Almost Optimal Algorithms for Linear Stochastic Bandits with Heavy-Tailed Payoffs

Han Shao* Xiaotian Yu* Irwin King Michael R. Lyu

Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{hshao, xtyu, king, lyu}@cse.cuhk.edu.hk

Introduction

Background: Why heavy tails in bandits

- Generally, payoffs in Multi-Armed Bandits (MAB) are assumed under sub-Gaussian noises.
- In practice, many scenarios contain heavy-tailed noises, e.g., extreme payoffs in financial markets.
- Existing work has not achieved the optimal regret of linear stochastic bandits with heavy tails.

Practical motivation: An illustration

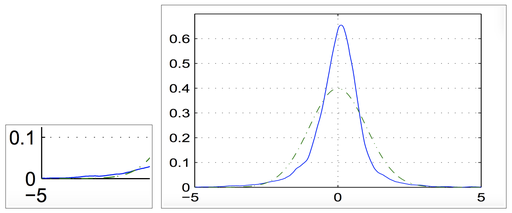


Figure 1: High-probability extreme returns in finance.

In the figure, the blue-solid line denotes the NASDAQ returns for the last 20 years, and the green-dashed line denotes a fitted Gaussian distribution.

The problem of LinBET

Definition 1. *Given a decision 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., $\mathbb{E}\left[|y_t|^{1+\epsilon}|\mathcal{F}_{t-1}\right] \leq b$ or $\mathbb{E}\left[|y_t - \langle x_t, \theta_*\rangle|^{1+\epsilon}|\mathcal{F}_{t-1}\right] \leq c$, where $\epsilon \in (0, 1]$, and $b, c \in (0, +\infty)$.*

Our contributions

- We provide the lower bound for the problem of LinBET as $\Omega(T^{\frac{1}{1+\epsilon}})$, where $\epsilon \in (0, 1]$.
- We develop two novel bandit algorithms, which are named as MENU and TOFU. Both algorithms achieve the regret $\widetilde{O}(T^{\frac{1}{1+\epsilon}})$ with high probability.
- We conduct experiments to demonstrate the effectiveness of our proposed algorithms.

Lower Bound of LinBET

Theorem for Lower Bound. *We define a set $S_d \triangleq \{(\theta_1, \cdots, \theta_d) : \forall i \in [d/2], (\theta_{2i-1}, \theta_{2i}) \in \{(2\Delta, \Delta), (\Delta, 2\Delta)\}\}$ with $\Delta \in (0, 1/d]$ and $D_{(d)} \triangleq \{(x_1, \cdots, x_d) \in \mathbb{R}_+^d : x_1 + x_2 = \cdots = x_{d-1} + x_d = 1\}$. If θ_* is chosen uniformly at random from S_d, and the payoff for each $x \in D_{(d)}$ is in $\{0, (1/\Delta)^{\frac{1}{\epsilon}}\}$ with mean $\theta_*^\top x$, we have $\forall \mathcal{A}$ and $\forall T \geq (d/12)^{\frac{\epsilon}{1+\epsilon}}$,*

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

TOFU and Upper Bound

Comparison between our TOFU and CRT

- TOFU truncates all historical payoffs in terms of $\{\mu_i\}_{i=1}^d$ for each round.
- CRT only truncates the payoff at time t and does not revisit historical information.

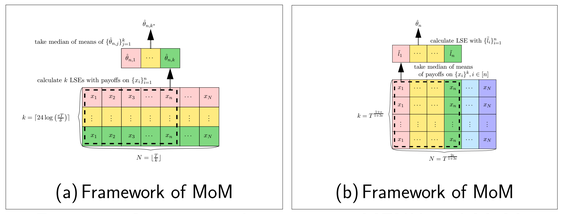
Theorem for TOFU. *Assume that for all t and $x_t \in D_t$ with $\|x_t\|_2 \leq D$, $\|\theta_*\|_2 \leq S$, $|x_t^\top \theta_*| \leq L$ and $\mathbb{E}[|y_t|^{1+\epsilon}|\mathcal{F}_{t-1}] \leq b$. Then, with probability at least $1-\delta$, for every $T \geq 1$, the regret of the TOFU algorithm satisfies*

$$R(\mathit{TOFU}, T) \leq \widetilde{O}\left(d b^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}\right).$$

Algorithm 1 TOFU

input d, b, ϵ, δ, λ, T, $\{D_t\}_{t=1}^T$
initialization: $V_0 = \lambda I_d$, $C_0 = \mathbf{B}(\mathbf{0}, S)$
for $t = 1, 2, \cdots, T$ **do**
  $b_t = \left(\frac{b}{\log\left(\frac{2T}{\delta}\right)}\right)^{\frac{1}{\epsilon}} t^{\frac{1-\epsilon}{2(1+\epsilon)}}$
  $(x_t, \tilde{\theta}_t) = \arg\max_{(x,\theta) \in D_t \times C_{t-1}} \langle x, \theta \rangle$ ▷ *to select an arm*
  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**
    $Y_i^\dagger = (y_1 \mathbb{1}_{u_{i,1} y_1 \leq b_t}, \cdots, y_t \mathbb{1}_{u_{i,t} y_t \leq b_t})$ ▷ *to truncate the payoffs*
  end for
  $\theta_t^\dagger = V_t^{-1/2}(u_1^\top Y_1^\dagger, \cdots, u_d^\top Y_d^\dagger)$
  $\beta_t = 4\sqrt{d} b^{\frac{1}{1+\epsilon}} \left(\log\left(\frac{2dT}{\delta}\right)\right)^{\frac{\epsilon}{1+\epsilon}} t^{\frac{1-\epsilon}{2(1+\epsilon)}} + \lambda^{\frac{1}{2}} S$
  Update $C_t = \{\theta : \|\theta - \theta_t^\dagger\|_{V_t} \leq \beta_t\}$
end for

MENU and Upper Bound



(a) Framework of MoM (b) Framework of MoM

Figure 2: Comparison between our MENU and MoM.

