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

- **Model:** sequential decision making to maximize cumulative rewards

**input:** the arm set \{1, \cdots, K\}, and the number of rounds $T \geq K$

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 $u_* \triangleq \max_{i=1,\ldots,K} u_i$
Structured Bandits

- Classical
- Unimodal
- Lipschitz
- Linear
Linear Stochastic Bandits (LSB)

Problem definition

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

a 2-d case: (1, 0) (0.1, 0.5) (0.8, 0.2) (0.5, 0.5) ...

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

- Remarks:
  - Usually, $\eta_t$ 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

- **Goal**: regret minimization, pure exploration
- **Agent**: single, multiple
- **Arm Set**: discrete, continuous
- **Feedback**: full information, semi-bandit, bandit
- **Bandits**: structured, unstructured, stochastic, adversarial, unimodal, linear, sub-Gaussian, heavy-tailed
A Taxonomy

- Goal: regret minimization, pure exploration
- Arm Set: discrete, continuous, structured, unstructured
- Feedback: stochastic, adversarial
- Agent: single, multiple
- Bandits: full information, semi-bandit, bandit
- Unimodal, linear, sub-Gaussian, heavy-tailed
Goal and Metric

Regret minimization

\[
\min R(A, T) \text{ (equivalent to rewards maximization)}
\]

\[
R(A, T) \triangleq \max_{i=1,\ldots,K} \mathbb{E} \left[ \sum_{t=1}^{T} y_t(i) - \sum_{t=1}^{T} y_t(I_t) \right] = T u_* - \sum_{t=1}^{T} u_{I_t} \tag{1}
\]
Goal and Metric

Pure exploration

Probability of error: \( P[x_T \neq \text{Opt}] \leq \delta \)

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

<table>
<thead>
<tr>
<th>round</th>
<th>arm 1</th>
<th>arm 2</th>
<th>arm 3</th>
<th>arm 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 4</td>
<td>$\frac{1}{1} = 1$</td>
<td>$\frac{0}{1} = 0$</td>
<td>$\frac{1}{1} = 1$</td>
<td>$\frac{1}{1} = 1$</td>
</tr>
<tr>
<td>5</td>
<td>$\frac{1+0}{2} = 0.5$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.5</td>
<td>0</td>
<td>$\frac{1+0}{2} = 0.5$</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>$\frac{1+0}{2} = 0.5$</td>
</tr>
<tr>
<td>8</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>$\frac{1+0}{3} = 0.3$</td>
</tr>
</tbody>
</table>
Heuristic Methods for Regret Minimization

Selecting the arm with largest empirical average + standard deviation

A four-armed case with Bernoulli distributions

True means: \( \{0.7, 0.8, 0.6, 0.5\} \)

<table>
<thead>
<tr>
<th>round</th>
<th>arm 1</th>
<th>arm 2</th>
<th>arm 3</th>
<th>arm 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – 4</td>
<td>( \frac{1}{1} + 1 = 2 )</td>
<td>( \frac{0}{1} + 1 = 1 )</td>
<td>( \frac{1}{1} + 1 = 2 )</td>
<td>( \frac{1}{1} + 1 = 2 )</td>
</tr>
<tr>
<td>5</td>
<td>( \frac{1+0}{2} + 0.7 = 1.2 )</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>1.2</td>
<td>1</td>
<td>( \frac{1+0}{2} + 0.7 = 1.2 )</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1.2</td>
<td>1</td>
<td>1.2</td>
<td>( \frac{1+0}{2} + 0.7 = 1.2 )</td>
</tr>
<tr>
<td>8</td>
<td>1.2</td>
<td>1</td>
<td>1.2</td>
<td>( \frac{1+0}{3} + 0.6 = 0.9 )</td>
</tr>
</tbody>
</table>
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 suprema 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

