

# Heavy-Tailed Linear Bandits: Huber Regression with One-Pass Update

Jing Wang<sup>1,2</sup>, Yu-Jie Zhang<sup>3</sup>, Peng Zhao<sup>1,2</sup>, Zhi-Hua Zhou<sup>1,2</sup>

<sup>1</sup> National Key Laboratory for Novel Software Technology, Nanjing University, China

<sup>2</sup> School of Artificial Intelligence, Nanjing University, China

<sup>3</sup> Center for Advanced Intelligence Project, RIKEN, Japan

## Abstract

We study the stochastic linear bandits with heavy-tailed noise. Two principled strategies for handling heavy-tailed noise, truncation and median-of-means, have been introduced to heavy-tailed bandits. Nonetheless, these methods rely on specific noise assumptions or bandit structures, limiting their applicability to general settings. The recent work [Huang et al., 2024] develops a soft truncation method via the adaptive Huber regression to address these limitations. However, their method suffers undesired computational cost: it requires storing all historical data and performing a full pass over these data at each round. In this paper, we propose a *one-pass* algorithm based on the online mirror descent framework. Our method updates using only current data at each round, reducing the per-round computational cost from  $\mathcal{O}(t \log T)$  to  $\mathcal{O}(1)$  with respect to current round  $t$  and the time horizon  $T$ , and achieves a near-optimal and variance-aware regret of order  $\tilde{\mathcal{O}}(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}})$  where  $d$  is the dimension and  $\nu_t^{1+\varepsilon}$  is the  $(1+\varepsilon)$ -th central moment of reward at round  $t$ .

## 1 Introduction

Stochastic Linear Bandits (SLB) models sequential decision-making process with linear structured reward distributions, which has been extensively studied in the literature [Dani et al., 2008, Abbasi-Yadkori et al., 2011, Li et al., 2021]. Specifically, in general SLB the observed reward at time  $t$  is the inner product of the arm’s feature vector  $X_t$  and an unknown parameter  $\theta_*$ , corrupted by sub-Gaussian noise  $\eta_t$ , namely,

$$r_t = X_t^\top \theta_* + \eta_t. \quad (1)$$

However, in many real-world scenarios, the noise in data often follows a heavy-tailed distribution, such as in financial markets [Cont and Bouchaud, 2000, Hull, 2012] and online advertising [Choi et al., 2020, Jebarajakirthi et al., 2021]. This motivates the study of heavy-tailed linear bandits (HvtLB) model [Medina and Yang, 2016, Shao et al., 2018, Xue et al., 2020, Huang et al., 2024], where the noise  $\eta_t$  in (1) satisfies that for some  $\varepsilon \in (0, 1]$ ,  $\nu > 0$ ,

$$\mathbb{E}[\eta_t] = 0, \quad \mathbb{E}[|\eta_t|^{1+\varepsilon}] \leq \nu^{1+\varepsilon}. \quad (2)$$

To address heavy-tailed noise, statistical methods in estimation and regression often rely on two key principles: truncation and median-of-means (MOM) [Lugosi and Mendelson, 2019]. Truncation

Table 1: Comparisons of our regret bounds and computational complexity to previous best-known results for heavy-tailed linear bandits. For the regret, the logarithmic dependence over  $T$  is hidden by  $\tilde{\mathcal{O}}(\cdot)$ -notation;  $d$  is the dimension,  $\nu_t$  is the moment bound. For the computational cost, we only keep the dependence on the time step  $t$  and overall time step  $T$ .

