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 $O(\sqrt{d}T^{\frac{1}{1+\delta}}(\log dT)^{\frac{\delta}{1+\delta}})$, 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 [Lei et al., 2017], clinical trial [Villar et al., 2015] 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[Rachev, 2003, Stehlík et al., 2010, Ibragimov et al., 2015]. 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, ϵ, has finite $(1+\delta)$-th moment, i.e., $\mathbb{E}[|\epsilon|^{1+\delta}] \leq \nu_\delta < \infty$ for some $\delta \in (0,1]$. This assumption is common in the bandit literature which consider heavy-tailed errors [Bubeck et al., 2013, Medina and Yang, 2016, Shao et al., 2018, Xue et al., 2020]. We propose to use the Huber estimator [Huber, 1964] 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. Sun et al. [2020] 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

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (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 $P_{\mathcal{X}}$ independently. We denote the number of arms as K. When the learning agent pulls the i-th arm at time t, the arm reveals a stochastic reward

$$y_{i,t} = X_t^T \beta_i + \epsilon_{i,t}, \tag{1}$$

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 observed one reward, $y_t := y_{a(t),t}$. We also assume $\mathbb{E}[\epsilon_{i,t} \mid \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}\left[|\epsilon_{i,t}|^{1+\delta} \mid \mathcal{F}_{t-1}\right] \leq \nu_\delta < \infty$, for $\nu_\delta > 0$ and for some $\delta \in (0, 1]$, where $\mathcal{F}_t = \sigma(X_1, ..., X_{t+1}, a(1), a(2), ..., a(t), y_1, ..., y_t)$ denotes the σ-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 $a^*(t) = \arg\max_{j \in [K]} X_t^T \beta_j$ 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.

$$y_{i,t} = X_{i,t}^T \beta + \epsilon_{i,t}. \tag{2}$$

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 $X_t = \{X_{1,t}^T \ \cdots \ X_{K,t}^T\}^T \in \mathbb{R}^{Kd}$ and $\beta_i = \{\mathbf{0}^T \ \cdots \ \mathbf{0}^T \ \beta^T \ \mathbf{0}^T \ \cdots \ \mathbf{0}^T\}^T \in \mathbb{R}^{Kd}$, where $\mathbf{0}$ 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 β_i. On the other hand, when we apply an algorithm designed for (2) on setting (1), we simply let $\beta = \{\beta_1^T \ \cdots \ \beta_K^T\}^T \in \mathbb{R}^{Kd}$ and $X_{i,t} = \{\mathbf{0}^T \ \cdots \ \mathbf{0}^T \ X^T \ \mathbf{0}^T \ \cdots \ \mathbf{0}^T\}^T \in \mathbb{R}^{Kd}$.

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. We let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimum and maximum eigenvalues of matrix A respectively. For any vector $v \in \mathbb{R}^d$, $v = (v_1, ..., v_d)^T$, L_2-norm is $||v||_2 = (\sum_{i=1}^d |v_i|^2)^{\frac{1}{2}}$ 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 $A \succcurlyeq B$ if $B - A$ is positive semi definite. For any set A, let $\mathbb{1}_A(\cdot)$ denote indicator function of a subset A of some set W. For $a, b \in \mathbb{R}$, $a \vee b := \max\{a, b\}$. The set $\mathbb{N}$ denotes the set of natural numbers. When $Z \subset \mathbb{N}$, we denote the Gram matrix as $\hat{\Sigma}(Z) := \frac{1}{|Z|} \sum_{r\in Z} X_r X_r^T$.

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 $\tilde{O}(d\sqrt{T})$ for setting (2), where $\tilde{O}(\cdot)$ ignores logarithmic terms. The strategy of LinOFUL is to construct a tight confidence region of the true parameter β 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 $C_t = \{\tilde{\beta} : ||\tilde{\beta} - \hat{\beta}_t||_{V_t} \leq \alpha_t\}$, in which the center $\hat{\beta}_t$ is the Ridge estimator of β 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 $||\beta - \hat{\beta}_t||_{V_t} \approx ||\sum_{r=1}^t X_{a(r),r}\epsilon_{a(r),r}||_{V_t^{-1}}$ is a self-normalized martingale, where $\sum_{r=1}^t X_{a(r),r}\epsilon_{a(r),r}$ forms a martingale and $V_t \approx \sum_{r=1}^t X_{a(r),r} X_{a(r),r}^T$ is the normalization term. Abbasi-Yadkori et al. [2011] proved that when ϵ is sub-Gaussian, C_t contains the true parameter with high probability for small enough positive value of α_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 Pena et al., 2004, Peña et al., 2009] instead.

