Bandits: Bayesian Bandits and Thompson Sampling

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning
Fall 2022

Today

- Feedback from last lecture
- Recap
- Instance-dependent regret of UCB
- Bayesian bandit
- Thompson sampling

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. Just a few people filled out form, and net zero on the pace

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2. Just a few people filled out form, and net zero on the pace
3. Added slide numbers

Today

- Feedback from last lecture
- Recap
- Instance-dependent regret of UCB
- Bayesian bandit
- Thompson sampling

Recap

Recap

- Pure greedy, pure exploration, ETC, ε-greedy achieve suboptimal worst-case regret

Recap

- Pure greedy, pure exploration, ETC, ε-greedy achieve suboptimal worst-case regret
- UCB uses Optimism in the Face of Uncertainty (OFU) principle and achieves *optimal* rate $\tilde{O}(\sqrt{T})$ of worst-case regret

Recap

- Pure greedy, pure exploration, ETC, ε-greedy achieve suboptimal worst-case regret
- UCB uses Optimism in the Face of Uncertainty (OFU) principle and achieves *optimal* rate $\tilde{O}(\sqrt{T})$ of worst-case regret
- Instance-dependent regret should be more informative than worst-case regret, but we haven't actually bounded it for UCB yet

Today

- Feedback from last lecture
- Recap
- Instance-dependent regret of UCB
- Bayesian bandit
- Thompson sampling

Instance-dependent regret for UCB: strategy

(Reminder from last time)

Instance-dependent regret for UCB: strategy

(Reminder from last time)

1. Now that we can incorporate information about the $\mu^{(k)}$, we'll try to precisely bound how often each suboptimal arm k is sampled, $N_T^{(k)}$

Instance-dependent regret for UCB: strategy

(Reminder from last time)

1. Now that we can incorporate information about the $\mu^{(k)}$, we'll try to precisely bound how often each suboptimal arm k is sampled, $N_T^{(k)}$
2. To do that, we'll use the uniform Hoeffding bound to see how often the UCB for $k^\star$ is (with high probability) higher than the UCB for k

Instance-dependent regret for UCB: strategy

(Reminder from last time)

1. Now that we can incorporate information about the $\mu^{(k)}$, we'll try to precisely bound how often each suboptimal arm k is sampled, $N_T^{(k)}$
2. To do that, we'll use the uniform Hoeffding bound to see how often the UCB for $k^\star$ is (with high probability) higher than the UCB for k
3. Then we'll multiply $N_T^{(k)}$ by the suboptimality of arm k, and sum this over the arms k to get the total regret

Instance-dependent regret for UCB

Instance-dependent regret for UCB

By uniform Hoeffding: w/p $\geq 1 - \delta$,

Instance-dependent regret for UCB

By uniform Hoeffding: w/p $\geq 1 - \delta$,

$\mathrm{UCB}_t^{(k^\star)} \geq \mu^{(k^\star)} = \mu^\star$,

Instance-dependent regret for UCB

By uniform Hoeffding: w/p $\geq 1 - \delta$,

$\text{UCB}_t^{(k^\star)} \geq \mu^{(k^\star)} = \mu^\star$, and $\forall k$, $\text{UCB}_t^{(k)} := \hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$

$$:= \hat{\mu}_t^{(k)} + B_t^{(k)}$$

Instance-dependent regret for UCB

By uniform Hoeffding: w/p $\geq 1 - \delta$,

$\mathrm{UCB}_t^{(k^\star)} \geq \mu^{(k^\star)} = \mu^\star$, and $\forall k$, $\mathrm{UCB}_t^{(k)} := \hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$

$$:= \hat{\mu}_t^{(k)} + B_t^{(k)} \leq \mu^{(k)} + 2B_t^{(k)}$$

Instance-dependent regret for UCB

By uniform Hoeffding: w/p $\geq 1 - \delta$,

$\text{UCB}_t^{(k^\star)} \geq \mu^{(k^\star)} = \mu^\star$, and $\forall k$, $\text{UCB}_t^{(k)} := \hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$

$:= \hat{\mu}_t^{(k)} + B_t^{(k)} \leq \mu^{(k)} + 2B_t^{(k)}$

Denote $g_k := \mu^\star - \mu^{(k)}$ the *gap* between the best arm and arm k's mean

