Bandits: Explore-Then-Commit, ε-greedy, UCB

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2023

Today

- Feedback from last lecture
- Recap
- Regret analysis of ETC
- ε-greedy algorithm
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm

Feedback from feedback forms

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

Today

- Feedback from last lecture
- Recap
- Regret analysis of ETC
- ε-greedy algorithm
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm

Recap

- Multi-armed bandits (or MAB or just bandits)
    - Online learning of a 1-state/1-horizon MDP
    - Exemplify exploration vs exploitation
    - Pure greedy & pure exploration achieve linear regret
    - Hoeffding's inequality
- Today: let's do better than linear regret!

Notes from last lecture

1. $$\text{Regret}_T = T\mu^\star - \sum_{t=0}^{T-1} \mu_{a_t} = \sum_{t=0}^{T-1} (\mu^\star - \mu_{a_t})$$

Expected regret at time t *given* that you chose arm a_t

2. Recall $\text{Regret}_T = \Omega(T)$, i.e., linear regret

$\Rightarrow$ for some $c > 0$ and T_0, $\text{Regret}_T \geq cT \quad \forall T \geq T_0$

3. Why is linear regret bad? $\Rightarrow$ *average regret* $:= \dfrac{\text{Regret}_T}{T} \not\to 0$

4. Hoeffding inequality: sample mean of N i.i.d. samples on $[0,1]$ satisfies

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

Explore-Then-Commit (ETC)

N_{e} = Number of explorations

Algorithm hyper parameter $N_{\mathsf{e}} < T/K$ (we assume T >> K)

For $k = 1,\ldots,K$: (Exploration phase)

Pull arm k N_{e} times to observe $\{r_i^{(k)}\}_{i=1}^{N_{\mathsf{e}}} \sim \nu_k$

Calculate arm k's empirical mean: $\hat{\mu}_k = \frac{1}{N_{\mathsf{e}}} \sum_{i=1}^{N_{\mathsf{e}}} r_i^{(k)}$

For $t = N_{\mathsf{e}}K, \ldots, (T-1)$: (Exploitation phase)

Pull the best empirical arm $a_t = \arg\max_{i \in [K]} \hat{\mu}_i$

Q: how to set N_{e}?

Regret Analysis Strategy

1. Calculate regret during exploration stage
2. Quantify error of arm mean estimates at end of exploration stage
3. Using step 2, calculate regret during exploitation stage

   (Actually, will only be able to upper-bound total regret in steps 1-3)

4. Minimize our upper-bound over N_{e}

Today

- Feedback from last lecture
- Recap
- Regret analysis of ETC
- ε-greedy algorithm
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm

Back to Regret Analysis of ETC

1. What is a bound for the regret during exploration stage?

$$\text{Regret}_{N_\text{e}K} \leq N_\text{e}K \text{ with probability } 1$$

2. Quantify error of arm mean estimates at end of exploration stage

a) Hoeffding $\Rightarrow \mathbb{P}\left(|\hat{\mu}_k - \mu_k| \leq \sqrt{\ln(2/\delta)/2N_\text{e}}\right) \geq 1 - \delta$

$\mathbb{P}(\forall k, A_1^c, \ldots, A_K^c) \geq 1 - \sum_{k=1}^{K} \mathbb{P}(A_k)$

$\Updownarrow$

b) Recall Union/Boole/Bonferroni bound: $\mathbb{P}(\text{any of } A_1, \ldots, A_K) \leq \sum_{k=1}^{K} \mathbb{P}(A_k)$

c) $\delta \to \delta/K$, Union bound with $A_k = \left\{|\hat{\mu}_k - \mu_k| > \sqrt{\ln(2K/\delta)/2N_\text{e}}\right\}$, and Hoeffding:

$$\Rightarrow \mathbb{P}\left(\forall k, |\hat{\mu}_k - \mu_k| \leq \sqrt{\ln(2K/\delta)/2N_\text{e}}\right) \geq 1 - \delta$$

Regret Analysis of ETC (cont'd)

