Safe Online Convex Optimization with Heavy-Tailed Observation Noises

Yunhao Yang[1], Bo Xue[2], Yunzhi Hao[3,4], Ying Li,[5] Yuanyu Wan[1,4,6,*]

[1]School of Software Technology, Zhejiang University, Ningbo, China
[2]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[3]Big Graph Center & School of Computer and Computing Science, Hangzhou City University, Hangzhou, China
[4]State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, China
[5]Bangsheng Technology Co., Ltd., Hangzhou, China
[6]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security, Hangzhou, China

Abstract

We investigate safe online convex optimization (SOCO), where each decision must satisfy a set of unknown linear constraints. Assuming that the unknown constraints can be observed with a sub-Gaussian noise for each chosen decision, previous studies have established a high-probability regret bound of $O(T^{2/3})$. However, this assumption may not hold in many practical scenarios. To address this limitation, in this paper, we relax the assumption to allow any noise that admits finite $(1+\epsilon)$-th moments for some $\epsilon \in (0, 1]$, and propose two algorithms that enjoy an $O(T^{c_\epsilon})$ regret bound with high probability, where T is the time horizon and $c_\epsilon = (1+\epsilon)/(1+2\epsilon)$. The key idea of our two algorithms is to respectively utilize the median-of-means and truncation techniques to achieve accurate estimation under heavy-tailed noises. To the best of our knowledge, these are the first algorithms designed to handle SOCO with heavy-tailed observation noises.

Introduction

Online learning (Shalev-Shwartz et al. 2012) has received ever-increasing attention in recent years, due to its ability to efficiently handle applications with large-scale streaming data, such as online spam filtering, portfolio selection, and online recommendation. It is generally formulated as a repeated game between a player and an adversary. In each round t, the player must select a decision x_t from a set $\mathcal{X} \subseteq \mathbb{R}^d$, after which the adversary chooses a loss function $c_t(\cdot) : \mathcal{X} \mapsto \mathbb{R}$. The player will suffer a loss $c_t(x_t)$ in each round and aims to minimize the regret over total T rounds, which is defined as:

$$R(T) = \sum_{t=1}^{T} c_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} c_t(x). \tag{1}$$

To achieve this goal, online convex optimization (OCO), a special case of online learning with convex functions and convex decision sets, has been extensively studied over the past decades (Zinkevich 2003; Hazan, Agarwal, and Kale 2007; Shalev-Shwartz and Singer 2007; Hazan and Kale 2012; Wan and Zhang 2021; Wan, Tu, and Zhang 2020; Wan et al. 2024). It is well-known that several algorithms, such as online gradient descent (OGD) (Zinkevich 2003), have been proposed to achieve an optimal regret bound of $O(\sqrt{T})$ (Abernethy et al. 2008).

However, in many real-world applications, beyond the basic set $\mathcal{X}$, the decisions of the player must also satisfy additional safety constraints (Balasubramanian and Ghadimi 2018; Usmanova, Krause, and Kamgarpour 2019a; Fereydounian et al. 2020). For example, in communication networks, the maximum allowable radiated power constrains the transmission rate to ensure human safety (Luong et al. 2019). In robotics applications, the control actions must satisfy certain safety constraints to ensure the closed-loop stability of the system (Åström and Murray 2008; Ferraguti et al. 2022). Moreover, as in these two examples, the safety constraints are determined by some system parameters, which are typically unknown to the player, and thus limit the applicability of standard OCO algorithms.

Motivated by these safety requirements, Chaudhary and Kalathil (2022) recently consider safe online convex optimization (SOCO), a variant of OCO with a set of unknown linear safety constraints. Compared with the standard OCO, the new challenge is that the player needs to estimate the unknown parameters that characterize the safe set. To this end, they propose a new algorithm called safe online projected gradient descent (SO-PGD), which divides the total T rounds into a safe exploration phase for conservatively estimating the safe set and an exploitation phase for minimizing the regret under the estimated safety constraints. Under the assumption that the unknown safety constraints can be observed with sub-Gaussian noises for each chosen decision, SO-PGD achieves a regret bound of $O(T^{2/3})$ while satisfying the safety constraints in all rounds with high probability. However, in many practical scenarios such as extreme returns in financial market investments (Cont and Bouchaud 2000) and fluctuations in neural oscillations (Roberts, Boonstra, and Breakspear 2015), the observation noises are not sub-Gaussian but heavy-tailed, which undermines the theoretical guarantees of SO-PGD.

To address this issue, we investigate SOCO under heavy-tailed observation noises that admit only finite $(1 + \epsilon)$-th moments for some $\epsilon \in (0, 1]$, and develop two novel algorithms that enjoy an $O(T^{c_\epsilon})$ regret bound where $c_\epsilon = (1 + \epsilon)/(1 + 2\epsilon)$, while satisfying the safety constraints in all rounds with high probability. Specifically, to estimate the

*Corresponding author, wanyy@zju.edu.cn

unknown safety constraints under heavy-tailed noises, our first algorithm divides the exploration phase of SO-PGD into several groups, and takes the median of means of the estimations within these groups. For the same goal, our second algorithm refines SO-PGD by simply truncating the extreme values of the safety constraints received during the exploration phase. Note that both the median-of-means and truncation techniques have been extensively utilized to address bandits with heavy-tailed feedback—another paradigm of online learning (Shao et al. 2018; Lu et al. 2019; Xue et al. 2021, 2023). However, to the best of our knowledge, this is the first work to demonstrate their benefits in SOCO with heavy-tailed observation noises.

Related Work

This section reviews the related work on optimization with safety constraints and learning with heavy-tailed noises.

Optimization with Safety Constraints

Safety has garnered significant interest in the field of optimization, encompassing both offline and online settings (Sui et al. 2015; Amani, Alizadeh, and Thrampoulidis 2019; Usmanova, Krause, and Kamgarpour 2019a,b; Khezeli and Bitar 2020; Fereydounian et al. 2020; Chaudhary and Kalathil 2022). Specifically, in the online setting, previous studies mainly focus on linear bandits or Gaussian processes with unknown safety constraints (Sui et al. 2015; Amani, Alizadeh, and Thrampoulidis 2019; Khezeli and Bitar 2020). Their loss functions are either linear or can be modeled by some regularity assumptions. To address this limitation, Chaudhary and Kalathil (2022) consider SOCO with unknown linear safety constraints, where the loss functions can be arbitrarily chosen. They develop the first algorithm for SOCO, namely SO-PGD, by extending the classical OGD method (Zinkevich 2003). If the unknown safety constraints are observed with sub-Gaussian noises, SO-PGD can safely enjoy an $O(T^{2/3})$ regret bound with high probability. However, as previously discussed, this assumption may not hold, because the observation noise could be heavy-tailed in many practical scenarios.

