Linear Stochastic Bandits with Heavy-Tailed Payoffs

Han Shao

Committee: Prof. ZHANG Shengyu
Prof. KING Kuo Chin Irwin
Prof. LYU Rong Tsong Michael
Prof. CHAN Siu On
Prof. YEUNG Dit-Yan

Department of Computer Science and Engineering
The Chinese University of Hong Kong

MPhil Oral Defense, April 8, 2019

Outline

- Introduction
- A Survey of Bandits
- Linear Stochastic Bandits with Heavy-Tailed Payoffs
- Conclusions and Future Directions

Outline

- Introduction
- A Survey of Bandits
- Linear Stochastic Bandits with Heavy-Tailed Payoffs
- Conclusions and Future Directions

Multi-Armed Bandits (MAB)

- An agent has T rounds to play bandits
- At each time, the agent pulls one arm and observes a reward
- There is an optimal arm

Multi-Armed Bandits (MAB)

⋮

Multi-Armed Bandits (MAB)

How to maximize cumulative rewards?

Multi-Armed Bandits (MAB)

Problem definition

- Scenario: K arms

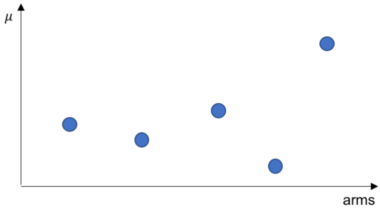
v_1 v_2 v_3 $\cdots$ v_K

- Model: sequential decision making to maximize cumulative rewards

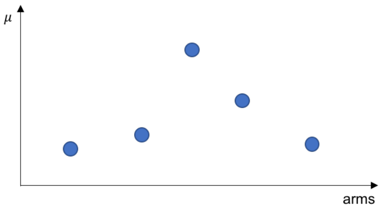
input: the arm set $\{1, \cdots, K\}$, and the number of rounds $T \geqslant K$
For time $t = 1, \cdots, T$,
  an agent selects an arm $I_t \in \{1, \cdots, K\}$
  observes a stochastic reward $y_t(I_t) \sim v_{I_t}$ of the chosen arm I_t

- Remarks: for $y \sim v_i$, $\mathbb{E}[y] = u_i$ and $u_* \triangleq \max_{i=1,\cdots,K} u_i$

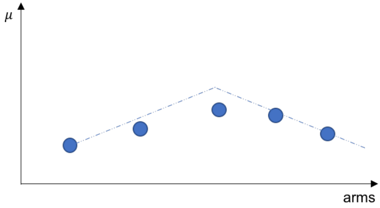
Structured Bandits



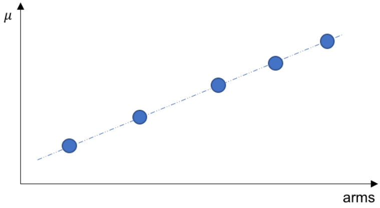
classical



unimodal



Lipschitz



linear

Linear Stochastic Bandits (LSB)

Problem definition

- Scenario:
  - Arms are represented by d-dimensional vectors

$\cdots$

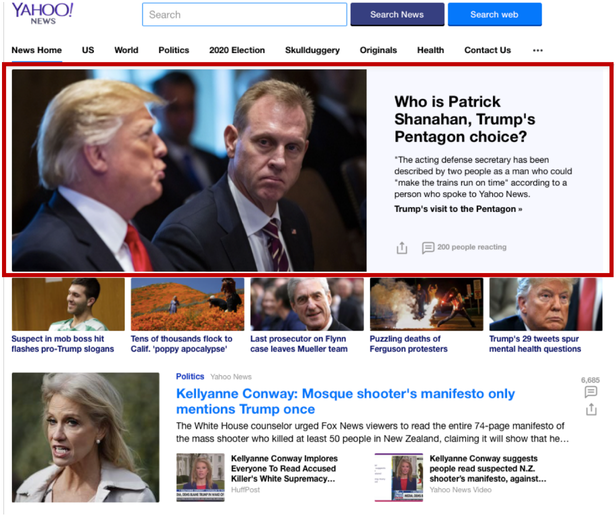
a 2-d case: $(1, 0)$ $(0.1, 0.5)$ $(0.8, 0.2)$ $(0.5, 0.5)$ $\cdots$

Input: the number of rounds T
for time $t = 1, \cdots, T$,
given the arm set $D_t \subseteq \mathbf{R}^d$, an agent selects an arm $x_t \in D_t$
observes a stochastic reward $y_t(x_t) = x_t^\top \theta_* + \eta_t$, where η_t is a stochastic noise

- Remarks:
  - Usually, η_t follows a sub-Gaussian distribution

Motivation

Personalized recommendations

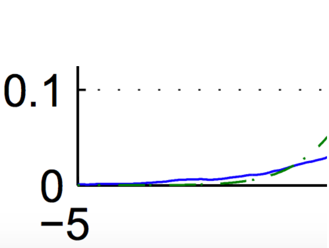


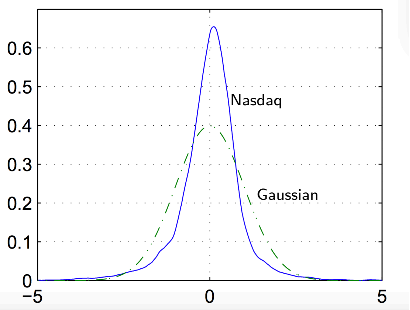
News recommendation (Li et al., 2010)

Motivation

Portfolio managements

- Sequentially invest T units of money in d financial products
- At each round, select a weight $w \in [0,1]^d$
- Returns in the investment are rewards in LSB
- High-probability extreme returns exist in financial markets

