Online Learning

Online Learning

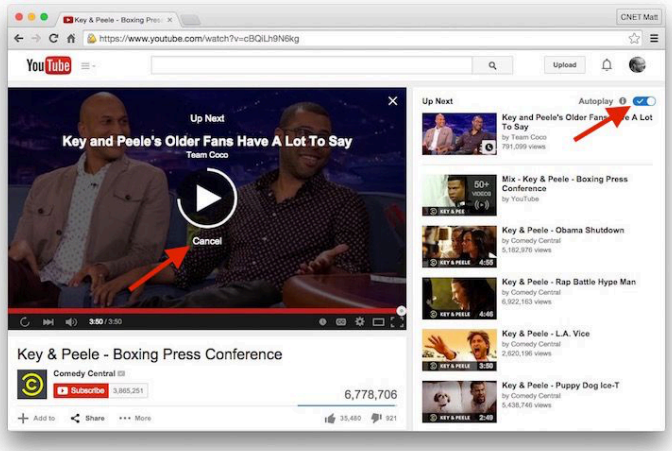
- No / little data initially
- In each time step
  - take an action based on current beliefs
  - get data / observations that depend on the action(s) taken

* Contrast to non-online setting of 1-time action based on available data.

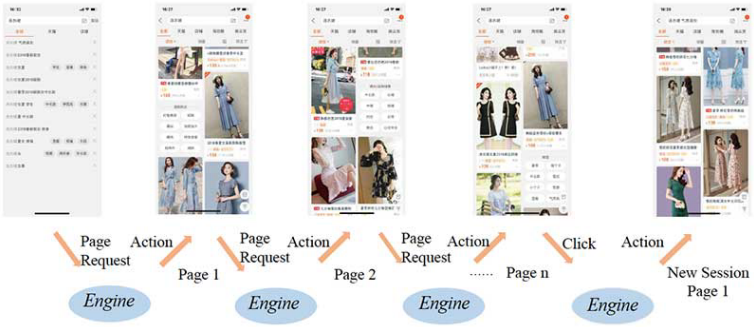
Online Learning Examples

- Recommendations

  eg Youtube Autoplay



- Shopping sites



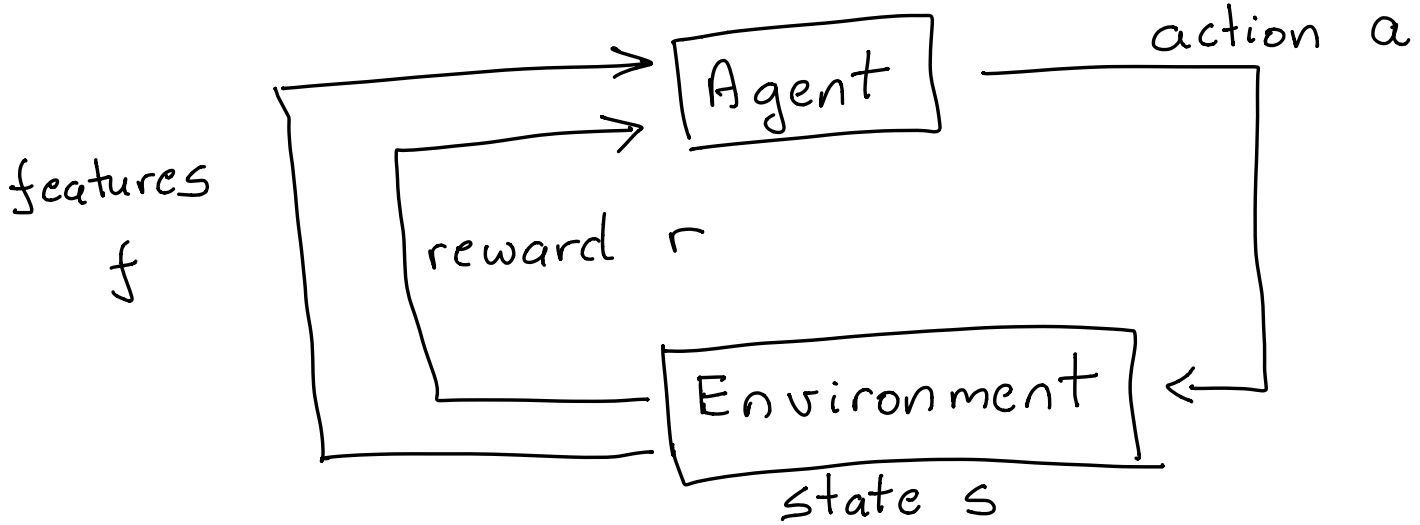
- Sequential strategy games

  eg: AlphaGo

- Self-driving Cars

...

