Online Learning
—Stochastic Bandit

Jie Tang

Tsinghua University

May 13, 2020

Overview

1. Introduction
2. Stochastic Bandit
3. Appendix

Outline

1 Introduction

2 Stochastic Bandit

3 Appendix

Introduction

- Many models we have introduced, such as SVMs, NNs, Adaboost, etc., fall into the category of the supervised learning paradigm. Their learning processes are usually dividing datasets into training, validation, test parts and training a model on batch data.
- A basic assumption behind this procedure is that distributions of all three dataset are the same, as well as the unseen data to come in the future.
- This paradigm works well in many scenario, for example, image classification, face detection, etc.

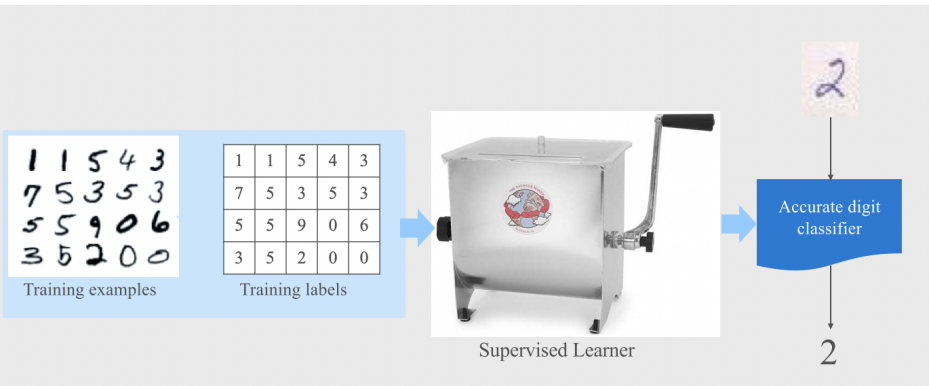


Figure 1: Illustration of supervised learning paradigm.

Introduction

- These assumptions may be violated in many cases.
- Consider the online news recommender system. It is a typical scenario when pure supervised learning may fail.
- The goal of the recommender system is to maximize its CTR(click through rate) in a given time.
- What will we do when taking a classical supervised learning approach?



Figure 2: A screen-shot of toutiao. Content in the purple eclipse comes from recommender system.

Bandit Model

- From our example of online news recommendation, we can spot several fundamental differences of bandit model compared to supervised learning.
- batch data v.s. sequential data.
  - Bandit model belongs to an online learning paradigm. Intuitively, online methods are those that "process one datum at a time".
  - Supervised learning digests much more data every time, and usually, the larger the size of the data, the better the model.
- i.i.d. assumption v.s. not necessarily i.i.d.
  - Samples in bandit come one by one and the latter will be affected by the choice of the former, which is not the typical i.i.d. case in supervised learning.
- Meanwhile, to be used in real online environment, bandit models, the same as many other online learning methods, favor computationally and spatially efficient solutions.

William R. Thompson
Pathologist



Richard S. Sutton
University of Alberta
Reinforcement Learning



Peter Auer
University of Leoben



Csaba Szepesvari
University of Alberta

Thompson sampling:
For clinic experiment
design(1933)



UCB1(2002)

$$a_t \in \arg\max_i \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log t}{T_i(t-1)}},$$

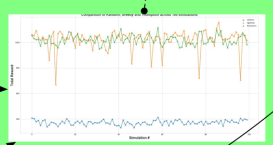
Generalized linear bandit(2010)

$$\mathbb{E}[R_t | A_t] = \mu(m'_{A_t}\theta_*)$$

Real Bandit In Las Vegas



Epsilon-greedy(1998)



Thompson sampling for linear
bandit:theory(2013)

$$\tilde{O}\left(\frac{d^2}{\epsilon}\sqrt{T^{1+\epsilon}}\right)$$



Shipra Agrawal
Columbia University

MAB: Animal
Experiment(1955)



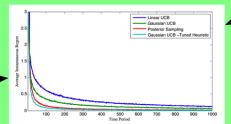
LinUCB on Yahoo!
News(2010)



LinUCB:Improved
Algorithm(2010)

$$(X_t, \tilde{\theta}_t) = \underset{(x,\theta)\in D_t \times C_{t-1}}{\operatorname{argmax}} \langle x, \theta \rangle ,$$

Thompson Sampling:
Bayesian regret bound for
many models(2014)





Yasin Abbasi-Yadkori
Queensland University of Technology

Frederick Mosteller & Robert Bush
Statisticians



Li Lihong
Google



Daniel Russo
Columbia University
Bussiness School

Outline

1 Introduction

2 Stochastic Bandit

3 Appendix

Stochastic Bandit

- A stochastic bandit is a collection of distributions $v = (P_a : a \in \mathcal{A})$ where $\mathcal{A}$ is the set of available arms. It is a repeated game over n rounds between the learner and the environment.
- Define initial history $\mathcal{H}_1 = \emptyset$. For each $t = 1, 2, ...n$:
  - The learner select an arm $a_t \in \mathcal{A}$ based on history $\mathcal{H}_t$.
  - The environment samples a reward $r_t \sim P_a$ and reveals it to the learner.
  - Expand the history $\mathcal{H}_{t+1} = \mathcal{H}_t \cup \{(a_t, r_t)\}$.
- How the learner selects a_t based on $\mathcal{H}_t$ is called a strategy, which is a mapping from $\mathcal{H}_t$ to $\mathcal{A}$ and usually denoted by π.

Learning Objective

- The learner's goal is to maximize the total reward

$$S_n = \sum_{i=1}^{n} r_t,$$

- A stochastic bandit instance is determined by the instantiation of $v = (P_a : a \in \mathcal{A})$, where $v \in \mathcal{E}$ and $\mathcal{E}$ is the environment class. The learner usually has partial information about v.
- There are two kinds of bandits based on the structure of $\mathcal{E}$: unstructured bandit and structured bandit.
- An environment class $\mathcal{E}$ is unstructured if $\mathcal{A}$ is finite and there exists sets of distributions $\mathcal{M}_a$ such that

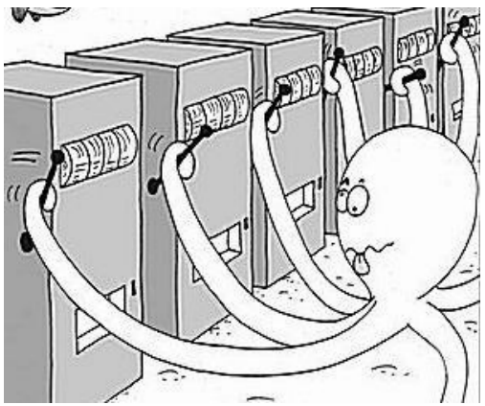
$$\mathcal{E} = \{v = (P_a : a \in \mathcal{A}) : P_a \in \mathcal{M}_a, \forall a \in \mathcal{A}\}.$$

- Any environment that is not unstructured is called structured.

