Heavy-tailed Linear Bandit with Huber Regression

Minhyun Kang\(^1\)  

Gi-Soo Kim\(^{1,2}\)  

\(^1\) Artificial Intelligence Graduate School, UNIST, Ulsan, Republic of Korea  
\(^2\) Department of Industrial Engineering, UNIST, Ulsan, Republic of Korea

Abstract

Linear bandit algorithms have been extensively studied and have shown successful in sequential decision tasks despite their simplicity. Many algorithms however work under the assumption that the reward is the sum of linear function of observed contexts and a sub-Gaussian error. In practical applications, errors can be heavy-tailed, especially in financial data. In such reward environments, algorithms designed for sub-Gaussian error may underexplore, resulting in suboptimal regret. In this paper, we relax the reward assumption and propose a novel linear bandit algorithm which works well under heavy-tailed errors as well. The proposed algorithm utilizes Huber regression. When contexts are stochastic with positive definite covariance matrix and the \((1 + \delta)\)-th moment of the error is bounded by a constant, we show that the high-probability upper bound of the regret is \(O(\sqrt{dT^{1+\delta}})\), where \(d\) is the dimension of context variables, \(T\) is the time horizon, and \(\delta \in (0, 1]\). This bound improves on the state-of-the-art regret bound of the Median of Means and Truncation algorithm by a factor of \(\sqrt{\log T}\) and \(\sqrt{d}\) for the case where the time horizon \(T\) is unknown. We also remark that when \(\delta = 1\), the order is the same as the regret bound of linear bandit algorithms designed for sub-Gaussian errors. We support our theoretical findings with synthetic experiments.

1 INTRODUCTION

Bandit algorithms are widely used in sequential decision-making problems such as mobile health \cite{lei2017bandit}, clinical trial \cite{villar2015design} where the goal of the learning agent is to select good actions successively out of many available actions at each time point. Linear bandits make use of contextual information when choosing the actions, or arms. Upon choosing an arm, a random reward is revealed to the agent. The agent then learns the reward model using rewards observed so far under assumption that the expected value of the reward is a linear function of context variables. Using the updated model and various exploration strategies, the agent chooses the next arm.

Most bandit research is studied under the assumption that error distribution is sub-Gaussian. However, the tails of data distribution might not decay as fast as that of sub-Gaussian in practical applications including financial markets and insurance \cite{rachev2003financial, stehlík2010extreme, ibragimov2015heavy}. Our work relaxes the sub-Gaussian assumption and proposes a new algorithm that is robust to heavy-tailed errors. We assume the noise of the reward, \(\epsilon\), has finite \((1 + \delta)\)-th moment, i.e., \(E[|\epsilon|^{1+\delta}] \leq \nu_\delta < \infty\) for some \(\delta \in (0, 1]\). This assumption is common in the bandit literature which consider heavy-tailed errors \cite{bubeck2013bandits, medina2016median, xue2020linear}. We propose to use the Huber estimator \cite{huber1964robust} to estimate the reward model parameters. The Huber loss works as a square loss when the input variable is small and works as an absolute loss otherwise. It suppresses the loss value when an observation deviated significantly from the mean, so that it does not dominates on other observations. \cite{sun2020adaptive} proposed an adaptive Huber estimator with robustification parameter adapted to dimension of covariates, size of sample and moment bound on the error. The upper bound of \(L_2\)-norm of the estimation error is tight but is inversely proportional to the minimum eigenvalue of the Gram matrix of the covariates. When the covariates are independently and identically distributed (i.i.d.) and are sampled from a distribution with positive definite covariance matrix, it can be shown that the minimum eigenvalue of their Gram matrix is strictly bigger than a positive constant with high probability for sufficiently large samples. However, data accumulated by bandit agents are not i.i.d. since the arm selection depends on the arms chosen before. Hence, the Gram matrix

\[\text{Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence} \ (\text{UAI} \ 2023), \ PMLR 216:1027–1036.\]
formed by contexts has no guarantee to have strictly positive minimum eigenvalue, and the estimation error can be higher. In this work, we propose to use the forced-sampling method of [Goldenshluger and Zeevi 2013] and [Bastani and Bayati 2020] to guarantee the minimum eigenvalue bound in the non-i.i.d. bandit setting.

In Section 2, we explain our problem settings. Then in Section 3, previous works on the linear bandit with heavy-tailed errors are reviewed. We introduce previous results needed for our theoretical analysis of estimator and the formal definition of the Huber regression in Section 4. In Section 5, a description of the proposed algorithm and its theoretical analysis are presented. Section 6 is devoted to the simulation study of the proposed algorithm compared with three existing algorithms.

2 PROBLEM FORMULATION

In linear bandits, we assume that the expected value of the reward is linear in the contexts. The bandit agent aims to learn the unknown linear parameter through consecutive arm pulls. We denote $d$-dimensional context vector at time $t$ by $X_t \in \mathbb{R}^d$. In each time step, $X_t$ is sampled from unknown distribution $P_X$ independently. We denote the number of arms as $K$. The agent observes the reward and arm-pulled at time $t$, the arm reveals a stochastic reward

$$y_{i,t} = X_t^T \beta_i + \epsilon_{i,t},$$

where $\beta_i \in \mathbb{R}^d$ is an arm-specific fixed parameter and each $\epsilon_{i,t}$ is a heavy-tailed noise. We denote the index of the arm pulled by the bandit agent at time $t$ as $a(t)$. Hence at each time $t$, the agent observes one reward, $y_t := y_{a(t),t}$. We also assume $E[\epsilon_{i,t} | F_{t-1}] = 0$ and $E[|\epsilon_{i,t}|^{1+\delta} | F_{t-1}] \leq \nu_\delta < \infty$, for $\nu_\delta > 0$ and for some $\delta \in (0,1)$, where $F_t = \sigma(X_1,...,X_{t+1},a(1),a(2),...,a(t),y_1,...,y_t)$ denotes the $\sigma$-algebra generated by observations $\{X_1,...,X_{t+1},a(1),a(2),...,a(t),y_1,...,y_t\}$. Note that sub-Gaussian errors satisfy this assumption with $\delta = 1$. The regret at time $t$ is defined by

$$r_t := X_t^T \beta_{a^*(t)} - X_t^T \beta_{a(t)}.$$

where $a^*(t) = \arg \max_j \{X_t^T \beta_j\}$ is the optimal arm.