2. Quantify error of arm mean estimates at end of exploration stage:

$$\mathbb{P}\left(\forall k,\ |\hat{\mu}_k - \mu_k| \le \sqrt{\ln(2K/\delta)/2N_{\mathsf{e}}}\right) \ge 1 - \delta$$

3. Using step 2, calculate regret during exploitation stage:

Denote (apparent) best arm after exploration stage by $\hat{k}$ and actual best arm by $k^\star$

$$\begin{aligned}
\text{regret at each step of exploitation phase} &= \mu_{k^\star} - \mu_{\hat{k}} \\
&= \mu_{k^\star} + (\hat{\mu}_{k^\star} - \hat{\mu}_{k^\star}) - \mu_{\hat{k}} + (\hat{\mu}_{\hat{k}} - \hat{\mu}_{\hat{k}}) \\
&= (\mu_{k^\star} - \hat{\mu}_{k^\star}) + (\hat{\mu}_{\hat{k}} - \mu_{\hat{k}}) + (\hat{\mu}_{k^\star} - \hat{\mu}_{\hat{k}}) \\
&\le \sqrt{\ln(2K/\delta)/2N_{\mathsf{e}}} + \sqrt{\ln(2K/\delta)/2N_{\mathsf{e}}} + 0 \qquad \text{w/p } 1 - \delta \\
&= \sqrt{2\ln(2K/\delta)/N_{\mathsf{e}}}
\end{aligned}$$

$$\Rightarrow \text{total regret during exploitation} \le T\sqrt{2\ln(2K/\delta)/N_{\mathsf{e}}} \qquad \text{w/p } 1 - \delta$$

Regret Analysis of ETC (cont'd)

4. From steps 1-3: with probability $1 - \delta$,

$$\text{Regret}_T \leq N_{\text{e}}K + T\sqrt{2\ln(2K/\delta)/N_{\text{e}}}$$

What's a choice of N_{e} that gives sublinear regret?

Any N_{e} so that $N_{\text{e}} \to \infty$ and $N_{\text{e}}/T \to 0$ (e.g., $N_{\text{e}} = \sqrt{T}$)

Minimize over N_{e}:

$$\text{optimal } N_{\text{e}} = \left(\frac{T\sqrt{\ln(2K/\delta)/2}}{K} \right)^{2/3}$$

(A bit more algebra to plug optimal N_{e} into Regret_T equation above)

$$\Rightarrow \text{Regret}_T \leq 3T^{2/3}(K\ln(2K/\delta)/2)^{1/3} = o(T)$$

Today

- Feedback from last lecture
- Recap
- Regret analysis of ETC
- ε-greedy algorithm
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm

ε-greedy

ETC very abrupt (huge difference between exploration and exploitation stages)

ε-greedy like a smoother version of ETC:
at *every* step, do pure greedy w/p $1 - \varepsilon$, and do pure exploration w/p ε

Initialize $\hat{\mu}_0 = \cdots = \hat{\mu}_K = 1$
For $t = 0,\ldots,T-1$:
Sample $E_t \sim$ Bernoulli(ε)
If $E_t = 1$, choose $a_t \sim$ Uniform($1,\ldots,K$) (pure explore)
If $E_t = 0$, choose $a_t = \arg\max_{k\in\{1,\ldots,K\}} \hat{\mu}_k$ (pure exploit)
Update $\hat{\mu}_{a_t}$

ε-greedy (cont'd)

Can also allow ε to depend on t; should it increase, decrease, or stay flat?

The more learned by time t, the less exploration needed at/after time t

It turns out that ε-greedy with $\varepsilon_t = \left(\frac{K \ln(t)}{t}\right)^{1/3}$ also achieves

$$\text{Regret}_t = \tilde{O}(t^{2/3}K^{1/3}),$$

where $\tilde{O}(\,\cdot\,)$ hides logarithmic factors

- Regret rate (ignoring log factors) is the same as ETC, but holds for all t, not just the full time horizon T
- Nothing in ε-greedy (including ε_t above) depends on T, so don't need to know horizon!