K-armed Bandit

- In K-armed bandit, there are K arms to be pulled by a learner in a n rounds repeated game.
- For each round $t = 1, 2, ...n$:
  - The learner chooses an arm $a_t \in \{1, 2, ...K\}$ based on $\mathcal{H}_t$.
  - The learner receives the reward $r_{a_t,t}$ of a_t.
  - $\mathcal{H}_{t+1} = \mathcal{H}_t \cup (a_t, r_t)$.
- The goal is to maximize the cumulative reward

$$R_n = \max_{i=1,2...K} \mathbb{E}[\sum_{t=1}^{n} r_{i,t}] - \mathbb{E}[\sum_{t=1}^{n} r_{a_t,t}]$$



Stochastic K-armed bandit

- The reward distributions $v_1, v_2, ...v_K$ of the arms with respective means $\mu_1, \mu_2, ...\mu_K$ are fixed but unknown to the learner. For simplicity, we assume $\mu_i \in [0, 1]$.
- Denote

$$i^* = \arg\max_{i=1,2,...K} \mu_i,$$

$$\mu^* = \max_{i=1,2,..K} \mu_i.$$

$$\Delta_i = \mu^* - \mu_i$$

$$T_i(n) = \sum_{t=1}^{n} \mathbf{1}[a_t = i]$$

- The cumulative regret can be written as

$$R_n = n\mu^* - \mathbb{E}[\sum_{t=1}^{n} \mu_{a_t}] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$$

Regret's Order

- Intuition: The smaller the regret, the better our strategy.
- How small? A linear regret — a regret increasing linearly w.r.t. t — means the learner performs no better than a random strategy which exploits no information of $\mathcal{H}_t$ as the game goes.
- A sub-linear regret is needed. It means near-zero average regret per round when the horizon n grows toward infinity, e.g., a regret of order $O(\sqrt{n})$.
- How to guarantee a sub-linear regret?
  - A good balance between exploitation and exploration.

Exploitation and Exploration

- Playing the arm with the highest empirical mean would not work — it does not leave a room for future adjustment.
  - If the arm with the highest empirical mean gets played every round, it is possible that the best arm may never be found, thus the learner suffers a linear regret.
  - Think about two arms a_1, a_2 with rewards $r_{a_1} \sim Bernoulli(0.2), r_{a_2} \sim Bernoulli(0.3)$. First two rounds, the agent pulls a_1 with reward 1 and pulls a_2 with reward 0.
  - Then under this strategy, only a_1 will be pulled in the future, which is the sub-optimal arm.
- Random strategy does not work either. It does exploration all the time.
- Question: how to balance exploitation and exploration to develop a good strategy.

Bandit Strategy

- ϵ-greedy algorithms
- UCB: Optimism in face of uncertainty (OIFU) and the UCB-type algorithms
- Thompson Sampling: also known as probability matching

ϵ-greedy

- Both pure exploitation and pure exploration lead to linear regrets.
- What if a certain probability ϵ is reserved for exploration, i.e., randomly selecting some arm from $\mathcal{A}$?
- ϵ-greedy algorithm:
  - With probability $1 - \epsilon$, it plays the arm with the highest estimated reward;
  - With probability ϵ, it plays a random arm.

ϵ-greedy

- Both pure exploitation and pure exploration lead to linear regrets.
- What if a certain probability ϵ is reserved for exploration, i.e., randomly selecting some arm from $\mathcal{A}$?
- ϵ-greedy algorithm:
  - With probability $1 - \epsilon$, it plays the arm with the highest estimated reward;
  - With probability ϵ, it plays a random arm.
- For any fixed $\epsilon > 0$, we have

$$\mathbb{E}(R_n) = \epsilon \frac{n}{K} \sum_{i=1}^{K} \Delta_i,$$

which is still linear w.r.t. n.

ϵ_n-Greedy

- Key idea: Use a decreasing ϵ w.r.t. t? The exploration probability decreases as t increases — (the estimates for the reward become more accurate)
  - At round t, with probability ϵ_t, pull a random arm.
  - At round t, with probability $1 - \epsilon_t$, pull the arm with the highest empirical mean.
- Theoretical guarantee (given by Auer et al.)
  - Let $\Delta = \min_{i:\Delta_i > 0} \Delta_i$ and consider $\epsilon_t = \min(\frac{CK}{\Delta^2 t}, 1)$.
  - When $t \geq \frac{CK}{\Delta^2 t}$, the probability of choosing a sub-optimal arm i is bounded $\frac{C}{\Delta^2 t}$ for some constant $C > 0$.
  - As a consequence, $\mathbb{E}[T_i(n)] \leq \frac{C}{\Delta^2} \log n$ and $R_n \leq \sum_{i:\Delta_i > 0} \frac{C\Delta_i}{\Delta^2} \log n$, which is a logarithmic regret.

Regret Analysis of ϵ_n-greedy

- The proof's hallmark is to upper-bound the probability of any arm being chosen at time t, and then add instantaneous regret over time t.
- Let $x_0 = \frac{1}{2K}\sum_{t=1}^{n}\epsilon_t$ and $\hat{\mu}_{i,t-1} = \frac{1}{T_i(t-1)}\sum_{s=1}^{T_i(t-1)} r_{i,s}$. We have $\mathbb{P}(I_n = j) \leq \frac{\epsilon_n}{K} + (1-\frac{\epsilon_n}{K})\mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \hat{\mu}^*_{T^*(n-1)})$ and

$$\begin{aligned}\mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \hat{\mu}^*_{T^*(n-1)}) \leq & \mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \mu_j + \frac{\Delta_j}{2}) \\ & + \mathbb{P}(\hat{\mu}^*_{T^*(n-1)} \leq \mu^* - \frac{\Delta_j}{2}).\end{aligned}$$

Let $T_j^R(n)$ be the the number of plays in which arm j was chosen at random in the first n plays. Then we have

$$\begin{aligned}&\mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \mu_j + \frac{\Delta_j}{2}) \\ &= \sum_{t=1}^{n}\mathbb{P}(T_j(n) = t \wedge \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2})\end{aligned}$$

Continuing Proof

$$= \sum_{t=1}^{n} \mathbb{P}(T_j(n) = t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) \mathbb{P}(\hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2})$$

$$\leq \sum_{t=1}^{n} \mathbb{P}(T_j(n) = t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) e^{-\Delta_j^2 t/2} \text{(Chernoff-Hoeffding bound)}$$

$$\leq \sum_{t=1}^{\lfloor x_0 \rfloor} \mathbb{P}(T_j(n) = t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) e^{-\Delta_j^2 t/2} + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2}$$

$$\text{(since } \sum_{t=x+1}^{\infty} e^{-Kt} \leq \frac{1}{K} e^{-Kx})$$

$$\leq \sum_{t=1}^{\lfloor x_0 \rfloor} \mathbb{P}(T_j^R(n) \leq t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) e^{-\Delta_j^2 t/2} + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2}$$

$$\leq x_0 \mathbb{P}(T_j^R(n) \leq x_0) + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2} (T_j^R(n) \text{ and } \hat{\mu}_{j,t} \text{ are independent.})$$

Continuing Proof

As $\mathbb{E}[T_j^R] = \frac{1}{K}\sum_{t=1}^{n}\epsilon_t$, $\mathrm{Var}[T_j^R] = \sum_{t=1}^{n}\frac{\epsilon_t}{K}(1-\frac{\epsilon_t}{K}) \le \frac{1}{K}\sum_{t=1}^{n}\epsilon_t$, by Bernstein's inequality we have $\mathbb{P}(T_j^R(n) < x_0) \le e^{-x_0/5}$.