Instance-dependent regret for UCB

By uniform Hoeffding: w/p $\geq 1 - \delta$,

$\text{UCB}_t^{(k^\star)} \geq \mu^{(k^\star)} = \mu^\star$, and $\forall k$, $\text{UCB}_t^{(k)} := \hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$

$$:= \hat{\mu}_t^{(k)} + B_t^{(k)} \leq \mu^{(k)} + 2B_t^{(k)}$$

Denote $g_k := \mu^\star - \mu^{(k)}$ the *gap* between the best arm and arm k's mean

$\Rightarrow$ if $B_t^{(k)} < g_k/2$, then $\text{UCB}_t^{(k^\star)} > \text{UCB}_t^{(k)}$

Instance-dependent regret for UCB

By uniform Hoeffding: w/p $\geq 1 - \delta$,

$\text{UCB}_t^{(k^\star)} \geq \mu^{(k^\star)} = \mu^\star$, and $\forall k$, $\text{UCB}_t^{(k)} := \hat{\mu}_t^{(k)} + \sqrt{\ln(2KT/\delta)/2N_t^{(k)}}$

$:= \hat{\mu}_t^{(k)} + B_t^{(k)} \leq \mu^{(k)} + 2B_t^{(k)}$

Denote $g_k := \mu^\star - \mu^{(k)}$ the *gap* between the best arm and arm k's mean

$\Rightarrow$ if $B_t^{(k)} < g_k/2$, then $\text{UCB}_t^{(k^\star)} > \text{UCB}_t^{(k)}$

When is $B_t^{(k)} < g_k/2$?

$$\sqrt{\ln(2KT/\delta)/2N_t^{(k)}} < \frac{g_k}{2}$$

$$\ln(2KT/\delta)/2N_t^{(k)} < \frac{g_k^2}{4}$$

$$2\ln(2KT/\delta)/g_k^2 < N_t^{(k)}$$

Instance-dependent regret for UCB (cont'd)

From last slide: w/p $\geq 1 - \delta$, $\forall t, k$ such that $N_t^{(k)} > 2\ln(2KT/\delta)/g_k^2$,

Instance-dependent regret for UCB (cont'd)

From last slide: w/p $\geq 1 - \delta$, $\forall t, k$ such that $N_t^{(k)} > 2\ln(2KT/\delta)/g_k^2$,

$\text{UCB}_t^{(k^\star)} > \text{UCB}_t^{(k)}$

Instance-dependent regret for UCB (cont'd)

From last slide: w/p $\geq 1 - \delta$, $\forall t, k$ such that $N_t^{(k)} > 2\ln(2KT/\delta)/g_k^2$,

$\mathrm{UCB}_t^{(k^\star)} > \mathrm{UCB}_t^{(k)}$ (arm k not pulled at time t)

Instance-dependent regret for UCB (cont'd)

From last slide: w/p $\geq 1 - \delta$, $\forall t, k$ such that $N_t^{(k)} > 2\ln(2KT/\delta)/g_k^2$,

$\mathrm{UCB}_t^{(k^\star)} > \mathrm{UCB}_t^{(k)}$ (arm k not pulled at time t) $\Rightarrow$ $1_{\{a_t = k\}} = 0$

Instance-dependent regret for UCB (cont'd)

From last slide: w/p $\geq 1 - \delta$, $\forall t, k$ such that $N_t^{(k)} > 2\ln(2KT/\delta)/g_k^2$,

$\text{UCB}_t^{(k^\star)} > \text{UCB}_t^{(k)}$ (arm k not pulled at time t) $\Rightarrow \quad \mathbb{1}_{\{a_t = k\}} = 0$

$$\text{Regret}_T = \sum_{k=1}^{K} (\mu^\star - \mu^{(k)}) N_T^{(k)}$$

$$N_T^{(k)} = \sum_{t=0}^{T-1} \mathbb{1}_{\{a_t = k\}} = \sum_{t=0}^{T-1} \Big(\mathbb{1}_{\{a_t = k\}} \mathbb{1}_{\{N_t^{(k)} \leq 2\ln(2KT/\delta)/g_k^2\}} + \overbrace{\mathbb{1}_{\{a_t=k\}} \mathbb{1}_{\{N_t^{(k)} > 2\ln(2KT/\delta)/g_k^2\}}}^{=0 \text{ w/p } 1-\delta} \Big)$$

$$\leq 2\ln(2KT/\delta)/g_k^2 \quad \text{w/p} \geq 1-\delta$$

$$\text{Regret}_T \leq \sum_{k=1}^{K} \cancel{g_k} \frac{2\ln(2KT/\delta)}{g_k^{\cancel{2}}} = \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \quad \text{wp } 1-\delta$$

