

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...



V_1



V_2



V_3



V_4



V_5

Goal: maximize ones' gains in a casino ?

Multi-Armed Bandit Game

- K Independent Arms: $a \in \{1, \dots, K\}$
- 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:



$B(\mu_1)$



$B(\mu_2)$



$B(\mu_3)$



$B(\mu_4)$



$B(\mu_5)$

- 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



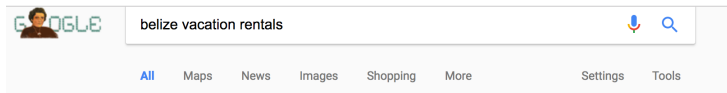
ν_4



ν_5

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

Ad placement...



Google logo with a profile picture of a person. Search bar contains "belize vacation rentals". Search icon and microphone icon are visible. Below the search bar are navigation links: All, Maps, News, Images, Shopping, More, Settings, Tools. The "All" link is underlined.

About 9,610,000 results (0.80 seconds)

Vacation Rentals Belize - Located On The Eastern Shore

Ad www.lowcaycottages.com/ ▾

4 Stunning Cottages In Tropical Belize, Amazing Views, All Amenities, Book Now!

Vacation Rentals - More Space & Privacy - homeaway.com

Ad www.homeaway.com/Vacation-Rental ▾

4.4 ★★★★★ rating for homeaway.com

Hotel Apartments, **Vacation Rentals**, Houses, Cabins, B&Bs, and more.

When booking a vacation rental, HomeAway is the go-to. – Forbes.com

Ratings: Selection 9/10 - Prices 8.5/10 - Fees 8.5/10 - Website 8.5/10 - Customer service 8/10

[Myrtle Beach](#) · [Key West](#) · [Hotel vs Vacation Rental](#) · [San Diego](#) · [Honolulu](#) · [Los Angeles](#)

Vacation Rental - Don't Limit Yourself To Hotels - vrbo.com

Ad www.vrbo.com/More_Space/More_Privacy ▾

4.4 ★★★★★ rating for vrbo.com

Why Rent A Room When You Can Have The Whole House! Book Online with VRBO™.

Search Listings Worldwide · Secure Online Payments · Earn Rental Income · 24/7 Customer Service

“Excellent Value” – Smarter Travel

[Pet-Friendly VRBO](#) · [Last Minute Rentals](#) · [Plan Your Best Vacation](#) · [Book Online Now](#)

Luxury Beach Villas in Belize - The Perfect Vacation Getaway

Ad www.sandypointresorts.com/Villa-Vacation/Rental-Belize ▾ (844) 360-1553

Luxurious Beachfront **Vacation Rentals**. Variety of Villas Available. Book Today!

Belize Vacation Rentals



The Goal

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

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

- In T rounds, our reward is:

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

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

- Suppose:

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

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

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

- **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 is this strategy? How do we set τ ?
- Let's look at confidence intervals.

Hoeffding's bound

- If we pull arm 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 $1 - \delta$,

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

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

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

Our regret

- (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:

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

- How do we choose τ ?

The Naive Strategy's Regret

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

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

Can we be more adaptive?

- 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 $1 - \delta$,

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

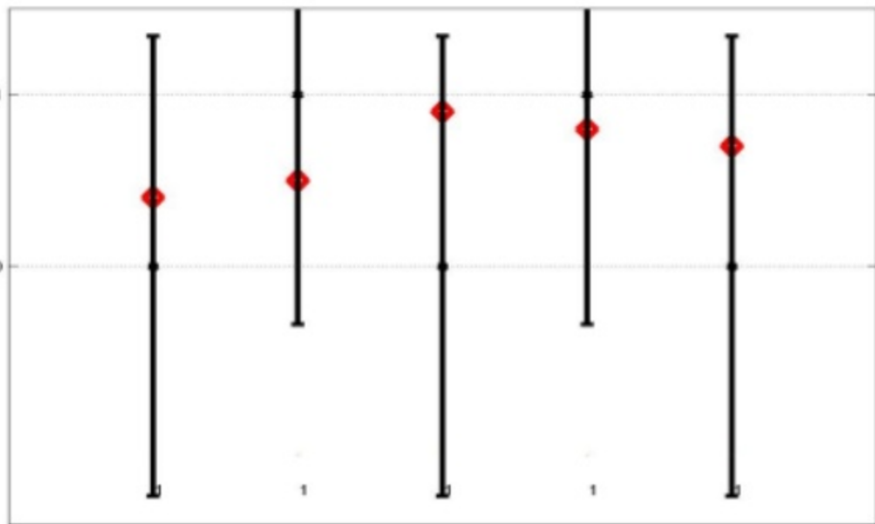
- with $\delta \rightarrow \delta/(TK)$, the above bound will hold for all time arms $a \in [K]$ and timesteps $t \leq T$.

Example

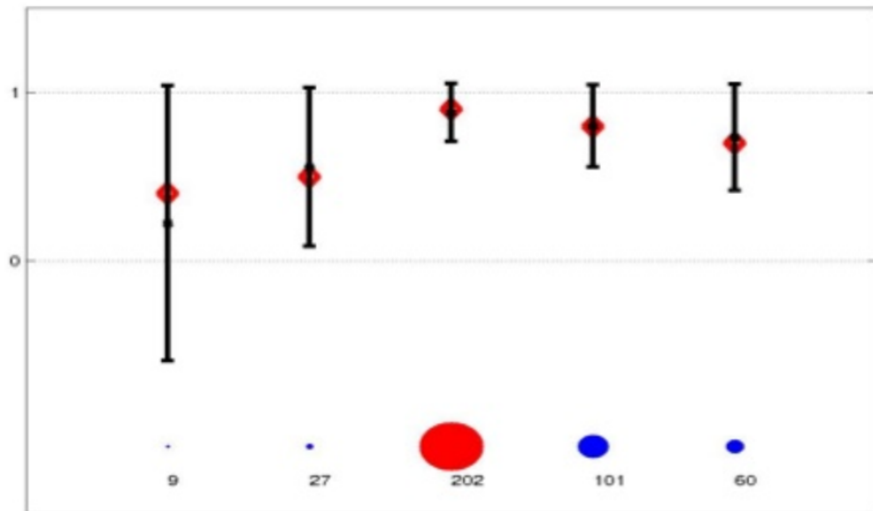
Example

Example

Confidence Bounds...



UCB: a reasonable state of our uncertainty...



Upper Confidence Bound (UCB) Algorithm

- At each time t ,
 - Pull arm:

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

(where $c \leq 10$ is a constant).

- Observe reward X_t .
- Update $\mu_{a,t}$, $N_{a,t}$, and $\operatorname{ConfBound}_{a,t}$.
- How well does this do?

Instantaneous Regret

- With probability greater than $1 - \delta$ all the confidence bounds will hold.
- Question: If

$$\operatorname{argmax} \hat{\mu}_{a,t} + \text{ConfBound}_{a,t} \leq \mu_*$$

could UCB pull arm a at time t ?

- Question: If pull arm a at time t , how much regret do we pay? i.e.

$$\mu_* - \mu_{a_t} \leq ??$$

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

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

- 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/>