Bandits and Exploration: How do we (optimally) gather information?

Sham M. Kakade

Machine Learning for Big Data
CSE547/STAT548

University of Washington

Announcements...

- HW 4 posted soon (short)
- Poster session: June 1, 9-11:30a; ask TA/CSE students for help printing
- Projects: the term is approaching the end....

Today:

- Quick overview: Parallelization and Deep learning
- Bandits:
  1. Vanilla k-arm setting
  2. Linear bandits and ad-placement
  3. Game trees?

The problem

- In unsupervised learning, we just have data...
- In supervised learning, we have inputs X and labels Y (often we spend resources to get these labels).
- In reinforcement learning (very general), we act in the world, there is "state" and we observe rewards.
- Bandit Settings: We have K decisions each round and we do only received feedback for the chosen decision...

Gambling in casino...

ν_1

ν_2

ν_3

ν_4

ν_5

Goal: maximize ones' gains in a casino ?

Multi-Armed Bandit Game

$R_a \sim \Pr$ (reward for arm a)

- K Independent Arms: $a \in \{1, \ldots K\}$
- Each arm a returns a random reward R_a if pulled.
  (simpler case) assume R_a is not time varying.
- Game:
  - You chose arm a_t at time t.
  - You then observe:

$$X_t = R_{a_t}$$

  where R_{a_t} is sampled from the underlying distribution of that arm.
- The distribution of R_a is not known.

More real motivations...

Clinical trials:

$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

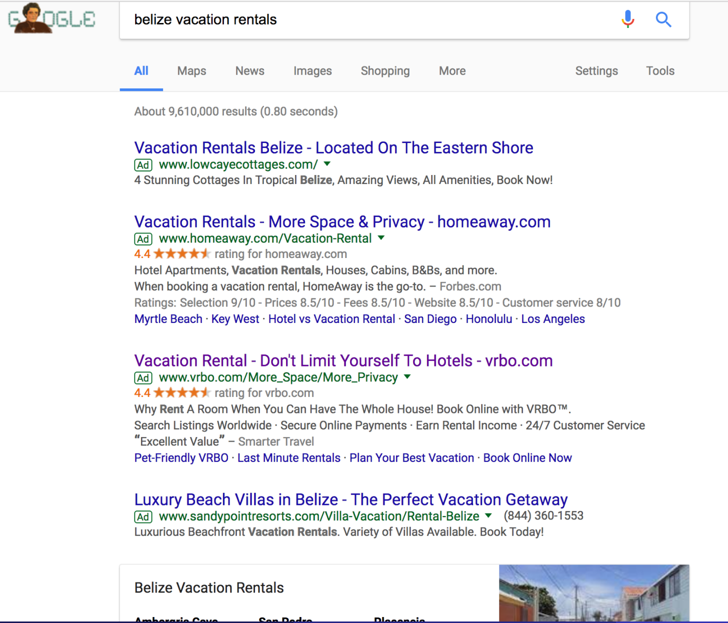
- choose a treatment A_t for patient t
- observe a response $X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t}$
- Goal: maximize the number of patient healed

Recommendation tasks:

ν_1 ν_2 ν_3 ν_4 ν_5

- recommend a movie A_t for visitor t
- observe a rating $X_t \sim \nu_{A_t}$ (e.g. $X_t \in \{1, \ldots, 5\}$)

Ad placement...



The Goal

- We would like to maximize our long term future reward.
- Our (possibly randomized) sequential strategy/algorithm $\mathcal{A}$ is:

$$a_t = \mathcal{A}(a_1, X_1, a_2, X_2, \ldots a_{t-1}, X_{t-1})$$

- In T rounds, our reward is:

$$\mathbb{E}[\sum_{t=1}^{T} X_t | \mathcal{A}]$$

  where the expectation is with respect to the reward process and our algorithm.
- Objective: What is a strategy which maximizes our long term reward?

Our Regret

$0 \le R_a \le 1$

- Suppose:

$$\mu_a = \mathbb{E}[R_a]$$

sublinear in T

- Assume $0 \le \mu_a \le 1$.
- Let $\mu_* = \max_a \mu_a$
- In expectation, the best we can do is obtain $\mu_* T$ reward in T steps.

expected

- In T rounds, our regret is:

$$\mu_* T - \mathbb{E}\left[\sum_{t=1}^{T} X_t | \mathcal{A}\right] \le ??$$

- Objective: What is a strategy which makes our regret small?

A Naive Strategy

- For the first τ rounds, sample each arm τ/K times.
- For the remainder of the rounds, choose the arm with best observed empirical reward.
- How ~~goes~~ good is this strategy? How do we set τ?
- Let's look at confidence intervals.