<table>
<thead>
<tr>
<th>work</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Thompson, 1933)</td>
<td>original formalization</td>
</tr>
</tbody>
</table>
| (Lai & Robbins, 1985) | the first theoretical analysis  
\[
\lim_{T \to \infty} \frac{R(A, T)}{\log(T)} \geq \sum_{\Delta_i > 0} \frac{\Delta_i}{\Delta_i} \\
\lim_{T \to \infty} \frac{R(UCB, T)}{\log(T)} \leq \sum_{\Delta_i > 0} \frac{\Delta_i}{\Delta_i} \\
\] |
| (Agrawal, 1995) | a simpler algorithm  
\[
\lim_{T \to \infty} \frac{R(SM, T)}{\log(T)} \leq \sum_{\Delta_i > 0} \frac{\Delta_i}{\Delta_i} \\
\] |
| (Auer et al., 2002) | finite-time analysis  
\[
R(UCB1, T) = O \left( \sum_{\Delta_i > 0} \frac{\log(T)}{\Delta_i} \right) \\
R(UCB1, T) = O \left( \sqrt{T} \right) \\
\] |
| (Agrawal et al., 2012) | Bernoulli payoffs  
\[
R(TS, T) = O \left( \left( \sum_{\Delta_i > 0} \frac{1}{\Delta_i} \right)^2 \log(T) \right) \\
\] |
| (Kaufmann et al., 2012) | Bernoulli payoffs  
\[
\lim_{T \to \infty} \frac{R(TS, T)}{\log(T)} \leq \sum_{\Delta_i > 0} \frac{\Delta_i}{\Delta_i} \\
\] |
| (Garivier et al., 2018) | finite-time lower bound  
small $T$: lower bound  
\[
R(A, T) \geq \sum_{\Delta_i > 0} \frac{\Delta_i T}{2K} \\
\]  
large $T$: lower bound  
\[
R(A, T) = \Omega \left( \sum_{\Delta_i > 0} \frac{\Delta_i \log(T)}{\Delta_i} \right) \\
\] |
## Theoretical Developments of Pure Exploration in MAB

<table>
<thead>
<tr>
<th>work</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Even-Dar et al., 2002)</td>
<td>bounded payoffs ( \mathbb{P} \left[ T \geq \sum_{k=1}^{K} \Delta_k^{-2} \log \left( \frac{K}{\delta \Delta_k} \right) \right] \leq \delta )</td>
</tr>
<tr>
<td>(Audibert &amp; Bubeck, 2010)</td>
<td>bounded payoffs ( \mathbb{P}[\text{Out} \neq \text{Opt}] \leq TK \exp \left( -\frac{T-K}{H_1} \right) )</td>
</tr>
<tr>
<td>(Karnin et al., 2013)</td>
<td>bounded payoffs ( \mathbb{P} \left[ T \geq \sum_{k=1}^{K} \Delta_k^{-2} \log \left( \frac{1}{\delta} \log \left( \frac{1}{\Delta_k} \right) \right) \right] \leq \delta ) ( \mathbb{P}[\text{Out} \neq \text{Opt}] \leq \log(K) \exp \left( -\frac{T}{\log(K)H_2} \right) )</td>
</tr>
<tr>
<td>(Jamieson et al., 2014)</td>
<td>sub-Gaussian noises ( \mathbb{P} \left[ T \geq H_1 \log \left( \frac{1}{\delta} \right) + H_3 \right] \leq 4\sqrt{c\delta} + 4c\delta )</td>
</tr>
<tr>
<td>(Kaufmann et al., 2016)</td>
<td>two-armed Gaussian bandits ( \lim_{\delta \to 0} \frac{\mathbb{E}[T]}{\log(\frac{1}{\delta})} \geq \frac{2(\sigma_1+\sigma_2)^2}{(u_1-u_2)^2} ) ( \lim_{\delta \to 0} \frac{\mathbb{E}[T]}{\log(\frac{1}{\delta})} \leq \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} \leq \frac{(u_1-u_2)^2}{2(\sigma_1+\sigma_2)^2} )</td>
</tr>
</tbody>
</table>
A Taxonomy

- **Goal**
  - Regret minimization
  - Pure exploration

- **Agent**
  - Single
  - Multiple

- **Arm Set**
  - Discrete
  - Continuous
    - Structured
    - Unstructured

- **Feedback**
  - Full information
  - Semi-bandit
  - Bandit

- **Bandits**
  - Stochastic
  - Adversarial
  - Linear
  - Sub-Gaussian
  - Heavy-tailed
## Theoretical Developments of LSB