Learning with Heavy-tailed Noises

Unlike SOCO, the heavy-tailed noise has already received considerable attention in previous studies on bandits (Liu and Zhao 2011; Bubeck, Cesa-Bianchi, and Lugosi 2013; Medina and Yang 2016; Shao et al. 2018; Lu et al. 2019; Cayci, Eryilmaz, and Srikant 2020; Xue et al. 2021, 2023; Gou, Yi, and Zhang 2023). The most related works to this paper are those on the heavy-tailed variant of stochastic linear bandits (Shao et al. 2018; Xue et al. 2021). Specifically, Shao et al. (2018) first consider a variant of stochastic linear bandits (Abbasi-Yadkori, Pál, and Szepesvári 2011) with heavy-tailed payoffs, in which the player selects an action $x_t \in \mathcal{X}$ and observes stochastic payoffs as $\theta_*^\top x_t + \omega_t$, where θ_* is an underlying parameter and ω_t is a random noise. Moreover, the noise distribution has finite moments of order $1+\epsilon$, where $\epsilon \in (0, 1]$. For this problem, Shao et al. (2018) develop two bandit algorithms by utilizing the median-of-means and truncation techniques respectively, which can enjoy nearly optimal regret bounds in terms of T. However, these regret bounds exhibit a linear dependence on the dimension d. Later, Xue et al. (2021) further investigate a heavy-tailed variant of stochastic linear bandits with finite arms (Chu et al. 2011), and propose two algorithms with regret bounds that are sublinear to the dimension d and nearly optimal in terms of T. Nonetheless, the primary techniques for handling heavy-tailed feedback remain the median-of-means and truncation. In this paper, we apply these two techniques to address SOCO with heavy-tailed observation noises.

Preliminaries

In this section, we first introduce necessary notations and the problem setting, and then recall the detailed procedures of the existing SO-PGD algorithm (Chaudhary and Kalathil 2022).

Notations

For any positive integer K, let $[K] = \{1, 2, \ldots, K\}$. For any two integers M_1 and M_2 satisfying $M_1 < M_2$, we express $[M_1, M_2] = \{M_1, M_1+1, \ldots, M_2\}$. For any random vector ζ, we define $Cov(\zeta) = \mathrm{E}[\zeta\zeta^\top]$. For any vector $x \in \mathbb{R}^d$, we use $x(i)$ to denote its i-th element, and use $|x|$, $\lceil x \rceil$, and $\lfloor x \rfloor$ to denote the corresponding element-wise operations. For any two vectors $x, y \in \mathbb{R}^d$, we use $x = y$, $x < y$, and $x > y$ to denote the generalized relationships. For any matrix A, we use A_i to denote its i-th column. If it is also positive semi-definite, we define $\|x\|_A = \sqrt{x^\top A x}$. For any convex set $\mathcal{X}$, we denote $\Pi_{\mathcal{X}}(x)$ as the projection of any vector x onto $\mathcal{X}$ with respect to the Euclidean norm, i.e., $\Pi_{\mathcal{X}}(x) = \mathrm{argmin}_{y\in\mathcal{X}} \|x - y\|$. Moreover, let $\mathbf{1}_{\{\cdot\}}$ represent an indicator function and $\mathrm{E}[X]$ denote the expectation of X.

Problem Setting

We investigate the SOCO problem with unknown linear constraints, which can be viewed as a repeated game between a player and an adversary. In each round t, the player selects an action x_t from a convex set $\mathcal{X}$ that is defined by m linear safety constraints:

$$\mathcal{X} = \{x \in \mathbb{R}^d : \theta^\top x \le b\},$$

with an unknown matrix $\theta \in \mathbb{R}^{d\times m}$ and a knowable vector $b \in \mathbb{R}^m$. Then, the adversary selects a convex function $c_t(\cdot) : \mathcal{X} \mapsto \mathbb{R}$, and the player suffers a loss $c_t(x_t)$. Simultaneously, the player observes the constraint feedback

$$y_t = \theta^\top x_t + \omega_t,$$

where ω_t represents a zero-mean random noise. The goal of SOCO is to select a sequence of actions to minimize the regret $R(T)$ defined in (1), while satisfying the safety constraints with high probability, i.e.,

$$\mathbb{P}(x_t \in \mathcal{X}, \forall t \in [T]) \ge (1 - \delta),$$

for some $\delta \in (0, 1)$.

Moreover, following Chaudhary and Kalathil (2022), we introduce some assumptions.

Assumption 1. *Each loss function $c_t(x)$ is convex and possesses a bounded gradient, i.e.,*

$$\max_{t\in[T]} \max_{x\in\mathcal{X}} \|\nabla c_t(x)\| \leq G.$$

Assumption 2. *Both the set $\mathcal{X}$ and the safety constraint parameters θ are bounded as specified below.*

- *$\mathcal{X}$ is convex, compact, and satisfies $\|x\| \leq L, \forall x \in \mathcal{X}$;*
- *θ satisfies $\max_{i\in[m]} \|\theta_i\| \leq L_\theta$.*

Assumption 3. *There exists a safe baseline action $x^s \in \mathcal{X}$ such that $\theta^\top x^s = b^s < b$. Moreover, the player knows x^s and b^s, which implies that the safety gap $\Delta^s = \min_{i\in[m]}(b(i) - b^s(i))$ is also knowable.*

Note that Assumptions 1 and 2 have been commonly utilized in previous studies on OCO. In contrast, Assumption 3 is specific to SOCO, which plays a critical role in enabling safe exploration during the initial phase of learning, as it is impossible to satisfy safety constraints without any prior in the beginning. In the following, we also assume that the observation noise is heavy-tailed, which is the key difference between this paper and the previous study of Chaudhary and Kalathil (2022).

Assumption 4. *The noise sequence $\{\omega_1, \omega_2, \ldots, \omega_T\}$ is heavy-tailed with respect to a filtration $\{F_1, F_2, \ldots, F_T\}$, which satisfies*

- $\mathrm{E}[\omega_t|F_{t-1}] = 0, \forall t \in [T]$;
- $\mathrm{E}[|y_t|^{1+\epsilon}|F_{t-1}] \leq q, \forall t \in [T]$;
- $\mathrm{E}[|y_t - \theta^\top x_t|^{1+\epsilon}|F_{t-1}] \leq c, \forall t \in [T]$;

where $\epsilon \in (0, 1]$, and $q, c \in (0, +\infty)$. With slight abuse of notations, here q and c denote uniform upper bounds for each dimension of the corresponding vectors.

SO-PGD

The detailed procedures of SO-PGD proposed by Chaudhary and Kalathil (2022) are outlined in Algorithm 1, where x_s is given by Assumption 3, and other inputs are parameters of this algorithm. Specifically, this algorithm can be divided into two phases. In the initial phase, it spends T_0 rounds to make a safe exploration, i.e., playing the following action:

$$x_t = (1-\gamma)x^s + \gamma\zeta_t,$$

at each round t, where x^s is a safe baseline, $\gamma \in [0, 1)$ denotes the exploration radius, and ζ_t is a random vector with zero mean such that

$$\|\zeta_t\| \leq \min\{1, L\}, Cov(\zeta_t) = \sigma^2 I, \tag{2}$$

for some constant σ. According to Assumptions 2 and 3, it is easy to verify that the above x_t satisfies the safety constraints if we set $\gamma = \Delta^s / L_\theta$ (see Lemma 1 of Chaudhary and Kalathil (2022) for details).