Hoeffding's bound

- If we pull arm $\overset{a}{\downarrow}$ N_a times, our empirical estimate for arm a is:

$$\hat{\mu}_a = \frac{1}{N_a} \sum_{t:a_t=a} X_t$$

- By Hoeffding's bound, with probability greater than $1-\delta$,

$$|\hat{\mu}_a - \mu_a| \le \mathcal{O}\sqrt{\frac{\log(1/\delta)}{N_a}}$$

$\delta \Leftarrow \delta/K$

- By the union bound, with probability greater than $1-\delta$,

$$\forall a,\ |\hat{\mu}_a - \mu_a| \le \mathcal{O}\sqrt{\frac{\log(K/\delta)}{N_a}}$$

Our regret

$$|\mu_{\hat{a}} - \mu_*| \le \sqrt{\frac{\log K/\delta}{\tau/K}}$$

$\hat{a}$: the arm we chose

- (Exploration rounds) What is our regret for the first τ rounds?
- (Exploitation rounds) What is our regret for the remainder τ rounds?
- Our total regret is: with $\Pr \ge 1-\delta$

$$\mu_* T - \sum_{t=1}^{T} X_t \le \tau + \mathcal{O}\sqrt{\frac{\log(K/\delta)}{\tau/K}}(T-\tau)$$

- How do we choose τ? find τ which min. expression

$$\tau = K^{-\frac{1}{3}} T^{2/3}$$

The Naive Strategy's Regret

Av. Reg $\approx O\left(\frac{1}{T^{\frac{1}{3}}}\right)$

$K^{-\frac{1}{3}}$

- Choose $\tau = K^{1/3} T^{2/3}$ and $\delta = 1/T$.
- Theorem: Our total (expected) regret is:

$$\mu_* T - \mathbb{E}[\sum_{t=1}^{T} X_t | \mathcal{A}] \leq \mathcal{O}(K^{1/3} T^{2/3} (\log(KT))^{1/3})$$

Can we be more adaptive?

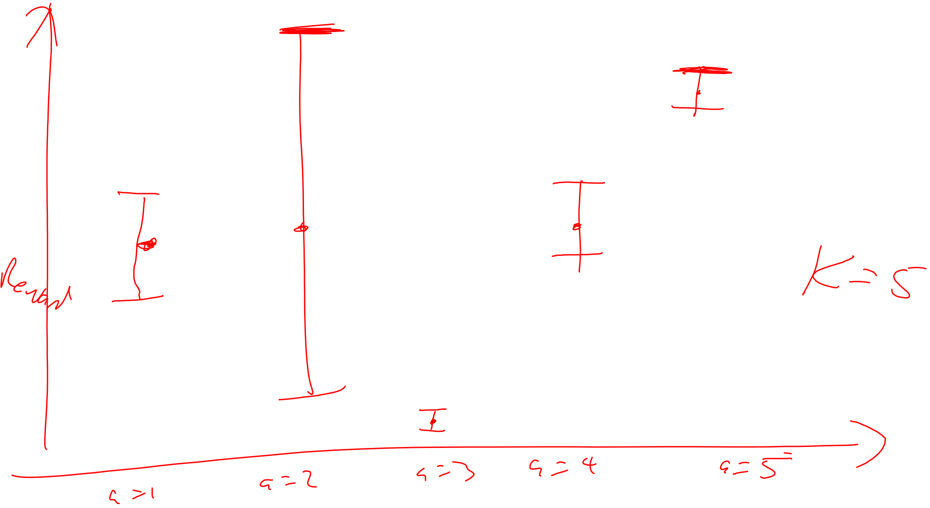
we might know not to pull certain arms!

- Are we still pulling arms that we know are sub-optimal? How do we know this??
- Let $N_{a,t}$ be the number of times we pulled arm a up to time t.
- Confidence interval at time t: with probability greater than $1-\delta$,

$$|\hat{\mu}_{a,t} - \mu_a| \le \mathcal{O}\sqrt{\frac{\log(1/\delta)}{N_{a,t}}} \qquad \sqrt{\frac{\log \frac{TK}{\delta}}{N_{a,t}}}$$

- with $\delta \to \delta/(TK)$, the above bound will hold for all time arms $a \in [K]$ and timesteps $t \le T$.