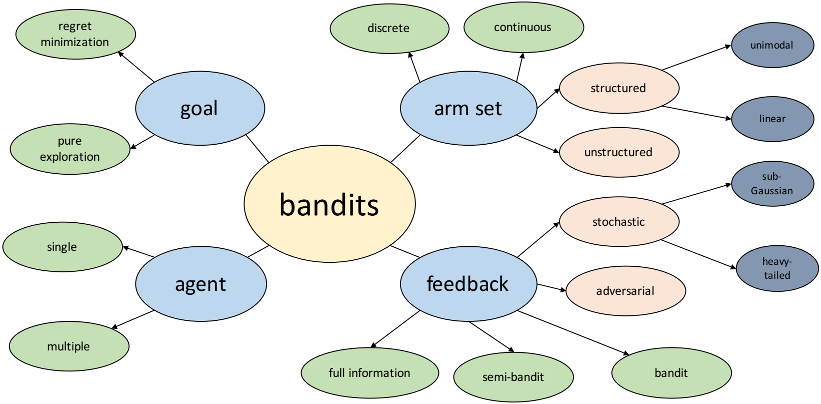




Outline

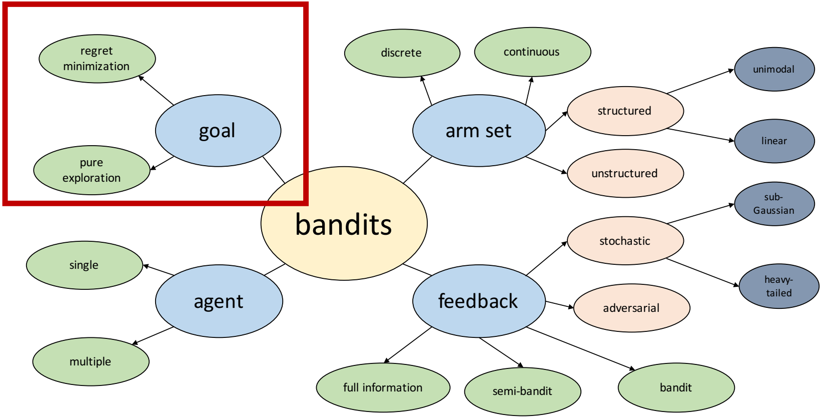
- Introduction
- A Survey of Bandits
- Linear Stochastic Bandits with Heavy-Tailed Payoffs
- Conclusions and Future Directions

A Taxonomy



bandits
goal
regret minimization
pure exploration
arm set
discrete
continuous
structured
unimodal
linear
unstructured
agent
single
multiple
feedback
stochastic
sub-Gaussian
heavy-tailed
adversarial
full information
semi-bandit
bandit

A Taxonomy

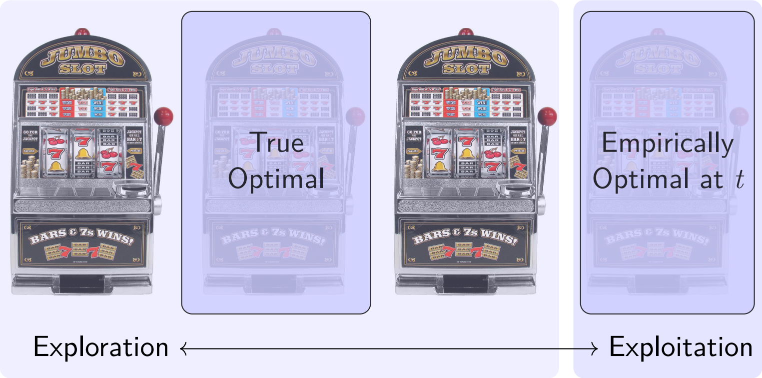

regret minimization
pure exploration
goal
discrete
continuous
arm set
structured
unstructured
unimodal
linear
bandits
sub-Gaussian
stochastic
heavy-tailed
adversarial
single
agent
multiple
feedback
full information
semi-bandit
bandit

Goal and Metric

Regret minimization

$\min \mathbf{R}(\mathcal{A}, T)$ (equivalent to rewards maximization)

$$\mathbf{R}(\mathcal{A}, T) \triangleq \max_{i=1,\cdots,K} \mathbb{E}\left[\sum_{t=1}^{T} y_t(i) - \sum_{t=1}^{T} y_t(I_t)\right] = Tu_* - \sum_{t=1}^{T} u_{I_t} \quad (1)$$



Goal and Metric

Pure exploration

Probability of error: $\mathbb{P}[x_T \neq \text{Opt}] \leqslant \delta$

- x_T is the output of $\mathcal{A}$ at time T and Opt is the optimal arm
- Two settings:
  - Fixed confidence: given δ, what is the smallest T?
  - Fixed budget: given T, what is the smallest δ?

Heuristic Methods for Regret Minimization

Selecting the arm with largest empirical average

A four-armed case with Bernoulli distributions
True means: $\{0.7, 0.8, 0.6, 0.5\}$

round	arm 1	arm 2	arm 3	arm 4
$1-4$	$\frac{1}{1}=1$	$\frac{0}{1}=0$	$\frac{1}{1}=1$	$\frac{1}{1}=1$
5	$\frac{1+0}{2}=0.5$	0	1	1
6	0.5	0	$\frac{1+0}{2}=0.5$	1
7	0.5	0	0.5	$\frac{1+0}{2}=0.5$
8	0.5	0	0.5	$\frac{1+0}{3}=0.3$

$\vdots$

Heuristic Methods for Regret Minimization

Selecting the arm with largest empirical average + standard deviation

A four-armed case with Bernoulli distributions

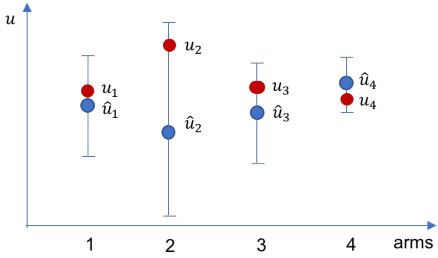
True means: $\{0.7, 0.8, 0.6, 0.5\}$

round	arm 1	arm 2	arm 3	arm 4
$1-4$	$\frac{1}{1}+1=2$	$\frac{0}{1}+1=1$	$\frac{1}{1}+1=2$	$\frac{1}{1}+1=2$
5	$\frac{1+0}{2}+0.7=1.2$	1	2	2
6	1.2	1	$\frac{1+0}{2}+0.7=1.2$	2
7	1.2	1	1.2	$\frac{1+0}{2}+0.7=1.2$
8	1.2	1	1.2	$\frac{1+0}{3}+0.6=0.9$

$\vdots$

Methodology for Stochastic Bandits

- Frequentist approach: Upper Confidence Bound (UCB)
  - Construct an estimate and confidence interval of u_i
  - Select the arm with the largest value among supremes of the confidence intervals



- Bayesian approach: Thompson sampling
  - Construct a posterior distribution of u_i
  - Sample from posterior distributions and select the arm with the largest sample value