In related literature on bandits for linear payoffs, the formulation is sometimes presented as follows, where instead of arm-specific parameters and single context, we have a single parameter $\beta \in \mathbb{R}^d$ and arm-specific contexts, $X_{i,t} \in \mathbb{R}^d$ for each arm $i$.

$$y_{i,t} = X_{i,t}^T \beta + \epsilon_{i,t}.$$  

However, we note that an algorithm designed for one setting can always be applied to the other setting. For example, when the algorithm designed for [1] needs to be applied on [2], we simply let $X_t = \{X_{i,t}^T \cdots X_{K,t}^T\}^T \in \mathbb{R}^{K\times d}$ and $X_t = \{0^T \cdots 0^T \beta^T \cdots 0^T\}^T \in \mathbb{R}^{K\times d}$, where $0$ is a $d$-dimensional zero vector and all the elements up to the $d\times(i-1)$-th element and after the $d\times i$-th element are all 0 in $\beta_i$. On the other hand, when we apply an algorithm designed for [2] on setting [1], we simply let $\beta = \{\beta_1^T \cdots \beta_K^T\}^T \in \mathbb{R}^{K\times d}$ and $X_{i,t} = \{0^T \cdots 0^T X_{i,t}^T 0^T \cdots 0^T\}^T \in \mathbb{R}^{K\times d}$.

Notations For any vector $v \in \mathbb{R}^d$ and positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, we let \(|v|_A := \sqrt{v^T A v}\). We let $[N] = \{1,2,\cdots ,N\}$ for any natural number $N$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimum and maximum eigenvalues of matrix $A$ respectively. For any vector $v \in \mathbb{R}^d$, $\|v\|_2 = (\sum_{i=1}^d |v_i|^2)^{\frac{1}{2}}$ and max norm is $\|v\|_{\infty} = \max_{1\leq i \leq d} |v_i|$. For any matrices $A,B \in \mathbb{R}^{d\times d}$, let $A \geq B$ if $A - B$ is positive semi definite. For any set $A$, let $1_A(\cdot)$ denote indicator function of a subset of some set $W$. For $a,b \in \mathbb{R}$, $a \lor b := \max\{a,b\}$. The set $\mathbb{N}$ denotes the set of natural numbers. When $Z \subset \mathbb{N}$, we denote the Gram matrix as $\Sigma(Z) := \sum_{t \in Z} X_t X_t^T$.

3 RELATED WORK

When errors are sub-Gaussian and the time horizon is unknown, LinOFUL [Abbasi-Yadkori et al. 2011] is the state-of-the-art algorithm, achieving a tight regret bound of order $O(d\sqrt{T})$ for setting [2], where $O(\cdot)$ ignores logarithmic terms. The strategy of LinOFUL is to construct a tight confidence region of the true parameter $\beta$ at each time step and pull the arm according to the Optimism in the Face of Uncertainty (OFU) principle. The confidence region at time $t$ is an ellipsoid of the form $C_t = \{\beta : \|\beta - \hat{\beta}\|_{V_t} \leq \alpha_t\}$, in which the center $\hat{\beta}$ is the Ridge estimator of $\beta$ using the contexts and rewards of the chosen arms up to time $t$ as covariate and outcome respectively and $V_t$ is the Gram matrix of the covariates. The Mahalanobis norm $\|\beta - \hat{\beta}\|_{V_t} \approx \sum_{r=1}^T X_{a(r),t} \epsilon_{a(r),t}|V_t^{-1}|$ is a self-normalized martingale, where $\sum_{r=1}^T X_{a(r),t} \epsilon_{a(r),t}$ forms a martingale and $V_t \approx \sum_{r=1}^T X_{a(r),t} X_{a(r),t}^T$ is the normalization term. [Abbasi-Yadkori et al. 2011] proved that when $\epsilon$ is sub-Gaussian, $C_t$ contains the true parameter with high probability for small enough positive value of $\alpha_t$. The main challenge in the proof is that the self-normalized martingale is not a martingale when arms are chosen adaptively. Hence, standard Azuma-Hoeffding inequalities cannot apply. [Abbasi-Yadkori et al. 2011] applied the inequality for self-normalized martingales [de la Peña et al. 2004; Peña et al. 2009] instead.

When errors $\epsilon_{i,t}$ are not sub-Gaussian but have only finite $(1 + \epsilon)$-th moments, the self-normalized inequality does not hold anymore. Therefore, we do not have a guarantee that the Ridge estimator of $\beta$ converges to the true value fast enough to allow a tight confidence region and optimal re-
Table 1: Regret order of linear bandit designed for heavy-tailed reward

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Regret Order</th>
<th>Fixed Context Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoM [Medina and Yang, 2016]</td>
<td>$O(\sqrt{dT \frac{1+\delta}{1+2\delta}} (\log T)^\frac{3+\delta}{2+2\delta})$</td>
<td>Yes</td>
</tr>
<tr>
<td>Truncation [Medina and Yang, 2016]</td>
<td>$O(dT \frac{1+\delta}{1+2\delta} \log T)$</td>
<td>No</td>
</tr>
<tr>
<td>MENU [Shao et al., 2018]</td>
<td>$O(dT \frac{1+\delta}{1+2\delta} (\log T)^\frac{3+\delta}{2+2\delta})$</td>
<td>Yes</td>
</tr>
<tr>
<td>TOFU [Shao et al., 2018]</td>
<td>$O(dT \frac{1+\delta}{1+2\delta} (\log T)^\frac{3+\delta}{2+2\delta})$</td>
<td>No</td>
</tr>
<tr>
<td>SupLinBMM [Xue et al., 2020]</td>
<td>$O(\sqrt{dT \frac{1+\delta}{1+2\delta}} (\log T)^\frac{3+\delta}{2+2\delta})$</td>
<td>No</td>
</tr>
<tr>
<td>SupLinBTC [Xue et al., 2020]</td>
<td>$O(\sqrt{dT \frac{1+\delta}{1+2\delta}} (\log T)^\frac{3+\delta}{2+2\delta})$</td>
<td>No</td>
</tr>
<tr>
<td>Huber Bandit (ours)</td>
<td>$O(\sqrt{dT \frac{1+\delta}{1+2\delta}} (\log T)^\frac{3+\delta}{2+2\delta})$</td>
<td>No</td>
</tr>
</tbody>
</table>