| Method     | Algorithm                       | Regret  | Comp. cost              | Remark   |
|------------|---------------------------------|---|-------------------------|--|
| MOM        | MENU [Shao et al., 2018]        | $\tilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right)$  | $\mathcal{O}(\log T)$   | fixed arm set and repeated pulling   |
|            | CRMM [Xue et al., 2023]         |   | $\mathcal{O}(1)$        |  |
| Truncation | TOFU [Shao et al., 2018]        | $\tilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right)$  | $\mathcal{O}(t)$        | absolute moment<br>$\mathbb{E}[ r_t ^{1+\varepsilon} \mid \mathcal{F}_{t-1}] \leq u$   |
|            | CRTM [Xue et al., 2023]         |   | $\mathcal{O}(1)$        |  |
| Huber      | HEAVY-OFUL [Huang et al., 2024] | $\tilde{\mathcal{O}}\left(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\right)$ | $\mathcal{O}(t \log T)$ | instance-dependent bound   |
| Huber      | Hvt-UCB (Corollary 1)           | $\tilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right)$  | $\mathcal{O}(1)$        | $\mathbb{E}[ \eta_t ^{1+\varepsilon} \mid \mathcal{F}_{t-1}] \leq \nu^{1+\varepsilon}$ |
| Huber      | Hvt-UCB (Theorem 1)             | $\tilde{\mathcal{O}}\left(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\right)$ | $\mathcal{O}(1)$        | instance-dependent bound   |

methods handle outliers by directly removing extreme data points. In the case of soft version truncation, such as Catoni’s M-estimator [Catoni, 2012], which reduces the impact of outliers by assigning them lower weights. This ensures robust estimates while maintaining potentially valuable information from extreme data points. Median-of-means (MOM) methods take a different approach by dividing the dataset into several groups, calculating the mean within each group, and then using the median of these group means as the final estimate. This method limits the influence of outliers to only a few groups, preventing them from affecting the overall dataset.

For HvtLB, Shao et al. [2018] constructed algorithms using truncation and MOM-based least squares approach for parameter estimation, and achieved an  $\tilde{\mathcal{O}}(dT^{\frac{1}{1+\varepsilon}})$  regret bound which is proven to be optimal. However, these approaches have some notable issues: truncation methods require the assumption of bounded absolute moments for rewards, which cannot vanish in deterministic case, making them suboptimal in noiseless scenarios; MOM methods rely on repeated arm pulling and assumption of a fixed arm set, which is only feasible in bandit settings. These limitations reveal that methods based on truncation and MOM heavily depend on specific assumptions or the bandit structure, making them challenging to adapt to broader scenarios, such as online MDPs and adaptive control. A more detailed discussion of these challenges is provided in Section 5.

Recently, Sun et al. [2020] made remarkable progress in the study of heavy-tailed statistics by proposing adaptive Huber regression, which leverages the Huber loss [Huber, 1964] to achieve non-asymptotic guarantees. Li and Sun [2023] further applied this technique to SLB, achieving an optimal and variance-aware regret bound under bounded variance conditions ( $\varepsilon = 1$ ). Subsequently, Huang et al. [2024] extended the adaptive Huber regression to handle heavy-tailed scenarios ( $\varepsilon \in (0, 1]$ ), achieving an optimal and instance-dependent regret bound of  $\tilde{\mathcal{O}}\left(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\right)$  for HvtLB. Notably, the adaptive Huber regression does not rely on additional noise assumptions or the repeated pulling of bandit settings. This generality allows the approach to be further adapted to other decision-making scenarios, such as reinforcement learning with function approximation.

In each round  $t$ , the adaptive Huber regression has to solve the following optimization problem:

$$\hat{\theta}_t = \arg \min_{\theta \in \Theta} \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} \ell_s(\theta), \quad (3)$$

where  $\Theta$  is the feasible set of the parameter,  $\lambda$  is the regularization parameter, and  $\ell_s(\theta)$  represents