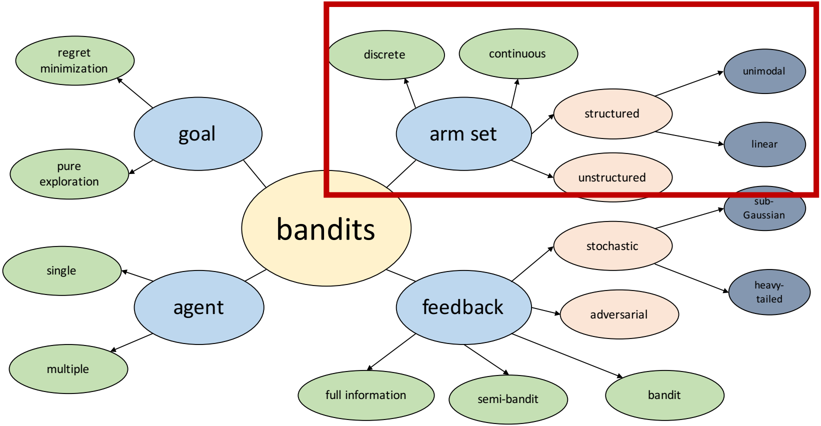
Theoretical Developments of Regret Minimization in MAB

work	results
(Thompson, 1933)	original formalization
(Lai & Robbins, 1985)	the first theoretical analysis $\lim_{T\to\infty} \frac{\mathbf{R}(\mathcal{A},T)}{\log(T)} \geqslant \sum_{\Delta_i>0} \frac{\Delta_i}{\mathsf{KL}(u_i,u_*)}$ $\lim_{T\to\infty} \frac{\mathbf{R}(\mathsf{UCB},T)}{\log(T)} \leqslant \sum_{\Delta_i>0} \frac{\Delta_i}{\mathsf{KL}(u_i,u_*)}$
(Agrawal, 1995)	a simpler algorithm $\lim_{T\to\infty} \frac{\mathbf{R}(\mathsf{SM},T)}{\log(T)} \leqslant \sum_{\Delta_i>0} \frac{\Delta_i}{\mathsf{KL}(u_i,u_*)}$
(Auer et al., 2002)	finite-time analysis $\mathbf{R}(\mathsf{UCB1},\, T) = O\left(\sum_{\Delta_i>0} \frac{\log(T)}{\Delta_i}\right)$ $\mathbf{R}(\mathsf{UCB1},\, T) = O\left(\sqrt{T}\right)$
(Agrawal et al., 2012)	Bernoulli payoffs $\mathbf{R}(\mathsf{TS},\, T) = O\left(\left(\sum_{\Delta_i>0} \frac{1}{\Delta_i}^2\right)^2 \log(T)\right)$
(Kaufmann et al., 2012)	Bernoulli payoffs $\lim_{T\to\infty} \frac{\mathbf{R}(\mathsf{TS},T)}{\log(T)} \leqslant \sum_{\Delta_i>0} \frac{\Delta_i}{\mathsf{KL}(u_i,u_*)}$
(Garivier et al., 2018)	finite-time lower bound small T: lower bound $\mathbf{R}(\mathcal{A},\, T) \geqslant \sum_{\Delta_i>0} \frac{\Delta_i T}{2K}$ large T: lower bound $\mathbf{R}(\mathcal{A},\, T) = \Omega\left(\sum_{\Delta_i>0} \frac{\Delta_i \log(T)}{\mathsf{KL}(u_i,u_*)}\right)$

Theoretical Developments of Pure Exploration in MAB

work	results
(Even-Dar et al., 2002)	bounded payoffs $\mathbb{P}\left[T \geqslant \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{K}{\delta \Delta_k}\right)\right] \leqslant \delta$
(Audibert & Bubeck, 2010)	bounded payoffs $\mathbb{P}[\text{Out} \neq \text{Opt}] \leqslant TK \exp\left(-\frac{T-K}{H_1}\right)$
(Karnin et al., 2013)	bounded payoffs $\mathbb{P}\left[T \geqslant \sum_{k=1}^{K} \Delta_k^{-2} \log\left(\frac{1}{\delta} \log\left(\frac{1}{\Delta_k}\right)\right)\right] \leqslant \delta$ $\mathbb{P}[\text{Out} \neq \text{Opt}] \leqslant \log(K) \exp\left(-\frac{T}{\log(K) H_2}\right)$
(Jamieson et al., 2014)	sub-Gaussian noises $\mathbb{P}\left[T \geqslant H_1 \log\left(\frac{1}{\delta}\right) + H_3\right] \leqslant 4\sqrt{c\delta} + 4c\delta$
(Kaufmann et al., 2016)	two-armed Gaussian bandits $\lim_{\delta \to 0} \frac{\mathbb{E}[T]}{\log(\frac{1}{\delta})} \geqslant \frac{2(\sigma_1+\sigma_2)^2}{(u_1-u_2)^2}$ $\lim_{\delta \to 0} \frac{\mathbb{E}[T]}{\log(\frac{1}{\delta})} \leqslant \frac{2(\sigma_1+\sigma_2)^2}{(u_1-u_2)^2}$ $\lim_{T \to \infty} \sup -\frac{\log(\mathbb{P}[\text{Out} \neq \text{Opt}])}{T} \leqslant \frac{(u_1-u_2)^2}{2(\sigma_1+\sigma_2)^2}$

A Taxonomy


regret minimization
pure exploration
goal
discrete
continuous
arm set
structured
unstructured
unimodal
linear
bandits
single
multiple
agent
feedback
stochastic
adversarial
sub-Gaussian
heavy-tailed
full information
semi-bandit
bandit

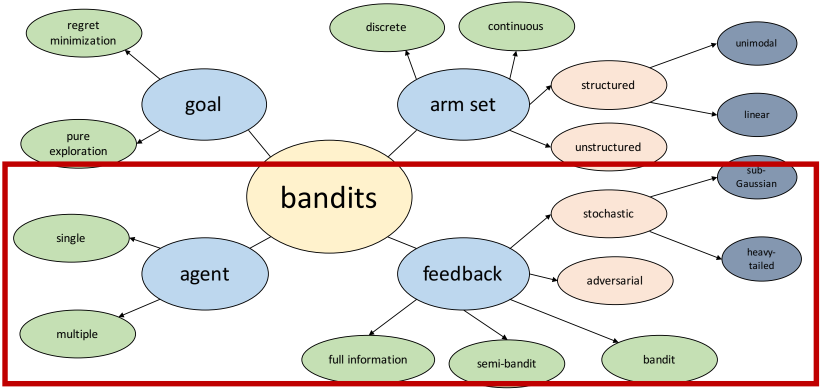
Theoretical Developments of LSB

work	results
(Abe & Long, 1999; Auer, 2000)	original formalization
(Auer, 2002)	first theoretical analysis; K arms $\mathbf{R}(\text{LinRel}, T) = O\left(\sqrt{Td}\log^{\frac{3}{2}}(KT\log(T))\right)$
(Dani et al., 2008)	compact arm set; bounded payoffs $\mathbf{R}(\mathcal{A}, T) = \Omega\left(d\sqrt{T}\right)$ $\mathbf{R}(\text{CB}_2, T) = O\left(d\sqrt{T}\log^{\frac{3}{2}}(T)\right)$
(Abbasi-Yadkori et al., 2011)	compact arm set; sub-Gaussian noises $\mathbf{R}(\text{OFUL}, T) = O\left(d\sqrt{T}\log(T)\right)$
(Agrawal & Goyal, 2013)	K arms; sub-Gaussian noises $\mathbf{R}(\text{TS}, T) = O\left(d^2\sqrt{T}\log(dT)\right)$
(Lattimore & Szepesvari, 2017)	K arms; Gaussian payoffs $\lim_{T\to\infty} \frac{\mathbf{R}(\mathcal{A},T)}{\log(T)} \geqslant c(\mathcal{A},\theta)$ $\lim_{T\to\infty} \frac{\mathbf{R}(\text{OA},T)}{\log(T)} \leqslant c(\mathcal{A},\theta)$