Regret. Hence, other estimators than the naive Ridge estimators should be considered. Recently, novel estimators [Medina and Yang, 2016] Shao et al. [2018] Xue et al. [2020] have been proposed motivated from the Median of Means (MoM) method and truncation method of [Bubeck et al., 2013] for multi-armed bandits without contexts. All these works follow [Abbasi-Yadkori et al., 2011] that they first construct a confidence region of $\hat{\beta}$ and then choose the arm according to the OFU principle.

[Medina and Yang, 2016] was the first to extend the MoM method to the estimation of the linear regression parameter, $\beta$. They proposed an algorithm which conducts in batches, where during each batch, the agent pulls the same arm with the same context variable repeatedly. At the end of each batch, the algorithm computes the MoM of the rewards which share the same context variable. Then the algorithm updates the Ridge estimator of $\beta$ using the context and MoM of rewards as new covariate-outcome pair. The caveat of the method is that while the errors of the individual rewards are heavy-tailed, the error of the MoM can be shown to be sub-Gaussian with high-probability. Therefore, the self-normalized inequality applies straightforwardly and a tight confidence region can be constructed. The paper derived a high-probability upper bound of the regret of order $O(\sqrt{dT \frac{1+\delta}{1+2\delta}} (\log T)^\frac{3+\delta}{2+2\delta})$. We remark that when $\delta = 1$, the order reduces to $O(dT \frac{1}{4})$ which is suboptimal when applied to sub-Gaussian rewards which have finite second moments. The MoM method does not recover the tight $O(\sqrt{T})$ regret bound despite the sub-Gaussianity of MoMs because we need multiple samples, up to $O(T \frac{1+\delta}{1+2\delta})$ samples, to construct a single MoM.

[Shao et al., 2018] refined the MoM method of [Medina and Yang, 2016] and proposed a novel algorithm called MENU which enjoys a tighter regret upper bound. MENU also executes in batches and requires to pull the same arm with the same context repeatedly in each batch. Instead of computing the MoM of rewards however, MENU updates multiple estimates of $\beta$ where each estimate is updated using only one context-reward pair. Among the different estimates $\beta^1, \beta^2, \cdots, \beta^k$, where $k$ is the size of the batch, MENU then selects the estimate which has the median value of $||\hat{\beta}^j - \beta||_V$, where $\beta$ is the true parameter. While $\beta^1, \beta^2, \cdots, \beta^k$ have not all a tight estimation error bound, it can be shown that $\beta^j$ which has the median Mahalanobis distance with $\beta$ achieves a tight estimation error bound. This refined estimator requires less samples $k$ in each batch to achieve the same estimation error bound as in the MoM method. Consequently, the regret of MENU is $O(\sqrt{dT \frac{1+\delta}{1+2\delta}} (\log T)^\frac{3+\delta}{2+2\delta})$. Now we observe that the bound is optimal with respect to $T$ when $\delta = 1$. The MENU algorithm is easy to employ and runs fast. However, the restriction that the same context variable should be observed during the same batch can be restrictive. In practice, context variables are often stochastic. Even for the same arm, the context information may change. In this case, we cannot run MENU.

[Medina and Yang, 2016] proposed an alternative estimator which does not require to pull the same arm with the same context variable in a row. The algorithm computes a Ridge estimator using truncated reward $\tilde{y}_t = y_t 1_{|y_t| \leq b_t}$, where the value of $b_t$ increases with time $t$. The quantity $\tilde{y}_t$ is bounded but biased. Hence, the error $\tilde{y}_t - X_{\alpha(t), \beta}^T \beta$ can be decomposed into a sub-Gaussian variable plus a bias term. The authors carefully choose the value of $b_t$ to trade off the self-normalized bound for the sub-Gaussian variables and the bound on the cumulative bias. Consequently, their algorithm achieves a high probability regret bound of $O(dT \frac{1+\delta}{1+2\delta} (\log T)^\frac{3+\delta}{2+2\delta})$. This bound also does not recover the $O(\sqrt{T})$ bound when $\delta = 1$.

[Shao et al., 2018] refined the truncation estimator of [Medina and Yang, 2016]. Instead of truncating the reward $y_t$, they truncated each element of $V_t^{-1/2} X_{\alpha(r), y_t}$, for every $r = 1, 2, \cdots, t$ by a time increasing threshold $b_t$. Hence, the truncation depends not only on the reward but also on the contexts of the chosen arms so far. Also, at each time $t$, the truncation is re-operated on all observations up to time $t$. This increases the time complexity of the algorithm but makes the estimator more accurate to obtain the tight regret upper bound of $O(dT \frac{1+\delta}{1+2\delta} (\log T)^\frac{3+\delta}{2+2\delta})$ which reproduces the optimal regret bound with respect to $T$ when $\delta = 1$. The analysis however relies on Bernstein’s inequalities for
We present the regret bounds of the aforementioned algorithm with Huber regression. Huber loss function [Huber, 1964] we propose a novel algorithm which does not require the prior knowledge of $\beta$ set of time steps, the Huber estimator of $\tau > 0$ for some robustification parameter $\tau$. In this paper, we propose to estimate the parameter $\beta$ in terms of $\nu$ and $T$. Removing these constraints in linear bandit algorithms in terms of $d$ and $T$ in Table 1.

In recent works in multi-armed bandits without contexts, there are algorithms which do not require the prior knowledge of $\nu$ [Lee et al., 2020] and even $\delta$ as well [Huang et al., 2022]. Removing these constraints in linear bandit would be promising.