the Huber loss at round  $s$ . Unlike the OFUL algorithm [Abbasi-Yadkori et al., 2011] in SLB with sub-Gaussian noise, which uses least squares (LS) for estimation and can update recursively based on the closed-form solution, the Huber loss is partially quadratic and partially linear. Solving (3) requires storing all historical data and performing a full pass over all historical data at each round  $t$  to update the parameter estimation. This results in a per-round storage cost of  $\mathcal{O}(t)$  and computational cost of  $\mathcal{O}(t \log T)$  with respect to the current round  $t$  and the time horizon  $T$ , making it computationally infeasible for large-scale or real-time applications. Thus, we need a *one-pass* algorithm for HvtLB that is as efficient as OFUL in SLB with sub-Gaussian noise, updating the estimation at each round  $t$  using only the current data, without storing historical data. While one-pass algorithms based on the truncation and MOM approach have been proposed by Xue et al. [2023], but as noted earlier, these methods heavily rely on specific assumptions or the bandit structure. Given the attractive properties of Huber loss-based method [Huang et al., 2024], a natural question arises: *Is it possible to design a one-pass Huber-loss-based algorithm that achieves optimal regret?*

**Our Results.** In this work, we propose a Huber loss-based one-pass algorithm Hvt-UCB utilizing the Online Mirror Descent (OMD) framework, which achieves the optimal regret bound of  $\tilde{\mathcal{O}}(dT^{\frac{1}{1+\varepsilon}})$  while eliminating the need to store historical data. Additionally, our algorithm can further achieve the instance-dependent regret bound of  $\tilde{\mathcal{O}}(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2 + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}})$ , where  $\nu_t^{1+\varepsilon}$  is the  $(1+\varepsilon)$ -th central moment of reward at round  $t$ . Our approach preserves both optimal and instance-dependent regret guarantees achieved by Huang et al. [2024] while significantly reducing the per-round computational cost from  $\mathcal{O}(t \log T)$  to  $\mathcal{O}(1)$  with respect to the current round  $t$  and the time horizon  $T$ . Moreover, our algorithm does not rely on any additional assumptions, making it more broadly applicable. A comprehensive comparison of our results and previous works on heavy-tailed linear bandits is presented in Table 1.

**Techniques.** Our main technical contribution is adapting the OMD framework, originally developed for regret minimization in online optimization, to parameter estimation in HvtLB. This adaptation requires linking the parameter gap (estimation error) with the loss value gap (regret) and involves two key components: (i) *handling heavy-tailed noise*: Zhang et al. [2016] have adopted a variant of Online Newton Step (ONS) to Logistic Bandits settings, but their method cannot be applied to HvtLB since their method relies on the global strong convexity of the logistic loss, whereas the Huber loss is partially linear. We generalize their approach to the OMD framework and introduce a recursive normalization factor that ensures normal data stays in the quadratic region of the Huber loss with high probability, maintaining robustness in estimation. This normalization factor also introduces a negative term in the estimation error, which is crucial for controlling stability; (ii) *supporting one-pass update*: The OMD framework introduces a stability term in estimation error, which represents the impact of online updates. The adaptive normalization factor adds a multiplicative bias to this term, resulting in a positive  $\tilde{\mathcal{O}}(\sqrt{T})$  term that weakens the bound. This requires a careful design of the normalization factor and learning rate of OMD to cancel out the positive term using the negative term, ensuring stability and maintaining a tight bound.

**Notations.** For a vector  $\mathbf{x} \in \mathbb{R}^d$  and postive semi-definite matrix  $V \in \mathbb{R}^{d \times d}$ , we denote  $\|\mathbf{x}\|_V = \sqrt{\mathbf{x}^\top V \mathbf{x}}$ . For a strictly convex and continuously differentiable function  $\psi : \mathcal{W} \rightarrow \mathbb{R}$ , the Bregman divergence is define as  $\mathcal{D}_\psi(\mathbf{w}_1, \mathbf{w}_2) = \psi(\mathbf{w}_1) - \psi(\mathbf{w}_2) - \langle \nabla \psi(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle$ .