After the exploration phase, SO-PGD has collected the following information:

$$X_{T_0} = [x_1, x_2, \ldots, x_{T_0}]^\top, Y_{T_0} = [y_1, y_2, \ldots, y_{T_0}]^\top, \tag{3}$$

and estimates the unknown matrix θ via the ℓ_2-regularized least square, i.e., computing

$$\hat{\theta} = V_{T_0}^{-1} X_{T_0}^\top Y_{T_0}, \tag{4}$$

Algorithm 1: SO-PGD

Input: $x^s, \gamma, \eta, T_0, \delta, \lambda, \beta_{T_0}(\delta)$
1: **for** $t = 1, 2, \ldots, T_0$ **do**
2: $\quad x_t = (1-\gamma)x^s + \gamma\zeta_t$
3: **end for**
4: $V_{T_0} = \lambda I + \sum_{t=1}^{T_0} x_t x_t^\top$
5: $\hat{\theta} = V_{T_0}^{-1} X_{T_0}^\top Y_{T_0}$, where X_{T_0} and Y_{T_0} are defined in (3)
6: $C_i(\delta) = \{\tilde{\theta}_i \in \mathbb{R}^d : \|\tilde{\theta}_i - \hat{\theta}_i\|_{V_{T_0}} \leq \beta_{T_0}(\delta)\}$
7: $\hat{\mathcal{X}} = \{x \in \mathbb{R}^d : \tilde{\theta}_i^\top x \leq b(i), \forall \tilde{\theta}_i \in C_i(\delta), \forall i \in [m]\}$
8: **for** $t = T_0 + 1, T_0 + 2, \ldots, T$ **do**
9: $\quad x_{t+1} \leftarrow \Pi_{\hat{\mathcal{X}}}(x_t - \eta\nabla c_t(x_t))$
10: **end for**
Output: $\{x_1, \ldots, x_T\}$

where $V_{T_0} = \lambda I + \sum_{t=1}^{T_0} x_t x_t^\top$ and the parameter $\lambda > 0$ is introduced to make V_{T_0} invertible. Moreover, under the sub-Gaussian assumption on the observation noise, Chaudhary and Kalathil (2022) have shown that $\hat{\theta}$ has a confidence radius $\beta_{T_0}(\delta)$ such that

$$\mathbb{P}(\theta_i \in C_i(\delta), \forall i \in [m]) \geq 1 - \delta, \tag{5}$$

where $C_i(\delta) = \{\tilde{\theta}_i \in \mathbb{R}^d : \|\tilde{\theta}_i - \hat{\theta}_i\|_{V_{T_0}} \leq \beta_{T_0}(\delta)\}$. This result implies that the i-th column of the true parameter θ is contained within the set $C_i(\delta)$ with probability at least $1-\delta$. Thus, they construct a conservative estimation of the safety constraints as

$$\hat{\mathcal{X}} = \{x \in \mathbb{R}^d : \tilde{\theta}_i^\top x \leq b(i), \forall \tilde{\theta}_i \in C_i(\delta), \forall i \in [m]\}. \tag{6}$$

Then, SO-PGD proceeds to the second phase by simply performing OGD (Zinkevich 2003) over the set $\hat{\mathcal{X}}$, i.e.,

$$x_{t+1} \leftarrow \Pi_{\hat{\mathcal{X}}}(x_t - \eta\nabla c_t(x_t))$$

where η is the learning rate.

Main Results

In this section, we first revisit SO-PGD under heavy-tailed noises, and then introduce our two algorithms as well as the corresponding theoretical guarantees.

Revisiting SO-PGD

According to the analysis of Chaudhary and Kalathil (2022), the regret of SO-PGD can be divided into three parts, including the regret of the safe exploration phase, the regret of the inaccuracy of $\hat{\mathcal{X}}$, and the regret of OGD during the exploitation phase. As a result, the new challenge for handling heavy-tailed noises is how to estimate $\mathcal{X}$ accurately. A naive idea is to reuse the least square estimation $\hat{\theta}$ in (4), but redefine its confidence radius as:

$$\beta_{T_0}(\delta) = (3dcm/\delta)^{\frac{1}{1+\epsilon}} T_0^{\frac{1-\epsilon}{2(1+\epsilon)}} + \lambda^{\frac{1}{2}} L_\theta. \tag{7}$$

Note that this change ensures that (5) holds under the heavy-tailed assumption, and thus $\hat{\mathcal{X}}$ defined in (6) is still a conservative estimation of the safety constraints. In this way, we show that SO-PGD has the following regret bound even under the heavy-tailed assumption.

Theorem 1. *Suppose Assumptions 1, 2, 3, and 4 hold, and ζ_t satisfies* (2) *for any $t \in [T_0]$. If Algorithm 1 is run with $\gamma = \Delta^s/L_\theta$, $\eta = 2L/(G\sqrt{T})$, $\beta_{T_0}(\delta)$ in* (7)*, and $T_0 \geq \frac{8L^2}{\gamma^2\sigma^2}\left(\frac{\beta^2_{T_0}(\delta)}{(\Delta^s)^2} + \log\left(\frac{d}{\delta}\right)\right)$, then with probability at least $1-2\delta$, it ensures that $x_t \in \mathcal{X}$ for any $t \in [T]$, and*

$$R(T) \leq 2LGT_0 + 2LG\sqrt{T} + \frac{LG\sqrt{8d}\beta_{T_0}(\delta)}{C(\theta,b)\sqrt{\gamma^2\sigma^2}}\frac{T}{\sqrt{T_0}}, \quad (8)$$

where $C(\theta, b)$ is a positive constant that depends only on the matrix θ and vector b.

Remark. Recall that $c_\epsilon = (1+\epsilon)/(1+2\epsilon)$. By substituting (7) and $T_0 \approx T^{c_\epsilon}$ into (8), with probability at least $1-2\delta$, SO-PGD can achieve the following regret bound:

$$R(T) = O\left(\delta^{-\frac{1}{1+\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}}\right), \quad (9)$$

while satisfying the safety constraints. At first glance, compared with our desired regret bounds, this regret bound has the same dependence on T. However, we want to emphasize that it suffers a sublinear dependence on δ^{-1}, and thus does not hold with high probability.

Our Algorithm Based on Median of Means

To address the above limitation, we first develop a variant of SO-PGD by utilizing the median of means to improve the estimation of the safety constraints. The complete procedures are summarized in Algorithm 2, which is named **SO**-PGD with **M**edian of **M**eans (**SOMM**).