UCB regret with large g_k

$$\text{Regret}_T \le \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \ge 1-\delta$$

UCB regret with large g_k

$$\text{Regret}_T \le \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \ge 1-\delta$$

<u>Logarithmic</u> in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

UCB regret with large g_k

$$\text{Regret}_T \leq \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \geq 1-\delta$$

<u>Logarithmic</u> in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

But need to think about g_k to be sure

UCB regret with large g_k

$$\text{Regret}_T \leq \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \geq 1-\delta$$

Logarithmic in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

But need to think about g_k to be sure

When all g_k are large relative to $\sqrt{1/T}$:

UCB regret with large g_k

$$\text{Regret}_T \leq \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \geq 1-\delta$$

<u>Logarithmic</u> in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

But need to think about g_k to be sure

When all g_k are large relative to $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k}$$

UCB regret with large g_k

$$\text{Regret}_T \leq \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \geq 1-\delta$$

Logarithmic in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

But need to think about g_k to be sure

When all g_k are large relative to $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \leq K \frac{2\ln(2KT/\delta)}{\min_k g_k}$$

UCB regret with large g_k

$$\text{Regret}_T \leq \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \geq 1-\delta$$

<u>Logarithmic</u> in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

But need to think about g_k to be sure

When all g_k are large relative to $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \leq K\frac{2\ln(2KT/\delta)}{\min_k g_k} \ll 2K\ln(2KT/\delta)\sqrt{T}$$

UCB regret with large g_k

$$\text{Regret}_T \leq \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \geq 1-\delta$$

<u>Logarithmic</u> in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

But need to think about g_k to be sure

When all g_k are large relative to $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \leq K\frac{2\ln(2KT/\delta)}{\min_k g_k} \ll 2K\ln(2KT/\delta)\sqrt{T}$$

Instance-dependent bound indeed much better!

UCB regret with large g_k

$$\text{Regret}_T \leq \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \geq 1-\delta$$

<u>Logarithmic</u> in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

But need to think about g_k to be sure

When all g_k are large relative to $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \leq K\frac{2\ln(2KT/\delta)}{\min_k g_k} \ll 2K\ln(2KT/\delta)\sqrt{T}$$

Instance-dependent bound indeed much better!

<u>Idea</u>: CLT says that with T steps, we'll easily find best arm if it's better by $\gg \sqrt{1/T}$

UCB regret with large g_k

$$\text{Regret}_T \leq \sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \text{ w/p } \geq 1-\delta$$

<u>Logarithmic</u> in T: seems *much* better than worst-case lower-bound of $\Omega(\sqrt{T})$

But need to think about g_k to be sure

When all g_k are large relative to $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \leq K \frac{2\ln(2KT/\delta)}{\min_k g_k} \ll 2K\ln(2KT/\delta)\sqrt{T}$$

Instance-dependent bound indeed much better!

<u>Idea</u>: CLT says that with T steps, we'll easily find best arm if it's better by $\gg \sqrt{1/T}$ so basically we make relatively few mistakes

UCB regret with small g_k

UCB regret with small g_k

If $\min_k g_k$ is much smaller than $\sqrt{1/T}$:

UCB regret with small g_k

If $\min_k g_k$ is much smaller than $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k}$$

UCB regret with small g_k

If $\min_k g_k$ is much smaller than $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \geq \frac{2\ln(2KT/\delta)}{\min_k g_k}$$

UCB regret with small g_k

If $\min_k g_k$ is much smaller than $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \geq \frac{2\ln(2KT/\delta)}{\min_k g_k} \gg 2\ln(2KT/\delta)\sqrt{T}$$

UCB regret with small g_k

If $\min_k g_k$ is much smaller than $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \geq \frac{2\ln(2KT/\delta)}{\min_k g_k} \gg 2\ln(2KT/\delta)\sqrt{T}$$

Way worse than worst-case upper-bound of $\tilde{O}(\sqrt{T})$...

UCB regret with small g_k

If $\min_k g_k$ is much smaller than $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \geq \frac{2\ln(2KT/\delta)}{\min_k g_k} \gg 2\ln(2KT/\delta)\sqrt{T}$$

Way worse than worst-case upper-bound of $\tilde{O}(\sqrt{T})$...