## 4 PRELIMINARIES

In this paper, we propose to estimate the parameter $\beta$, in [1] with Huber regression. Huber loss function [Huber, 1964] is defined by

$$L_r(x) = \begin{cases} \frac{x^2}{2}, & \text{if } |x| \leq \tau, \\ \tau|x| - \tau^2/2, & \text{if } |x| > \tau, \end{cases}$$

for some robustification parameter $\tau > 0$. When $Z$ is a set of time steps, the Huber estimator of $\hat{\beta}$ fitted on data observed at time steps in $Z$ is defined as

$$\hat{\beta}(Z) = \arg \min_{\beta \in \mathbb{R}^d} \mathcal{L}_r(\tilde{\beta}, Z)$$

where

$$\mathcal{L}_r(\hat{\beta}, Z) = \frac{1}{|Z|} \sum_{r \in Z} l_r(y_r - X_r^T \hat{\beta}).$$

The Huber loss works as a square loss when the difference $|y_r - X_r^T \hat{\beta}|$ is smaller than $\tau$ and works as an absolute loss otherwise. Sun et al. [2020] analyzed the estimation error bound of Huber estimator in the fixed-design setting. They derived the value of $\tau$ which minimizes the estimation error bound in the following theorem.

**Theorem 1** (Theorem 1 of [Sun et al., 2020]). For any $\alpha > 0$, $\tau_0 \geq \nu_\delta$, the estimator $\hat{\beta}(t)$ with $\tau = \tau_0(t/\alpha)^{1/(1+\delta)}$ satisfies the bound

$$\|\hat{\beta}(t) - \beta\|_2 \leq 4c_1^{-1}L\tau_0d^{1/2} \left(\frac{\alpha}{T}\right)^{1+\delta}$$

with probability at least $1 - (2d + 1)e^{-\alpha}$, provided that

$$t \geq \max\{8M^4c_1^{-1}\alpha, 2^{4+\delta}M^2c_1^{-1}\alpha, 16\sqrt{2}c_1^{-1}LMd^{1/2}\alpha\}.$$  

Here, $c_1 \leq \lambda_{\min}(\Sigma([t]))$, $\tilde{\Sigma}([t]) = \frac{1}{\tau} \sum_{r \in [t]} X_r X_r^T$, $M = \max_{1 \leq r \leq t} \|X_r\|_2$, $L = \max_{1 \leq r \leq t} ||X_r||_{\infty}$.

In Theorem 1 we observe that the estimation error bound is proportional to the inverse of $c_1$, the minimum eigenvalue of the Gram matrix $\Sigma([t])$. The following theorem shows that when $X_r$’s are sampled independently and identically from a distribution with positive-definite covariance matrix, then $\lambda_{\min}(\Sigma([t]))$ is larger than a positive constant with high probability for sufficiently large $t$.

**Theorem 2** (Theorem 1.1 of [Tropp, 2012]). Consider a finite sequence $\{B_k\}$ of independent, random, self-adjoint matrix with dimension $d$. Assume $B_k \geq 0$ and $\lambda_{\max}(B_k) \leq M$ a.s. Then

$$\mathbb{P}(\lambda_{\min}(\sum_k B_k) \leq \alpha \mu_{\min}) \leq d \exp \left(-\frac{(1 - \alpha)^2 \mu_{\min}}{2M}\right),$$

where $\alpha \in [0, 1]$, $\mu_{\min} := \lambda_{\min}(\mathbb{E}[\sum_k B_k])$. In bandits however, the problem is that we do not have i.i.d. data due to adaptivity in the choice of arms. In [1], even if $X_i$’s are sampled i.i.d. from $P_X$, we can only use a subset of $X_i$’s for the estimation of $\hat{\beta}$, for each arm $i$. Due to adaptivity in the choice of arms, the subset is not a random sample from the full set of $X_i$’s. We propose to use the forced-sampling strategy of Goldenshluger and Zeevi [2013] to tackle this problem. The main idea behind the strategy is to maintain two estimators for each $\beta_i$, one using...
a small number of observations forcibly sampled in an i.i.d.

Without loss of generality, we assume Assumption 2.

Assumption 4 states that the expected Gram matrix of con-

At each time $t$, a context variable $X_t \in \mathcal{X} \subset \mathbb{R}^d$ is sampled i.i.d. from $P_X$. When arm $i$ is pulled, the arm returns a stochastic reward $y_t$ as in equation (7).

Assumption 1. At each time $t$, a context variable $X_t \in \mathcal{X} \subset \mathbb{R}^d$ is sampled i.i.d. from $P_X$. When arm $i$ is pulled, the arm returns a stochastic reward $y_t$ as in equation (7).

Assumption 2. Without loss of generality, we assume

||X_t||_2 \leq 1, ||\beta_i||_2 \leq 1, \forall t \in [T], i \in [K].

Assumption 3. [Arm optimality, Assumption 3 of Bastani and Bayati [2020]] The arm set is partitioned into two sets, optimal arms and sub optimal arms.

$[K] = K_{opt} \cup K_{sub}$, where $K_{opt} \cap K_{sub} = \emptyset$.

$i \in K_{sub}$ satisfies for $h > 0$,

$X^T \beta_i \leq \max_{j \neq i} X^T \beta_j - h, \forall X \in \mathcal{X}$.

For $i \in K_{opt}$, $\exists$ non-empty set

$U_i = \{x \in \mathcal{X} | X^T \beta_i > \max_{j \neq i} X^T \beta_j + h\}$

such that $P_X(X \in U_i) \geq \gamma > 0$.

Assumption 4. For all $i \in K_{opt}$ defined in Assumption 3, $\lambda_{min}(E[X_i X_i^T | X \in U_i]) \geq \gamma$, for $\gamma > 0$, where expectation is taken with respect to the distribution $P_X$.

Assumption 4 states that the expected Gram matrix of contexts in $U_i$ is positive definite, for each $i \in K_{opt}$. Assumptions 3 and 4 also guarantee a positive minimum eigenvalue for $E[X_i X_i^T]$ via the following lemma.