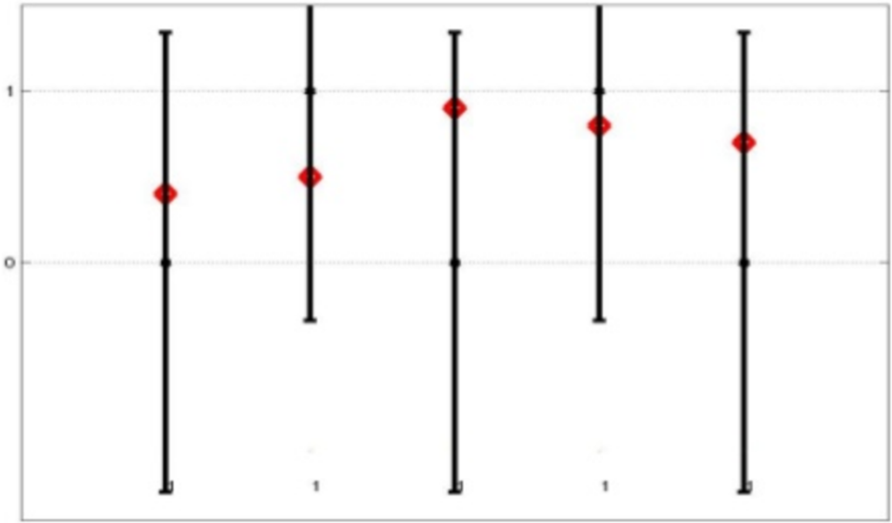
Example



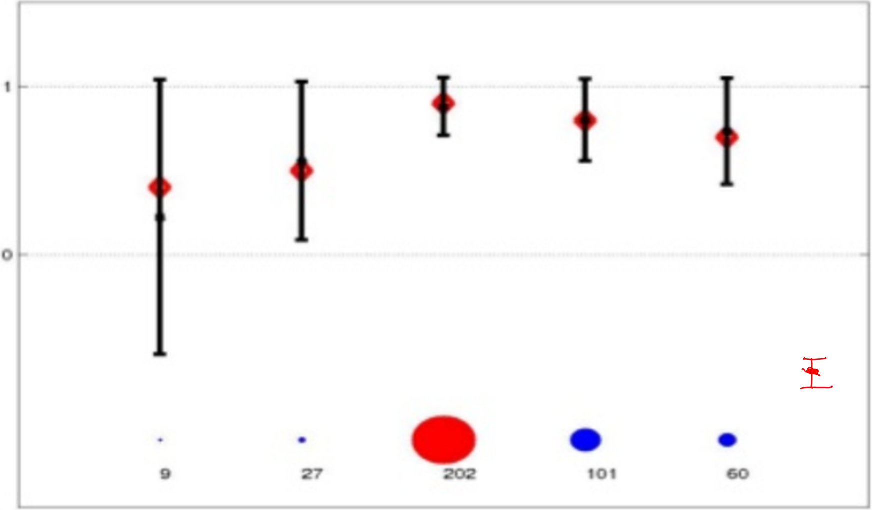
Example

Example

Confidence Bounds...



UCB: a reasonable state of our uncertainty...



Upper Confidence Bound (UCB) Algorithm

- At each time t,
  - Pull arm:

$$
\begin{aligned}
a_t &= \mathrm{argmax} \hat{\mu}_{a,t} + c\sqrt{\frac{\log(KT/\delta)}{N_{a,t}}} \\
&:= \mathrm{argmax} \hat{\mu}_{a,t} + \mathrm{ConfBound}_{a,t}
\end{aligned}
$$

    (where $c \leq 10$ is a constant).
  - Observe reward X_t.
  - Update $\mu_{a,t}$, $N_{a,t}$, and $\mathrm{ConfBound}_{a,t}$.
- How well does this do?

Instantaneous Regret

- With probability greater than $1-\delta$ all the confidence bounds will hold.
- Question: If
$$\text{argmax}\hat{\mu}_{a,t} + \text{ConfBound}_{a,t} \leq \mu_*$$
could UCB pull arm a at time t?
- Question: If pull arm a at time t, how much regret do we pay? i.e.
$$\mu_* - \mu_{a_t} \leq ??$$

Total Regret

- Theorem: The total (expected) regret of UCB is:

$$\mu_* T - \mathbb{E}[\sum_{t=1}^{T} X_t | \mathcal{A}] \leq \sqrt{KT \log(KT)}$$

- This better than the Naive strategy.
- Up to log factors, it is optimal.
- Practical algorithm?

Proof Idea: for $K = 2$

- Suppose arm $a = 2$ is not optimal.
- Claim 1: All confidence intervals will be valid (with Pr $\geq 1 - \delta$).
- Claim 2: If we pull arm $a = 1$, then no regret.
- Claim 3: If we pull $a = 2$, then we pay $2C_{a,t}$ regret. To see this:
  - Why?
  $$\hat{\mu}_{a,t} + C_{a,t} \geq \hat{\mu}_{1,t} + C_{1,t} \geq \mu_*$$
  - Why?
  $$\mu_a \geq \hat{\mu}_{a,t} - C_{a,t}$$
- The total regret is:
$$\sum_t C_{a,t} \leq \sum_t \frac{1}{\sqrt{N_{a,t}}}$$
- Note that $N_{a,t} \leq t$ (and increasing).

Acknowledgements

- http://gdrro.lip6.fr/sites/default/files/JourneeCOSdec2015-Kaufman.pdf
- https://sites.google.com/site/banditstutorial/