But can match worst-case upper-bound by splitting arms into two groups:

UCB regret with small g_k

If $\min_k g_k$ is much smaller than $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \geq \frac{2\ln(2KT/\delta)}{\min_k g_k} \gg 2\ln(2KT/\delta)\sqrt{T}$$

Way worse than worst-case upper-bound of $\tilde{O}(\sqrt{T})$...

But can match worst-case upper-bound by splitting arms into two groups:

$\{k : g_k \leq \sqrt{1/T}\}$ and $\{k : g_k > \sqrt{1/T}\}$

UCB regret with small g_k

If $\min_k g_k$ is much smaller than $\sqrt{1/T}$:

$$\sum_{k=1}^{K} \frac{2\ln(2KT/\delta)}{g_k} \geq \frac{2\ln(2KT/\delta)}{\min_k g_k} \gg 2\ln(2KT/\delta)\sqrt{T}$$

Way worse than worst-case upper-bound of $\tilde{O}(\sqrt{T})$...

But can match worst-case upper-bound by splitting arms into two groups:

$$\{k : g_k \leq \sqrt{1/T}\} \quad \text{and} \quad \{k : g_k > \sqrt{1/T}\}$$

$$\text{Regret}_T = \sum_{\{k: g_k \leq \sqrt{1/T}\}} g_k N_T^{(k)} + \sum_{\{k: g_k > \sqrt{1/T}\}} g_k N_T^{(k)}$$

$$\leq K \frac{1}{\sqrt{T}} \max_k N_T^{(k)} \leq K\sqrt{T}$$

$$\leq \sum_{\{k: g_k > \sqrt{1/T}\}} \frac{2\ln(2KT/\delta)}{g_k} \leq 2K\ln(2KT/\delta)\sqrt{T}$$

$$\text{Regret}_T = \tilde{O}(\sqrt{T})$$

UCB regret with VERY small g_k

UCB regret with VERY small g_k

Of course, if $\nu^{(1)} = \cdots = \nu^{(K)}$ and hence $\mu^{(1)} = \cdots = \mu^{(K)}$, then $\text{Regret}_T = 0$…

UCB regret with VERY small g_k

Of course, if $\nu^{(1)} = \cdots = \nu^{(K)}$ and hence $\mu^{(1)} = \cdots = \mu^{(K)}$, then $\text{Regret}_T = 0\ldots$

neither bound is tight

UCB regret with VERY small g_k

Of course, if $\nu^{(1)} = \cdots = \nu^{(K)}$ and hence $\mu^{(1)} = \cdots = \mu^{(K)}$, then $\text{Regret}_T = 0\ldots$

neither bound is tight

$$\text{Regret}_T = \sum_{k=1}^{K} g_k N_t^{(k)}$$

UCB regret with VERY small g_k

Of course, if $\nu^{(1)} = \cdots = \nu^{(K)}$ and hence $\mu^{(1)} = \cdots = \mu^{(K)}$, then $\text{Regret}_T = 0\ldots$

neither bound is tight

$$\text{Regret}_T = \sum_{k=1}^{K} g_k N_t^{(k)} \leq \max_k g_k \sum_{k=1}^{K} N_t^{(k)}$$

UCB regret with VERY small g_k

Of course, if $\nu^{(1)} = \cdots = \nu^{(K)}$ and hence $\mu^{(1)} = \cdots = \mu^{(K)}$, then $\text{Regret}_T = 0$…

neither bound is tight

$$\text{Regret}_T = \sum_{k=1}^{K} g_k N_t^{(k)} \leq \max_k g_k \sum_{k=1}^{K} N_t^{(k)} = T \max_k g_k$$

UCB regret with VERY small g_k

Of course, if $\nu^{(1)} = \cdots = \nu^{(K)}$ and hence $\mu^{(1)} = \cdots = \mu^{(K)}$, then $\text{Regret}_T = 0\ldots$

neither bound is tight

$$\text{Regret}_T = \sum_{k=1}^{K} g_k N_t^{(k)} \leq \max_k g_k \sum_{k=1}^{K} N_t^{(k)} = T \max_k g_k$$

Tighter than other bounds when $\max_k g_k \ll \frac{\ln(T)}{T}$, i.e., for small g_k and/or small T

UCB regret with VERY small g_k

Of course, if $\nu^{(1)} = \cdots = \nu^{(K)}$ and hence $\mu^{(1)} = \cdots = \mu^{(K)}$, then $\text{Regret}_T = 0\ldots$

neither bound is tight

$$\text{Regret}_T = \sum_{k=1}^{K} g_k N_t^{(k)} \leq \max_k g_k \sum_{k=1}^{K} N_t^{(k)} = T \max_k g_k$$

Tighter than other bounds when $\max_k g_k \ll \frac{\ln(T)}{T}$, i.e., for small g_k and/or small T

Reasonable to expect Regret_T to scale like T times worst arm regret for *any algorithm* when it's too hard to distinguish the arms!