When errors $\epsilon_{i,t}$ are not sub-Gaussian but have only finite $(1+\delta)$-th moments, the self-normalized inequality does not hold anymore. Therefore, we do not have a guarantee that the Ridge estimator of β 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

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

gret. 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] in that they first construct a confidence region of β 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, β. 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 β 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 $O(\sqrt{d}T^{\frac{1+2\delta}{1+3\delta}}(\log T)^{\frac{3\delta+1}{2(1+\delta)}})$. We remark that when $\delta = 1$, the order reduces to $\tilde{O}(dT^{\frac{3}{4}})$ 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 $O(T^{\frac{1+\delta}{1+3\delta}})$ 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 β where each estimate is updated using only one context-reward pair. Among the different estimates $\hat{\beta}^1, \hat{\beta}^2, \cdots, \hat{\beta}^k$, where k is the size of the batch, MENU then selects the estimate which has the median value of $||\hat{\beta}^j - \beta||_{V_t}$ where β is the true parameter. While $\hat{\beta}^1, \hat{\beta}^2, \cdots, \hat{\beta}^k$ have not all a tight estimation error bound, it can be shown that $\hat{\beta}^j$ which has the median Mahalanobis distance with β 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 $O(d^{\frac{3+\delta}{2(1+\delta)}}T^{\frac{1}{1+\delta}}(\log T)^{\frac{3\delta+1}{2(1+\delta)}})$. 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 $\hat{y}_t = y_t \mathbb{1}_{|y_t| \le b_t}$, where the value of b_t increases with time t. The quantity $\hat{y}_t$ is bounded but biased. Hence, the error $\hat{y}_t - X_{a(t),t}^T\beta$ 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 $O(dT^{\frac{2+\delta}{2(1+\delta)}}\log T)$. 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 $V_t^{-1/2}X_{a(r),r}y_r$ 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 $O(dT^{\frac{1}{1+\delta}}(\log T)^{\frac{3\delta+1}{2(1+\delta)}})$ which reproduces the optimal regret bound with respect to T when $\delta = 1$. The analysis however relies on Bernstein's inequalities for

martingales, which can be applied to self-normalized martingales under restrictive conditions only. We note that a self-normalized martingale becomes a martingale only when the contexts of chosen arms constitute a fixed design, i.e., when the covariates used in the estimator are fixed prior to observing the outcomes(rewards). In adaptively collected data, we usually do not have a fixed design since the covariate at time t is chosen based on the rewards up to time $t-1$.

Xue et al. [2020] blended the median of means and truncation method with the SupLinUCB algorithm [Chu et al., 2011] and achieved the regret bounds $O(\sqrt{d}(\log T)^{\frac{3}{2}}T^{\frac{1}{1+\delta}})$ and $O(\sqrt{d}(\log T)^2 T^{\frac{1}{1+\delta}})$, respectively. They refined the estimators of Medina and Yang [2016] and Shao et al. [2018] so that the contexts at the current time point are also considered when taking median of means and truncation. However, the derivation of the confidence interval for β based on their estimators is valid under fixed design only. Therefore, the authors adopted the phased structure of SupLinUCB[Chu et al., 2011] which ensures that the contexts of arms chosen at time points in the same phase constitute a fixed design [Auer, 2002]. Therefore, the arms that are chosen at time points in the same phase are only correlated with rewards from precedent phases and not correlated with rewards in the same phase. Hence, when an estimator $\hat{\beta}$ based on MoM or truncation is computed from observations in the same phase, the Hoeffding's inequality can be applied to the self-normalized martingales. The bounds of Xue et al. [2020] are state-of-the-art, shaving off $\sqrt{d}$ factor from the bounds of Shao et al. [2018]. However, their algorithm requires to know the time horizon T prior to running the algorithm to determine the optimal number of phases. In this paper, we propose a novel algorithm which does not require the knowledge of T.