## 2 Related Work

**Heavy-Tailed Bandits.** The multi-armed bandits (MAB) problem with heavy-tailed rewards, characterized by finite  $(1 + \varepsilon)$ -moment, was first studied by [Bubeck et al. \[2013\]](#). They introduced two widely used robust estimation and regression methods from statistics into the heavy-tailed bandits setting: truncation and median-of-means (MOM). [Medina and Yang \[2016\]](#) extended these methods to heavy-tailed linear bandits (HvtLB) and proposed two algorithms: the truncation-based CRT algorithm and the MOM-based MoM algorithm, achieving  $\tilde{O}(dT^{\frac{2+\varepsilon}{2(1+\varepsilon)}})$  and  $\tilde{O}(dT^{\frac{1+2\varepsilon}{1+3\varepsilon}})$  regret bounds, respectively. Later, [Shao et al. \[2018\]](#) established an  $\Omega(dT^{\frac{1}{1+\varepsilon}})$  lower bound for HvtLB and introduced the truncation-based TOFU algorithm and the MOM-based MENU algorithm, both achieving an  $\tilde{O}(dT^{\frac{1}{1+\varepsilon}})$  regret bound. For HvtLB with finite arm set, [Xue et al. \[2020\]](#) established an  $\Omega(d^{\frac{\varepsilon}{1+\varepsilon}} T^{\frac{1}{1+\varepsilon}})$  lower bound, and proposed the truncation-based BTC algorithm and MOM-based BMM algorithm, both of which achieve an  $\tilde{O}(\sqrt{dT}^{\frac{1}{1+\varepsilon}})$  regret bound, reducing the dependence on the dimension  $d$  under the finite arm set setting. To the best of our knowledge, the two works [[Kang and Kim, 2023](#), [Li and Sun, 2023](#)] first introduce Huber loss to linear bandits. Specifically, [Kang and Kim \[2023\]](#) studied the heavy-tailed linear contextual bandits (HvtLCB) with fixed and finite arm set, where the context  $X_t$  at each round is independently and identically distributed from a fixed distribution. They proposed the Huber-Bandit algorithm based on Huber regression, achieving an  $\tilde{O}(\sqrt{dT}^{\frac{1}{1+\varepsilon}})$  regret bound. On the other hand, [Li and Sun \[2023\]](#) were the first to apply adaptive Huber regression technique [[Sun et al., 2020](#)] to SLB, focusing on noise with finite variance (namely, HvtLB with  $\varepsilon = 1$ ). They proposed AdaOFUL algorithm, achieving an optimal variance-aware regret of  $\tilde{O}(d\sqrt{\sum_{t=1}^T \nu_t^2})$ , where  $\nu_t$  is the upper bound of the variance. This result is subsequently improved by [Huang et al. \[2024\]](#), which is also the most related work to ours. [Huang et al. \[2024\]](#) considered general heavy-tailed scenarios with  $\varepsilon \in (0, 1]$  and proposed HEAVY-OFUL with an instance-dependent regret of  $\tilde{O}(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}})$  where  $\nu_t^{1+\varepsilon}$  is the  $(1 + \varepsilon)$ -th central moment of reward at round  $t$ . We emphasize that all the above Huber regression-based algorithms for heavy-tailed bandits are not updated in a one-pass manner.