Denote $n' = CK/\Delta^2$. For $n \ge n'$, we can lower-bound x_0 by

$$x_0 = \frac{1}{2K}\sum_{t=1}^{n}\epsilon_t = \frac{1}{2K}\sum_{t=1}^{n'-1}\epsilon_t + \frac{1}{2K}\sum_{t=n'}^{n}\epsilon_t$$

$$\ge \frac{n'-1}{2K} + \frac{C}{2\Delta^2}\log\frac{n+1}{n'} \ge \frac{C}{\Delta^2}\log\frac{n\Delta^2 e^{1/2}}{CK}.$$

Thus,

$$\mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \ge \mu_j + \frac{\Delta_j}{2}) \le x_0\mathbb{P}(T_j^R(n) \le x_0) + \frac{2}{\Delta_j^2}e^{-\Delta_j^2\lfloor x_0\rfloor/2}$$

$$\le x_0 e^{-x_0/5} + \frac{2}{\Delta_j^2}e^{-\Delta_j^2\lfloor x_0\rfloor/2} = O(\frac{C}{\Delta^2 n}).$$

By a symmetric analysis we can get similar results for $\mathbb{P}(\hat{\mu}^*_{T^*(n-1)} \le \mu^* - \frac{\Delta_j}{2})$. Summing over $\frac{1}{n}$ leads to log function.

Discussion

- The regret analysis establishes a bound on the instantaneous regret and a bound on cumulative regret over finite time n. It's called finite-time regret analysis, which is a stronger bound than an asymptotic one.
- Recall the bound is a result of $\epsilon_t = \min(\frac{CK}{\Delta^2 t}, 1)$, which requires the knowledge of Δ. In practice, it's usually unknown.
- Another drawback of ϵ_n-greedy is the exploration treats all arms equally and makes no distinction of sub-optimal arms.

UCB

- Recall the exploitation of ϵ-greedy is guided by indexes of arms, i.e., their estimated reward. This kind of policy is called index-based policy.
- ϵ-greedy gives the following strategy.
  - At time t, we have

$$\text{with probability } 1-\epsilon,\ a_t = \arg\max_i \hat{\mu}_{i,T_i(t)}.$$

- Based on OIFU[1], we can design an index-based policy based not only on estimated mean but also uncertainty of the mean. This leads to the upper-confidence bound (UCB) algorithm.
  - At time t, we have

$$a_t = \arg\max_i \hat{\mu}_{i,T_i(t)} + \text{Uncertainty}(\hat{\mu}_{i,T_i(t)}).$$

[1] Optimism In Face of Uncertainty (OIFU) says that when we are selecting arms, we should consider both the estimated mean and its uncertainty (in most case, variance)

UCB1

- Hoeffding's inequality: Let $X_1, X_2, ...X_m$ be i.i.d. random variable taking values in $[0, 1]$. For any $\epsilon > 0$, with probability at least $1 - \epsilon$, we have

$$\mathbb{E}(X) \leq \frac{1}{m} \sum_{s=1}^{m} X_s + \sqrt{\frac{\log \epsilon^{-1}}{2m}}.$$

- We can exploit the inequality to construct upper-confidence bound. The important question is how to choose ϵ properly.
- Could ϵ be constant over t?
- UCB1 policy exploits Hoeffding's inequality and play at time t

$$a_t \in \arg\max_i \hat{\mu}_{i,t-1} + \sqrt{\frac{2 \log t}{T_i(t-1)}},$$

  where $\hat{\mu}_{i,t-1} = \frac{1}{T_i(t-1)} \sum_{s=1}^{T_i(t-1)} r_{i,s}$.
- Comparing UCB and Hoeffding's inequality, we can see that $\epsilon_t = \frac{1}{t^4}$.
- UCB1 has sub-linear regret.

Regret Analysis of UCB1

Let $c_{t,s} = \sqrt{2 \log t / s}$, For any arm i, we upper-bound $T_i(n)$ on any sequence of plays. Let N be an arbitrary positive integer. We have

$$
\begin{aligned}
T_i(n) &= 1 + \sum_{t=K+1}^{n} \mathbf{1}(I_t = i) \\
&\leq N + \sum_{t=K+1}^{n} \mathbf{1}(I_t = i, T_i(t-1) \geq N) \\
&\leq N + \sum_{t=K+1}^{n} \mathbf{1}(\mu^*_{T^*(t-1)} + c_{t-1,T^*(t-1)} \leq \hat{\mu}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)}, \\
&\quad T_i(t-1) \geq N) \\
&\leq N + \sum_{t=K+1}^{n} \mathbf{1}(\min_{0<s<t} \hat{\mu}_s + c_{t-1,s} \leq \max_{N \leq s_i < t} \hat{\mu}_{i,s_i} + c_{t-1,s_i}) \\
&\leq N + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=N}^{t-1} \mathbf{1}(\mu^*_s + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i})
\end{aligned}
$$

Continuing Proof

Note the fact that $\hat{\mu}_s^* + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}$ implies at least one of the following must hold (Proofs by Contradiction)

$$\hat{\mu}_s^* \leq \mu^* - c_{t,s}, \hat{\mu}_{i,s_i} \geq \mu_i + c_{t,s_i}, \mu^* < \mu_i + 2c_{t,s_i}$$

By Chernoff-Hoeffding bound, we have

$$\mathbb{P}(\hat{\mu}_s^* \leq \mu^* - c_{t,s}) \leq e^{-4\log t} = t^{-4}, \mathbb{P}(\hat{\mu}_{i,s_i} \geq c_{t,s_i}) \leq e^{-4\log t} = t^{-4}$$

When $s_i \geq 8\log n/\Delta_i^2$, $\mu^* < \mu_i + 2c_{t,s_i}$ doesn't hold. Therefore,

$$\mathbb{E}[T_i(n)] \leq \lceil \frac{8\log n}{\Delta_i^2} \rceil + \sum_{t=1}^{\infty}\sum_{s=1}^{t}\sum_{s_i=1}^{t} 2t^{-4} \leq \frac{8\log n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

Recall that $R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$, we can conclude that UCB1's regret is logarithmic thus sub-linear.