Lemma 1. Let $U$ be a set with $P(X \in U) \geq \gamma$. If $\lambda_{min}(E[X_i X_i^T | X \in U]) \geq \gamma$ for $\gamma > 0$, then

$\lambda_{min}(E[X_i X_i^T]) \geq \gamma p$.

5 PROPOSED ESTIMATOR AND ALGORITHM

Let the set $T_i := \{(2^n - 1)Kq + j | n \in \mathbb{N} \cup \{0\}, j \in \{q(i - 1) + 1, q(i - 1) + 2, \ldots, qi\}, q \in \mathbb{N}\}$ be the set of predetermined forced sampling time steps for arm $i$ and $T_{i,t} = T_i \cap [t]$ be the set of forced sampling time steps until time $t$. Since $|T_{i,t}| = O(\log T)$, the regret at forced sampling steps is $O(K\log T)$ at maximum. Let the set of time steps where arm $i$ is pulled, either forcibly or adaptively, until time $t$ be $S_{i,t} = \{r | a(r) = i, r \leq t\}$ and we call it all-sample set of the arm $i$. We have $T_{i,t} \subset S_{i,t}$.

Algorithm 1 Huber bandit

1: Input: $h, \nu, \alpha$
2: $\hat{\beta}(T_{i,0}) = \hat{\beta}(S_{i,0}) = 0$
3: for $t \in [T]$ do
4: Observe $X_t \sim P_X$
5: if $t = T_{i,t}$ then
6: $a(t) = i$
7: else
8: $D = \{i \in [K] | \max_{j \in [K]} X^T_i \hat{\beta}(T_{j,t-1}) - X^T_i \hat{\beta}(T_{i,t-1}) \leq \frac{1}{2}\}$
9: $a(t) = \arg \max_{i \in D} X^T_i \hat{\beta}(S_{i,t-1})$
10: end if
11: Update $S_{a(t), t} = S_{a(t), t-1} \cup \{t\}$
12: Observe reward $y_t = X^T \hat{\beta}(S_{i,t})$
13: if $t = T_{i,t}$ then
14: $\tau(T_{i,t}) = \nu d / (\log (t^2(2d+1)/\alpha))^{1/(1+\delta)}$
15: $\hat{\beta}(T_{i,t}) = \arg \min_{\beta \in \mathbb{R}^d} \frac{1}{|T_{i,t}|} \sum_{r \in T_{i,t}} l(y_r - X^T \hat{\beta})$
16: else
17: $\tau(S_{i,t}) = \nu d / (\log (t^2(2d+1)/\alpha))^{1/(1+\delta)}$
18: $\hat{\beta}(S_{i,t}) = \arg \min_{\beta \in \mathbb{R}^d} \frac{1}{|S_{i,t}|} \sum_{r \in S_{i,t}} l(y_r - X^T \hat{\beta})$
19: end if
20: end for

The proposed algorithm (Algorithm 1) works as follows. At each time $t \geq 1$, we observe context $X_t$. If $t \in T_{i,t}$, we pull the arm $i$. If $t \notin T_{i,t}$, we choose the arm using a two step procedure. First, we eliminate arms that are supposed to be suboptimal using the forced-sampling estimator $\hat{\beta}(T_{i,t-1})$. Afterward, we choose the arm $i$ which has maximum value of $X^T_i \hat{\beta}(S_{i,t})$ among the arms that survived the first step.

The following theorem shows that the proposed algorithm has regret upper bound of order $O(\sqrt{dT \gamma / \alpha} \log dT \gamma / \alpha)$. 

1031
Theorem 3. Suppose Assumptions 1-4 hold. When
\[ q \geq 6 \left( \frac{32(\tau_0 \lor 1)d^{1/2}}{h_\gamma p} \right)^{1/4} + t \geq \frac{(Kq)^2}{\phi}, \]
\[ C_1 = (1 + \delta) 64\tau_0 \frac{4}{\gamma_\tau} \left( \begin{array}{c}
\frac{\delta}{(1+\delta)} \\
\frac{\delta}{(2d+1)}
\end{array} \right) \]
and \( \tau_0 \geq \nu_\delta \), the cumulative regret \( R(T) \) is bounded by
\[ R(T) \leq C_1 T^{1+\delta} \left( \log(T^2(2d+1)10K/\phi) \right)^{1/1+\delta} \sqrt{d} 
+ 4Kq \log T + \frac{2(Kq)^2}{\phi} \]
with probability at least \( 1 - \phi \).

Proof. We first need to adapt Theorem 1, which is originally proved for independent data, to work for adaptively collected data. In the original version of the theorem, the i.i.d. assumption is exploited in two parts, (i) bounding the norm of the gradient of the loss function which can be written as the sum of nonlinear function of i.i.d. errors and (ii) guaranteeing the minimum eigenvalue of the Gram matrix to be bounded below by a positive constant. For (i), since the errors are no more i.i.d., we derive a novel proof using martingale inequalities in Lemma 2. As for (ii), we borrow idea of forced sampling strategy used in existing works which we see later.

Lemma 2 (All sample estimator bound). Let \( \tau = \tau_0(|S_{i,t}|/\log(t^2(2d+1)/|\alpha|))^{1/(1+\delta)} \), \( \tau_0 \geq \nu_\delta \). If \( t \geq (Kq)^2 \) and \( c_1 \leq \lambda_{\min}(\Sigma(S_{i,t})) \), we have with probability at least \( 1 - \frac{\delta}{t} \),
\[ \| \hat{\theta}(S_{i,t}) - \theta_i \|_2 \leq \left( \frac{4}{pt} \log(t^2(2d+1)/|\alpha|) \right)^{1/(1+\delta)} 4\tau_0 d^{1/2} c_1^{-1} \]
for \( i \in [K] \).