Today

- Feedback from last lecture
- Recap
- Regret analysis of ETC
- ε-greedy algorithm
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm

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

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

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

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:

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

<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

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

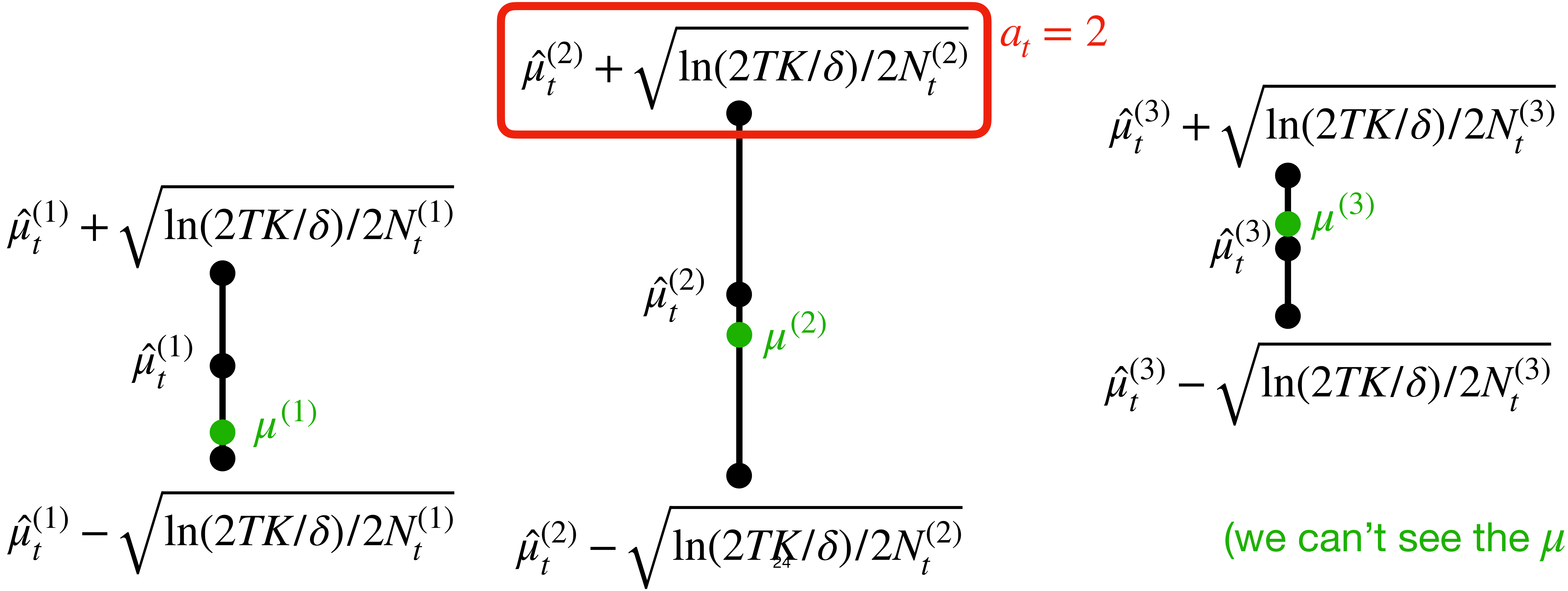
- Feedback from last lecture
- Recap
- Regret analysis of ETC
- ε-greedy algorithm
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm

Upper Confidence Bound (UCB) algorithm

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

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

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



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

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
- Regret analysis of ETC
- ε-greedy algorithm
- Confidence intervals for the arms
- Upper Confidence Bound (UCB) algorithm

Summary:

- ETC and ε-greedy, achieve sublinear regret $\tilde{O}(T^{2/3})$
- Hoeffding can be used to provide (uniform) bounds on the arm means
- UCB algorithm follows "optimism in the face of uncertainty" principle

Attendance:
bit.ly/3RcTC9T

Feedback:
bit.ly/3RHtlxy