Discussion

- UCB-type policy does not require knowledge of gap between optimal arm and other arms, while ϵ_t-greedy does.
- UCB-type policy's index contains two parts, the former for exploitation, the latter for exploration.
- Recall UCB1's strategy is

$$a_t \in \arg\max_i \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log t}{T_i(t-1)}},$$

- UCB1's exploration is based on Hoeffding's inequality, and ϵ is set to $\frac{1}{t^4}$. This choice is justified in the three-fold summation over t while we are upper-bounding $\mathbb{E}[T_i(n)]$.
- UCB(ρ) adjusts the choice of ϵ by a constant factor, which does not influence the order the regret and controls the tendency to exploration. Formally, UCB(ρ) pulls arm by

$$a_t \in \arg\max_i \hat{\mu}_{i,t-1} + \sqrt{\frac{\rho\log t}{T_i(t-1)}}.$$

Bayesian K-armed Bandit

- So far we have made no assumptions about the reward distribution of any arm.
- Bayesian bandit exploits prior knowledge of rewards $\mathbb{P}(\mu)$.
- If a prior is assumed, we should be able to guide our exploration based on the prior and posterior.
- Formally, the strategy is that arm a is selected with probability

$$\mathbb{P}(a_t = a|\mathcal{H}_t) = \mathbb{P}(\mu_a > \mu_{a'}, \forall a' \neq a|\mathcal{H}_t),$$

  where $\mathcal{H}_t$ is the history up to time t, i.e., the sequence $\{(a_i, r_i)\}_{i=1}^{t-1}$.
- Note that since a prior has been assumed for μ_i for any arm i, this formula is indeed a well-defined probability. The condition on $\mathcal{H}_t$ can be explained as posterior

Thompson Sampling

- Thompson sampling is based on this idea. It was first designed by the clinic Thompson to help medicine experiment.
- Thompson sampling at time t is described as follows.
  - Use Bayes rule to compute posterior $\mathbb{P}(\mu_i|\mathcal{H}_t)$.
  - Sample a reward $\tilde{r}_i$ for each arm i from $\mathbb{P}(\mu_i|\mathcal{H}_t)$.
  - Select arm by

$$a_t = \arg\max_i \tilde{r}_i,$$

    where tie breaks even.
  - Observe the real reward r_t.
  - $\mathcal{H}_{t+1} = \mathcal{H}_t \cup \{(a_t, r_t)\}$.
- Thompson Sampling avoids probability computation totally via sampling, making it usually more computational efficient.
  - Sampling from K Gaussian is far faster than numerical computation of K complex probability formula.

Regret of Thompson Sampling

- From a Bayesian perspective, there should be a prior on the parameters of the bandit model, and the Bayesian regret is defined as

$$R_n^{\text{Bayesian}} = \mathbb{E}[R_n].$$

- The Bayesian regret of Thompson sampling is shown to be $O(\sqrt{Kn}\log n)$.
- The proof needs some advanced math. Check Agrawal et al.[2] for a full treatment.

[2] Agrawal, Shipra, and Navin Goyal. "Thompson sampling for contextual bandits with linear payoffs." International Conference on Machine Learning. 2013.

Discussion

- All three algorithms do exploration in some way, which could be explained as gaining more information from the information theory perspective.
- We have shown three algorithms, each with a sub-linear regret. A natural question is what is the best regret we can achieve?
- We have to clarify our usage of best regret here.

Minimax Regret

- Minimax regret is an indicator of the hardness of the underlying bandit in the worst-case sense.
  - The worst-case regret of a policy π (the learner's strategy to take actions) on a set of stochastic bandit environments $\mathcal{E}$ is

$$R_n(\pi, \mathcal{E}) = \sup_{v \in \mathcal{E}} R_n(\pi, v)$$

    .
  - The minimax regret is

$$R_n^*(\mathcal{E}) = \inf_{\pi \in \Pi} R_n(\pi, \mathcal{E}) = \inf_{\pi \in \Pi} \sup_{v \in \mathcal{E}} R_n(\pi, v),$$

    where Π is the set of all policies.

Minimax Regret Lower Bound

Theorem

Let $\mathcal{E}^k$ be the the set of K-armed Gaussian bandits with unit variance and means $\mu \in [0, 1]$. Then there exists a universal constant $c > 0$ such that for all $k > 1$ and $n \geq k$, it holds that $R_n^*(\mathcal{E}^k) \geq c\sqrt{kn}$.

- Check Sébastien's paper [3] for proof of it.
- Minimax regret's bound is a useful measure of the robustness of a set of policies, but are often excessively conservative, since they always consider the worst case.
- Minimax regret's bound is also instance-independent, i.e., it does not depend on Δ_i. We have derived instance-dependent regret upper-bound for UCB1. UCB1's instance-independent upper-bound is $O(\sqrt{kn \log n})$, thus only a log factor away from this lower-bound.

[3] Bubeck, Sébastien, Vianney Perchet, and Philippe Rigollet. "Bounded regret in stochastic multi-armed bandits." Conference on Learning Theory. 2013.

Contextual Bandit

- We have been focusing on K-armed bandit whose environment is stationary and whose arm set is finite.
- In real world, this is hardly the scenario.
  - Large arm set: Millions of items to be recommended.
  - Non-stationary: The world is very messy.
- Contextual bandits generalize the finite-armed setting by allowing the learner to make use of side information.
  - With arms represented by its side information, or features, contextual bandits could generalize knowledge of one arm to others due to interplay between features of different arm.
  - Stochastic K-armed bandit is equivalent to a Contextual bandit whose arm is encoded by one-hot encoding.
- The most simple, and most studied contextual bandit is stochastic linear bandit.

Stochastic Linear Bandit

- Linear bandit adds an additional information about the reward's distribution compared to K-armed bandit. It assumes that

$$r_{a_t} = \langle \theta^*, a_t \rangle + \eta_t,$$

  where $\langle *, * \rangle$ is the dot product operator, η_t is the noise under mild condition and θ^* is the linear model's parameter.
- The setting is depicted as follows. For each round $t = 1, 2, ...n$:
  - Agent chooses an arm whose feature is $a_t \in \mathcal{A}$, where $\mathcal{A}$ is the set of all arms' features.
  - Agent receives the reward $r_{a_t,t} = \langle \theta^*, a_t \rangle + \eta_t$ of a_t.
  - Note that only reward of a_t is observed by the agent.
- The goal is again minimizing the cumulative regret

$$R_T = \max_{a_i \in \mathcal{A}} \mathbb{E}[\sum_{t=1}^{T} r_{a_i,t}] - \mathbb{E}[\sum_{t=1}^{T} r_{a_t,t}]$$

  up to the horizon T.

LinUCB

- Recall that UCB implements the 'optimism in the face of uncertainty (OIFU)' principle.
- In linear bandit, the upper confidence bound needs some work to derive since rewards received give information of not only the pulled arm but also others.
- It is natural to take the following step to construct such upper confidence bound for each arm.
  - At time t, we construct a confidence set $\mathcal{C}_t$ so that θ^* lies in it with high probability.
  - We compute $\text{UCB}(a) = \arg\max_{\theta \in \mathcal{C}_t} \langle \theta, a \rangle$ for arm a.
- Once we have $\text{UCB}(a)$ for all $a \in \mathcal{A}$, we can choose

$$a_t = \arg\max_{a \in \mathcal{A}} \text{UCB}(a).$$

- The problem left is how to construct $\mathcal{C}_t$.

LinUCB

- One way to construct $\mathcal{C}_t$ is exploiting regularized least-squares estimator and a carefully designed ellipsoid centered at the estimator.
- The regularized least-squares estimator at time t is given by

$$\hat{\theta}_t = \arg\min_{\theta} \sum_{i=1}^{t} (r_{a_i} - \langle \theta, a_i \rangle)^2 + \lambda \|\theta\|_2^2.$$

- Denote $V_0 = \lambda I$ and $V_t = V_0 + \sum_{i=1}^{t} a_i a_i^T$, it can be shown that

$$\hat{\theta}_t = V_t^{-1} \sum_{i=1}^{t} a_i r_{a_i}.$$

- We can choose $\mathcal{C}_t$ in the form of

$$\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}\|_{V_{t-1}}^2 \leq \beta_t\}$$

where β_t is is an increasing sequence of constants.