We present the regret bounds of the aforementioned 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 ν_δ [Lee et al., 2020] and even δ 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 β_i in (1) with Huber regression. Huber loss function [Huber, 1964] is defined by

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

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

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

where

$$\mathcal{L}_\tau(\tilde{\beta}, Z) = \frac{1}{|Z|}\sum_{r\in Z} l_\tau(y_r - X_r^T\tilde{\beta}).$$

The Huber loss works as a square loss when the difference $|y_r - X_r^T\tilde{\beta}|$ is smaller than τ 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 τ which minimizes the estimation error bound in the following theorem.

Theorem 1 (Theorem1 of Sun et al. [2020])**.** *For any $\alpha > 0$, $\tau_0 \ge \nu_\delta$, the estimator $\hat{\beta}([t])$ with $\tau = \tau_0(t/\alpha)^{1/(1+\delta)}$ satisfies the bound*

$$||\hat{\beta}([t]) - \beta||_2 \le 4c_l^{-1}L\tau_0 d^{1/2}\left(\frac{\alpha}{t}\right)^{\frac{\delta}{1+\delta}}$$

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

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

Here, $c_l \le \lambda_{\min}(\hat{\Sigma}([t]))$, $\hat{\Sigma}([t]) = \frac{1}{t}\sum_{r\in[t]} X_rX_r^T$, $M = \max_{1\le r\le t}||X_r||_2$, $L = \max_{1\le r\le t}||X_r||_\infty$.

In Theorem 1, we observe that the estimation error bound is proportional to the inverse of c_l, the minimum eigenvalue of the Gram matrix $\hat{\Sigma}([t])$. The following theorem shows that when X_r's are sampled independently and identically from a distribution with positive-definite covariance matrix, then $\lambda_{\min}(\hat{\Sigma}([t]))$ 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 $B_k \succcurlyeq 0$ and $\lambda_{\max}(B_k) \le M$ $a.s.$ Then*

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

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_t's are sampled i.i.d. from $P_\mathcal{X}$, we can only use a subset of X_t's for the estimation of β_i 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_t'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 β_i, one using

a small number of observations forcibly sampled in an i.i.d. fashion and another estimator which uses all observations, both forcibly sampled and adaptively sampled. The forced-sample estimator is then used to discard the suboptimal arms and concentrate the arm choices to optimal arms. Consequently, a constant portion of the adaptively sampled data is guaranteed to follow an i.i.d. distribution over a specific region in $\mathcal{X}$. Hereby, the all-sample estimator which uses both forcibly and adaptively sampled data enjoys a tight estimation error bound.

Goldenshluger and Zeevi [2013] first proposed the forced sampling strategy for a contextual bandit setting. Bastani and Bayati [2020] extended this strategy to the multiple-arm setting and used it to prove a tight estimation error bound of the Lasso estimator in bandit settings which also requires a similar minimum eigenvalue bound as the one Huber estimator requires.

Before proceeding, we state few assumptions here.

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

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.*

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

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

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

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

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

such that $\mathbb{P}_{\mathcal{X}}(X \in U_i) \geq p > 0$.

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

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 $\mathbb{E}[XX^T]$ via the following lemma.

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

$$\lambda_{\min}\left(\mathbb{E}[XX^T]\right) \geq \gamma p.$$

5 PROPOSED ESTIMATOR AND ALGORITHM

Let the set $T_i := \{(2^n - 1)Kq + j | n \in \mathbb{N} \cup \{0\}, j \in \{q(i-1)+1, q(i-1)+2, ..., qi\}, q \in \mathbb{N}\}$ 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 forcedly 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