**One-Pass Bandit Learning.** In SLB, [Abbasi-Yadkori et al. \[2011\]](#) proposed the OFUL algorithm, achieving an  $\mathcal{O}(d\sqrt{T})$  regret bound. It uses Least Squares (LS) for parameter estimation, naturally supporting one-pass updates with a closed-form solution. As an extension of SLB, Generalized Linear Bandits (GLB) were first introduced by [Filippi et al. \[2010\]](#). Their proposed GLM-UCB algorithm achieves an  $\mathcal{O}(\kappa^{-1}d\sqrt{T})$  regret bound, where  $\kappa$  measures the non-linearity of the generalized linear model. This algorithm relies on Maximum Likelihood Estimation (MLE) which doesn't have a closed-form solution like LS in SLB and requires storing all historical data, resulting in a per-round computational cost of  $\mathcal{O}(t)$ . To this end, [Zhang et al. \[2016\]](#) proposed OL2M, the first one-pass algorithm for Logistic Bandits (a specific class of GLB), achieving an  $\mathcal{O}(\kappa^{-1}d\sqrt{T})$  regret bound with per-round computational cost of  $\mathcal{O}(1)$ . Later, [Jun et al. \[2017\]](#) explored one-pass algorithms for GLB by introducing the online-to-confidence-set conversion. Using this conversion, they developed GLOC, a one-pass algorithm that achieves the same theoretical guarantees and per-round cost as OL2M but is applicable to all GLB. [Xue et al. \[2023\]](#) considered heavy-tailed GLB and proposed two one-pass algorithms: truncation-based CRTM and mean-of-medians-based CRMM, both achieving an  $\tilde{O}(\kappa^{-1}dT^{\frac{1}{1+\varepsilon}})$  regret bound. Subsequent works have tackled more challenging topics, such

as reducing the dependence on  $\kappa$  with one-pass algorithms or designing one-pass algorithms for more complex models like Multinomial Logistic Bandits and Multinomial Logit MDPs [Zhang and Sugiyama, 2023, Li et al., 2024, Lee and Oh, 2024, Li et al., 2025].

### 3 One-Pass Heavy-tailed Linear Bandits

In this section, we present our one-pass method for handling linear bandits with heavy-tailed noise. Section 3.1 introduces the problem setup. Section 3.2 reviews the latest work and identifies the inefficiency of the previous method. Section 3.3 introduces the OMD-type estimator based on Huber loss, along with the UCB-based arm selection strategy and the regret guarantee.

#### 3.1 Problem Setup

We investigate the heavy-tailed linear bandits. At round  $t$ , the learner chooses an arm  $X_t$  from the feasible set  $\mathcal{X}_t \subseteq \mathbb{R}^d$ , and then the environment reveals a reward  $r_t \in \mathbb{R}$  such that

$$r_t = X_t^\top \theta_* + \eta_t, \quad (4)$$

where  $\theta_* \in \mathbb{R}^d$  is the unknown parameter and  $\eta_t \in \mathbb{R}$  is the random noise. We define the filtration  $\{\mathcal{F}_t\}_{t \geq 1}$  as  $\mathcal{F}_t \triangleq \sigma(\{X_1, r_1, \dots, X_{t-1}, r_{t-1}, X_t\})$ . The noise  $\eta_t \in \mathbb{R}$  is  $\mathcal{F}_t$ -measurable and satisfies for some  $\varepsilon \in (0, 1]$ ,  $\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$  and  $\mathbb{E}[|\eta_t|^{1+\varepsilon} | \mathcal{F}_{t-1}] = \nu_t^{1+\varepsilon}$  with  $\nu_t$  being  $\mathcal{F}_{t-1}$ -measurable. The goal of the learner is to minimize the following (pseudo) regret,

$$\text{REG}_T = \sum_{t=1}^T \max_{\mathbf{x} \in \mathcal{X}_t} \mathbf{x}^\top \theta_* - \sum_{t=1}^T X_t^\top \theta_*.$$

We work under the following assumption as prior works [Xue et al., 2023, Huang et al., 2024].

**Assumption 1** (Boundedness). The feasible set and unknown parameter are assumed to be bounded:  $\forall \mathbf{x} \in \mathcal{X}_t, \|\mathbf{x}\|_2 \leq L$ ,  $\theta_* \in \Theta$  holds where  $\Theta \triangleq \{\theta \mid \|\theta\|_2 \leq S\}$ .

#### 3.2 Reviewing Previous Efforts

In this part, we review the best-known algorithm of Huang et al. [2024]. They leveraged the adaptive Huber regression method [Sun et al., 2020] to tackle the challenges of heavy-tailed noise. Specifically, adaptive Huber regression is based on the following Huber loss [Huber, 1964].