LinUCB's Regret

- With carefully chosen β_t, as t grows, the volume of the ellipse is shrinking since V_t's eigenvalues are increasing.
- The following theorem gives a clue about the choosing of β_t.

Theorem

Let $\delta \in (0,1)$. With probability $1-\delta$ it holds that for all $t \in \mathbb{N}$,

$$\|\hat{\theta}_t - \theta^*\|_{V_t(\lambda)} < \sqrt{\lambda}\|\theta^*\|_2 + \sqrt{2\log\frac{1}{\delta} + \log\frac{detV_t(\lambda)}{\lambda^d}}.$$

- It can be shown that if
  - $|\langle \theta^*, a\rangle| \leq 1$ for all $a \in \mathcal{A}$,
  - $\|a\|_2 \leq L$ for all $a \in \mathcal{A}$,
  - $\sqrt{\beta_n} = \sqrt{\lambda}L + \sqrt{2\log\frac{1}{\delta} + d\log\frac{d\lambda+nL^2}{d\lambda}}$,

  then with $\delta = 1/n$ it holds that $R_n \leq Cd\sqrt{n}\log(nL)$, where C is a constant and d is the dimension of arm features.

Computation Issue

- LinUCB requires the solution to the following optimization problem

$$(a_t, \tilde{\theta}) = \arg\max_{a \in \mathcal{A}, \theta \in \mathcal{C}_t} \langle \theta, a \rangle$$

  to obtain a_t. This is a bi-linear optimization problem which could be very hard.

- When $\mathcal{C}_t$ is chosen to be the ellipsoid above, however, this problem is equivalent to

$$a_t = \arg\max_{a \in \mathcal{A}} \langle \hat{\theta}_t, a \rangle + \sqrt{\beta_t} \|a\|_{V_{t-1}^{-1}},$$

  which can be solved by iterating over arm set $\mathcal{A}$ and in time $O(|\mathcal{A}|)$. Here $\|a\|_{V_{t-1}^{-1}} = \sqrt{a^\top V_{t-1}^{-1} a}$

Thompson Sampling for Linear Bandit

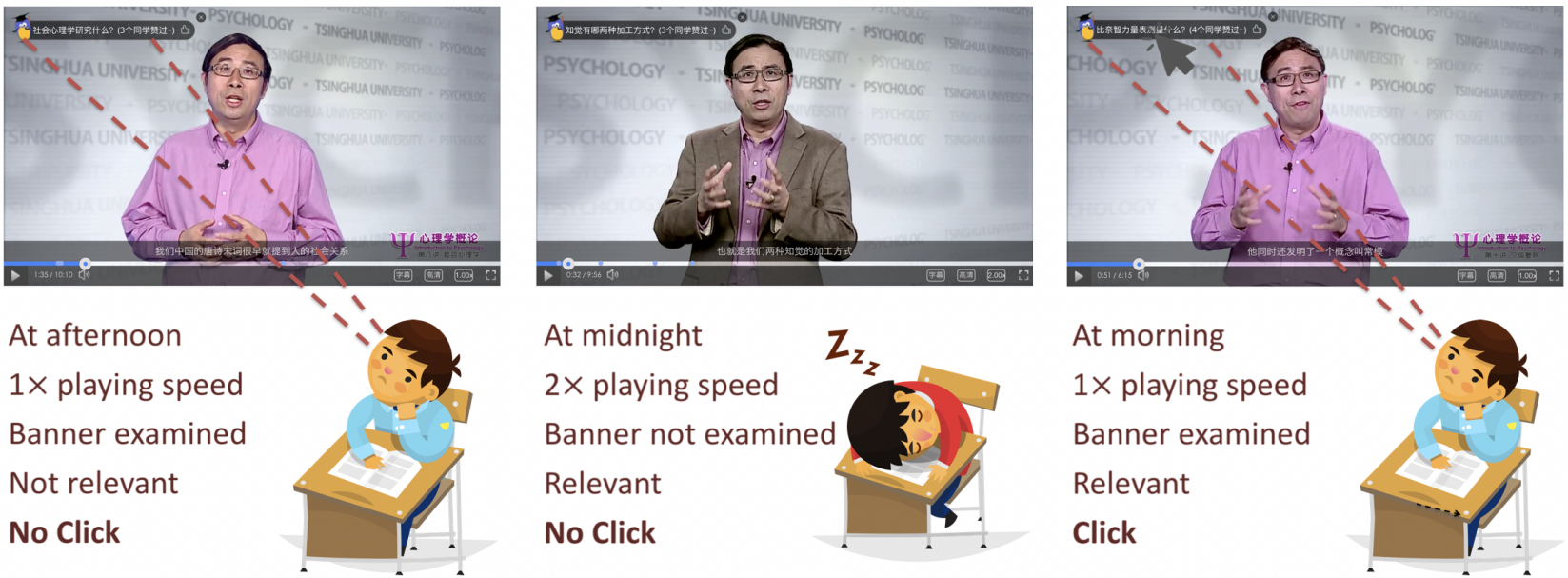
- Thompson sampling, or probability matching, has been shown as a good strategy for K-armed bandit.
- We can design similar strategy based on Thompson sampling for linear bandit.
- Assume $\theta_1 \sim N(0, \lambda I)$. It can be shown that $\theta_t \sim N(V_{t-1}^{-1} \sum_{i=1}^{t-1} a_i r_{a_i}, V_{t-1}^{-1})$. Then Thompson sampling for linear bandit is realized as follows. At time t,
  - Sample $\tilde{\theta}$ from $N(V_{t-1}^{-1} \sum_{i=1}^{t-1} a_i r_{a_i}, V_{t-1}^{-1})$.
  - Play arm
$$a_t = \arg\max_{a \in \mathcal{A}} \langle \tilde{\theta} \rangle.$$
- Thompson sampling for linear bandit has been shown to have regret of order $O(d^{3/2}\sqrt{T} \log T)$, which has a additional $\sqrt{d}$ factor compared to LinUCB.
- However, Thompson sampling has shown better performance than LinUCB empirically.

E-C Bandit with Implicit Feedback

- Clicks are implicit feedbacks, which imply biased preference of users.
  - In a page returned by search engine, users always prefer items positioned higher.
  - An item not being click may not decide that item is bad.
- Examination hypothesis: an item is clicked if and only if it is examinated and relevant.
- Based on examination hypothesis we proposed E-C bandit model[4].

[4] Yi Qi, Qingyun Wu, Hongning Wang, Jie Tang and Maosong Sun. "Bandit Learning with Implicit Feedback." NeurIPS (2018).

E-C Bandit



Question: Does no click indicates poor content recommendation?

Examination Hypothesis: Item is Clicked ***if and only if*** it's both examined and relevant.

Figure 3: An illustration of implicit feedback. This is an application we deployed on XuetangX.