Proof. We mainly provide the part we solve differently from the proof of [Sun et al. 2020]. If we know the minimum eigenvalue bound of the Gram matrix, we can obtain \( L_2 \)-norm bound
\[ \| \hat{\theta}(S_{i,t}) - \theta_i \|_2 \leq 2|S_{i,t}|^{-1} c_1^{-1} \| \sum_{r \in S_{i,t}} \psi_r(\epsilon_r) X_r \|_2 \]
where we use \( \max_{r \in S_{i,t}} |X_r|_\infty \leq 1 \) and \( X_r \) is \( j \)-th element of the context \( X_r \). We observe that in bandit settings, the right-hand side of (3) is the sum of adapted data. To bound this quantity, we invoke a bound on supermartingales. We can first construct a supermartingale as follows. Let
\[ M_t := \exp \left( \sum_{r=1}^t \left[ (X_{r,j}/\tau) \psi_r(\epsilon_r) - \nu_\delta \frac{\epsilon_r}{\tau} \right] \right). \]

Then
\[ \mathbb{E}[M_t | F_{t-1}] \]
\[ = M_{t-1} \mathbb{E}\left[ \exp \left( (X_{r,j}/\tau) \psi_r(\epsilon_r) \right) | F_{t-1} \right] \exp \left( \frac{\nu_\delta}{\tau^2} \right) \]
\[ \leq M_{t-1} \left( 1 + X_{r,j}(\epsilon_r/\tau) + |\epsilon_r/\tau|^{1+\delta} \right) / \exp \left( \frac{\nu_\delta}{\tau^2} \right) \]
\[ \leq M_{t-1} \left( 1 + \frac{\nu_\delta}{\tau^2} \right) / \exp \left( \frac{\nu_\delta}{\tau^2} \right) \]
\[ = M_{t-1} \]
shows that \( M_t \) is a supermartingale with \( M_0 := 1 \), where the second inequality is derived from
\[ - \log(1 - u + |u|^{1+\delta}) \leq \frac{1}{\tau} \psi_r(\tau u) \leq \log(1 + u + |u|^{1+\delta}) \]
\( \forall u \in \mathbb{R} \) and
\[ \exp((X_{r,j}/\tau) \psi_r(\epsilon_r)) \]
\[ \leq \left( 1 + \epsilon_r/\tau + |\epsilon_r/\tau|^{1+\delta} \right) X_{r,j} 1_{X_{r,j} > 0} \]
\[ \cdot \left( 1 - \epsilon_r/\tau + |\epsilon_r/\tau|^{1+\delta} \right) X_{r,j} 1_{X_{r,j} < 0} \]
\[ \leq 1 + X_{r,j}(\epsilon_r/\tau) + |\epsilon_r/\tau|^{1+\delta}. \]

Iteratively applying the law of total expectation on \( M_t \) gives \( \mathbb{E}[M_t] \leq 1 \) and hence,
\[ \mathbb{E} \left[ \exp \left( \sum_{r \in S_{i,t}} (X_{r,j}/\tau) \psi_r(\epsilon_r) \right) \right] \leq \exp(|S_{i,t}|\nu_\delta \tau^{-1-\delta}). \]

Markov’s inequality gives
\[ \mathbb{P} \left( \sum_{r \in S_{i,t}} (X_{r,j}/\tau) \psi_r(\epsilon_r) > \nu_\delta |S_{i,t}|z \right) \]
\[ \leq \exp(-\nu_\delta |S_{i,t}|z) \mathbb{E} \left[ \exp \left( \sum_{r \in S_{i,t}} (X_{r,j}/\tau) \psi_r(\epsilon_r) \right) \right] \]
\[ \leq \exp(\nu_\delta |S_{i,t}|(\tau^{-1-\delta} - z)) \]
for some \( z > 0 \). Then when \( \tau \geq (2/z)^{1/(1+\delta)} \) and \( z = 2\nu_\delta^{-1} |S_{i,t}|^{-1} \log(t^2(2d+1)/|\alpha|) \), with probability at least \( 1 - \alpha/t^2 \), the same bound as in Theorem 1 can be obtained. \( \square \)
We note that in Lemma 2 one does not need to invoke a bound for self-normalized supermartingales if the minimum eigenvalue of the Gram matrix is lower bounded by a positive constant. However in bandit settings, guaranteeing a positive constant lower bound for the minimum eigenvalue of Gram matrices is challenging. As mentioned earlier, we utilize the forced-sampling strategy to address this challenge. We prove below through a sequence of lemmas that by means of forced-sampling, the Gram matrix $\Sigma(S_{i,t})$ of the all-sample estimator of any arm $i \in K_{opt}$ has a minimum eigenvalue greater than a positive constant.

**Lemma 3** (Lemma EC.23 of Bastani and Bayati [2020]). Let $A$ be a set of random variables. Consider a subset $A' \subset A$ of i.i.d. random variables. If $\lambda_{\min}(\Sigma(A')) \geq \gamma$ for some $\gamma > 0$, then

$$\lambda_{\min}(\Sigma(A)) \geq \frac{|A'|}{|A|}\gamma.$$ 

Lemma 3 states that we can get the minimum eigenvalue bound of the Gram matrix for a possibly non i.i.d. set $A$, if we know the bound for an i.i.d. subset of the set $A$. We will construct the i.i.d. subset of the set $S_{i,t}$ and show that the size of the subset is proportional to $|S_{i,t}|$ in terms of $t$.

We first establish the estimation error bound of the forced sampling estimator, which plays a crucial role in constructing a sufficiently big i.i.d. subset of $S_{i,t}$ for every $i \in K_{opt}$.

**Lemma 4** (Forced-sampling estimator bound). Define an event at time $t$ as

$$A_t := \left\{ \|\hat{\beta}(T_{i,t}) - \beta_i\|_2 \leq \frac{h}{4}, \forall i \in [K] \right\}.$$ 

For $\alpha \in (0, 1)$, we have $\mathbb{P}(A_t) \geq 1 - 2K\alpha^2/t^2$, provided that $t \geq (Kq)^2$.

**Proof.** The lemma is direct application of Theorem 1. The proof follows the lines in Proposition 2 of Bastani and Bayati [2020]. Since the forced sampling set $T_{i,t}$ is set deterministically prior to running the algorithm, the samples in $T_{i,t}$ are i.i.d. Therefore, the minimum eigenvalue bound of the Gram matrix can be derived from Lemma 1 and Theorem 2.