Note that the main idea of the median of means is to perform the ℓ_2-regularized least square estimation multiple times and select the median among these estimations. In this way, we can be less reliant on the accuracy of any single estimate. Specifically, to perform the median of means during the safe exploration phase, we first generate $N = T_0/k$ random actions, i.e., $\hat{x}_n = (1-\gamma)x^s + \gamma\zeta_n, \forall n \in [N]$, and repeat each random action k times to obtain k observations about the safety constraints, i.e., $\{y^1_n, \ldots, y^k_n\}$. Here, both k and T_0 are adjustable parameters, and we assume that $N = T_0/k$ is an integer without loss of generality. Then, we can calculate $V_N = \lambda I + \sum_{n=1}^N \hat{x}_n\hat{x}_n^\top$ and conduct the least square estimation for each sequence of observations to get k estimators, i.e.,

$$[\hat{\theta}_{1,j}, \hat{\theta}_{2,j}, \ldots, \hat{\theta}_{m,j}] = V_N^{-1}X_N^\top Y_N^j,$$

for each $j \in [k]$, where

$$X_N = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N]^\top, Y_N^j = [y_1^j, y_2^j, \ldots, y_N^j]^\top. \quad (10)$$

Moreover, we calculate the distance between these estimators, denoted as "means": $\|\hat{\theta}_{i,j} - \hat{\theta}_{i,s}\|_{V_N}$, and obtain the median of means $r_{i,j}$ for any $i \in [m]$ and $j \in [k]$.

Now, we are ready to refine the set $C_i(\delta)$ used in SO-PGD and the corresponding confidence radius. Let $i^* = \operatorname{argmin}_{j\in[k]} r_{i,j}$ denote the minimum of the median distance for each $i \in [m]$. To be precise, we can prove that (5) still holds under the heavy-tailed assumption by using a suitable $k = \lceil 24\log(m/\delta)\rceil$ and redefining

$$C_i(\delta) = \{\tilde{\theta}_i \in \mathbb{R}^d : \|\tilde{\theta}_i - \hat{\theta}_{i,i^*}\|_{V_N} \leq \beta_N(\delta)\}, \quad (11)$$

Algorithm 2: SOMM

Input: $x^s, \gamma, \eta, T_0, \delta, \lambda$
Initialization: $k = \lceil 24\log(m/\delta)\rceil$, $N = T_0/k$
for $n = 1, 2, \ldots, N$ **do**
  $\hat{x}_n = (1-\gamma)x^s + \gamma\zeta_n$, $\tau = 1$
  for $t = (n-1)*k+1, (n-1)*k+2, \ldots, nk$ **do**
    Play $x_t = \hat{x}_n$ and observe $y_n^\tau = y_t$
    Update $\tau = \tau + 1$
  end for
end for
Compute $V_N = \lambda I + \sum_{t=1}^N x_t x_t^\top$ and k estimators:

$$[\hat{\theta}_{1,j}, \hat{\theta}_{2,j}, \ldots, \hat{\theta}_{m,j}] = V_N^{-1}X_N^\top Y_N^j, \forall j \in [k],$$

where X_N and Y_N^j are defined in (10)
for $i = 1, 2, \ldots, m$ **do**
  for $j = 1, 2, \ldots, k$ **do**
    $r_{i,j} = \text{median of } \{\|\hat{\theta}_{i,j} - \hat{\theta}_{i,s}\|_{V_N} : s \in [k]\backslash\{j\}\}$
  end for
end for
$i^* = \operatorname{argmin}_{j\in[k]} r_{i,j}$ for all $i \in [m]$
$\beta_N(\delta) = 3((12dc)^{\frac{1}{1+\epsilon}} N^{\frac{1-\epsilon}{2(1+\epsilon)}} + \lambda^{\frac{1}{2}} L_\theta)$
$C_i(\delta) = \{\tilde{\theta}_i \in \mathbb{R}^d : \|\tilde{\theta}_i - \hat{\theta}_{i,i^*}\|_{V_N} \leq \beta_N(\delta)\}$
$\hat{\mathcal{X}} = \{x \in \mathbb{R}^d : \tilde{\theta}_i^\top x \leq b(i), \forall \tilde{\theta}_i \in C_i(\delta), \forall i \in [m]\}$
for $t = T_0+1, T_0+2, \ldots, T$ **do**
  $x_{t+1} = \Pi_{\hat{\mathcal{X}}}(x_t - \eta\nabla c_t(x_t))$
end for
Output: $\{x_1, ..., x_T\}$

where $\beta_N(\delta) = 3((12dc)^{\frac{1}{1+\epsilon}} N^{\frac{1-\epsilon}{2(1+\epsilon)}} + \lambda^{\frac{1}{2}} L_\theta)$, and L_θ, ϵ, c are given by Assumptions 2 and 4. Compared with the confidence radius in (7), here $\beta_N(\delta)$ has a tighter dependence on δ, which is critical for achieving the desired high-probability regret bound.

Next, following SO-PGD, we only need to construct $\hat{\mathcal{X}}$ based on $C_i(\delta)$ defined in (11), and perform OGD (Zinkevich 2003) over $\hat{\mathcal{X}}$ during the exploitation phase. By refining the original analysis of SO-PGD, we establish the following theoretical guarantee for our SOMM.

Theorem 2. *Suppose Assumptions 1, 2, 3, and 4 hold, and ζ_n satisfies* (2) *for any $n \in [N]$. If Algorithm 2 is run with $\gamma = \Delta^s/L_\theta$, $\eta = 2L/(G\sqrt{T})$, and $N \geq \frac{8L^2}{\gamma^2\sigma^2}\left(\frac{\beta^2_N(\delta)}{(\Delta^s)^2} + \log\left(\frac{d}{\delta}\right)\right)$, then with probability at least $1-2\delta$, it ensures that $x_t \in \mathcal{X}$ for any $t \in [T]$, and*

$$R(T) = O\left(T \cdot N^{\frac{-\epsilon}{1+\epsilon}} + T_0 + \sqrt{T}\right). \quad (12)$$

Remark. By substituting $T_0 \approx T^{c_\epsilon}$, $N = T_0/k$, and $k = \lceil 24\log(m/\delta)\rceil$ into (12), with probability at least $1-2\delta$, our SOMM can enjoy the following regret bound:

$$R(T) = O\left((\log(1/\delta))^{\frac{1}{1+\epsilon}} T^{\frac{1+\epsilon}{1+2\epsilon}}\right), \quad (13)$$

while satisfying the safety constraints. Compared to the regret bound in (9), it reduces the sublinear dependence on δ^{-1} to a polylogarithmic dependence, and thus can hold with high probability.

Our Algorithm Based on Truncated Mean

Furthermore, inspired by previous studies on heavy-tailed bandits (Shao et al. 2018; Xue et al. 2021), we also develop a variant of SO-PGD based on the truncated means. The detailed procedures are outlined in Algorithm 3, and the algorithm is named **SO**-PGD with **T**runcated **M**eans (**SOTM**).