Input: $h,\ \nu_\delta,\ \alpha$
$\hat{\beta}(T_{i,0}) = \hat{\beta}(S_{i,0}) = 0^d$
for $t \in [T]$ **do**
  Observe $X_t \sim \mathcal{P}_{\mathcal{X}}$
  if $t \in T_i$ **then**
    $a(t) = i$
  else
    $\mathcal{D} = \{i \in [K] | \max_{j \in [K]} X_t^T \hat{\beta}(T_{j,t-1}) - X_t^T \hat{\beta}(T_{i,t-1}) \leq \frac{h}{2}\}$.
    $a(t) = \arg\max_{i \in \mathcal{D}} X_t^T \hat{\beta}(S_{i,t-1})$
  end if
  Update $S_{a(t),t} = S_{a(t),t-1} \cup \{t\}$
  Observe reward $y_t = X_t^T \beta_{a(t)} + \epsilon_{a(t),t}$
  if $t \in T_i$ **then**
    $\tau(T_{i,t}) = \nu_\delta(|T_{i,t}| / \log(t^2(2d+1)/\alpha))^{1/(1+\delta)}$
    $\hat{\beta}(T_{i,t}) = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{|T_{i,t}|} \sum_{r \in T_{i,t}} l_\tau(y_r - X_r^T\beta)$
  else
    $\tau(S_{i,t}) = \nu_\delta(|S_{i,t}| / \log(t^2(2d+1)/\alpha))^{1/(1+\delta)}$
    $\hat{\beta}(S_{i,t}) = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{|S_{i,t}|} \sum_{r \in S_{i,t}} l_\tau(y_r - X_r^T\beta)$
  end if
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 $X_t^T \hat{\beta}(S_{i,t})$ among the arms that survived the first step.

The following theorem shows that the proposed algorithm has regret upper bound of order $O(\sqrt{d}T^{\frac{1}{1+\delta}}(\log dT)^{\frac{\delta}{1+\delta}})$.

Theorem 3. *Suppose Assumptions 1-4 hold. When*

$$q \geq 6\left(\frac{32(\tau_0 \vee 1)d^{1/2}}{h\gamma p}\right)^{\frac{1+\delta}{\delta}},\ t \geq \frac{(Kq)^2}{\phi},$$
$$C_1 = (1+\delta)\frac{64\tau_0}{\gamma p}\left(\frac{4}{p}\right)^{\delta/(1+\delta)}$$

and $\tau_o \geq \nu_\delta$, the cumulative regret $R(T)$ is bounded by

$$R(T) \leq C_1 T^{\frac{1}{1+\delta}}\left(\log(T^2(2d+1)10K/\phi)\right)^{\delta/(1+\delta)}\sqrt{d} + 4Kq\log T + \frac{2(Kq)^2}{\phi}$$

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 $c_l \leq \lambda_{\min}(\hat{\Sigma}(S_{i,t}))$, we have with probability at least $1-\frac{\alpha}{t^2}$,*

$$||\hat{\beta}(S_{i,t}) - \beta_i||_2 \leq \left(\frac{4}{pt}\log(t^2(2d+1)/\alpha)\right)^{\delta/(1+\delta)} 4\tau_0 d^{1/2} c_l^{-1}$$

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

$$||\hat{\beta}(S_{i,t}) - \beta_i||_2 \leq 2|S_{i,t}|^{-1}c_l^{-1}||\sum_{r\in S_{i,t}}\psi_\tau(\epsilon_r)X_r||_2$$

$\psi_\tau(\epsilon_r) := l'_\tau(\epsilon_r)$ is the derivative of the Huber loss. The detailed explanation can be found in Section C.3 of Sun et al. [2020]. The L_2-norm is bounded by max-norm.

$$\left\|\sum_{r\in S_{i,t}}\psi_\tau(\epsilon_r)X_r\right\|_2 \leq d^{1/2}\left\|\sum_{r\in S_{i,t}}\psi_\tau(\epsilon_r)X_r\right\|_\infty = d^{1/2}\tau\max_{1\leq j\leq d}\left|\sum_{r\in S_{i,t}}(X_{rj}/\tau)\psi_\tau(\epsilon_r)\right|, \quad (3)$$

where we use $\max_{r\in S_{i,t}}||X_r||_\infty \leq 1$ and X_{rj} 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

$$M_t := \exp\left(\sum_{r=1}^{t}\left[(X_{rj}/\tau)\psi_\tau(\epsilon_r) - \frac{\nu_\delta}{\tau^{1+\delta}}\right]\right).$$