Other Classes of Structured Bandits

- Lipschitz (Magureanu et al., 2014): continuum-armed bandit problems
- Convex (Agarwal et al., 2011): continuum-armed bandit problems
- Unimodal (Combes & Proutiere, 2014): single-peak preferences economics and voting theory
- Dueling (Yue et al., 2012): intranet-search systems
- General (Combes et al., 2017)

A Taxonomy



regret minimization
pure exploration
goal
discrete
continuous
arm set
structured
unstructured
unimodal
linear
bandits
sub-Gaussian
stochastic
heavy-tailed
single
agent
multiple
feedback
adversarial
full information
semi-bandit
bandit

Some Important Variants of Bandits

- Agent
  - More than one agents $\rightarrow$ multi-player bandits
  - Application: cognitive radio systems
- Feedback
  - Rewards are not stochastic $\rightarrow$ adversarial bandits
  - Observe feedback about more arms $\rightarrow$
    - online learning with full information
    - online learning with semi-bandit feedback
  - Distributions of noises are non-sub-Gaussian $\rightarrow$ bandits with heavy-tailed distributions

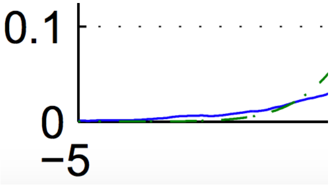
Outline

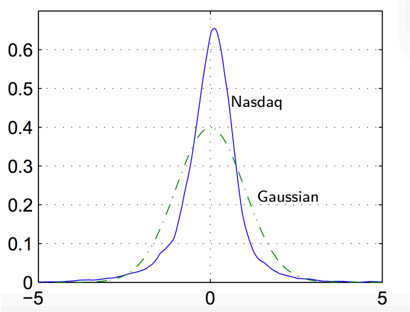
- Introduction
- A Survey of Bandits
- Linear Stochastic Bandits with Heavy-Tailed Payoffs
- Conclusions and Future Directions

What Is A Heavy-Tailed Distribution?

Practical scenarios

- High-probability extreme returns in financial markets



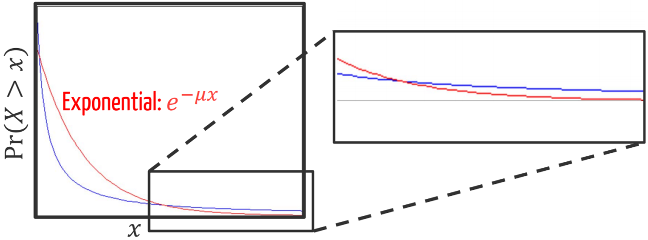


- Many other real cases
  1. Delays in communication networks (Liebeherr et al., 2012)
  2. Analysis of biological data (Burnecki et al., 2015)
  3. ...

Heavy-Tailed Distributions

Intuition and definition

- A distribution with a "tail" that is "heavier" than an exponential



http://users.cms.caltech.edu/~adamw/papers/2013-SIGMETRICS-heavytails.pdf

- Mathematically, a random variable X is said to be heavy-tailed if $\lim_{x\to\infty} e^{\phi x}\mathbb{P}[|X| > x] = \infty$ for all $\phi > 0$

Heavy-Tailed Distributions in Bandits

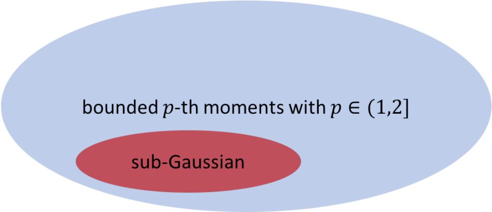
- Heavy-tailed distributions in bandits (Bubeck et al., 2013)

$$\mathbb{E}[X^p] < +\infty, \tag{2}$$

  where X is a stochastic reward/noise, and $p \in (1, 2]$
- Remarks
  - Eq. (2) is a subcase of the general definition of heavy tails
  - $p > 1$ is necessary for bandits as the expected payoff of each arm should be finite
  - The bounded p-th moments with $p \in (2, +\infty)$ can reduce to the case of $p = 2$ (Jensen's inequality)
  - Payoffs with sub-Gaussian noises are light-tailed with finite 2-nd moment

Weaker Assumption: Bounded p-th Moments

Examples



- Standard *Student's t-Distribution* with 3 degrees of freedom
  - The 2-nd central moment is bounded by 3
  - The 2-nd raw moment (with a constant shift a) is bounded by $3 + a^2$
- Pareto distribution with shape parameter α and scale parameter x_m
  - The p-th raw moments are bounded by $\alpha x_m^p / (\alpha - p)$, for all $p \in (1, \alpha)$
  - The p-th central moments are not directly available

LSB with Heavy-Tailed Payoffs

Problem definition

input: the arm set $\{D_t\}_{t=1}^{T}$, and the number of rounds T
For time $t = 1, \cdots, T$,
given the arm set $D_t \subseteq \mathbf{R}^d$, an agent selects an arm $x_t \in \mathbb{D}_t$
observes a stochastic reward $y_t = x_t^\top \theta + \eta_t$, where η_t is a stochastic noise

- Previous assumption (Abbasi-Yadkori et al., 2011): η_t is sub-Gaussian conditional on $\mathcal{F}_{t-1}$
- Our assumption: y_t or η_t is heavy-tailed conditional on $\mathcal{F}_{t-1}$
  - Bounded raw moments
  - Bounded central moments
- A connection in regret:
  $\widetilde{O}\left(\sqrt{T}\right)$ (sub-Gaussian) $\to \widetilde{O}\left(\sqrt{T}\right)$ (2-nd moment bounded)

Linear Stochastic Bandits with Heavy-Tailed Payoffs (LinBET)

LinBET

Given a arm set D_t for time step $t = 1, \cdots, T$, an algorithm $\mathcal{A}$, of which the goal is to maximize cumulative payoffs over T rounds, chooses an arm $x_t \in D_t$. With $\mathcal{F}_{t-1}$, the observed stochastic payoff $y_t(x_t)$ is conditionally heavy-tailed, i.e., $\mathbb{E}\left[|y_t|^p|\mathcal{F}_{t-1}\right] \leqslant b$ or $\mathbb{E}\left[|y_t - \langle x_t, \theta_* \rangle|^p|\mathcal{F}_{t-1}\right] \leqslant c$, where $p \in (1, 2]$, and $b, c \in (0, +\infty)$.