Theorem for MENU. *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 $\mathbb{E}[|\eta_t|^{1+\epsilon}|\mathcal{F}_{t-1}] \leq c$. Then, with probability at least $1-\delta$, for every $T \geq 256 + 24\log(e/\delta)$, the regret of the MENU algorithm satisfies*

$$R(\mathit{MENU}, T) \leq \widetilde{O}\left(d^{\frac{1}{1+\epsilon}+\frac{1}{2}} c^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}\right).$$

Algorithm 2 MENU

input d, c, ϵ, δ, λ, S, T, $\{D_n\}_{n=1}^N$
initialization: $k = \lceil 24 \log\left(\frac{eT}{\delta}\right) \rceil$, $N = \lfloor \frac{T}{k} \rfloor$, $V_0 = \lambda I_d$, $C_0 = \mathbf{B}(\mathbf{0}, S)$
for $n = 1, 2, \cdots, N$ **do**
  $(x_n, \tilde{\theta}_n) = \arg\max_{(x,\theta) \in D_n \times C_{n-1}} \langle x, \theta \rangle$ ▷ *to select an arm*
  Play x_n with k times and observe payoffs $y_{n,1}, y_{n,2}, \cdots, y_{n,k}$
  $V_n = V_{n-1} + x_n x_n^\top$
  For $j \in [k]$, $\hat{\theta}_{n,j} = V_n^{-1} \sum_{i=1}^n y_{i,j} x_i$ ▷ *to calculate LSE for the j-th group*
  For $j \in [k]$, let r_j be the median of $\{\|\hat{\theta}_{n,j} - \hat{\theta}_{n,s}\|_{V_n} : s \in [k] \backslash j\}$
  $k^* = \arg\min_{j \in [k]} r_j$ ▷ *to take median of means of estimates*
  $\beta_n = 3\left((9dc)^{\frac{1}{1+\epsilon}} n^{\frac{1-\epsilon}{2(1+\epsilon)}} + \lambda^{\frac{1}{2}} S\right)$
  $C_n = \{\theta : \|\theta - \hat{\theta}_{n,k^*}\|_{V_n} \leq \beta_n\}$ ▷ *to update the confidence region*
end for

Discussions on Theoretical Results

Future directions

- The lower bound can be further improved by the dimension d or by the moment parameter b or c.
- The regret of MENU on d can be improved.
- New algorithms achieve problem-dependent regret upper bounds with polylog(T).

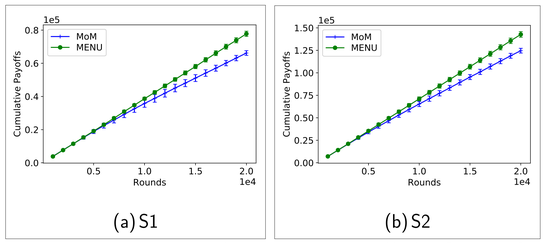
Experiments and Conclusion

Datasets

Table 1: Statistics of synthetic datasets in experiments. For Student's t-distribution, ν denotes the degree of freedom, l_p denotes the location, s_p denotes the scale. For Pareto distribution, α denotes the shape and s_m denotes the scale. NA denotes not available.

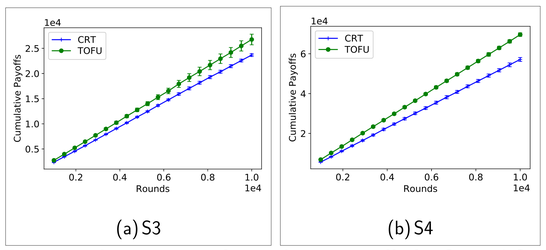
dataset	distribution {parameters}	$\{\epsilon, b, c\}$
S1	Student's t-distribution $\{\nu = 3, l_p = 0, s_p = 1\}$	{1.00, NA, 3.00}
S2	Student's t-distribution $\{\nu = 3, l_p = 0, s_p = 1\}$	{1.00, NA, 3.00}
S3	Pareto distribution $\{\alpha = 2, s_m = \frac{x_t^\top \theta_*}{2}\}$	{0.50, 7.72, NA}
S4	Pareto distribution $\{\alpha = 2, s_m = \frac{x_t^\top \theta_*}{2}\}$	{0.50, 54.37, NA}

Results (`https://github.com/Aaronyxt/LinBET_nips2018`)



(a) S1 (b) S2

Figure 3: Comparison of cumulative payoffs for S1 and S2.



(a) S3 (b) S4

Figure 4: Comparison of cumulative payoffs for S3 and S4.

Conclusion

We broke the assumption of sub-Gaussian noises in payoffs of bandits. We rigorously analyzed the lower bound of LinBET, and developed two novel bandit algorithms with regret upper bounds matching the lower bound up to polylogarithmic factors.