Using Lemma 4 we can prove the following lemma which states that under the high-probability event $A_t$, the set $D$ of arms after discarding the suboptimal arms using $\hat{\beta}(T_{i,t})$ contains arms from $K_{opt}$ only. Therefore under this high-probability event, the algorithm makes action choices using the estimates $\hat{\beta}(S_{i,t})$’s of arms $i \in K_{opt}$ only. Hence, we need to guarantee sufficiently large i.i.d. subset of contexts for arms in $K_{opt}$ only.

**Lemma 5** (Lemma EC.18 of Bastani and Bayati [2020]). If $A_{t-1}$ holds, then $D = \{ i \in [K] | \max_{j \in [K]} X_{T_{j,t-1}}^t \hat{\beta}(T_{j,t-1}) - X_{T_{i,t-1}}^t \hat{\beta}(T_{i,t-1}) \leq \frac{h}{2} \}$ contains the optimal arm $a^*(t)$ and no arms from $K_{sub}$.

Now we derive i.i.d. subset of the set $S_{i,t}$ for any arm $i \in K_{opt}$. We first define the following subset of $[t]$ which describes the time steps where contexts are sampled from the optimal regions $U_i$’s and where the most recently updated forced sampling estimators have low estimation error bound.

$$A_{i,t} := \{ r \in [t] | A_{r-1} \text{ holds, } X_r \in U_i \text{ and } r \notin \bigcup_{j \in [K]} T_{j,t} \}.$$ 

Observe that the random variables $\{x_r | r \in A_{i,t}\}$ are i.i.d. in $U_i$. This is because the event $\{x_r \in U_i\}$ is independent of the event $A_{r-1}$ while the event $\{r \notin T_{j,t}\}$ is deterministic. Therefore, we can prove $\lambda_{\min}(\hat{\Sigma}(A_{i,t}))$ is strictly positive with high probability via Assumption 4 and Theorem 2.

**Lemma 6** states that $A_{i,t}$ is a subset of $S_{i,t}$ and Lemma 7 states that the size of $A_{i,t}$ is large enough. Particularly, Lemma 7 guarantees that the size of $A_{i,t}$ is at least proportional to $t$.

**Lemma 6** (All sample set, Lemma EC.11 of Bastani and Bayati [2020]). For $i \in K_{opt}$, if $t \in A_{i,t}$, then $a(t) = i$.

**Lemma 7** (All sample set, Lemma EC.14 of Bastani and Bayati [2020]). If $t \geq (Kq)^2$, for $i \in K_{opt}$, 

$$\mathbb{P}(|A_{i,t}| \geq tp/4) \geq 1 - \frac{\alpha}{t^2}.$$ 

Substituting $S_{i,t}$ into $A$ and $A_{i,t}$ into $A'$ in Lemma 3, we get the minimum eigenvalue bound of $\hat{\Sigma}(S_{i,t})$. Since $|A_{i,t}|$ is proportional to $t$ and $|S_{i,t}| \leq t$, the ratio $|A'|/|A|$ is of constant order. Theorem 2 states that

$$\mathbb{P}\left( \lambda_{\min}(\hat{\Sigma}(S_{i,t})) \leq \frac{\gamma}{2} \right) \leq d \exp\left( -\frac{|A_{i,t}| \gamma}{8} \right),$$

for $\gamma$ in Assumption 1. The size of the set $A_{i,t}$ is guaranteed by Lemma 6. When $t \geq \frac{d}{\alpha}$ and $q \geq \frac{192}{\alpha^2}$, with probability at least $1 - \frac{\alpha}{t^2}$,

$$|A_{i,t}| > \frac{tp}{4} \geq \frac{|T_{i,t}|p}{4} \geq \frac{8}{\gamma} (\log(d/\alpha) + \log t^2),$$

and hence

$$\lambda_{\min}(\hat{\Sigma}(S_{i,t})) \geq \frac{\gamma}{2}.$$ 

Therefore, together with Lemma 3 with probability at least $1 - \frac{\alpha}{t^2}$,

$$\lambda_{\min}(\hat{\Sigma}(S_{i,t})) \geq \frac{\gamma |A_{i,t}|}{2|S_{i,t}|} > \frac{\gamma |A_{i,t}|}{2t} > \frac{p \gamma}{8}. \quad (4)$$

Finally, we are ready to prove the Theorem 2.

**Proof of Theorem 2**. We consider two cases; (a) when $t < (Kq)^2/\phi$ or $t \notin \bigcup_{i \in [K]} T_{i,t}$, (b) when $t \geq (Kq)^2/\phi$ and $t \notin \bigcup_{i \in [K]} T_{i,t}$. For (a), we know that $|\bigcup_{i \in [K]} T_{i,t}| \leq 2Kq \log T$ by definition of $T_{t,t}$. Hence, the time occupied by the event...
With probability at least $1 - 2\alpha K/t^2$, $A_t$ holds. When $A_t$ holds, due to Lemma 4, we choose an arm from $K_{opt}$ only. Then we can obtain minimum eigenvalue bound in Lemma 2 holds. Then with probability at least $\rho_i$ of (4) $\forall i \in D$ with probability at least $1 - 2\alpha K/t^2$. The regret at time $t$ is

$$r_t = X^T \beta_{a^*(t)} - X^T \beta_{a(t)} = X^T \beta_{a^*(t)} - X^T \beta(S_{a(t)}, t) + X^T \hat{\beta}(S_{a(t)}, t) + X^T \beta_{a^*(t)} - X^T \beta_{a(t)}$$

$$\leq \|X_t\|_2 \|\beta_{a^*(t)} - \hat{\beta}(S_{a(t)}, t)\|_2 + \|X_t\|_2 \|\beta_{a(t)} - \hat{\beta}(S_{a(t)}, t)\|_2.$$

With probability at least $1 - \alpha K/t^2$, $\forall i \in D$, $L_2$-norm bound in Lemma 2 holds. Then with probability at least $1 - 10\alpha K$,

$$\sum_{t=1}^{T} r_t \leq 2 \sum_{t=1}^{T} \left( \frac{4}{p t \log(t^2(2d + 1)/\alpha)} \right)^{\delta/(1+\delta)} \frac{32\tau_0 d^{1/2}}{\gamma p}$$

$$\leq \sum_{t=1}^{T} \left( \frac{4}{p t \log(T^2(2d + 1)/\alpha)} \right)^{\delta/(1+\delta)} \frac{64\tau_0 d^{1/2}}{\gamma p}$$

$$\leq (T \tau_0^{1/2} - 1)(1 + \delta) \left( \frac{4}{p \log(T^2(2d + 1)/\alpha)} \right)^{\delta/(1+\delta)} \frac{64\tau_0 d^{1/2}}{\gamma p}.$$

Then together with (a), let $\phi = 10\alpha K$ gives the desired result.