UCB regret with VERY small g_k

Of course, if $\nu^{(1)} = \cdots = \nu^{(K)}$ and hence $\mu^{(1)} = \cdots = \mu^{(K)}$, then $\text{Regret}_T = 0\ldots$

neither bound is tight

$$\text{Regret}_T = \sum_{k=1}^{K} g_k N_t^{(k)} \leq \max_k g_k \sum_{k=1}^{K} N_t^{(k)} = T \max_k g_k$$

Tighter than other bounds when $\max_k g_k \ll \frac{\ln(T)}{T}$, i.e., for small g_k and/or small T

Reasonable to expect Regret_T to scale like T times worst arm regret for *any algorithm* when it's too hard to distinguish the arms!

Summary: instance-dependent analysis gives more nuanced bounds on regret

Questions about UCB

Questions about UCB

1. Can we get rid of T in the algorithm so we don't have to know the time horizon?

Questions about UCB

1. Can we get rid of T in the algorithm so we don't have to know the time horizon?

   Yes: a more careful analysis allows to essentially replace T with t.

Questions about UCB

1. Can we get rid of T in the algorithm so we don't have to know the time horizon?

   Yes: a more careful analysis allows to essentially replace T with t.

2. How to choose δ, since it impacts the algorithm <u>and</u> the regret bound?

Questions about UCB

1. Can we get rid of T in the algorithm so we don't have to know the time horizon?

    Yes: a more careful analysis allows to essentially replace T with t.

2. How to choose δ, since it impacts the algorithm <u>and</u> the regret bound?

    No satisfying answer that I know of to this.

Questions about UCB

1. Can we get rid of T in the algorithm so we don't have to know the time horizon?

   Yes: a more careful analysis allows to essentially replace T with t.

2. How to choose δ, since it impacts the algorithm and the regret bound?

   No satisfying answer that I know of to this.

3. What if we have prior information about the arms before collecting the data?

Questions about UCB

1. Can we get rid of T in the algorithm so we don't have to know the time horizon?

   Yes: a more careful analysis allows to essentially replace T with t.

2. How to choose δ, since it impacts the algorithm <u>and</u> the regret bound?

   No satisfying answer that I know of to this.

3. What if we have prior information about the arms before collecting the data?

   There are heuristics for incorporating such information into UCB, but no single obvious and natural way to do so; Thompson sampling will though!

Questions about UCB

1. Can we get rid of T in the algorithm so we don't have to know the time horizon?
   Yes: a more careful analysis allows to essentially replace T with t.
2. How to choose δ, since it impacts the algorithm <u>and</u> the regret bound?
   No satisfying answer that I know of to this.
3. What if we have prior information about the arms before collecting the data?
   There are heuristics for incorporating such information into UCB, but no single obvious and natural way to do so; Thompson sampling will though!
4. OFU principle seems reasonable, but why does it work?

Questions about UCB

1. Can we get rid of T in the algorithm so we don't have to know the time horizon?
   Yes: a more careful analysis allows to essentially replace T with t.
2. How to choose δ, since it impacts the algorithm <u>and</u> the regret bound?
   No satisfying answer that I know of to this.
3. What if we have prior information about the arms before collecting the data?
   There are heuristics for incorporating such information into UCB, but no single obvious and natural way to do so; Thompson sampling will though!
4. OFU principle seems reasonable, but why does it work?
   We will try to answer this today.