E-C Bandit

- The model is formulated as

$$\begin{aligned} \mathbb{P}(C_t = 1 | E_t = 0, \mathbf{x}_{C,t}) &= 0 \\ \mathbb{P}(C_t = 1 | E_t = 1, \mathbf{x}_{C,t}) &= \rho(\mathbf{x}_{C,t}^\top \boldsymbol{\theta}_C^*) \\ \mathbb{P}(E_t = 1 | \mathbf{x}_{E,t}) &= \rho(\mathbf{x}_{E,t}^\top \boldsymbol{\theta}_E^*) \end{aligned} \tag{1}$$

  where E_t denotes the examination variable, $\mathbf{x}_{C,t}, \mathbf{x}_{E,t}$ are the contextual feature of relevance and examination respectively.
- Learning this model requires approximation since the max-likelihood estimator is non-convex.
- Based on variational inference and Thompson sampling, this model can be learnt on the fly.

E-C Bandit

Algorithm 1 Thompson sampling for E-C Bandit

1: Initiate $\mathbf{\Sigma}_C = \lambda I, \mathbf{\Sigma}_E = \lambda I, \hat{\boldsymbol{\theta}}_C = \boldsymbol{\theta}_{C,0}, \hat{\boldsymbol{\theta}}_E = \boldsymbol{\theta}_{E,0}$.
2: **for** $t = 0, 1, 2...$ **do**
3: Observe the available arm set $\mathcal{A}_t \subset \mathcal{A}$ and its corresponding context set $\mathcal{X}_t := \{(\mathbf{x}_C^a, \mathbf{x}_E^a) : a \in \mathcal{A}_t\}$.
4: Randomly sample $\tilde{\boldsymbol{\theta}}_C \sim N(\hat{\boldsymbol{\theta}}_C, \mathbf{\Sigma}_C)$, $\tilde{\boldsymbol{\theta}}_E \sim N(\hat{\boldsymbol{\theta}}_E, \mathbf{\Sigma}_E)$.
5: Select:

$$a_t = \arg\max_{a \in \mathcal{A}_t} \rho((\mathbf{x}_C^a)^\top \tilde{\boldsymbol{\theta}}_C) \rho((\mathbf{x}_E^a)^\top \tilde{\boldsymbol{\theta}}_E)$$

6: Play the selected arm a_t and Observe the reward C_t.
7: Update $\mathbf{\Sigma}_C, \hat{\boldsymbol{\theta}}_C, \mathbf{\Sigma}_E, \hat{\boldsymbol{\theta}}_E$ according to Eq (3) to Eq (6) respectively.
8: **end for**

Figure 4: The Thompson sampling algorithm for E-C bandit.

$$\mathbf{\Sigma}_{C,\text{post}}^{-1} = \mathbf{\Sigma}_C^{-1} + 2q^{1-C}\lambda(\xi_C)\mathbf{x}_C\mathbf{x}_C^\top \tag{3}$$

$$\hat{\boldsymbol{\theta}}_{C,\text{post}} = \mathbf{\Sigma}_{C,\text{post}}(\mathbf{\Sigma}_C^{-1}\hat{\boldsymbol{\theta}}_C + \frac{1}{2}(-q)^{1-C}\mathbf{x}_C) \tag{4}$$

$$\mathbf{\Sigma}_{E,\text{post}}^{-1} = \mathbf{\Sigma}_E^{-1} + 2\lambda(\xi_E)\mathbf{x}_E\mathbf{x}_E^\top \tag{5}$$

$$\hat{\boldsymbol{\theta}}_{E,\text{post}} = \mathbf{\Sigma}_{E,\text{post}}(\mathbf{\Sigma}_E^{-1}\hat{\boldsymbol{\theta}}_E + \frac{1}{2}(2q-1)^{1-C}\mathbf{x}_E) \tag{6}$$

Figure 5: Update formula of E-C bandit.

Result

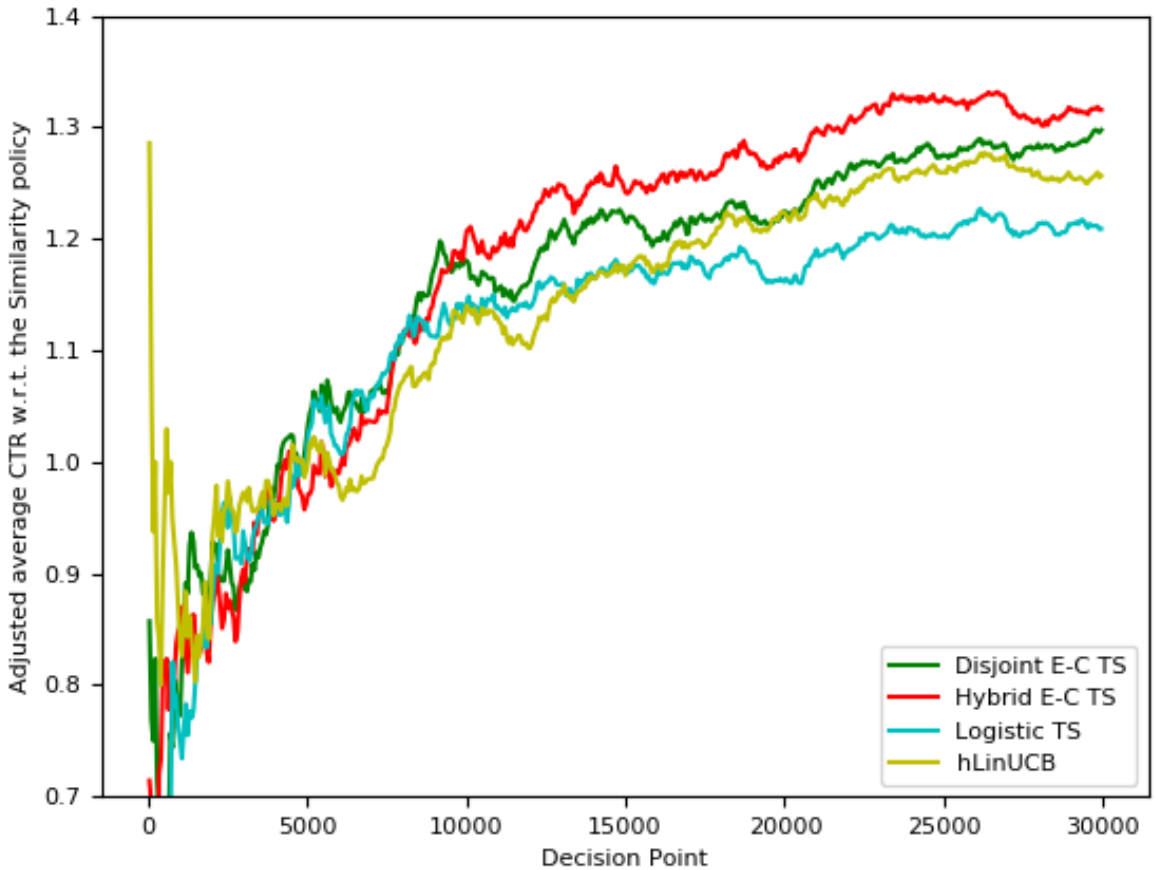


Figure 6: Performance comparison on MOOC's data.

