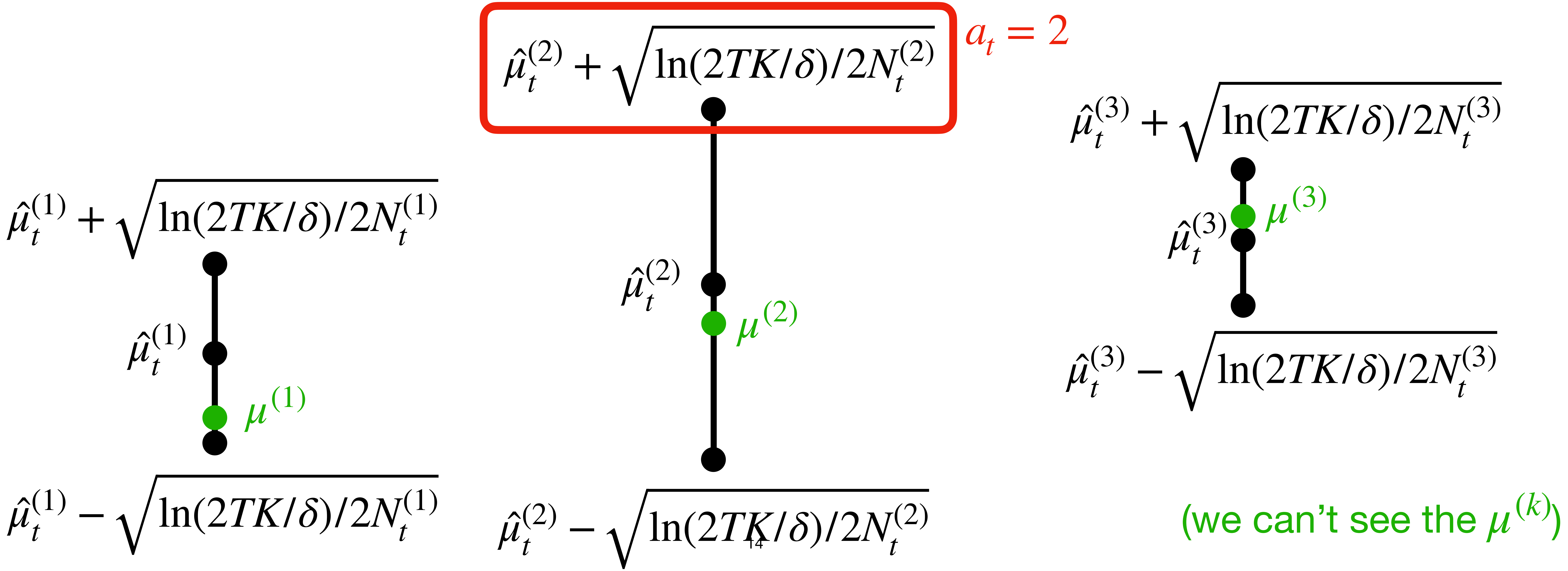
- ✓ Feedback from last lecture
- ✓ Recap
- ✓ Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm
- UCB regret analysis

Upper Confidence Bound (UCB) algorithm

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

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

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

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:

$\mu^{(k^\star)} - \mu^{(a_t)} \leq \hat{\mu}_t^{(k^\star)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k^\star)}} - \mu^{(a_t)}$ (CI coverage on arm $k^\star$) Next step?

$\leq \hat{\mu}_t^{(a_t)} + \sqrt{\ln(2KT/\delta)/2N_t^{(a_t)}} - \mu^{(a_t)}$ (a_t maximizes UCB by definition)

$\leq \sqrt{\ln(2KT/\delta)/2N_t^{(a_t)}} + \sqrt{\ln(2KT/\delta)/2N_t^{(a_t)}}$ (CI coverage on arm a_t)

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

all lines above hold simultaneously for all t w/p $1-\delta$ because of *uniform* Hoeffding

Sum of UCB per-time-step regrets

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

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

$$\sum_{t=0}^{T-1} \sqrt{\frac{1}{N_t^{(a_t)}}} = \sum_{t=0}^{T-1} \sum_{k=1}^{K} 1_{\{a_t=k\}} \sqrt{\frac{1}{N_t^{(k)}}} = \sum_{k=1}^{K} \sum_{n=1}^{N_T^{(k)}} \sqrt{\frac{1}{n}} \leq K \sum_{n=1}^{T} \sqrt{\frac{1}{n}} \leq 2K\sqrt{T}$$

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

UCB total regret

Finally, putting it all together, we get:

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

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

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

Can we do better than $\Omega(\sqrt{T})$ regret?

Short answer: no

But how can we know that?

A *lower bound* on the achievable regret

So far we our theoretical analysis has always considered a fixed algorithm and analyzed it (by deriving a regret upper bound with high probability)

To get a lower bound, we would need to consider what regret could be achieved by *any* algorithm, and show it can't be better than some rate

Intuition for lower bound

1. CLT tells us that with T i.i.d. samples from a distribution ν, we can only learn ν's mean μ to within $\Omega(1/\sqrt{T})$
2. Then since in a bandit, we get at most T samples total, certainly we can't learn any of the arm means better than to within $\Omega(1/\sqrt{T})$
3. This means that if an arm $\tilde{k}$ is about $1/\sqrt{T}$ away from the best arm $k^\star$, then at no point during the bandit can we confidently tell them apart
4. Thus, we should expect to sample $\tilde{k}$ roughly as often as $k^\star$, which is at best roughly $T/2$ times (if we ignore any other arms)
5. Finally, since the regret incurred each time we pull arm $\tilde{k}$ is $1/\sqrt{T}$, and we pull it $T/2$ times, we get a regret lower bound of $(1/\sqrt{T}) \times T/2 = \Omega(\sqrt{T})$

Today

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

Summary:

Upper Confidence Bound (UCB) algorithm:

- Uses uncertainty quantification *inside* algorithm
- Performs adaptive exploration via the principle of *optimism in the face of uncertainty* (OFU)
- Achieves regret of $\tilde{O}(\sqrt{TK})$
- A regret lower-bound exists that says one can't do better than $\Omega(T)$ regret

Attendance:

bit.ly/3RcTC9T

Feedback:

bit.ly/3RHtlxy