<table>
<thead>
<tr>
<th>work</th>
<th>results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Abe &amp; Long, 1999; Auer, 2000)</td>
<td>original formalization</td>
</tr>
<tr>
<td>(Auer, 2002)</td>
<td>first theoretical analysis; $K$ arms</td>
</tr>
<tr>
<td></td>
<td>$\mathbf{R}(\text{LinRel}, T) = O\left(\sqrt{T}d\log^\frac{3}{2}(KT\log(T))\right)$</td>
</tr>
<tr>
<td>(Dani et al., 2008)</td>
<td>compact arm set; bounded payoffs</td>
</tr>
<tr>
<td></td>
<td>$\mathbf{R}(\mathcal{A}, T) = \Omega\left(d\sqrt{T}\right)$</td>
</tr>
<tr>
<td></td>
<td>$\mathbf{R}(\text{CB}_2, T) = O\left(d\sqrt{T}\log^\frac{3}{2}(T)\right)$</td>
</tr>
<tr>
<td>(Abbasi-Yadkori et al., 2011)</td>
<td>compact arm set; sub-Gaussian noises</td>
</tr>
<tr>
<td></td>
<td>$\mathbf{R}(\text{OFUL}, T) = O\left(d\sqrt{T}\log(T)\right)$</td>
</tr>
<tr>
<td>(Agrawal &amp; Goyal, 2013)</td>
<td>$K$ arms; sub-Gaussian noises</td>
</tr>
<tr>
<td></td>
<td>$\mathbf{R}(\text{TS}, T) = O\left(d^2\sqrt{T}\log(dT)\right)$</td>
</tr>
<tr>
<td>(Lattimore &amp; Szepesvari, 2017)</td>
<td>$K$ arms; Gaussian payoffs</td>
</tr>
<tr>
<td></td>
<td>$\lim_{T \to \infty} \frac{\mathbf{R}(\mathcal{A}, T)}{\log(T)} \geq c(\mathcal{A}, \theta)$</td>
</tr>
<tr>
<td></td>
<td>$\lim_{T \to \infty} \frac{\mathbf{R}(\text{OA}, T)}{\log(T)} \leq c(\mathcal{A}, \theta)$</td>
</tr>
</tbody>
</table>
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
- single
- multiple
- discrete
- continuous
- structured
- unstructured
- unimodal
- linear
- sub-Gaussian
- heavy-tailed
- stochastic
- 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

Mathematically, a random variable \( X \) is said to be heavy-tailed if

\[
\lim_{x \to \infty} e^{\phi x} \mathbb{P}[|X| > x] = \infty \quad \text{for all } \phi > 0
\]

Heavy-Tailed Distributions in Bandits

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

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

(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 $\alpha$ 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 \( t = 1, \ldots, T \),

given the arm set \( D_t \subseteq \mathbb{R}^d \), an agent selects an arm \( x_t \in D_t \)

observes a stochastic reward \( y_t = x_t^\top \theta + \eta_t \), where \( \eta_t \) is a stochastic noise

- Previous assumption (Abbasi-Yadkori et al., 2011): \( \eta_t \) is sub-Gaussian conditional on \( F_{t-1} \)
- Our assumption: \( y_t \) or \( \eta_t \) is heavy-tailed conditional on \( F_{t-1} \)
  - Bounded raw moments
  - Bounded central moments

- A connection in regret:

\[
\tilde{O}\left(\sqrt{T}\right) \text{ (sub-Gaussian)} \rightarrow \tilde{O}\left(\sqrt{T}\right) \text{ (2-nd moment bounded)}
\]
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., $\mathbb{E} [ |y_t|^p |\mathcal{F}_{t-1}] \leq b$ or $\mathbb{E} [ |y_t - \langle x_t, \theta_* \rangle|^p |\mathcal{F}_{t-1}] \leq c$, 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
Assume $d \geq 2$ is even. For $D_t \in \mathbb{R}^d$, we fix the arm set as $D_t = D(d)$, where $D(d) \triangleq \{ (x_1, \ldots, x_d) \in \mathbb{R}^d_+ : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1 \}$. Let $S_d \triangleq \{ (\theta_1, \ldots, \theta_d) : \forall i \in [d/2], (\theta_{2i-1}, \theta_{2i}) \in \{ (2\Delta, \Delta), (\Delta, 2\Delta) \} \}$ 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 $\theta_*$ 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 $T \geq (d/12)^{\frac{p-1}{p}}$, we have

$$\mathbb{E}[R(\mathcal{A}, T)] \geq \frac{d}{192} T^{\frac{1}{p}}.$$
Lower Bound of LinBET

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

- Arm set: \( D_2(2) \triangleq \{(x_1, x_2) \in \mathbb{R}^2_+: x_1 + x_2 = 1\} \)
- \( \theta^* \) is chosen uniformly at random from \( \{\mu_1, \mu_2\} \), where \( \mu_1 = (2\Delta, \Delta) \) and \( \mu_2 = (\Delta, 2\Delta) \)
- \( \Delta \) 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 \( \mathbb{E} [ |y_t|^p | \mathcal{F}_{t-1} ] \leq d \) case