Similar to SOMM, we only improve the safe exploration phase of SO-PGD. Moreover, compared to SOMM, a critical point of SOTM is to truncate the observed information during the exploration phase delicately. Intuitively, the truncation can filter the extreme value of heavy-tailed noises while keeping useful information for estimating the unknown parameters. To be precise, we introduce a truncation criterion $\hat{q}$, and set it as

$$\hat{q} = \left(\log\left(2dm/\delta\right)/q\right)^{-\frac{1}{1+\epsilon}} T_0^{\frac{1-\epsilon}{2(1+\epsilon)}} \tag{14}$$

where q is given by Assumption 4. Recall that X_{T_0} and Y_{T_0} defined in (3) have been collected. For each constraint $i \in [m]$ and dimension $j \in [d]$, SOTM individually truncates the collected information as

$$Y_{i,j}^{\dagger} = [y_1(i)\mathbf{1}_{|u_j(1)y_1(i)|\le\hat{q}}; \ldots; y_{T_0}(i)\mathbf{1}_{|u_j(T_0)y_{T_0}(i)|\le\hat{q}}], \tag{15}$$

where $[u_1, \ldots, u_d]^\top = V_{T_0}^{-1/2} X_{T_0}^\top$ and V_{T_0} follows the definition in Algorithm 1.

Then, for the i-th column of the unknown matrix θ, SOTM computes an estimator as

$$\hat{\theta}_i^{\dagger} = V_{T_0}^{-1/2}[u_1^\top Y_{i,1}^{\dagger}; \ldots; u_d^\top Y_{i,d}^{\dagger}].$$

Under the heavy-tailed assumption, we can prove that (5) also holds with the following set:

$$C_i(\delta) = \{\tilde{\theta}_i \in \mathbb{R}^d : \|\tilde{\theta}_i - \hat{\theta}_i^{\dagger}\|_{V_{T_0}} \le \beta_{T_0}(\delta)\}$$

and the confidence radius of

$$\beta_{T_0}(\delta) = 4\sqrt{d}q^{\frac{1}{1+\epsilon}} \left(\log\left(2dm/\delta\right)\right)^{\frac{\epsilon}{1+\epsilon}} T_0^{\frac{1-\epsilon}{2(1+\epsilon)}} + \lambda^{\frac{1}{2}} L_\theta. \tag{16}$$

As a result, we continue to construct a conservative estimation $\hat{\mathcal{X}}$ as in SO-PGD and SOMM, and then simply perform OGD (Zinkevich 2003) over $\hat{\mathcal{X}}$ for the exploitation phase.

By incorporating the property of the redefined $C_i(\delta)$ into the original analysis of SO-PGD, we establish the following theoretical guarantee for our SOTM.

Theorem 3. *Suppose Assumptions 1, 2, 3, and 4 hold, and ζ_t satisfies* (2) *for any $t \in [T_0]$. If Algorithm 3 is run with $\gamma = \Delta^s/L_\theta$, $\eta = 2L/(G\sqrt{T})$, and $T_0 \ge \frac{8L^2}{\gamma^2\sigma^2}\left(\frac{\beta_{T_0}^2(\delta)}{(\Delta^s)^2} + \log\left(\frac{d}{\delta}\right)\right)$, then with probability at least $1-2\delta$, it ensures that $x_t \in \mathcal{X}$ for any $t \in [T]$, and*

$$R(T) = O\left(T \cdot T_0^{\frac{-\epsilon}{1+\epsilon}} \left(\log\left(1/\delta\right)\right)^{\frac{\epsilon}{1+\epsilon}} + T_0\right). \tag{17}$$

Remark. Similar to SOMM, by substituting $T_0 \approx T^{c_\epsilon}$ into (17), with probability at least $1-2\delta$, our SOTM can enjoy the same regret bound as in (13), while satisfying the safety constraints. However, we want to emphasize that SOTM only needs to know the constant q in Assumption 4, rather than the constant c required by SOMM. Moreover, it is worth noting that the requirements of these algorithms on the minimum T_0 are also slightly different.

Algorithm 3: SOTM

Input: $x^s, \gamma, \eta, T_0, \delta, \lambda$

for $t = 1, 2, \ldots, T_0$ **do**
  $x_t = (1-\gamma)x^s + \gamma\zeta_t$
end for
$\hat{q} = \left(\log\left(2dm/\delta\right)/q\right)^{-\frac{1}{1+\epsilon}} T_0^{\frac{1-\epsilon}{2(1+\epsilon)}}$
$V_{T_0} = \lambda I + \sum_{t=1}^{T_0} x_t x_t^\top$
$[u_1, \ldots, u_d]^\top = V_{T_0}^{-1/2} X_{T_0}^\top$
Compute $Y_{i,j}^{\dagger}$ as in (15) for any $i \in [m]$ and $j \in [d]$
$\hat{\theta}_i^{\dagger} = V_{T_0}^{-1/2}[u_1^\top Y_{i,1}^{\dagger}; \ldots; u_d^\top Y_{i,d}^{\dagger}], \forall i \in [m]$
Compute $\beta_{T_0}(\delta)$ as in (16)
$C_i(\delta) = \{\tilde{\theta}_i \in \mathbb{R}^d : \|\tilde{\theta}_i - \hat{\theta}_i^{\dagger}\|_{V_{T_0}} \le \beta_{T_0}(\delta)\}$
$\hat{\mathcal{X}} = \{x \in \mathbb{R}^d : \tilde{\theta}_i^\top x \le b(i), \forall \tilde{\theta}_i \in C_i(\delta), \forall i \in [m]\}$
for $t = T_0+1, T_0+2, \ldots, T$ **do**
  $x_{t+1} = \Pi_{\hat{\mathcal{X}}}(x_t - \eta\nabla c_t(x_t))$
end for

Output: $\{x_1, \ldots, x_T\}$

Analysis

In this section, we prove Theorems 1, 2, and 3. The main novelty of our proofs lies in identifying an appropriate method for measuring the confidence interval of the estimated safety constraints under the heavy-tailed assumption.

Proof of Theorem 1

The safe set is defined by m linear constraints. Here, we examine a more stringent scenario, in which each constraint in the safe set is satisfied with probability at least $1 - \frac{\delta}{m}$. Inspired by Shao et al. (2018), we decompose X_{T_0} as

$$X_{T_0} = U\Sigma_{T_0}V^\top,$$

where U is a $T_0 \times d$ matrix with orthonormal columns, V is a $d \times d$ unitary matrix, and Σ_{T_0} is an $d \times d$ diagonal matrix with non-negative entries. Let $u_j^\top$ denote the j-th row of

$$V_{T_0}^{-1/2} X_{T_0}^\top = V(\Sigma_{T_0}^2 + \lambda I)^{-\frac{1}{2}}\Sigma_{T_0}U^\top. \tag{18}$$

Thus, we have $\|u_j\| \le 1$, and can bound the gap between the estimation $\hat{\theta}_i$ and θ_i as below

$$\begin{aligned}
&\|\hat{\theta}_i - \theta_i\|_{V_{T_0}} = \|V_{T_0}^{-1}X_{T_0}^\top Y_{T_0,i} - \theta_i\|_{V_{T_0}} \\
=&\|V_{T_0}^{-1}X_{T_0}^\top Y_{T_0,i} - (V_{T_0}^{-1}(X_{T_0}^\top X_{T_0} + \lambda I))\theta_i\|_{V_{T_0}} \\
=&\|V_{T_0}^{-1}X_{T_0}^\top(Y_{T_0,i} - X_{T_0}\theta_i) - \lambda V_{T_0}^{-1}\theta_i\|_{V_{T_0}} \\
\le&\|V_{T_0}^{-1/2}X_{T_0}^\top(Y_{T_0,i} - X_{T_0}\theta_i)\| + \lambda\|\theta_i\|_{V_{T_0}^{-1}} \\
\le&\sqrt{\sum_{j=1}^{d}\left(u_j^\top(Y_{T_0,i} - X_{T_0}\theta_i)\right)^2} + \lambda^{1/2}L_\theta
\end{aligned} \tag{19}$$

where the last inequality is due to (18) and Assumption 2.