**Definition 1** (Huber loss). *Huber loss is defined as*

$$f_\tau(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \leq \tau, \\ \tau|x| - \frac{\tau^2}{2} & \text{if } |x| > \tau, \end{cases} \quad (5)$$

where  $\tau > 0$  is the robustification parameter.

The Huber loss is a robust modification of the squared loss that preserves convexity. It behaves as a quadratic function when  $|x|$  is less than the threshold  $\tau$ , ensuring strong convexity in this range. When  $|x|$  exceeds  $\tau$ , the loss transitions to a linear function to reduce the influence of outliers.

At round  $t$ , building on the Huber loss, adaptive Huber regression estimates the unknown parameter  $\theta_*$  of linear model (4) by using all past data up to  $t$  as

$$\hat{\theta}_t = \arg \min_{\theta \in \Theta} \left\{ \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{s=1}^t f_{\tau_s} \left( \frac{r_s - X_s^\top \theta}{\sigma_s} \right) \right\}, \quad (6)$$

where  $\lambda$  is the regularization parameter,  $f_{\tau_s}(\cdot)$  represents the Huber loss with dynamic robustification parameter  $\tau_s$ , and  $\sigma_s$  denotes the normalization factor. Based on the adaptive Huber regression, their proposed Heavy-OFUL algorithm achieves the optimal and instance-dependent regret bound.

**Efficiency Concern.** When the robustification parameter  $\tau_s = +\infty$ , Huber regression (6) reduces to least square, allowing one-pass updates with a closed-form solution. However, for finite  $\tau_s$ , which is essential for handling heavy-tailed noise, (6) requires solving a strongly convex and smooth optimization problem, which needs to store all past data resulting in a storage complexity of  $\mathcal{O}(t)$ . Using Gradient Descent (GD) to achieve an  $\epsilon$ -accurate solution requires  $\mathcal{O}(\log(1/\epsilon))$  iterations, with per-iteration computational cost of  $\mathcal{O}(t)$  to compute the gradient over all past data. To ensure an optimal regret,  $\epsilon$  is typically set to  $1/\sqrt{T}$ , leading to a total computational cost of  $\mathcal{O}(t \log T)$  at round  $t$ . This would be infeasible for a large-scale time horizon of  $T$ .

### 3.3 Hvt-UCB: Huber loss-based One-Pass Algorithm

The adaptive Huber regression method proposed by Huang et al. [2024] incurs significant computational burden in each round and requires storing all past data, which is impractical in real-world applications. Thus, we aim to design a one-pass algorithm for HvtLB based on the Huber loss (5).

For simplicity, let  $z_t(\theta) \triangleq \frac{r_t - X_t^\top \theta}{\sigma_t}$ , we introduce the loss function  $\ell_t(\theta)$  based on (5) as follows,

$$\ell_t(\theta) = \begin{cases} \frac{1}{2} \left( \frac{r_t - X_t^\top \theta}{\sigma_t} \right)^2 & \text{if } |z_t(\theta)| \leq \tau_t, \\ \tau_t \left| \frac{r_t - X_t^\top \theta}{\sigma_t} \right| - \frac{\tau_t^2}{2} & \text{if } |z_t(\theta)| > \tau_t, \end{cases} \quad (7)$$

where  $\sigma_t$  is the normalization factor and  $\tau_t$  is the robustification parameter, both for controlling the impact of heavy-tailed noise.