Today

- Feedback from last lecture
- Recap
- Instance-dependent regret of UCB
- Bayesian bandit
- Thompson sampling

Bayesian bandit

Bayesian bandit

A Bayesian bandit augments the bandit environment we've been working in so far with a prior distribution on the unknown reward distributions: $\pi(\nu^{(1)}, \ldots, \nu^{(K)})$

Bayesian bandit

A Bayesian bandit augments the bandit environment we've been working in so far with a prior distribution on the unknown reward distributions: $\pi(\nu^{(1)}, \ldots, \nu^{(K)})$

E.g., in a Bernoulli bandit, each $\nu^{(k)}$ is entirely characterized by its mean $\mu^{(k)} = \mathbb{P}_{r \sim \nu^{(k)}}(r = 1)$, so a prior on the $\nu^{(k)}$ is equivalent to a prior on the $\mu^{(k)}$

Bayesian bandit

A Bayesian bandit augments the bandit environment we've been working in so far with a prior distribution on the unknown reward distributions: $\pi(\nu^{(1)}, \ldots, \nu^{(K)})$

E.g., in a Bernoulli bandit, each $\nu^{(k)}$ is entirely characterized by its mean $\mu^{(k)} = \mathbb{P}_{r \sim \nu^{(k)}}(r = 1)$, so a prior on the $\nu^{(k)}$ is equivalent to a prior on the $\mu^{(k)}$

One such prior, since all the $\mu^{(k)}$ are bounded between 0 and 1, is the prior that is *Uniform* on the unit hypercube, i.e., $(\mu^{(1)}, \ldots, \mu^{(K)}) =: \boldsymbol{\mu} \sim \text{Uniform}([0,1]^K)$

Bayesian bandit

A Bayesian bandit augments the bandit environment we've been working in so far with a prior distribution on the unknown reward distributions: $\pi(\nu^{(1)}, \ldots, \nu^{(K)})$

E.g., in a Bernoulli bandit, each $\nu^{(k)}$ is entirely characterized by its mean $\mu^{(k)} = \mathbb{P}_{r \sim \nu^{(k)}}(r = 1)$, so a prior on the $\nu^{(k)}$ is equivalent to a prior on the $\mu^{(k)}$

One such prior, since all the $\mu^{(k)}$ are bounded between 0 and 1,
is the prior that is *Uniform* on the unit hypercube, i.e.,
$(\mu^{(1)}, \ldots, \mu^{(K)}) =: \boldsymbol{\mu} \sim \mathrm{Uniform}([0,1]^K)$

Note that the Bernoulli bandit reduced everything unknown about the bandit system to a K-dimensional vector $\boldsymbol{\mu}$

Bayesian bandit

A Bayesian bandit augments the bandit environment we've been working in so far with a prior distribution on the unknown reward distributions: $\pi(\nu^{(1)}, \ldots, \nu^{(K)})$

E.g., in a Bernoulli bandit, each $\nu^{(k)}$ is entirely characterized by its mean $\mu^{(k)} = \mathbb{P}_{r \sim \nu^{(k)}}(r = 1)$, so a prior on the $\nu^{(k)}$ is equivalent to a prior on the $\mu^{(k)}$

One such prior, since all the $\mu^{(k)}$ are bounded between 0 and 1, is the prior that is *Uniform* on the unit hypercube, i.e., $(\mu^{(1)}, \ldots, \mu^{(K)}) =: \boldsymbol{\mu} \sim \text{Uniform}([0,1]^K)$

Note that the Bernoulli bandit reduced everything unknown about the bandit system to a K-dimensional vector $\boldsymbol{\mu}$

Without the Bernoulli assumption, we may need many more dimensions to describe the possible distributions, and hence have to define a much higher-dimensional prior

Bayesian Bernoulli bandit

Bayesian Bernoulli bandit

The really nice thing about a Bayesian bandit is that we can use Bayes rule to exactly characterize our uncertainty about the reward distributions at every time step.

Bayesian Bernoulli bandit

The really nice thing about a Bayesian bandit is that we can use Bayes rule to exactly characterize our uncertainty about the reward distributions at every time step.

Example: Bayesian Bernoulli bandit

Bayesian Bernoulli bandit

The really nice thing about a Bayesian bandit is that we can use Bayes rule to exactly characterize our uncertainty about the reward distributions at every time step.

Example: Bayesian Bernoulli bandit

1. At $t = 0$, we have no data, and the distribution of the reward distributions is simply given by the prior on the reward parameters $\boldsymbol{\mu}$:
$$\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$$

Bayesian Bernoulli bandit

The really nice thing about a Bayesian bandit is that we can use Bayes rule to exactly characterize our uncertainty about the reward distributions at every time step.