Then

$$\begin{aligned}
&\mathbb{E}[M_t|\mathcal{F}_{t-1}] \\
&= M_{t-1}\mathbb{E}\left[\exp\left((X_{tj}/\tau)\psi_\tau(\epsilon_t)\right)|\mathcal{F}_{t-1}\right]/\exp\left(\frac{\nu_\delta}{\tau^{1+\delta}}\right) \\
&\leq M_{t-1}\mathbb{E}[1 + X_{tj}(\epsilon_t/\tau) + |\epsilon_t/\tau|^{1+\delta}|\mathcal{F}_{t-1}]/\exp\left(\frac{\nu_\delta}{\tau^{1+\delta}}\right) \\
&\leq M_{t-1}\left[1+\frac{\nu_\delta}{\tau^{1+\delta}}\right]/\exp\left(\frac{\nu_\delta}{\tau^{1+\delta}}\right) \\
&\leq M_{t-1}\exp\left(\frac{\nu_\delta}{\tau^{1+\delta}}\right)/\exp\left(\frac{\nu_\delta}{\tau^{1+\delta}}\right) \\
&= M_{t-1}
\end{aligned}$$

shows that M_t is a supermartingale with $M_0 := 1$, where the second inequality is derived from

$$-\log(1-u+|u|^{1+\delta}) \leq \frac{1}{\tau}\psi_\tau(\tau u) \leq \log(1+u+|u|^{1+\delta})$$

$\forall u \in \mathbb{R}$ and

$$\begin{aligned}
&\exp((X_{rj}/\tau)\psi_\tau(\epsilon_r)) \\
&\leq (1+\epsilon_r/\tau+|\epsilon_r/\tau|^{1+\delta})^{X_{rj}\mathbb{1}_{X_{rj}>0}} \\
&\cdot (1-\epsilon_r/\tau+|\epsilon_r/\tau|^{1+\delta})^{-X_{rj}\mathbb{1}_{X_{rj}<0}} \\
&\leq 1 + X_{rj}(\epsilon_r/\tau) + |\epsilon_r/\tau|^{1+\delta}.
\end{aligned}$$

Iteratively applying the law of total expectation on M_t gives $\mathbb{E}[M_t] \leq 1$ and hence,

$$\mathbb{E}\left[\exp\left(\sum_{r\in S_{i,t}}(X_{rj}/\tau)\psi_\tau(\epsilon_r)\right)\right] \leq \exp(|S_{i,t}|\nu_\delta\tau^{-1-\delta}).$$

Markov's inequality gives

$$\begin{aligned}
&\mathbb{P}\left(\sum_{r\in S_{i,t}}(X_{rj}/\tau)\psi_\tau(\epsilon_r) > \nu_\delta|S_{i,t}|z\right) \\
&\leq \exp(-\nu_\delta|S_{i,t}|z)\mathbb{E}\left[\exp\left(\sum_{r\in S_{i,t}}(X_{rj}/\tau)\psi_\tau(\epsilon_r)\right)\right] \\
&\leq \exp(\nu_\delta|S_{i,t}|(\tau^{-1-\delta}-z))
\end{aligned}$$

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. □

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 $\hat{\Sigma}(S_{i,t})$ 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 $\lambda_{\min}(\hat{\Sigma}(A')) \geq \gamma$ for some $\gamma > 0$, then*

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

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 $\mathbb{P}(A_t) \geq 1 - 2K\alpha/t^2$, 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 determinisitcally 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 5 which states that under the high-probability event A_t, the set $\mathcal{D}$ 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 $\mathcal{D} = \{i \in [K] | \max_{j\in[K]} X_t^T \hat{\beta}(T_{j,t-1}) - X_t^T \hat{\beta}(T_{i,t-1}) \leq \frac{h}{2}\}$ 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.

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

Observe that the random variables $\{x_r | r \in \mathcal{A}_{i,t}\}$ 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 $\lambda_{\min}(\hat{\Sigma}(\mathcal{A}_{i,t}))$ is strictly positive with high probability via Assumption 4 and Theorem 2.