Challenges and Contributions

Challenges

- The lower bound of LinBET
- How to develop a robust estimator and bandit algorithms for LinBET
- Regret analysis for the proposed bandit algorithms

Contributions

- The first to provide the lower bound for LinBET
- Develop two novel bandit algorithms to solve LinBET
- Conduct experiments to demonstrate the effectiveness of the algorithms

Lower Bound of LinBET

Results

Assume $d \geqslant 2$ is even. For $D_t \in \mathbf{R}^d$, we fix the arm set as $D_t = D_{(d)}$, where $D_{(d)} \triangleq \{(x_1, \cdots, x_d) \in \mathbf{R}_+^d : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1\}$. Let $S_d \triangleq \{(\theta_1, \cdots, \theta_d) : \forall i \in [d/2], (\theta_{2i-1}, \theta_{2i}) \in \{(2\Delta, \Delta), (\Delta, 2\Delta)\}\}$ with $\Delta \in (0, 1/d]$. Payoffs are in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ such that, for every $x \in D_{(d)}$, the expected payoff is $\theta_*^\top x$.

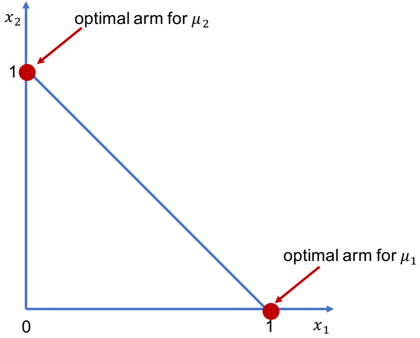
Theorem 1. If θ_* is chosen uniformly at random from S_d, and the payoff for each $x \in D_{(d)}$ is in $\{0, (1/\Delta)^{\frac{1}{p-1}}\}$ with mean $\theta_*^\top x$, then for any algorithm $\mathcal{A}$ and every $T \geqslant (d/12)^{\frac{p-1}{p}}$, we have

$$\mathbb{E}\left[R(\mathcal{A}, T)\right] \geqslant \frac{d}{192} T^{\frac{1}{p}}.$$

Lower Bound of LinBET

$d = 2$ and $\mathbb{E}[|y_t|^p|\mathcal{F}_{t-1}] \leqslant d$ case

- Arm set: $D_{(2)} \triangleq \{(x_1, x_2) \in \mathbf{R}_+^2 : x_1 + x_2 = 1\}$
- θ_* is chosen uniformly at random from $\{\mu_1, \mu_2\}$, where $\mu_1 = (2\Delta, \Delta)$ and $\mu_2 = (\Delta, 2\Delta)$
- Δ will be set as a small value dependent on T
- Change of measure (through $\mu_0 = (\Delta, \Delta)$)



Lower Bound of LinBET

$d = 2$ and $\mathbb{E}\left[|y_t|^p | \mathcal{F}_{t-1}\right] \leqslant d$ case

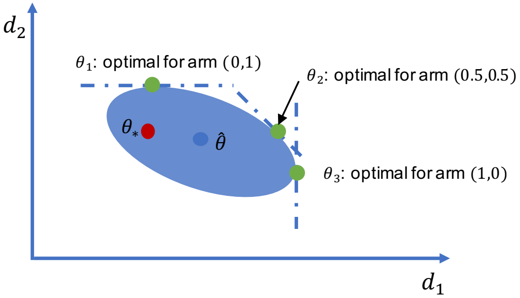
- Payoff distribution of x:

$$y(x) = \begin{cases} \left(\frac{1}{\Delta}\right)^{\frac{1}{p-1}} & \text{with a probability of } \Delta^{\frac{1}{p-1}} \theta_*^\top x, \\ 0 & \text{with a probability of } 1 - \Delta^{\frac{1}{p-1}} \theta_*^\top x \end{cases}$$

  - $\mathbb{E}\left[y(x)^p\right] \leqslant 2$
  - $\mathbb{E}\left[y(x)^q\right] \geqslant \left(\frac{1}{\Delta}\right)^{\frac{p-q}{p-1}}$, $q < p$

An Algorithm for LSB

Optimism in face of uncertainty (OFU) (Abbasi-Yadkori et al., 2011)



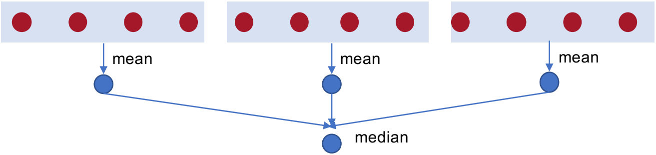
- At time t, select arm x_t by
  - $(x_t, \tilde{\theta}_t) = \arg\max_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$
  - $C_t = \{\theta : \|\theta - \hat{\theta}_{t,k^*}\|_{V_t} \leqslant \beta_t\},\ V_t = \lambda I + \sum_{\tau=1}^{t} x_\tau x_\tau^\top$

- The regret is bounded by $\widetilde{O}\left(\max_{t\in[T]} \beta_{t-1}\sqrt{T}\right)$
- For sub-Gaussian case, LSE $\rightarrow$ $\beta_t = \Theta\left(\sqrt{\log t}\right)$
- For heavy-tailed case, LSE $\rightarrow$ β_t is polynomial of t

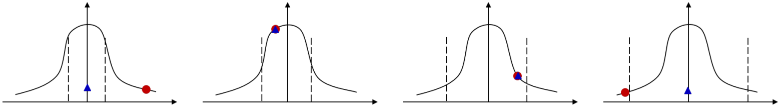
Techniques for Designing Algorithms

Median of means and truncation (Bubeck et al., 2013)

- Median of means



- Truncation



$t = 1$ $t = 2$ $t = 3$ $t = 4$

● sample drawn from the chosen arm
▲ sample after truncation

Previous Results

MoM and CRT by Medina & Yang (2016)

- Medina & Yang (2016) proposed two algorithms MoM (based on median of means) and CRT (based on truncation)
- Both achieved the regret of $\widetilde{O}(T^{\frac{3}{4}})$ when $p = 2$
- Is it possible to design algorithms to achieve the regret of $\widetilde{O}(\sqrt{T})$ when $p = 2$?