- Payoff distribution of \( x \):

\[
y(x) = \begin{cases} 
(\frac{1}{\Delta})^{\frac{1}{p-1}} & \text{with a probability of } \Delta^{\frac{1}{p-1}} \theta^\top \star x, \\
0 & \text{with a probability of } 1 - \Delta^{\frac{1}{p-1}} \theta^\top \star x
\end{cases}
\]

- \( \mathbb{E} [y(x)^p] \leq 2 \)
- \( \mathbb{E} [y(x)^q] \geq (\frac{1}{\Delta})^{\frac{p-q}{p-1}}, q < p \)
An Algorithm for LSB

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

\[ \begin{align*}
\theta_1 & : \text{optimal for arm } (0,1) \\
\theta_2 & : \text{optimal for arm } (0.5,0.5) \\
\theta_3 & : \text{optimal for arm } (1,0)
\end{align*} \]

At time \( t \), select arm \( x_t \) by

\[ \begin{align*}
(x_t, \tilde{\theta}_t) &= \arg \max_{(x, \theta) \in D_t \times C_{t-1}} \langle x, \theta \rangle \\
C_t &= \{ \theta : \| \theta - \hat{\theta}_{t,k*} \| V_t \leq \beta_t \}, \ V_t = \lambda I + \sum_{\tau=1}^{t} x_{\tau} x_{\tau}^T
\end{align*} \]

The regret is bounded by

\[ \tilde{O}\left( \max_{t \in [T]} \beta_{t-1} \sqrt{T} \right) \]

For sub-Gaussian case, LSE → \( \beta_t = \Theta \left( \sqrt{\log t} \right) \)

For heavy-tailed case, LSE → \( \beta_t \) is polynomial of \( t \)
Techniques for Designing Algorithms

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

▶ Median of means

▶ Truncation

- 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 $\tilde{O}(T^{3/4})$ when $p = 2$
- Is it possible to design algorithms to achieve the regret of $\tilde{O}(\sqrt{T})$ when $p = 2$?
Algorithms: Median of means under OFU (MENU)

Algorithm 1 MENU

1: **input** $d$, $c$, $p$, $\delta$, $\lambda$, $S$, $T$, $\{D_n\}_{n=1}^N$
2: **initialization:** $k = \lceil 24 \log \left( \frac{eT}{\delta} \right) \rceil$, $N = \lceil \frac{T}{k} \rceil$, $V_0 = \lambda I_d$, $C_0 = B(0, S)$
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: $V_n = V_{n-1} + x_n x_n^T$
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 $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_V : s \in [k] \backslash j\}$
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: $C_n = \{\theta : \|\theta - \hat{\theta}_{n,k^*}\|_V \leq \beta_n\}$
12: **end for**
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)

MENU

\[ \hat{\theta}_{n,k} \]

- Take median of means of \( \{ \hat{\theta}_{n,j} \}_{j=1}^{k} \)
- Calculate \( k \) LSEs with payoffs on \( \{ x_i \}_{i=1}^{n} \)
- \( k = \lceil 24 \log \left( \frac{\varepsilon}{\delta} \right) \rceil \)
- \( N = \lfloor T^{2p-2} \rfloor \)

MoM

\[ \hat{\theta}_{n} \]

- Take median of means of payoffs on \( \{ x_i \}_{i=1}^{k}, i \in [n] \)
- Calculate LSE with \( \{ \tilde{l}_i \}_{i=1}^{n} \)

\[ \tilde{l}_1 \quad \cdots \quad \tilde{l}_n \]

- \( k = T^{2p-2} \)
- \( N = T^{2p-2} \)
Understanding of MENU
Result comparison with MoM by Medina & Yang (2016)

For MoM by Medina & Yang (2016)

- The regret is bounded by \( \tilde{O}\left(\max_{n=1,\ldots,N} \beta_{n-1} k \sqrt{N}\right) \), where
  \[ \beta_n = \Theta\left(k^{-\frac{p-1}{p}} \sqrt{n}\right) \]
- The value of \( k \) and \( N \) is constrained by \( \max_{n=1,\ldots,N} \beta_n = \Omega(1) \)
- The regret of the MoM algorithm is \( \tilde{O}(c^{\frac{1}{p}} dT^{\frac{2p-1}{3p-2}}) \)

For our MENU

- Make each group contain the same playing history to compute regret easily
  \[ k = \Theta(\log(T)) \]
  \[ \beta_n = \Theta(n^{\frac{2-p}{2p}}) \]
Upper Bound Analysis: MENU