Lemma 6 states that $\mathcal{A}_{i,t}$ is a subset of $S_{i,t}$ and Lemma 7 states that the size of $\mathcal{A}_{i,t}$ is large enough. Particularly, Lemma 7 guarantees that the size of $\mathcal{A}_{i,t}$ 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 $t \in \mathcal{A}_{i,t}$, then $a(t) = i$.*

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

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

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

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

for γ in Assumption 4. The size of the set $\mathcal{A}_{i,t}$ is guaranteed by Lemma 7. When $t \geq \frac{d}{\alpha}$ and $q \geq \frac{192}{\gamma p}$, with probability at least $1 - \frac{\alpha}{t^2}$,

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

and hence

$$\lambda_{\min}\left(\hat{\Sigma}(\mathcal{A}_{i,t})\right) \geq \frac{\gamma}{2}.$$

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

$$\lambda_{\min}\left(\hat{\Sigma}(S_{i,t})\right) > \frac{\gamma|\mathcal{A}_{i,t}|}{2|S_{i,t}|} > \frac{\gamma|\mathcal{A}_{i,t}|}{2t} > \frac{p\gamma}{8}. \quad (4)$$

Finally, we are ready to prove the Theorem 3.

proof of theorem 3. We consider two cases; (a) when $t < (Kq)^2/\phi$ or $t \in \cup_{i\in[K]} T_{i,t}$, (b) when $t \geq (Kq)^2/\phi$ and $t \notin \cup_{i\in[K]} T_{i,t}$. For (a), we know that $|\cup_{i\in[K]} T_{i,t}| \leq 2Kq \log T$ by definition of $T_{i,t}$. Hence, the time occupied by the event

(a) is at most $2Kq \log T + (Kq)^2/\phi$. We roughly bound the regret at this time steps using Assumption 2. For (b), due to Lemma 4, with probability at least $1 - 2\alpha K/t^2$, A_t holds. When A_t holds, due to lemma 5, we choose an arm from K_{opt} only. Then we can obtain minimum eigenvalue bound in (4) $\forall i \in \mathcal{D}$ with probability at least $1 - 2\alpha K/t^2$. The regret at time t is

$$\begin{aligned}
r_t &= X_t^T \beta_{a^*(t)} - X_t^T \beta_{a(t)} \\
&= X_t^T \beta_{a^*(t)} - X_t^T \hat{\beta}(S_{a(t),t}) + X_t^T \hat{\beta}(S_{a(t),t}) - X_t^T \beta_{a(t)} \\
&\leq X_t^T \beta_{a^*(t)} - X_t \hat{\beta}(S_{a^*(t),t}) + X_t^T \hat{\beta}(S_{a(t),t}) - X_t^T \beta_{a(t)} \\
&\leq ||X_t||_2 ||\beta_{a^*(t)} - \hat{\beta}(S_{a^*(t),t})||_2 \\
&+ ||X_t||_2 ||\beta_{a(t)} - \hat{\beta}(S_{a(t),t})||_2.
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{t=1}^{T} r_t \\
\leq& 2 \sum_{t=1}^{T} \left(\frac{4}{pt} \log(t^2(2d+1)/\alpha) \right)^{\delta/(1+\delta)} \frac{32\tau_0 d^{1/2}}{\gamma p} \\
\leq& \sum_{t=1}^{T} \left(\frac{4}{pt} \log(T^2(2d+1)/\alpha) \right)^{\delta/(1+\delta)} \frac{64\tau_0 d^{1/2}}{\gamma p} \\
\leq& (T^{\frac{1}{1+\delta}} - 1)(1+\delta) \left(\frac{4}{p} \log(T^2(2d+1)/\alpha) \right)^{\delta/(1+\delta)} \cdot \\
& \frac{64\tau_0 d^{1/2}}{\gamma p}.
\end{aligned}$$

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 $\tilde{O}(\log T)$ 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* $i \neq j$*,*

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

Theorem 4. *Suppose Assumptions 1-5 hold. When*

$$q \geq 6 \left(\frac{32(\tau_0 \vee 1) d^{1/2}}{h\gamma p} \right)^{\frac{1+\delta}{\delta}}, \ t \geq \frac{(Kq)^2}{\phi},$$
$$C_2 = 2^{14} \frac{\tau_0^2 C_0}{\gamma^2 p^3}, \ C_3 = 2^{12 + \frac{4\delta}{1+\delta}} \left(\frac{1+\delta}{1-\delta} \right) \frac{\tau_0^2 C_0}{\gamma^2 p^{\frac{2+4\delta}{1+\delta}}}$$

and $\tau_o \geq \nu_\delta$*, the expected regret is*

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[r_t] \\
&= \sum_{t}^{T} \mathbb{E}[X^T \beta_{a^*(t)} - X^T \beta_{a(t)}] \\
&\leq C_2 (\log T)(\log T + 1) dK + (\log T + 1) 12K(2d+1)
\end{aligned}$$