Conversational Contextual Bandit[5]

- Recommendation: contextual bandit
- Conversation: occasionally asks questions and gets answers
- Goal: to improve learning speed of contextual bandit
- Example:
  - Q: Are you interested in basketball?
  - A: Yes
- News recommendation: arm = article, reward = click, context = user + article

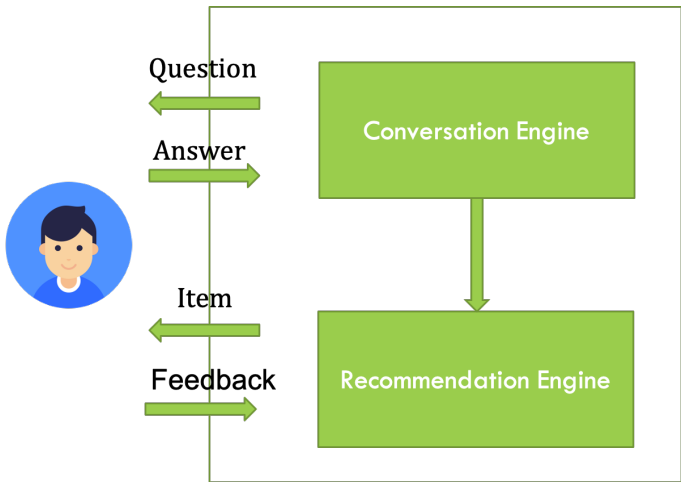


Figure 7: Conversation Recommendation.

[5] Xiaoying Zhang, Hong Xie, Hang Li, and John CS Lui. "Conversational Contextual Bandit: Algorithm and Application." In Proceedings of The Web Conference 2020, pp. 662-672. 2020.

Key-term and Article Bipartite Graph

- Question: key-term, Answer: Yes/No
- Bipartite graph of key-terms and articles
- Key-terms represent topics, entities, etc.
- Edges represent association, and weights on edges represent strengths of association
- Function of conversation frequency
  - $b(t) = t$, converse at every round
  - $b(t) = \lfloor \frac{t}{m} \rfloor$, converse at every m rounds
- Indicator function of conversation
  - $q(t) = b(t) - b(t-1)$

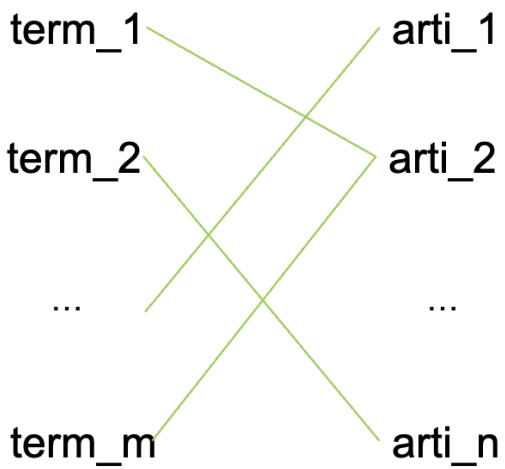


Figure 8: Key-term and article bipartite graph.

ConUCB Algorithm

- Given context vector $x_{t,a}$ for each arm a
- If $q(t) = 1$
  - Select key-term k_t
  - Receive Answer $\widetilde{r}_{t,k_t}$
  - Update parameters w.r.t. the conversation
- Estimate reward $r_{t,a}$ for each context $x_{t,a}$ based on history of both conversational and behavioral feedbacks
- Estimate confidence interval $c_{t,a}$
- Select arm $a_t = \text{argmax}_{a \in A_t} r_{t,a} + c_{t,a}$
- Receive reward r_{t,a_t}
- Update parameters w.r.t. the recommendation

The regret upper bound of ConUCB is better than LinUCB (see the paper for detailed results).

Experiment Results

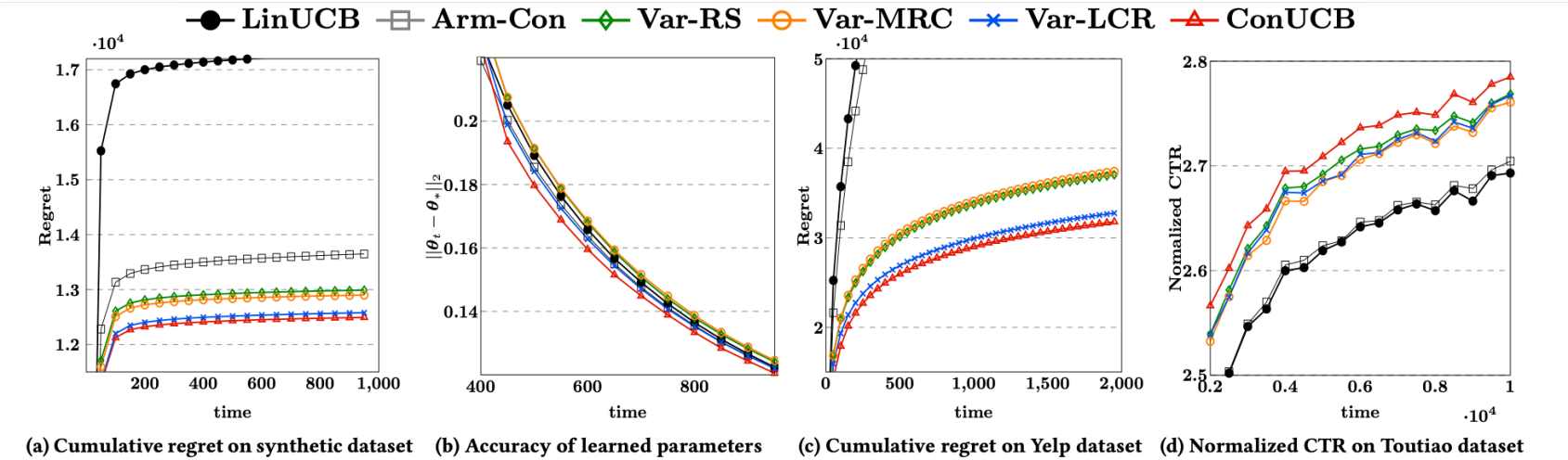


Figure 9: ConUCB outperforms baselines like LinUCB on synthetic, Yelp, and Toutiao datasets.

Conclusion

- Bandit provides a flexible framework to balance exploration and exploitation, which is the essential characteristic of many online learning environment.
- Both UCB-type algorithms and Thompson sampling algorithms lead to satisfying empirical results, yet their theoretical soundness is established mostly in simple case, for example, linear model. Meanwhile, ϵ-greedy is simple and effective in many cases.
- Bandit model has been used in many applications involving sequential decision process, for example, mobile health, online recommender system, etc.

Outline

1 Introduction

2 Stochastic Bandit

3 Appendix

Appendix: Detailed Proofs for Several Regret Analyses

Regret Analysis of ϵ-greedy

Show that $\mathbb{E}(R_n) = \epsilon \frac{n}{K} \sum_{i=1}^{K} \Delta_i$, where n is the round number, K is the number of arms and Δ_i is defined in pp.12.