To bound the first term of the right side of (19), we define

$$\psi_{j,\tau} = u_j(\tau)(y_\tau(i) - x_\tau^\top\theta_i),\ \phi_{j,\tau} = \psi_{j,\tau}\mathbf{1}_{|\psi_{j,\tau}|<\xi},$$

for some constant $\xi > 0$. It is not hard to verify that

$$\begin{aligned}
&\mathbb{P}\left(\sum_{j=1}^{d}\left(u_j^\top(Y_{T_0,i}-X_{T_0}\theta_i)\right)^2>\xi^2\right)\\
=&\mathbb{P}\left(\sum_{j=1}^{d}\left(\sum_{\tau=1}^{T_0}\psi_{j,\tau}\right)^2>\xi^2\right)\\
\leq&\mathbb{P}\left(\exists j,\tau,|\psi_{j,\tau}|>\xi\right)+\mathbb{P}\left(\sum_{j=1}^{d}\left(\sum_{\tau=1}^{T_0}\phi_{j,\tau}\right)^2>\xi^2\right)\\
\leq&\frac{2dcT_0^{\frac{1-\epsilon}{2}}}{\xi^{1+\epsilon}}+d\left(\frac{cT_0^{\frac{1-\epsilon}{2}}}{\xi^{1+\epsilon}}\right)^2
\end{aligned}\tag{20}$$

where the second inequality follows from (21) in Shao et al. (2018), and the last inequality follows from (23) and (24) in Shao et al. (2018) and depends on Assumption 4.

Then, to ensure that the true value θ_i lies in the set $C_i(\delta)$ estimated via $\hat{\theta}_i$ with probability at least $1-\frac{\delta}{m}$, we need to find a constant ξ satisfying

$$\frac{2dcT_0^{\frac{1-\epsilon}{2}}}{\xi^{1+\epsilon}}+d\left(\frac{cT_0^{\frac{1-\epsilon}{2}}}{\xi^{1+\epsilon}}\right)^2\leq\frac{\delta}{m}.$$

Notice that the above inequality can be derived from

$$\frac{T_0^{\frac{1-\epsilon}{2}}c}{\xi^{1+\epsilon}}+1\leq\sqrt{1+\frac{\delta}{md}}.\tag{21}$$

Moreover, due to the fact that $\delta<1$ and $md\geq 1$, (21) can be derived from

$$\frac{T_0^{\frac{1-\epsilon}{2}}c}{\xi^{1+\epsilon}}\leq\frac{\delta}{3dm},$$

which implies that we can set

$$\xi=\left(\frac{3dcm}{\delta}\right)^{\frac{1}{1+\epsilon}}T_0^{\frac{1-\epsilon}{2(1+\epsilon)}}.$$

Therefore, by substituting the above ξ into (20) and combining with (19), with a probability at least $1-\delta/m$, we have

$$\|\hat{\theta}_i-\theta_i\|_{V_{T_0}}\leq\left(\frac{3dcm}{\delta}\right)^{\frac{1}{1+\epsilon}}T_0^{\frac{1-\epsilon}{2(1+\epsilon)}}+\lambda^{1/2}L_\theta=\beta_{T_0}(\delta).\tag{22}$$

Then, we can derive (5) by using the union bound. From this result, it is easy to verify that $\hat{\mathcal{X}}$ constructed in Algorithm 1 satisfies

$$\mathbb{P}(\hat{\mathcal{X}}\subseteq\mathcal{X})\geq 1-\delta.$$

Note that Algorithm 1 ensures that $x_t\in\mathcal{X}$ for any $t\in[T_0]$ and $x_t\in\hat{\mathcal{X}}$ for any $t\in[T_0+1,T]$. By combining with the above result, with a probability at least $1-\delta$, we have $x_t\in\mathcal{X}$ for any $t\in[T]$.

Now, we are ready to prove the regret bound for Algorithm 1. According to the two phases used in Algorithm 1, the regret can be rewritten as

$$R(T)=\sum_{t=1}^{T_0}\left(c_t(x_t)-c_t(x^*)\right)+\sum_{t=T_0+1}^{T}\left(c_t(x_t)-c_t(x^*)\right)$$

where $x^*\in\operatorname{argmin}_{x\in\mathcal{X}}\sum_{t=1}^{T}c_t(x)$.

Under Assumptions 1 and 2, we can bound the first term as follows

$$\sum_{t=1}^{T_0}c_t(x_t)-c_t(x^*)\leq\sum_{t=1}^{T_0}G\|x_t-x^*\|\leq 2LGT_0.\tag{23}$$

Next, we relax the second term as

$$\begin{aligned}
&\sum_{t=T_0+1}^{T}\left(c_t(x_t)-c_t(x^*)\right)\\
=&\sum_{t=T_0+1}^{T}\left(c_t(x_t)-c_t(\hat{x}^*)\right)+\sum_{t=T_0+1}^{T}\left(c_t(\hat{x}^*)-c_t(x^*)\right),
\end{aligned}$$

where $\hat{x}^*\in\operatorname{argmin}_{x\in\hat{\mathcal{X}}}\sum_{t=1}^{T}c_t(x)$.

If $\theta_i\in C_i(\delta),\forall i\in[m]$, by adopting the standard OGD analysis (Hazan 2016) with Assumptions 1 and 2, we have

$$\sum_{t=T_0+1}^{T}\left(c_t(x_t)-c_t(\hat{x}^*)\right)\leq 2LGT^{1/2}.\tag{24}$$

Then, according to the proof of Proposition 2 in Chaudhary and Kalathil (2022), if $\theta_i\in C_i(\delta),\forall i\in[m]$, ζ_t satisfies (2) for any $t\in[T_0]$, and $T_0\geq\frac{8L^2}{\gamma^2\sigma^2}\left(\frac{\beta_{T_0}^2(\delta)}{(\Delta^s)^2}+\log\left(\frac{d}{\delta}\right)\right)$, with probability at least $1-\delta$, we have

$$\sum_{t=T_0+1}^{T}\left(c_t(\hat{x}^*)-c_t(x^*)\right)\leq\frac{LG\sqrt{8d}\beta_{T_0}(\delta)}{C(\theta,b)\sqrt{\gamma^2\sigma^2}}\frac{T}{\sqrt{T_0}}.\tag{25}$$

By combining (5), (23), (24), (25), with probability at least $1-2\delta$, we finally get the following regret bound:

$$R(T)\leq 2LGT_0+2LG\sqrt{T}+\frac{LG\sqrt{8d}\beta_{T_0}(\delta)}{C(\theta,b)\sqrt{\gamma^2\sigma^2}}\frac{T}{\sqrt{T_0}}.\tag{26}$$

Proof of Theorem 2

We start this proof by introducing a lemma from Shao et al. (2018), which will be used to determine a proper confidence radius of the median estimation.