Online Learning Setting



At time t,

- Agent determines a_t using $\{r_i, f_i, a_i\}_{i=1}^{t-1}$
- Environment state $s_{t-1} \to s_t$
- Reward r_t & features f_t to agent

} depend on a_t & history

Multi-armed Bandits

- Simplest online learning setting

- There are n "arms" / possible actions with unknown rewards
- At each time, one arm "played" & its reward is seen. No other features.

Simplest M·A·B Setting

(Stochastic Bernoulli Banclits)

- n coins w/ $Pr(\text{heads})$ $p_1, \ldots, p_k$ (unknown)
- Action: At each time, pick one coin to toss
  $a(t) \in \{1, \ldots, k\}$
- Reward: \$1 if Heads, \$0 if Tails

✱ No other information available

Total Reward $R(T) = \sum_{t=1}^{T} \mathbb{1}\{a(t) \text{ gave heads}\}$

($\mathbb{1}$: indicator func.)

$R(T)$: Random variable

Stochastic Bernoulli Banclits

Expected Reward $= E[R(T)]$

↓

Maximized by choosing arm $i^* = \arg\max_i p_i$

(but we do not know what that is)

Regret $(T) = T p_{i^*} - E[R(T)]$

↳ ideally, we want this to grow sub-linearly in T

Exploration vs Exploitation

Fundamental issue : should we

(a) Explore :- Play "new" arms to get good estimates of their p_i's

(b) Exploit :- Play the best arm so far to maximize current reward

Simple Idea : Explore then commit

- First play every arm τ times
- After these $n\tau$ time slots, find best estimated arm $\hat{i}$
- Play $\hat{i}$ forever after

Q: What are problems w/ this idea?

- How to pick τ?
- If we make mistake, "wrong forever"

Simple Exercise: 2 arms

Suppose we have 2 coins w/ p_1 & p_2, with $p_1 > p_2$

Strategy:- toss each one once, pick the higher one forever (break ties randomly)

Regret $R(T) = T p_1 - E\left[\sum_{s=1}^{T} X_s\right]$

Now, prob. arm 1 is chosen forever

$$= p_1 (1 - p_2) + \frac{p_1 p_2}{2}$$

& prob arm 2 is chosen $= p_2 (1 - p_1) + \frac{p_1 p_2}{2}$

Simple Exercise

$$R(T) = T p_1 - \left(\left[p_1(1-p_2) + \frac{p_1 p_2}{2} \right] p_1 + \left[p_2(1-p_1) + \frac{p_1 p_2}{2} \right] p_2 \right) T$$

$$= T p_1 - \left(\left[p_1 - \frac{p_1 p_2}{2} \right] p_1 + \left[p_2 - \frac{p_1 p_2}{2} \right] p_2 \right) T$$

$$= T p_1 - T \left(p_1^2 + p_2^2 - \frac{p_1^2 p_2}{2} - \frac{p_1 p_2^2}{2} \right)$$

↳ linear in T !