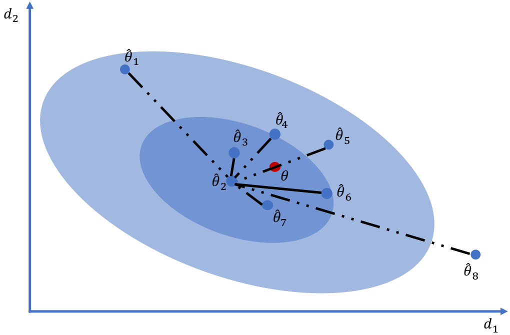
Algorithms: Median of means under OFU (MENU)

Algorithm 1 MENU

1: **input** d, c, p, δ, λ, S, T, $\{D_n\}_{n=1}^N$
2: **initialization:** $k = \lceil 24 \log\left(\frac{eT}{\delta}\right) \rceil$, $N = \lfloor \frac{T}{k} \rfloor$, $V_0 = \lambda I_d$, $C_0 = \mathbb{B}(\mathbf{0}, S)$
3: **for** $n = 1, 2, \cdots, N$ **do**
4: $(x_n, \tilde{\theta}_n) = \arg\max_{(x,\theta) \in D_n \times C_{n-1}} \langle x, \theta \rangle$
5: Play x_n for k times and observe payoffs $y_{n,1}, y_{n,2}, \cdots, y_{n,k}$
6: $V_n = V_{n-1} + x_n x_n^\top$
7: For $j \in [k]$, $\hat{\theta}_{n,j} = V_n^{-1} \sum_{i=1}^n y_{i,j} x_i$
8: For $j \in [k]$, let r_j be the median of $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} : s \in [k] \backslash j\}$
9: $k^* = \arg\min_{j \in [k]} r_j$
10: $\beta_n = 3\left((9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right)$
11: $C_n = \{\theta : \|\theta - \hat{\theta}_{n,k^*}\|_{V_n} \leqslant \beta_n\}$
12: **end for**

Understanding of MENU

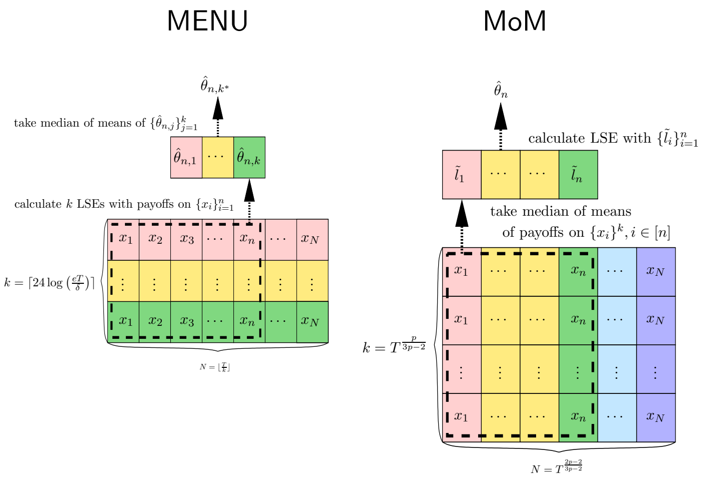
Median of means over linear parameters by Hsu & Sabato (2014)



- For each estimate, compute the distances between the estimate and estimates of other groups
- Take the median of the distances as the index of the estimate
- Select the estimate with the smallest index

Understanding of MENU

Framework comparison with MoM by Medina & Yang (2016)



Understanding of MENU

Result comparison with MoM by Medina & Yang (2016)

- For MoM by Medina & Yang (2016)
  - The regret is bounded by $\widetilde{O}\left(\max_{n=1,\cdots,N} \beta_{n-1} k\sqrt{N}\right)$, where $\beta_n = \Theta\left(k^{-\frac{p-1}{p}}\sqrt{n}\right)$
  - The value of k and N is constrained by $\max_{n=1,\cdots,N} \beta_n = \Omega(1)$
  - The regret of the MoM algorithm is $\widetilde{O}(c^{\frac{1}{p}} dT^{\frac{2p-1}{3p-2}})$
- For our MENU
  - Make each group contain the same playing history to compute regret easily
  - $k = \Theta\left(\log(T)\right)$
  - $\beta_n = \Theta(n^{\frac{2-p}{2p}})$

Upper Bound Analysis: MENU

Results

Theorem 2. Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \leqslant D$, $\|\theta_*\|_2 \leqslant S$, $|x_t^\top \theta_*| \leqslant L$ and $\mathbb{E}[|\eta_t|^p|\mathcal{F}_{t-1}] \leqslant c$. Then, with probability at least $1-\delta$, for every $T \geqslant 256 + 24\log(e/\delta)$, the regret of the MENU algorithm satisfies

$$R(\text{MENU}, T) \leqslant \widetilde{O}(c^{\frac{1}{p}}\, d^{\frac{1}{2}+\frac{1}{p}}\, T^{\frac{1}{p}}).$$

- The regret is $\widetilde{O}\left(\sqrt{T}\right)$ when $p = 2$

Algorithms: Truncation under OFU (TOFU)

Algorithm 2 TOFU

input d, b, p, δ, λ, T, $\{D_t\}_{t=1}^{T}$
initialization: $V_0 = \lambda I_d$, $C_0 = \mathbb{B}(\mathbf{0}, S)$
for $t = 1, 2, \cdots, T$ **do**
    $b_t = \left(\frac{b}{\log\left(\frac{2T}{\delta}\right)}\right)^{\frac{1}{p-1}} t^{\frac{2-p}{2p}}$
    $(x_t, \tilde{\theta}_t) = \arg\max_{(x,\theta)\in D_t \times C_{t-1}} \langle x, \theta \rangle$
    Play x_t and observe a payoff y_t
    $V_t = V_{t-1} + x_t x_t^\top$ and $X_t^\top = [x_1, \cdots, x_t]$
    $[u_1, \cdots, u_d]^\top = V_t^{-1/2} X_t^\top$
    for $i = 1, \cdots, d$ **do**
        $Y_i^\dagger = (y_1 \mathbb{1}_{u_{i,1} y_1 \leqslant b_t}, \cdots, y_t \mathbb{1}_{u_{i,t} y_t \leqslant b_t})$
    end for
    $\theta_t^\dagger = V_t^{-1/2}(u_1^\top Y_1^\dagger, \cdots, u_d^\top Y_d^\dagger)$
    $\beta_t = 4\sqrt{d} b^{\frac{1}{p}} \left(\log\left(\frac{2dT}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S$
    Update $C_t = \{\theta : \|\theta - \theta_t^\dagger\|_{V_t} \leqslant \beta_t\}$
end for