To enable a one-pass update, instead of solving Huber regression (6) using all historical data, we will adopt the Online Mirror Descent (OMD) framework [Orabona, 2019]. As a general and versatile framework for online optimization, OMD has been widely applied to various challenging *adversarial* online learning problems, such as online games [Rakhlin and Sridharan, 2013, Syrgkanis et al., 2015], adversarial bandits [Abernethy et al., 2008, Wei and Luo, 2018], and dynamic regret minimization [Jacobsen and Cutkosky, 2022, Zhao et al., 2024], among others. In contrast, we leverage its potential to address *stochastic* bandit problems, drawing inspiration from recent advancements in the study of logistic bandits [Zhang and Sugiyama, 2023]. Specifically, we propose the following OMD-based update to estimate the unknown parameter,

$$\hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \langle \theta, \nabla \ell_t(\hat{\theta}_t) \rangle + \mathcal{D}_{\psi_t}(\theta, \hat{\theta}_t) \right\}, \quad (8)$$

where  $\Theta$  is the feasible set defined in Assumption 1 and  $\mathcal{D}_{\psi_t}(\cdot, \cdot)$  is the Bregman divergence associated to the regularizer  $\psi_t : \mathbb{R}^d \mapsto \mathbb{R}$ . The choice of the regularizer is crucial because, in stochastic bandits,



parameter estimation is subsequently used for confidence set construction. Therefore, we define the regularizer using the local information of the current round  $t$  as

$$\psi_t(\theta) = \frac{1}{2} \|\theta\|_{V_t}^2, \text{ with } V_t \triangleq \lambda I + \frac{1}{\alpha} \sum_{s=1}^t \frac{X_s X_s^\top}{\sigma_s^2}, V_0 \triangleq \lambda I_d. \quad (9)$$

The choice of this local-norm regularizer is both practical and compatible with the self-normalized concentration used to construct the Upper Confidence Bound (UCB) of the estimation error. Here, the coefficient  $\alpha$  in  $V_t$  essentially represents the step size of OMD, which will be specified later.

**Computational Efficiency.** Clearly, the estimator (8) can be solved with a single projected gradient step, which can be equivalently reformulated as:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - V_t^{-1} \nabla \ell_t(\hat{\theta}_t), \quad \hat{\theta}_{t+1} = \arg \min_{\theta \in \Theta} \|\theta - \hat{\theta}_{t+1}\|_{V_t},$$

where the main computational cost lies in calculating the inverse matrix  $V_t^{-1}$ , which can be efficiently updated using the Sherman-Morrison-Woodbury formula with a time complexity of  $\mathcal{O}(d^2)$ . The projection step adds an additional time complexity of  $\mathcal{O}(d^3)$ . Focusing on one-pass updates and considering only the dependence on  $t$  and  $T$ , our estimator achieves a per-round computational complexity of  $\mathcal{O}(1)$ , representing a significant improvement over previous methods with a complexity of  $\mathcal{O}(t \log T)$ . Moreover, it eliminates the need to store all historical data, requiring only  $\mathcal{O}(1)$  storage cost throughout the learning process.

**Statistical Efficiency.** As mentioned earlier, OMD is a general framework that has demonstrated its effectiveness in adversarial online learning for regret minimization. Nonetheless, in stochastic bandits, a more critical requirement is the accuracy of parameter estimation. In the following, we demonstrate that OMD also performs exceptionally well in this regard. The key lemma below presents the estimation error of estimator (8) based on OMD-typed update.

**Lemma 1** (Estimation error). *If  $\sigma_t, \tau_t, \tau_0$  are set as*

$$\sigma_t = \max \left\{ \nu_t, \sigma_{\min}, \sqrt{\frac{2\beta_{t-1}}{\tau_0 \sqrt{\alpha} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}} \|X_t\|_{V_{t-1}^{-1}} \right\}, \quad \tau_t = \tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}, \quad \tau_0 = \frac{\sqrt{2\kappa} (\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}{\left(\log \frac{2T^2}{\delta}\right)^{\frac{1}{1+\varepsilon}}},$$

where  $\sigma_{\min}$  is a small positive constant which will be specified later and  $w_t \triangleq \frac{1}{\sqrt{\alpha}} \left\| \frac{X_t}{\sigma_t} \right\|_{V_{t-1}^{-1}}$ . Let the step size  $\alpha = 4$ , then with probability at least  $1 - 4\delta$ ,  $\forall t \geq 1$ , we have  $\left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t} \leq \beta_t$  with

$$\beta_t \triangleq 107 \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} + \sqrt{\lambda(2 + 4S^2)}, \quad \text{where } \kappa \triangleq d \log \left( 1 + \frac{L^2 T}{4\sigma_{\min}^2 \lambda d} \right). \quad (10)$$