Simple Idea : Explore then commit

Theorem

If mk slots for explore, then commit

$$\text{Regret}(T) \leq m \sum_{i=1}^{k} \Delta_i + (T - mk) \sum_{i=1}^{k} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)$$

suboptimality $p_{i^*} - p_i$

2 arms &

For fixed T, this minimized by $m = \frac{4}{\Delta^2} \log\left(\frac{T\Delta^2}{4}\right)$

which gives $\text{Regret}(T) \leq 1 + C\sqrt{T}$

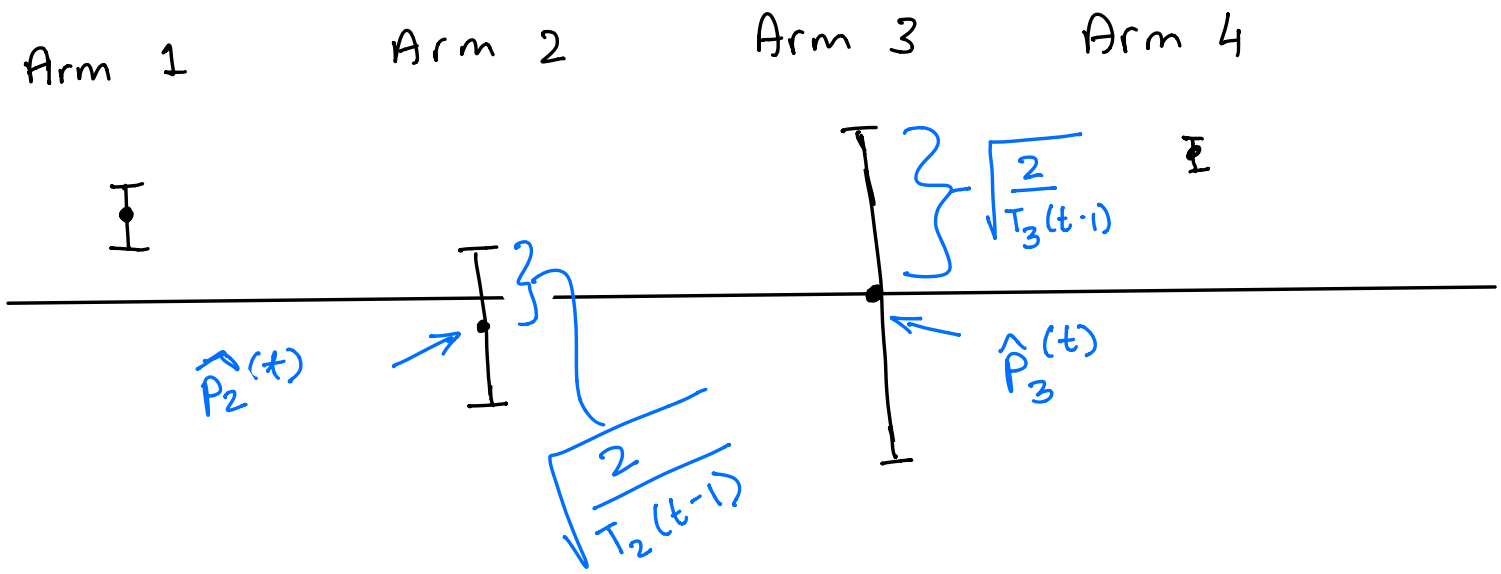
Beyond Explore & Commit

* Typically, no fixed T
* Best choice of m requires knowledge of Δ

Alternative :- Upper Confidence Bound (UCB)

Principle: Optimism. At each time

- Each arm has an uncertainty set for what its p_i could be
- Pick the arm with highest value of upper confidence bound u_i