Understanding of TOFU

Comparison with CRT by Medina & Yang (2016)

- For CRT, the payoff at time t is truncated by α_t
  - $y_t^\dagger = y_t \mathbb{1}_{y_t \leqslant \alpha_t}$
  - The regret of the CRT algorithm is $\widetilde{O}(bdT^{\frac{1}{2}+\frac{1}{2p}})$
- For TOFU, at time t, all of the historical payoffs are truncated by b_t for each u_i
  - u_i is the i-th row of $V_t^{-\frac{1}{2}} X_t^\top$
  - $Y_i^\dagger = (y_1 \mathbb{1}_{u_{i,1} y_1 \leqslant b_t}, \cdots, y_t \mathbb{1}_{u_{i,t} y_t \leqslant b_t})$
  - $\theta_t^\dagger = V_t^{-\frac{1}{2}} (u_1^\top Y_1^\dagger, \cdots, u_d^\top Y_d^\dagger)$
- A 2-d example

arms	$(0,1)$	$(1,0)$
#pulls	50	1

Upper Bound Analysis: TOFU

Results

Theorem 3. Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \leqslant D$, $\|\theta_*\|_2 \leqslant S$, $|x_t^\top \theta_*| \leqslant L$ and $\mathbb{E}[|y_t|^p|\mathcal{F}_{t-1}] \leqslant b$. Then, with probability at least $1-\delta$, for every $T \geqslant 1$, the regret of the TOFU algorithm satisfies

$$R(\text{TOFU}, T) \leqslant \widetilde{O}(b^{\frac{1}{p}}\, d T^{\frac{1}{p}}).$$

- The regret is $\widetilde{O}\left(\sqrt{T}\right)$ when $p = 2$

Experimental Results

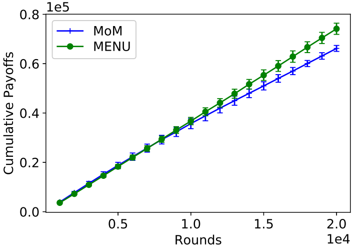
- Datasets
  - Four synthetic datasets
  - Metric: Cumulative payoffs
  - Baselines: MoM and CRT by Medina & Yang (2016)
- Setting
  - Run experiments in a personal computer with Intel CPU@3.70GHz and 16 GB memory
  - Run Independently ten times for each epoch
  - Show cumulative payoffs with one standard variance

Experimental Results

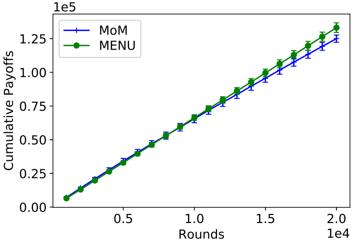
Synthetic Datasets

dataset	{#arms,#dims}	distribution {parameters}	$\{p, b, c\}$	optimal arm
S1	{20,10}	Student's t-distribution $\{\nu = 3, l_p = 0, s_p = 1\}$	{2.00, NA, 3.00}	4.00
S2	{100,20}	Student's t-distribution $\{\nu = 3, l_p = 0, s_p = 1\}$	{2.00, NA, 3.00}	7.40
S3	{20,10}	Pareto distribution $\{\alpha = 2, s_m = \frac{x_t^\top \theta_*}{2}\}$	{1.50, 7.72, NA}	3.10
S4	{100,20}	Pareto distribution $\{\alpha = 2, s_m = \frac{x_t^\top \theta_*}{2}\}$	{1.50, 54.37, NA}	11.39

Experimental Results



(a) S1



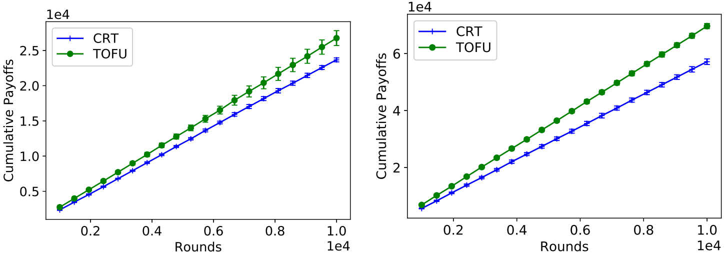
(b) S2

Figure 1: Comparison of cumulative payoffs for synthetic datasets S1-S2 with four algorithms.

Observation

- For S1-S2, our algorithm MENU beats MoM by Medina & Yang (2016)

Experimental Results



(a) S3 (b) S4

Figure 2: Comparison of cumulative payoffs for synthetic datasets S3-S4 with four algorithms.

Observation

- For S3-S4, our algorithm TOFU beats CRT by Medina & Yang (2016)

Summary

Contributions

- Derive lower bound for LinBET
- Develop two almost optimal bandit algorithms MENU and TOFU to solve LinBET
- Theoretical analysis of two algorithms

Publication: "Almost Optimal Algorithms for Linear Stochastic Bandits with Heavy-Tailed Payoffs" (NIPS 2018, Spotlight).

Discussions

- Efficiency of TOFU
- Problem-dependent bounds
- The impact of d

Outline

Conclusions

- Introduce the problem of bandits
- Conduct a brief survey
- Introduce our results in LinBET

Publication

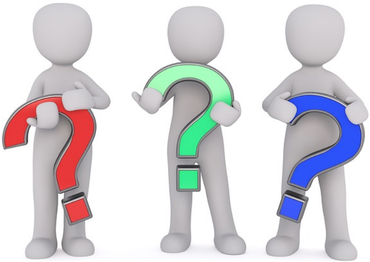
1. **Han Shao**, Xiaotian Yu, Irwin King and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 8430–8439, 2018. **Spotlight presentation**.
2. Xiaotian Yu, **Han Shao**, Michael R. Lyu and Irwin King. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 937–946, 2018.