Example: Bayesian Bernoulli bandit

1. At $t = 0$, we have no data, and the distribution of the reward distributions is simply given by the prior on the reward parameters $\boldsymbol{\mu}$:
$$\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$$
($\mathbb{P}$ will sometimes denote a continuous density instead of a true probability, e.g., for $\boldsymbol{\mu} \sim$ Uniform$([0,1]^K)$, we would write $\mathbb{P}(\boldsymbol{\mu}) = 1_{\{0 \leq \mu^{(k)} \leq 1 \ \forall k\}}$)

Bayesian Bernoulli bandit (cont'd)

1. At $t = 0$, $\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$

Bayesian Bernoulli bandit (cont'd)

1. At $t = 0$, $\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$
2. At $t = 1$, we have one data point $r_0 \sim \text{Bernoulli}(\mu^{(a_0)})$, and the distribution of $\boldsymbol{\mu}$ gets updated via Bayes rule:

$$\mathbb{P}(\vec{\mu} \mid r_0, a_0) = \frac{\mathbb{P}(r_0, a_0 \mid \vec{\mu})\,\mathbb{P}(\vec{\mu})}{\mathbb{P}(r_0, a_0)} = \frac{\mathbb{P}(r_0, a_0 \mid \vec{\mu})\,\mathbb{P}(\vec{\mu})}{\int_{\tilde{\mu} \in [0,1]^K} \mathbb{P}(r_0, a_0 \mid \tilde{\mu})\,\mathbb{P}(\tilde{\mu})\,d\tilde{\mu}}$$

$$= \frac{\mathbb{P}(r_0 \mid a_0, \vec{\mu})\,\cancel{\mathbb{P}(a_0 \mid \vec{\mu})}\,\mathbb{P}(\vec{\mu})}{\int \quad \text{''} \quad (\tilde{\mu})}$$

$$\mathbb{P}(a_0 \mid \vec{\mu}) = \mathbb{P}(a_0)$$

$$= \frac{\mathbb{P}(r_0 \mid a_0, \vec{\mu})\,\mathbb{P}(\vec{\mu})}{\int \quad \text{''} \quad (\tilde{\mu})}$$

Bayesian Bernoulli bandit (cont'd)

1. At $t = 0$, $\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$
2. At $t = 1$, we have one data point $r_0 \sim$ Bernoulli$(\mu^{(a_0)})$, and the distribution of $\boldsymbol{\mu}$ gets updated via Bayes rule:

$$\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0) = \frac{\mathbb{P}(r_0 \mid a_0, \boldsymbol{\mu})\mathbb{P}(\boldsymbol{\mu})}{\int_{\tilde{\boldsymbol{\mu}} \in [0,1]^K} \mathbb{P}(r_0 \mid a_0, \tilde{\boldsymbol{\mu}})\mathbb{P}(\tilde{\boldsymbol{\mu}}) d\tilde{\boldsymbol{\mu}}} = \frac{(\mu^{(a_0)})^{r_0}(1-\mu^{(a_0)})^{1-r_0}\pi(\mu)}{\int (\quad)(\tilde{\mu})}$$

if $\pi(\mu) = 1$

$$= \frac{(\mu^{(a_0)})^{r_0}(1-\mu^{(a_0)})^{1-r_0}}{\int (\quad)(\tilde{\mu})} = 2(\mu^{(a_0)})^{r_0}(1-\mu^{(a_0)})^{1-r_0}$$

Bayesian Bernoulli bandit (cont'd)

1. At $t = 0$, $\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$
2. At $t = 1$, we have one data point $r_0 \sim$ Bernoulli$(\mu^{(a_0)})$, and the distribution of $\boldsymbol{\mu}$ gets updated via Bayes rule:

$$\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0) = 2(\mu^{(a_0)})^{r_0}(1 - \mu^{(a_0)})^{1-r_0}$$

Bayesian Bernoulli bandit (cont'd)

1. At $t = 0$, $\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$
2. At $t = 1$, we have one data point $r_0 \sim$ Bernoulli($\mu^{(a_0)}$), and the distribution of $\boldsymbol{\mu}$ gets updated via Bayes rule:
$$\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0) = 2(\mu^{(a_0)})^{r_0}(1 - \mu^{(a_0)})^{1-r_0}$$
3. At $t = 2$, we have another data point $r_1 \sim$ Bernoulli($\mu^{(a_1)}$), and we can update the distribution of $\boldsymbol{\mu}$ again via Bayes rule, treating $\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0)$ as the prior

Bayesian Bernoulli bandit (cont'd)