Lemma 1. *(Lemma 3 of Shao et al. (2018)) Recall $\hat{\theta}_{i,j},\theta_{i,i^*}$ and V_N in Algorithm 2. If there exists a $\xi>0$, such that*

$$\mathbb{P}\left(\|\hat{\theta}_{i,j}-\theta_i\|_{V_N}\leq\xi\right)\geq\frac{3}{4}$$

holds for all $j\in[k]$, then it holds that

$$\|\hat{\theta}_{i,i^*}-\theta_i\|_{V_N}\leq 3\xi$$

with probability at least $1-e^{-\frac{k}{24}}$.

Note that $\hat{\theta}_{i,j}$ in Algorithm 2 is equivalent to $\hat{\theta}_i$ generated by Algorithm 1 with $T_0=N$. Therefore, by combining (22) with $\frac{\delta}{m}=\frac{1}{4}$ and $T_0=N$, it is easy to verify that we have

$$\mathbb{P}\left(\|\hat{\theta}_{i,j}-\theta_i\|_{V_N}\leq\xi\right)\geq\frac{3}{4}$$

for $\xi = (12dc)^{\frac{1}{1+\epsilon}} N^{\frac{1-\epsilon}{2(1+\epsilon)}} + \lambda^{\frac{1}{2}} L_\theta$. By combining the above result with Lemma 1 and setting $k = \lceil 24 \log(m/\delta) \rceil$, with probability at least $1-\delta$, we have

$$\|\hat{\theta}_{i,i^*} - \theta_i\|_{V_N} \leq 3\xi = \beta_N(\delta), \forall i \in [m].$$

In other words, we have proved that (5) holds with $C_i(\delta)$ defined in Algorithm 2.

Next, following the proof of Theorem 1, it is easy to verify that Algorithm 2 ensures $x_t \in \mathcal{X}$ for any $t \in [T]$ with a probability at least $1-\delta$. Moreover, similar to (26), it is not hard to verify that if ζ_n satisfies (2) for any $n \in [N]$, and $N \geq \frac{8L^2}{\gamma^2\sigma^2}\left(\frac{\beta_N^2(\delta)}{(\Delta^s)^2} + \log\left(\frac{d}{\delta}\right)\right)$, with a probability at least $1-2\delta$, Algorithm 2 has

$$R(T) \leq 2LGT_0 + 2LGT^{1/2} + \frac{LG\sqrt{8d}\beta_N(\delta)}{C(\theta,b)\sqrt{\gamma^2\sigma^2}} \frac{T}{\sqrt{N}},$$

which can be further simplified as

$$R(T) = O\left(T_0 + \sqrt{T} + T \cdot N^{\frac{-\epsilon}{1+\epsilon}}\right).$$

Proof of Theorem 3

We start this proof by demonstrating that the confidence radius defined in (16) is appropriate for $C_i(\delta)$ defined in Algorithm 3. First, similar to (19) in the proof of Theorem 1, we have

$$\|\hat{\theta}_i^\dagger - \theta_i\|_{V_{T_0}} \leq \sqrt{\sum_{j=1}^{d} \left(u_j^\top (Y_{i,j}^\dagger - X_{T_0}\theta_i)\right)^2} + \lambda^{\frac{1}{2}} L_\theta. \tag{27}$$

To bound the first term on the right side of above inequality, we notice that

$$\begin{aligned} u_j^\top (Y_{i,j}^\dagger - X_{T_0}\theta_i) &= \sum_{\tau=1}^{T_0} u_j(\tau)(Y_{i,j}^\dagger(\tau) - \mathrm{E}[y_\tau(i)]) \\ &= \sum_{\tau=1}^{T_0} u_j(\tau)\left(Y_{i,j}^\dagger(\tau) - \mathrm{E}[Y_{i,j}^\dagger(\tau)]\right) \\ &\quad - \sum_{\tau=1}^{T_0} u_j(\tau)\mathrm{E}[y_\tau(i)\mathbf{1}_{|u_j(\tau)y_\tau(i)|>\hat{q}}] \end{aligned} \tag{28}$$

where the last equality is due to (15).

Then, based on Bernstein's inequality (Seldin et al. 2012), with probability at least $1 - \frac{\delta}{md}$, we have

$$\begin{aligned} &\left|\sum_{\tau=1}^{T_0} u_j(\tau)\left(Y_{i,j}^\dagger(\tau) - \mathrm{E}[Y_{i,j}^\dagger(\tau)]\right)\right| \leq 2\hat{q}\log\left(\frac{2dm}{\delta}\right) \\ &+ \underbrace{\frac{1}{2\hat{q}}\sum_{\tau=1}^{T_0} \mathrm{E}\left[u_j(\tau)^2\left(Y_{i,j}^\dagger(\tau) - \mathrm{E}[Y_{i,j}^\dagger(\tau)]\right)^2\right]}_{:=A}. \end{aligned} \tag{29}$$

Following (35) in Shao et al. (2018), we can utilize Assumption 4 to derive an upper bound on the term A:

$$A \leq \frac{\sum_{\tau=1}^{T_0} |u_j(\tau)|^{1+\epsilon} q}{2\hat{q}^\epsilon} \tag{30}$$

and an upper bound on the last term on the right side of (28):

$$\sum_{\tau=1}^{T_0} \left|u_j(\tau)\mathrm{E}[y_\tau(i)\mathbf{1}_{|u_j(\tau)y_\tau(i)|>\hat{q}}]\right| \leq \frac{\sum_{\tau=1}^{T_0} |u_j(\tau)|^{1+\epsilon} q}{\hat{q}^\epsilon}. \tag{31}$$

Then, by combining (28) with (29), (30) and (31), with probability at least $1 - \frac{\delta}{md}$, we have

$$\begin{aligned} &u_j^\top (Y_{i,j}^\dagger - X_{T_0}\theta_i) \\ \leq &\frac{\sum_{\tau=1}^{T_0} |u_j(\tau)|^{1+\epsilon} q}{2\hat{q}^\epsilon} + \frac{\sum_{\tau=1}^{T_0} |u_j(\tau)|^{1+\epsilon} q}{\hat{q}^\epsilon} \\ &+ 2\hat{q}\log\left(\frac{2dm}{\delta}\right) \\ \leq &4q^{\frac{1}{1+\epsilon}}\left(\log\left(\frac{2dm}{\delta}\right)\right)^{\frac{\epsilon}{1+\epsilon}} T_0^{\frac{1-\epsilon}{2(1+\epsilon)}} \end{aligned}$$

where the last inequality is due to the definition of $\hat{q}$ in (14) and the fact $\sum_{\tau=1}^{T_0} |u_j(\tau)|^{1+\epsilon} \leq T_0^{\frac{1-\epsilon}{2}}$.