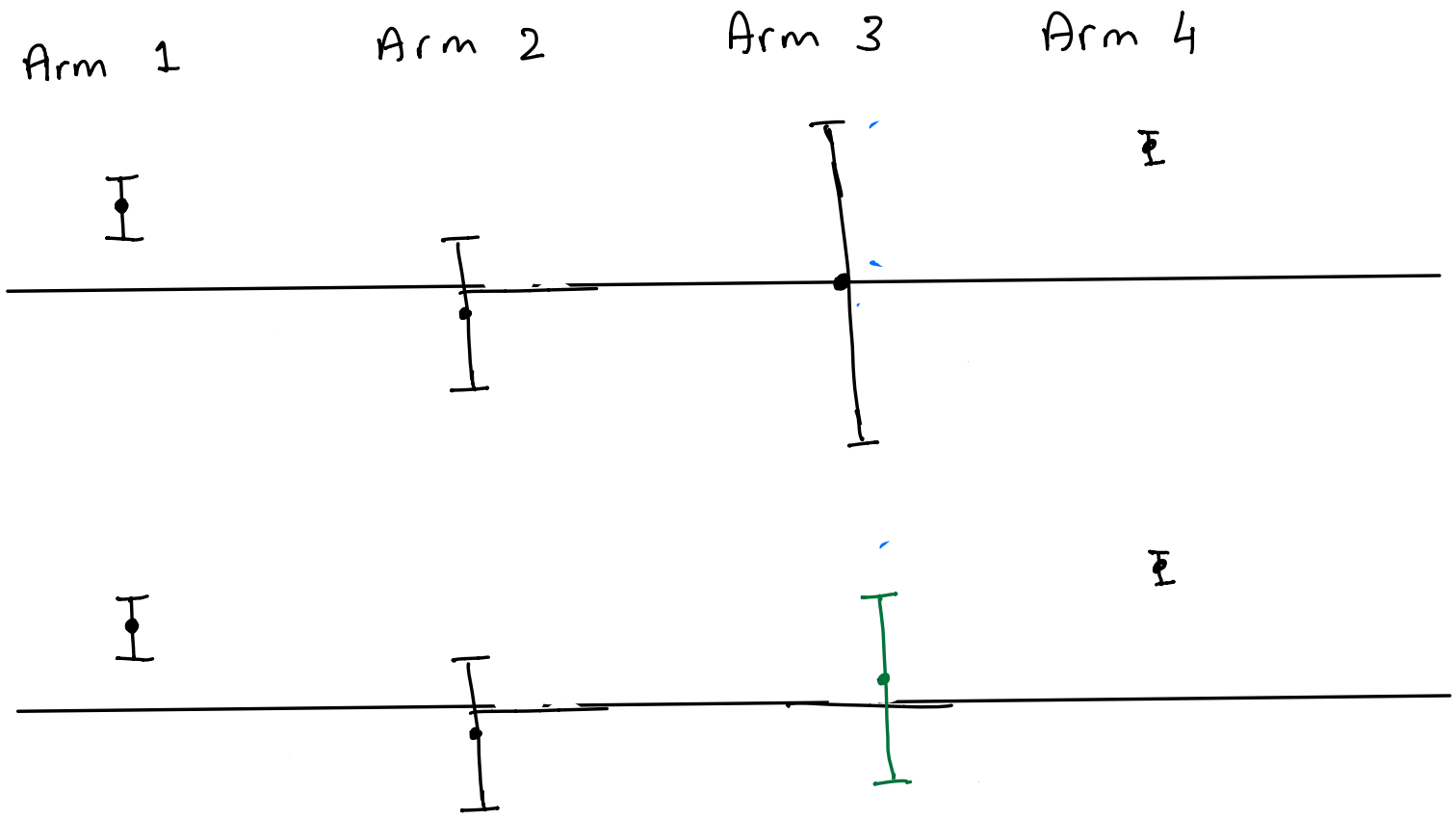
$\hat{p}_2(t)$ = empirical estimate of p_2 so far

$T_2(t-1)$ = # times arm 2 was played so far

In this case,

so arm 3 picked by UCB algorithm

UCB



only $\hat{p}_3^{(t)}$ & $T_3(t-1)$ change.

others as is

UCB

Theorem

$$\text{Regret}(T) \leq 3 \sum_{i=1}^{k} \Delta_i + \sum_i \frac{16 \log T}{\Delta_i}$$

* do not need to know T a-priori
* " " " " " Δ "

main idea

- Either we play best arm,
- Or we ensure we know played arm is bad.

Thompson sampling

- Maintain a pdf $q_i^{(t)}$ for each arm i

  $q_i^{(t)}$: current belief on what p_i is

- $q_i^{(0)} =$ Uniform $[0,1]$. for example
- At each time, draw a sample from each arm

  $a_i^{(t)} \sim q_i^{(t)}$ — a number in $[0,1]$

- Play the arm with highest $a_i^{(t)}$
- Update q_i for that arm. Others unchanged.