Results

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

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

- The regret is $\tilde{O}\left(\sqrt{T}\right)$ when $p = 2$
Algorithms: _Truncation under OFU (TOFU)_(1)

**Algorithm 2 TOFU**

1: **input** $d, b, p, \delta, \lambda, T, \{D_t\}_{t=1}^T$
2: **initialization:** $V_0 = \lambda I_d$, $C_0 = \mathbb{B}(0, S)$
3: for $t = 1, 2, \ldots, T$ do
4:   $b_t = \left( \frac{b}{\log \left( \frac{2T}{\delta} \right)} \right)^{\frac{1}{p-1}} t^{\frac{2-p}{2p}}$
5:   $(x_t, \tilde{\theta}_t) = \text{arg max}_{(x, \theta) \in D_t \times C_{t-1}} \langle x, \theta \rangle$
6:   Play $x_t$ and observe a payoff $y_t$
7:   $V_t = V_{t-1} + x_t x_t^\top$ and $X_t^\top = [x_1, \ldots, x_t]$
8:   $[u_1, \ldots, u_d]^\top = V_t^{-1/2} X_t^\top$
9:   for $i = 1, \ldots, d$ do
10:      $Y_i^\dagger = (y_1 \mathbf{1}_{u_i,1 y_1 \leq b_i}, \ldots, y_t \mathbf{1}_{u_i,t y_t \leq b_t})$
11:   end for
12:   $\theta_t^\dagger = V_t^{-1/2} (u_1^\top Y_1^\dagger, \ldots, u_d^\top Y_d^\dagger)$
13:   $\beta_t = 4\sqrt{d b^{\frac{1}{p}} \left( \log \left( \frac{2dT}{\delta} \right) \right)^{\frac{p-1}{p}}} t^{\frac{2-p}{2p}} + \lambda \frac{1}{2} S$
14:   Update $C_t = \{ \theta : \| \theta - \theta_t^\dagger \| V_t \leq \beta_t \}$
15: end for
Understanding of TOFU
Comparison with CRT by Medina & Yang (2016)

- For CRT, the payoff at time $t$ is truncated by $\alpha_t$
  - $y_t = y_t 1_{y_t \leq \alpha_t}$
  - The regret of the CRT algorithm is $\widetilde{O}(bdT^{1/2} + t)$

- 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 $V_t^{1/2} X_t^\top$
  - $Y_i^\dagger = (y_1 1_{u_i,1 y_1 \leq b_t}, \ldots, y_t 1_{u_i,t y_t \leq b_t})$
  - $\theta_t^\dagger = V_t^{1/2} (u_1^\top Y_1^\dagger, \ldots, u_d^\top Y_d^\dagger)$

- A 2-d example

<table>
<thead>
<tr>
<th>arms</th>
<th>(0, 1)</th>
<th>(1, 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#pulls</td>
<td>50</td>
<td>1</td>
</tr>
</tbody>
</table>
Upper Bound Analysis: TOFU

Results

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

$$R(\text{TOFU}, T) \leq \tilde{O}(b^{\frac{1}{p}} dT^{\frac{1}{p}}).$$

- The regret is $\tilde{O}\left(\sqrt{T}\right)$ 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

<table>
<thead>
<tr>
<th>dataset</th>
<th>{#arms,#dims}</th>
<th>distribution {parameters}</th>
<th>{p, b, c}</th>
<th>optimal arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>{20,10}</td>
<td>Student’s ( t )-distribution ( \nu = 3, l_p = 0, s_p = 1 )</td>
<td>{2.00, NA, 3.00}</td>
<td>4.00</td>
</tr>
<tr>
<td>S2</td>
<td>{100,20}</td>
<td>Student’s ( t )-distribution ( \nu = 3, l_p = 0, s_p = 1 )</td>
<td>{2.00, NA, 3.00}</td>
<td>7.40</td>
</tr>
<tr>
<td>S3</td>
<td>{20,10}</td>
<td>Pareto distribution ( \alpha = 2, s_m = \frac{x_t^\top \theta_*}{2} )</td>
<td>{1.50, 7.72, NA}</td>
<td>3.10</td>
</tr>
<tr>
<td>S4</td>
<td>{100,20}</td>
<td>Pareto distribution ( \alpha = 2, s_m = \frac{x_t^\top \theta_*}{2} )</td>
<td>{1.50, 54.37, NA}</td>
<td>11.39</td>
</tr>
</tbody>
</table>
Experimental Results