when $\delta = 1$ *and*

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

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 $\beta \in R^d$ 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 $\beta \in R^{dK}$ and context X_t into $X_{i,t} \in \mathbb{R}^{dK}, i \in [K]$ such that $X_{i,t}^T \beta = X_t^T \beta_i$. We randomly generate context $X_t \in \mathbb{R}^d$ 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 ϵ_t is generated from the student's t-distribution with a degree of freedom(df) $\{1.1, 1.5, 2\}$ and multiplied by 0.1 to balance with the norm condition of β_i and X_t. 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 $conf$ 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 $conf = 0.5, 1, 1.5, 2$ 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

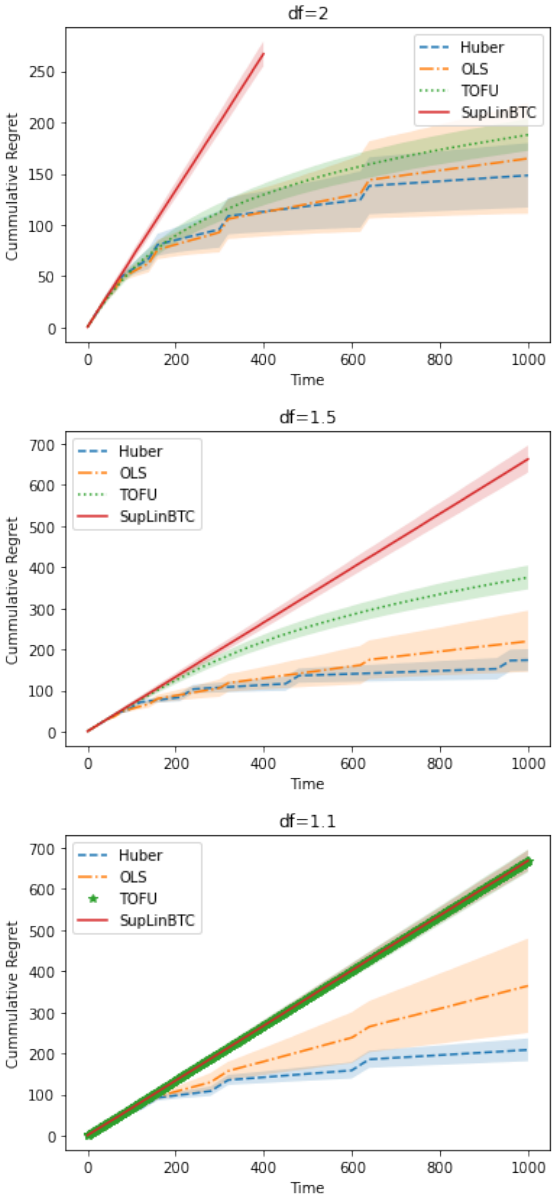

Figure 1: Cumulative regret of Huber bandit, OLS bandit, TOFU.

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 $O(\sqrt{d}T^{\frac{1}{1+\delta}}(\log dT)^{\frac{\delta}{1+\delta}})$ 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

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. 2004.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *International Conference on Machine Learning*, pages 9173–9200. PMLR, 2022.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

Marat Ibragimov, Rustam Ibragimov, and Johan Walden. *Heavy-tailed distributions and robustness in economics and finance*, volume 214. Springer, 2015.

Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. *Advances in Neural Information Processing Systems*, 33:8452–8462, 2020.

Huitian Lei, Yangyi Lu, Ambuj Tewari, and Susan A Murphy. An actor-critic contextual bandit algorithm for personalized mobile health interventions. *arXiv preprint arXiv:1706.09090*, 2017.

Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650. PMLR, 2016.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

Svetlozar Todorov Rachev. *Handbook of heavy tailed distributions in finance: Handbooks in finance, Book 1*. Elsevier, 2003.

Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

Milan Stehlík, Rastislav Potockỳ, Helmut Waldl, and Zdeněk Fabián. On the favorable estimation for fitting heavy tailed data. *Computational Statistics*, 25:485–503, 2010.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12: 389–434, 2012.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2): 199, 2015.

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. *arXiv preprint arXiv:2004.13465*, 2020.