Proof:

$$\begin{aligned} \mathbb{E}(R_n) &= \sum_{t=1}^{n} \sum_{i=1}^{k} \mathbb{P}(A_t = i) \Delta_i \\ &= \sum_{t=1}^{n} \sum_{i=1, i \neq i^*}^{k} \frac{\epsilon}{k} \Delta_i = \frac{n\epsilon}{k} \sum_{i=1}^{k} \Delta_i \end{aligned} \tag{2}$$

Chernoff-Hoeffding bound and Bernstein inequality

- **Chernoff-Hoeffding bound**. Let $X_1, X_2, ..., X_n$ be random variables with common range $[0, 1]$ and such that $\mathbb{E}(X_t|X_1, ..., X_{t-1}) = \mu$. Let $S_n = X_1 + ... + X_n$. Then for all $a \geq 0$

$$\mathbb{P}(S_n \geq n\mu + a) \leq \exp\left(-\frac{2a^2}{n}\right) \text{ and } \mathbb{P}(S_n \leq n\mu - a) \leq \exp\left(-\frac{2a^2}{n}\right)$$

- **Bernstein inequality**. Let $X_1, X_2, ..., X_n$ be random variables with range $[0, 1]$, and $\sum_{t=1}^{n} \text{Var}[X_t|X_{t-1}, ..., X_1] = \sigma^2$. Let $S_n = X_1 + ... + X_n$. Then for all $a \geq 0$

$$\mathbb{P}(S_n \geq \mathbb{E}(S_n) + a) \leq \mathbb{E}\left(-\frac{a^2/2}{\sigma^2 + a/2}\right)$$

Regret Analysis of ϵ_n-greedy

Show that ϵ_n-greedy has logarithmic regret.

Proof:

Recall that, for $n \geq \frac{CK}{\Delta^2}$, $\epsilon_n = \frac{CK}{\Delta^2 n}$, where C is a constant, K is the number of arms, Δ is defined in pp.17. Let

$$x_0 = \frac{1}{2K}\sum_{t=1}^{n}\epsilon_t, \quad \hat{\mu}_{i,t-1} = \frac{1}{T_i(t-1)}\sum_{s=1}^{T_i(t-1)} r_{i,s}$$

The probability that machine j is chosen at time n is

$$\mathbb{P}(I_n = j) \leq \frac{\epsilon_n}{K} + (1 - \frac{\epsilon_n}{K})\mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \hat{\mu}^*_{T^*(n-1)})$$

Continuing Proof

Note that $\hat{\mu}_{j,T_j(n-1)} \geq \hat{\mu}^*_{T^*(n-1)}$ holds means that at least one of the following holds

$$\hat{\mu}_{j,T_j(n-1)} \geq \mu_j + \frac{\Delta_j}{2}, \quad \hat{\mu}^*_{T^*(n-1)} \leq \mu^* - \frac{\Delta_j}{2}$$

(otherwise $\hat{\mu}_{j,T_j(n-1)} < \hat{\mu}^*_{T^*(n-1)}$)
Therefore,

$$\begin{aligned}\mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \hat{\mu}^*_{T^*(n-1)}) \leq & \mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \mu_j + \frac{\Delta_j}{2}) \\ & + \mathbb{P}(\hat{\mu}^*_{T^*(n-1)} \leq \mu^* - \frac{\Delta_j}{2}).\end{aligned}$$

($\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$)

Continuing Proof

Now the analysis for both terms on the right-hand side is the same. Let $T_j^R(n)$ be the the number of plays in which arm j was chosen at random in the first n plays. Then we have

$$\mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \mu_j + \frac{\Delta_j}{2})$$

$$= \sum_{t=1}^{n} \mathbb{P}(T_j(n) = t \wedge \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2})$$

$$= \sum_{t=1}^{n} \mathbb{P}(T_j(n) = t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) \mathbb{P}(\hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2})$$

$$\leq \sum_{t=1}^{n} \mathbb{P}(T_j(n) = t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) e^{-\Delta_j^2 t/2} \text{(Chernoff-Hoeffding bound)}$$

Continuing Proof

$$\leq \sum_{t=1}^{\lfloor x_0 \rfloor} \mathbb{P}(T_j(n) = t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) e^{-\Delta_j^2 t/2} + \sum_{\lfloor x_0 \rfloor}^{n} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2}$$

$$\leq \sum_{t=1}^{\lfloor x_0 \rfloor} \mathbb{P}(T_j(n) = t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2}$$

$$(\text{since } \sum_{t=x+1}^{\infty} e^{-Kt} \leq \frac{1}{K} e^{-Kx})$$

$$\leq \sum_{t=1}^{\lfloor x_0 \rfloor} \mathbb{P}(T_j^R(n) \leq t | \hat{\mu}_{j,t} \geq \mu_j + \frac{\Delta_j}{2}) + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2}$$

$$\leq x_0 \mathbb{P}(T_j^R(n) \leq x_0) + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor/2} (T_j^R(n) \text{ and } \hat{\mu}_{j,t} \text{ are independent.})$$

Continuing Proof

As $\mathbb{E}[T_j^R] = \frac{1}{K}\sum_{t=1}^{n}\epsilon_t$, $\mathrm{Var}[T_j^R] = \sum_{t=1}^{n}\frac{\epsilon_t}{K}(1-\frac{\epsilon_t}{K}) \leq \frac{1}{K}\sum_{t=1}^{n}\epsilon_t$, by Bernstein's inequality we have

$$\mathbb{P}(T_j^R(n) < x_0) \leq e^{-x_0/5}.$$

Denote $n' = CK/\Delta^2$. For $n \geq n' and$, we can lower-bound x_0 by

$$\begin{aligned} x_0 &= \frac{1}{2K}\sum_{t=1}^{n}\epsilon_t = \frac{1}{2K}\sum_{t=1}^{n'-1}\epsilon_t + \frac{1}{2K}\sum_{t=n'}^{n}\epsilon_t \\ &\geq \frac{n'-1}{2K} + \frac{C}{2\Delta^2}\log\frac{n+1}{n'}(\frac{1}{n} > \log(\frac{n+1}{n}) \text{ as } \log(n) \text{ is concave}) \\ &\geq \frac{C}{\Delta^2}\log\frac{n\Delta^2 e^{1/2}}{cK}. \end{aligned}$$

Continuing Proof

Combining the lower-bound of x_0, we have,

$$\mathbb{P}(\hat{\mu}_{j,T_j(n-1)} \geq \mu_j + \frac{\Delta_j}{2}) \leq x_0 \mathbb{P}(T_j^R(n) \leq x_0) + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor / 2}$$

$$\leq x_0 e^{-x_0/5} + \frac{2}{\Delta_j^2} e^{-\Delta_j^2 \lfloor x_0 \rfloor / 2}$$

$$= O(\frac{c}{\Delta^2 n}).$$

By a symmetric analysis we can get similar results for $\mathbb{P}(\hat{\mu}^*_{T^*(n-1)} \leq \mu^* - \frac{\Delta_j}{2})$. Summing over $\frac{1}{n}$ leads to log function.

Thanks.

HP: http://keg.cs.tsinghua.edu.cn/jietang/
Email: jietang@tsinghua.edu.cn