1. At $t = 0$, $\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$
2. At $t = 1$, we have one data point $r_0 \sim$ Bernoulli($\mu^{(a_0)}$), and the distribution of $\boldsymbol{\mu}$ gets updated via Bayes rule:
$$\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0) = 2(\mu^{(a_0)})^{r_0}(1 - \mu^{(a_0)})^{1-r_0}$$
3. At $t = 2$, we have another data point $r_1 \sim$ Bernoulli($\mu^{(a_1)}$), and we can update the distribution of $\boldsymbol{\mu}$ again via Bayes rule, treating $\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0)$ as the prior

$\vdots$

Bayesian Bernoulli bandit (cont'd)

1. At $t = 0$, $\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$
2. At $t = 1$, we have one data point $r_0 \sim$ Bernoulli($\mu^{(a_0)}$), and the distribution of $\boldsymbol{\mu}$ gets updated via Bayes rule:
$$\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0) = 2(\mu^{(a_0)})^{r_0}(1 - \mu^{(a_0)})^{1-r_0}$$
3. At $t = 2$, we have another data point $r_1 \sim$ Bernoulli($\mu^{(a_1)}$), and we can update the distribution of $\boldsymbol{\mu}$ again via Bayes rule, treating $\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0)$ as the prior

$\vdots$

Bayes rule at time step t gives us a distribution (called the posterior distribution)
$$\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0, r_1, a_1, \ldots, r_{t-1}, a_{t-1})$$
that exactly characterizes our uncertainty about $\boldsymbol{\mu}$.

Bayesian Bernoulli bandit (cont'd)

1. At $t = 0$, $\mathbb{P}(\boldsymbol{\mu}) = \pi(\boldsymbol{\mu})$
2. At $t = 1$, we have one data point $r_0 \sim$ Bernoulli($\mu^{(a_0)}$), and the distribution of $\boldsymbol{\mu}$ gets updated via Bayes rule:
$$\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0) = 2(\mu^{(a_0)})^{r_0}(1 - \mu^{(a_0)})^{1-r_0}$$
3. At $t = 2$, we have another data point $r_1 \sim$ Bernoulli($\mu^{(a_1)}$), and we can update the distribution of $\boldsymbol{\mu}$ again via Bayes rule, treating $\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0)$ as the prior

$\vdots$

Bayes rule at time step t gives us a distribution (called the posterior distribution)
$$\mathbb{P}(\boldsymbol{\mu} \mid r_0, a_0, r_1, a_1, \ldots, r_{t-1}, a_{t-1})$$
that exactly characterizes our uncertainty about $\boldsymbol{\mu}$. We can use this to choose a_t!

Today

- Feedback from last lecture
- Recap
- Instance-dependent regret of UCB
- Bayesian bandit
- Thompson sampling

Today's summary:

Today's summary:

Instance-dependent regret

- More descriptive than worst-case analysis
- UCB can do much better than worst-case $\Omega(\sqrt{T})$ regret in many cases

Today's summary:

Instance-dependent regret

- More descriptive than worst-case analysis
- UCB can do much better than worst-case $\Omega(\sqrt{T})$ regret in many cases

Bayesian bandit

- Adds an additional assumption of prior on reward distributions
- Bayes rule gives exact running uncertainty quantification for any algorithm

Today's summary:

Instance-dependent regret

- More descriptive than worst-case analysis
- UCB can do much better than worst-case $\Omega(\sqrt{T})$ regret in many cases

Bayesian bandit

- Adds an additional assumption of prior on reward distributions
- Bayes rule gives exact running uncertainty quantification for any algorithm

Thompson sampling

- Samples optimal arm from its (posterior) distribution
- Achieves excellent performance in practice

Today's summary:

Instance-dependent regret

- More descriptive than worst-case analysis
- UCB can do much better than worst-case $\Omega(\sqrt{T})$ regret in many cases

Bayesian bandit

- Adds an additional assumption of prior on reward distributions
- Bayes rule gives exact running uncertainty quantification for any algorithm

Thompson sampling

- Samples optimal arm from its (posterior) distribution
- Achieves excellent performance in practice

Next time:

- Gittins index

Today's summary:

Instance-dependent regret

- More descriptive than worst-case analysis
- UCB can do much better than worst-case $\Omega(\sqrt{T})$ regret in many cases

Bayesian bandit

- Adds an additional assumption of prior on reward distributions
- Bayes rule gives exact running uncertainty quantification for any algorithm

Thompson sampling

- Samples optimal arm from its (posterior) distribution
- Achieves excellent performance in practice

Next time:

- Gittins index

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