![Comparison of cumulative payoffs for synthetic datasets S1-S2 with four algorithms.](image)

**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

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


Discussions

▶ Efficiency of TOFU
▶ Problem-dependent bounds
▶ The impact of $d$
Outline

▶ Introduction

▶ A Survey of Bandits

▶ Linear Stochastic Bandits with Heavy-Tailed Payoffs

▶ Conclusions and Future Directions
Conclusions

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

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

<table>
<thead>
<tr>
<th>algorithm</th>
<th>MoM</th>
<th>MENU</th>
<th>CRT</th>
<th>TOFU</th>
</tr>
</thead>
<tbody>
<tr>
<td>regret</td>
<td>$\tilde{O}(T^{3p-2})$</td>
<td>$\tilde{O}(T^p)$</td>
<td>$\tilde{O}(T^{2+\frac{1}{2p}})$</td>
<td>$\tilde{O}(T^p)$</td>
</tr>
<tr>
<td>complexity</td>
<td>$O(T)$</td>
<td>$O(T\log T)$</td>
<td>$O(T)$</td>
<td>$O(T^2)$</td>
</tr>
<tr>
<td>storage</td>
<td>$O(1)$</td>
<td>$O(\log T)$</td>
<td>$O(1)$</td>
<td>$O(T)$</td>
</tr>
</tbody>
</table>
Upper Bound Analysis: MENU

Proof sketch

Lemma 1. [Confidence Ellipsoid of LSE] Let $\hat{\theta}_n$ denote the LSE of $\theta_*$ 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 $x_\tau \in D_\tau \subseteq \mathbb{R}^d$, $\mathbb{E}[|\eta_\tau|^p|\mathcal{F}_{\tau-1}] \leq c$ and $\|\theta_*\|_2 \leq S$. Then $\hat{\theta}_n$ satisfies

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

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

Proof sketch of Lemma 1

- Let $u_i$ denote the $i$-th row of $V_t^{-1/2}X_t^\top$

\[ \|\hat{\theta}_n - \theta_*\|_V \leq \sqrt{\sum_{i=1}^d (u_i^\top (Y_n - X_n\theta_*))^2} + \lambda \|\theta_*\|_V^{-1} \]

- Union bound

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

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

where $1_{\{|\cdot|\}}$ is the indicator function

- Both terms could be bounded by Markov’s inequality

- Set $\gamma = (9dc)^{\frac{1}{p}}n^{\frac{2-p}{2p}}$
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 $\mathbb{B}_V(\theta_*, \gamma) \triangleq \{\theta : \|\theta - \theta_*\| V_n \leq \gamma\}$.

$r_j$ be the median of $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\| V_n : s \in [k]\setminus j\}$.

Select arm $\text{arg min}_{j\in[k]} r_j$

- If $\hat{\theta}_{n,j} \in \mathbb{B}_V(\theta_*, \gamma)$, $\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\| V_n \leq 2\gamma$ for all $\hat{\theta}_{n,s} \in \mathbb{B}_V(\theta_*, \gamma)$ by triangle inequality. Therefore, $r_j \leq 2\gamma$.

- If $\hat{\theta}_{n,j} \notin \mathbb{B}_V(\theta_*, 3\gamma)$, $\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\| V_n > 2\gamma$ for all $\hat{\theta}_{n,s} \in \mathbb{B}_V(\theta_*, \gamma)$ by triangle inequality. Therefore, $r_j > 2\gamma$. 
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 \( x_\tau \in D_\tau \subseteq \mathbb{R}^d \), \( \mathbb{E}[|y_\tau|^p | \mathcal{F}_{\tau-1}] \leq b \) and \( \|\theta_*\|_2 \leq S \). With probability at least \( 1 - \delta \), we have

\[
\|\theta_t^\dagger - \theta_*\|_{V_t} \leq 4\sqrt{db} \frac{1}{p} \left( \log \left( \frac{2d}{\delta} \right) \right)^{\frac{p-1}{p}} t^{\frac{2-p}{2p}} + \lambda \frac{1}{2} S,
\]

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,

\[ \|\theta^*_t - \theta\|_{V_t} \leq \sqrt{\sum_{i=1}^{d} \left( u_i^T (Y_i^\dagger - X_t \theta) \right)^2} + \lambda \|\theta\|_{V_n^{-1}} \]

- For each \( i \)

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

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

- The first term is bounded by Bernstein’s inequality

- Set \( b_t = \left( b / \log(2d/\delta) \right)^{\frac{1}{p}} t^{\frac{2-p}{2p}} \)