Future Directions

1. Automatically learning in bandits
   - Setting: distributional parameter learning
   - Challenge: index learning and error control in distributional parameters
   - Motivation: unknown b or c information in real-world datasets

2. Removing forced exploration in structured bandits
   - Challenge: how to design an efficient adaptive learning framework
   - Motivation: the state-of-the-art algorithms use forced exploration

End



Comparison on Regret, Complexity and Storage of Four Algorithms

algorithm	MoM	MENU	CRT	TOFU
regret	$\widetilde{O}(T^{\frac{2p-1}{3p-2}})$	$\widetilde{O}(T^{\frac{1}{p}})$	$\widetilde{O}(T^{\frac{1}{2}+\frac{1}{2p}})$	$\widetilde{O}(T^{\frac{1}{p}})$
complexity	$O(T)$	$O(T\log T)$	$O(T)$	$O(T^2)$
storage	$O(1)$	$O(\log T)$	$O(1)$	$O(T)$

Upper Bound Analysis: MENU

Proof sketch

Lemma 1. [Confidence Ellipsoid of LSE] Let $\hat{\theta}_n$ denote the LSE of θ_* with the sequence of decisions $x_1, \cdots, x_n$ and observed payoffs $y_1, \cdots, y_n$. Assume that for all $\tau \in [n]$ and all $x_\tau \in D_\tau \subseteq \mathbf{R}^d$, $\mathbb{E}[|\eta_\tau|^p|\mathcal{F}_{\tau-1}] \leqslant c$ and $\|\theta_*\|_2 \leqslant S$. Then $\hat{\theta}_n$ satisfies

$$\Pr\left(\|\hat{\theta}_n - \theta_*\|_{V_n} \leqslant (9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S\right) \geqslant \frac{3}{4},$$

Lemma 2. Recall $\hat{\theta}_{n,j}$, $\hat{\theta}_{n,k^*}$ and V_n in MENU. If there exists a $\gamma > 0$ such that $\Pr\left(\|\hat{\theta}_{n,j} - \theta_*\|_{V_n} \leqslant \gamma\right) \geqslant \frac{3}{4}$ holds for all $j \in [k]$ with $k \geqslant 1$, then with probability at least $1 - e^{-\frac{k}{24}}$, $\|\hat{\theta}_{n,k^*} - \theta_*\|_{V_n} \leqslant 3\gamma$.

Upper Bound Analysis: MENU

Proof sketch of Lemma 1

- Let u_i denote the i-th row of $V_t^{-1/2} X_t^\top$
- $\|\hat{\theta}_n - \theta_*\|_{V_n} \leqslant \sqrt{\sum_{i=1}^d \left(u_i^\top (Y_n - X_n\theta_*)\right)^2} + \lambda\|\theta_*\|_{V_n^{-1}}$
- Union bound

$$\Pr\left(\sum_{i=1}^{d}\left(\sum_{\tau=1}^{n} u_{i,\tau}\eta_\tau\right)^2 > \gamma^2\right)$$

$$\leqslant \Pr\left(\exists i, \tau, |u_{i,\tau}\eta_\tau| > \gamma\right) + \Pr\left(\sum_{i=1}^{d}\left(\sum_{\tau=1}^{n} u_{i,\tau}\eta_\tau \mathbb{1}_{|u_{i,\tau}\eta_\tau| \leqslant \gamma}\right)^2 > \gamma^2\right),$$

  where $\mathbb{1}_{\{\cdot\}}$ is the indicator function
- Both terms could be bounded by Markov's inequality
- Set $\gamma = (9dc)^{\frac{1}{p}} n^{\frac{2-p}{2p}}$

Upper Bound Analysis: MENU

Proof sketch of Lemma 2

- By Azuma-Hoeffding's inequality, we have with prob. at least $1 - e^{-\frac{k}{24}}$, more than $2/3$ of $\{\hat{\theta}_{n,1}, \cdots, \hat{\theta}_{n,k}\}$ are contained in $\mathbb{B}_{V_n}(\theta_*, \gamma) \triangleq \{\theta : \|\theta - \theta_*\|_{V_n} \leqslant \gamma\}$
- r_j be the median of $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} : s \in [k] \backslash j\}$
- Select arm $\arg\min_{j \in [k]} r_j$
  - If $\hat{\theta}_{n,j} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$, $\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} \leqslant 2\gamma$ for all $\hat{\theta}_{n,s} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$ by triangle inequality. Therefore, $r_j \leqslant 2\gamma$
  - If $\hat{\theta}_{n,j} \notin \mathbb{B}_{V_n}(\theta_*, 3\gamma)$, $\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} > 2\gamma$ for all $\hat{\theta}_{n,s} \in \mathbb{B}_{V_n}(\theta_*, \gamma)$ by triangle inequality. Therefore, $r_j > 2\gamma$

Upper Bound Analysis: TOFU

Proof sketch

Lemma 3. [Confidence Ellipsoid of Truncated Estimate] With the sequence of decisions $x_1, \cdots, x_t$, the truncated payoffs $\{Y_i^\dagger\}_{i=1}^d$ and the parameter estimate $\theta_t^\dagger$ are defined in TOFU (i.e., Algorithm 2). Assume that for all $\tau \in [t]$ and all $x_\tau \in D_\tau \subseteq \mathbf{R}^d$, $\mathbb{E}[|y_\tau|^p|\mathcal{F}_{\tau-1}] \leqslant b$ and $\|\theta_*\|_2 \leqslant S$. With probability at least $1-\delta$, we have

$$\|\theta_t^\dagger - \theta_*\|_{V_t} \leqslant 4\sqrt{d} b^{\frac{1}{p}} \left(\log\left(\frac{2d}{\delta}\right)\right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}} + \lambda^{\frac{1}{2}} S, \quad (3)$$

where $\lambda > 0$ is a regularization parameter and $V_t = \lambda I_d + \sum_{\tau=1}^t x_\tau x_\tau^\top$.

Upper Bound Analysis: TOFU

Proof sketch of Lemma 3

- Like before,

$$\|\theta_t^\dagger - \theta_*\|_{V_t} \leqslant \sqrt{\sum_{i=1}^{d} \left(u_i^\top (Y_i^\dagger - X_t\theta_*)\right)^2} + \lambda \|\theta_*\|_{V_n^{-1}}$$

- For each i

$$u_i^\top \left(Y_i^\dagger - X_t\theta_*\right) = \sum_{\tau=1}^{t} u_{i,\tau} \left(Y_{i,\tau}^\dagger - \mathbb{E}[Y_{i,\tau}|\mathcal{F}_{\tau-1}]\right)$$

$$\leqslant \left|\sum_{\tau=1}^{t} u_{i,\tau}(Y_{i,\tau}^\dagger - \mathbb{E}[Y_{i,\tau}^\dagger|\mathcal{F}_{\tau-1}])\right| + \left|\sum_{\tau=1}^{t} u_{i,\tau}\mathbb{E}[Y_{i,\tau}\mathbb{1}_{|u_{i,\tau} Y_{i,\tau}|>b_t}|\mathcal{F}_{\tau-1}]\right|$$

- The first term is bounded by Bernstein's inequality
- Set $b_t = (b/\log(2d/\delta))^{\frac{1}{p}} \, t^{\frac{2-p}{2p}}$