By further combining the above result with (27), we can verify that (5) holds with $C_i(\delta)$ defined in Algorithm 3. Following the proof of Theorem 1, it is easy to verify that Algorithm 3 ensures $x_t \in \mathcal{X}$ for any $t \in [T]$ with a probability at least $1-\delta$. Finally, if ζ_t satisfies (2) for any $t \in [T_0]$, and $T_0 \geq \frac{8L^2}{\gamma^2\sigma^2}\left(\frac{\beta_{T_0}^2(\delta)}{(\Delta^s)^2} + \log\left(\frac{d}{\delta}\right)\right)$, it is not hard to verify that (26) also holds with probability at least $1-2\delta$, which can be further simplified as

$$R(T) = O\left(T_0 + \sqrt{T} + T \cdot T_0^{\frac{-\epsilon}{1+\epsilon}} \left(\log\left(1/\delta\right)\right)^{\frac{\epsilon}{1+\epsilon}}\right)$$

due to the definition of $\beta_{T_0}(\delta)$ in (16).

Conclusion and Future Work

In this work, we study safe online convex optimization with heavy-tailed observation noises. We first revisit SO-PGD—an existing algorithm developed for sub-Gaussian noises, but fail to derive a high-probability regret bound under the heavy-tailed assumption. To tackle this limitation, we propose two heavy-tailed variants of SO-PGD, namely SOMM and SOTM, by combining it with the median-of-means and truncation techniques, respectively. Both of them can enjoy a regret bound of $O(T^{c_\epsilon})$ while satisfying the unknown safety constraints with high probability. One potential limitation of our algorithms is that some prior information on the heavy-tailed noise is required. Note that recent advances in heavy-tailed bandits (Huang, Da, and Huang 2022; Genalti et al. 2024) have proposed adaptive algorithms for the challenging case without prior information. Thus, an open problem is whether their techniques can be applied to make our algorithms adaptive.

Acknowledgments

This work was partially supported by the Pioneer R&D Program of Zhejiang (No.2024C01021), and the National Natural Science Foundation of China (62306275). The authors would like to thank the anonymous reviewers for their helpful comments.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. In *Advances in neural information processing systems 24*, 2312–2320.

Abernethy, J.; Bartlett, P.; Rakhlin, A.; and Tewari, A. 2008. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the 19th Annual Conference on Computational Learning Theory*, 415–429.

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2019. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems 32*, 12544–12554.

Åström, K. J.; and Murray, R. M. 2008. *Feedback systems: An introduction for scientists and engineers*. Princeton University Press.

Balasubramanian, K.; and Ghadimi, S. 2018. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Advances in Neural Information Processing Systems 31*, 3455–3464.

Bubeck, S.; Cesa-Bianchi, N.; and Lugosi, G. 2013. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11): 7711–7717.

Cayci, S.; Eryilmaz, A.; and Srikant, R. 2020. Budget-Constrained Bandits over General Cost and Reward Distributions. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 4388–4398.

Chaudhary, S.; and Kalathil, D. 2022. Safe online convex optimization with unknown linear safety constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6175–6182.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 208–214.

Cont, R.; and Bouchaud, J.-P. 2000. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic dynamics*, 4(2): 170–196.

Fereydounian, M.; Shen, Z.; Mokhtari, A.; Karbasi, A.; and Hassani, H. 2020. Safe learning under uncertain objectives and constraints. *arXiv preprint arXiv:2006.13326*.

Ferraguti, F.; Landi, C. T.; Singletary, A.; Lin, H.-C.; Ames, A.; and Secchi, C. 2022. Safety and efficiency in robotics: The control barrier functions approach. *IEEE Robotics & Automation Magazine*, 29(3): 139–151.

Genalti, G.; Marsigli, L.; Gatti, N.; and Metelli, A. M. 2024. (ϵ, u)-Adaptive regret minimization in heavy-tailed bandits. In *Proceedings of 37th Conference on Learning Theory*, 1882–1915.

Gou, Y.; Yi, J.; and Zhang, L. 2023. Stochastic graphical bandits with heavy-tailed rewards. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, 734–744.

Hazan, E. 2016. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 157–325.

Hazan, E.; Agarwal, A.; and Kale, S. 2007. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69: 169–192.

Hazan, E.; and Kale, S. 2012. Projection-free online learning. In *Proceedings of the 29th International Conference on Machine Learning*, 1843–1850.

Huang, J.; Da, Y.; and Huang, L. 2022. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *Proceedings of the 39th International Conference on Machine Learning*, 9173–9200.

Khezeli, K.; and Bitar, E. 2020. Safe linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10202–10209.

Liu, K.; and Zhao, Q. 2011. Multi-armed bandit problems with heavy-tailed reward distributions. In *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*, 485–492.

Lu, S.; Wang, G.; Hu, Y.; and Zhang, L. 2019. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *Proceedings of the 36th International Conference on Machine Learning*, 4154–4163.

Luong, N. C.; Hoang, D. T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.-C.; and Kim, D. I. 2019. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials*, 21(4): 3133–3174.

Medina, A. M.; and Yang, S. 2016. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 1642–1650.

Roberts, J. A.; Boonstra, T. W.; and Breakspear, M. 2015. The heavy tail of the human brain. *Current opinion in neurobiology*, 31: 164–172.

Seldin, Y.; Laviolette, F.; Cesa-Bianchi, N.; Shawe-Taylor, J.; and Auer, P. 2012. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12): 7086–7093.

Shalev-Shwartz, S.; and Singer, Y. 2007. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69: 115–142.

Shalev-Shwartz, S.; et al. 2012. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2): 107–194.

Shao, H.; Yu, X.; King, I.; and Lyu, M. R. 2018. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems 31*, 2125–2135.

Sui, Y.; Gotovos, A.; Burdick, J. W.; and Krause, A. 2015. Safe Exploration for Optimization with Gaussian Processes. In *Proceedings of the 32nd International Conference on Machine Learning*, 997–1005.

Usmanova, I.; Krause, A.; and Kamgarpour, M. 2019a. Log barriers for safe non-convex black-box optimization. *arXiv preprint arXiv:1912.09478*.

Usmanova, I.; Krause, A.; and Kamgarpour, M. 2019b. Safe convex learning under uncertain constraints. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2106–2114.

Wan, Y.; Tu, W.-W.; and Zhang, L. 2020. Projection-free distributed online convex optimization with $O(\sqrt{T})$ communication complexity. In *Proceedings of the 37th International Conference on Machine Learning*, 9818–9828.

Wan, Y.; Wei, T.; Song, M.; and Zhang, L. 2024. Nearly optimal regret for decentralized online convex optimization. In *Proceedings of the 37th Annual Conference on Learning Theory*, 4862–4888.

Wan, Y.; and Zhang, L. 2021. Projection-free online learning over strongly convex sets. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 10076–10084.

Xue, B.; Wang, G.; Wang, Y.; and Zhang, L. 2021. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, 2936–2942.

Xue, B.; Wang, Y.; Wan, Y.; Yi, J.; and Zhang, L. 2023. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems 36*, 70880–70891.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning*, 928–936.