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. Review: Vanilla k-arm setting,UCB
  2. Today: UCB (continued), Thompson, Linear bandits and ad-placement

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...

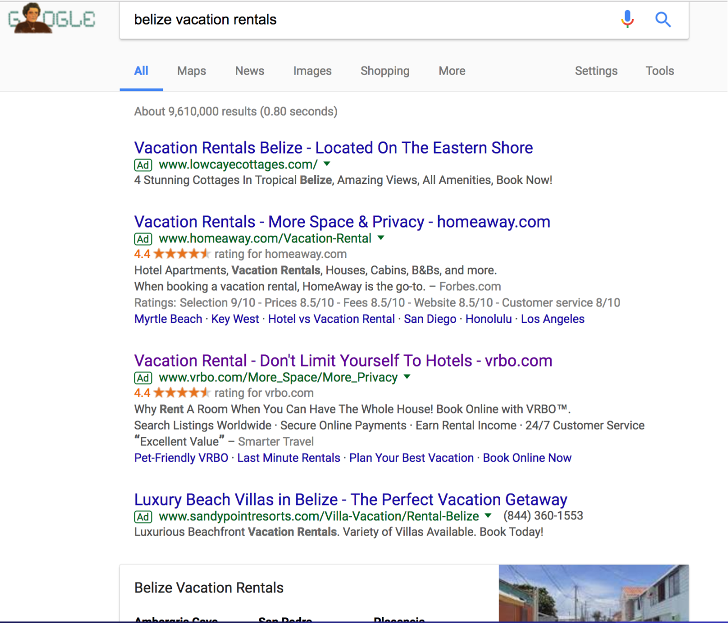
Review

Multi-Armed Bandit Game

- 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.


belize vacation rentals
All
Maps
News
Images
Shopping
More
Settings
Tools
About 9,610,000 results (0.80 seconds)
Vacation Rentals Belize - Located On The Eastern Shore
Ad www.lowcayecottages.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:

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

- 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

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

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:

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

Can we be more adaptive?

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

Upper Confidence Bound (UCB) Algorithm

- At each time t,
  - Pull arm:

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

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

Today

Instantaneous Regret

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

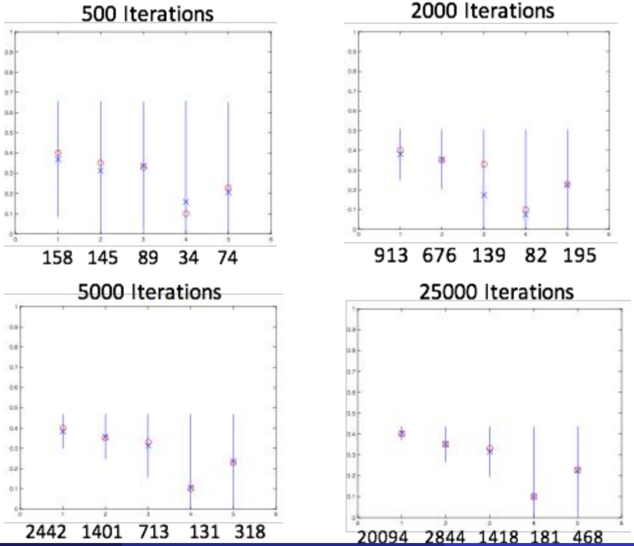
Total Regret

- 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?

Simulation


500 Iterations
158 145 89 34 74
2000 Iterations
913 676 139 82 195
5000 Iterations
2442 1401 713 131 318
25000 Iterations
20094 2844 1418 181 468

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).

Aside: Logarithmic regret

- The previous rates are not a function of problem dependent parameters.
- On any given problem, we expect to eventually start pulling the best arm.
- Define the "gap" as:

$$\Delta = \mu_* - \max_{a \neq a_*} \mu_a$$

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

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

  (same algorithm enjoys this bound.)
- Question: How is the "naive" algorithm different?

Thompson sampling

- Practical issues:
  - how to obtain good confidence intervals?
  - variants with "similar" performance?
- Suppose we are "Bayesian". We have a posterior distribution

$$\Pr(\mu_a | \mathrm{History}_{<t})$$

- Thompson sampling:
  - Sample from each posterior:

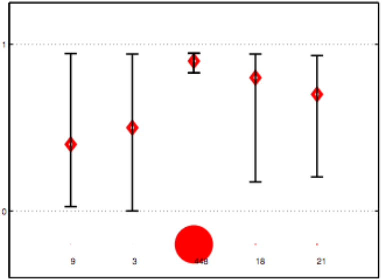
$$\nu_a \sim \Pr(\mu_a | \mathrm{History}_{<t})$$

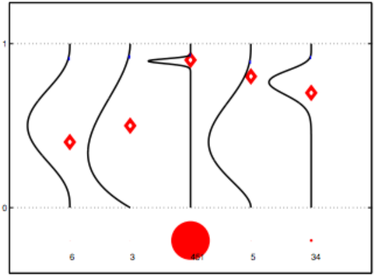
  - take action

$$a_t = \mathrm{argmax}_a \nu_a$$

  - update posteriors

Thompson sampling and Confidence intervals





Acknowledgements

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