We provide additional Theorem 4 which shows that if we add additional Assumption 5 of marginal condition, we can get $O(\log T)$ expected regret.

**Assumption 5** (Assumption 2 of Bastani and Bayati [2020]).

$\exists C_0 \in \mathbb{R}^+$ such that $\forall i, j \in [K]$ where $i \neq j$,

$$\mathbb{P}(0 < |X_i^T (\beta_i - \beta_j)| \leq \kappa) \leq C_0 K \forall \kappa \in \mathbb{R}^+.$$

**Theorem 4.** Suppose Assumptions 7 and 5 hold. When

$$q \geq 6 \left( \frac{32(\tau_0 \vee 1) d^{1/2}}{h \gamma p} \right)^{1/\delta}, t \geq \frac{(Kq)^2}{\phi},$$

$$C_2 = 2^{14} \frac{\tau_0^2 C_0}{\gamma^2 p^2}, C_3 = 2^{1/2 + 12/\delta} \left( \frac{1 + \delta}{1 - \delta} \right) \frac{\tau_0^2 C_0}{\gamma^2 p^{1/\delta}},$$

and $\tau_0 \geq \nu_3$, the expected regret is

$$\sum_{t=1}^{T} \mathbb{E}[r_t] \leq C_2(\log T)(\log T + 1)dK + (\log T + 1)12K(2d + 1)$$

when $\delta = 1$ and

$$\sum_{t=1}^{T} \mathbb{E}[r_t] \leq C_3 T^{\frac{1-\delta}{4}} (\log T)^{\frac{2\delta}{1-\delta}} dK + (\log T + 1)12K(2d + 1)$$

when $0 < \delta < 1$.

The full proof of Theorem 4 is deferred to the Supplementary Material.

**6 EXPERIMENTS**

We compare the Huber bandit algorithm with OLS bandit [Bastani and Bayati [2020], TOFU [Shao et al. [2018]] and SupLinBTC [Xue et al., 2020]]. OLS bandit has the same structure of the forced-sampling as the Huber bandit but the Ordinary Least Squares(OLS) estimator substitutes the Huber estimator. TOFU and SupLinBTC are designed for the single parameter $\beta \in \mathbb{R}^d$ and arm-specific contexts, $X_{i,t} \in \mathbb{R}^d$, $i \in [K]$ setting. To bring the algorithms to our setting, we reshape arm parameters $\beta_i \in \mathbb{R}^d$, $i \in [K]$ into one parameter $\beta \in \mathbb{R}^{dk}$ and context $X_i$ into $X_{i,t} \in \mathbb{R}^{dk}$, $i \in [K]$ such that $X_{i,t}^T \beta = X_i^T \beta_{i_1}$. We randomly generate context $X_i \in \mathbb{R}^d$ and arm parameters $\beta_i \in \mathbb{R}^d$, $i \in [K]$ from a uniform distribution in $[-1, 1]$ and normalize them to satisfy the assumption 2. Due to the random generations of the samples, the assumption 3 might not be satisfied. Instead, we arbitrarily set $h$ to 0.2. Nevertheless, the Huber bandit shows a good performance. The error term $\epsilon_t$ is generated from the student’s t-distribution with a degree of freedom(df) $\{1, 1.5, 2\}$ and multiplied by 0.1 to balance with the norm of condition of $\beta_i$ and $X_i$. If $df > n$, finite $n$-th moment of the student’s t-distribution exists. In the experiment, we assume $1 + \delta = df - 0.05$ moments exist. We take the context dimension $d = 5$, the number of arms $K = 10$, the time horizon $T = 1000$. The TOFU algorithm has one hyperparameter $cond$ which controls the size of the confidence interval of the regression parameter. Our algorithm and OLS bandit both have the hyperparameter $q$ which controls the number of forced sampling steps. We run the algorithms with $cond = 0.5, 1, 1.5, 2$ and $q = 2, 3, 4$ and report the results of the values that resulted in the lowest average regret. We run 100 iterations each with a new data set. The
Figure 1: Cumulative regret of Huber bandit, OLS bandit, TOFU.

The results of cumulative regret of Huber, OLS, TOFU and SupLinBTC averaged over 100 iterations are shown in figure 1. The shade of the graph shows the standard deviation of 100 iterations. We use $\alpha = 0.01$ for all algorithms.

7 CONCLUSION

In this paper, we proposed the Huber bandit algorithm robust to heavy-tailed error. The theoretical analysis shows that when contexts are stochastic with positive definite covariance matrix, the algorithm achieves the regret bound of $O(\sqrt{dT} \log^2 dT)$ which matches the state-of-the-art regret upper bound for linear bandits with sub-Gaussian errors in terms of the time horizon $T$ when $\delta = 1$. The practical performance was proved by comparing it with OLS bandit and two existing bandit algorithms designed for heavy-tailed data.

ACKNOWLEDGEMENTS

This work was supported by the Institute of Information & communications Technology Planning & evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2020-0-01336, Artificial Intelligence Graduate School Program (UNIST); No. 2022-0-00469, Development of Core Technologies for Task-oriented Reinforcement Learning for Commercialization of Autonomous Drones) and the “Research on multi-armed bandit methodologies for online sequential decision” Project Fund (1.200107.01) of UNIST, South Korea.

References