The proof of Lemma 1 is provided in Appendix B.5. Notably, both of our method and the HeavyOFUL algorithm of Huang et al. [2024] require the input of  $\sigma_t, \tau_t$ . In fact, our parameter setting for  $\sigma_t$  is simpler. HeavyOFUL requires maintaining two complex lower bounds for  $\sigma_t$  to satisfy:  $\forall \theta \in \Theta, |(X_t^\top \theta - X_t^\top \theta_*) / \sigma_t| \leq \tau_t$  ensuring that the loss remains quadratic for noiseless data. In contrast, our analysis focuses on a refined condition to guarantee the quadratic loss property

involving only the current estimate  $\hat{\theta}_t$ :  $|(X_t^\top \hat{\theta}_t - X_t^\top \theta_*)/\sigma_t| \leq \tau_t$ . This refinement enables us to design a recursive parameter setting that directly incorporates the upper confidence bound  $\beta_{t-1}$  from the previous round into the configuration of  $\sigma_t$  for the current round. Furthermore, this parameter setting can be directly applied to the adaptive Huber regression. Similar to [Huang et al. \[2024\]](#), our method requires the knowledge of the moment  $\nu_t$  of each round  $t$ . If  $\nu_t$  is not available but a general bound  $\nu$  is known, such that  $\mathbb{E}[|\eta_t|^{1+\varepsilon} | \mathcal{F}_{t-1}] \leq \nu^{1+\varepsilon}$ , we can still trivially achieve the same estimation error bound by replacing  $\nu_t$  with  $\nu$  in the setting of  $\sigma_t$ .

**UCB Construction.** Based on the one-pass estimator (8) and Lemma 1, we can specify the arm selection criterion as

$$X_t = \arg \max_{\mathbf{x} \in \mathcal{X}_t} \left\{ \langle \mathbf{x}, \hat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}. \quad (11)$$

Algorithm 1 summarizes the overall algorithm, and enjoys following instance-dependent regret.

**Theorem 1.** *By setting  $\sigma_t$ ,  $\tau_t$ ,  $\tau_0$ ,  $\alpha$  as in Lemma 1, and let  $\lambda = d$ ,  $\sigma_{\min} = \frac{1}{\sqrt{T}}$ ,  $\delta = \frac{1}{8T}$ , with probability at least  $1 - 1/T$ , the regret of Hvt-UCB is bounded by*

$$\text{REG}_T \leq \tilde{\mathcal{O}} \left( dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \right).$$

The proof of Theorem 1 can be found in Appendix C.1. Compared to [Huang et al. \[2024\]](#), our approach achieves the same order of instance-dependent regret bounds, while significantly reducing the per-round computational complexity from  $\mathcal{O}(t \log T)$  to  $\mathcal{O}(1)$ .

When the moment  $\nu_t$  of each round  $t$  is unknown, but a general bound  $\nu$  is known, such that  $\mathbb{E}[|\eta_t|^{1+\varepsilon} | \mathcal{F}_{t-1}] \leq \nu^{1+\varepsilon}$ , we can achieve the following regret bound.

**Corollary 1.** *Follow the parameter setting in Theorem 1, and replace the  $\nu_t$  with  $\nu$  in the  $\sigma_t$  setting, the regret of Hvt-UCB is bounded with probability at least  $1 - 1/T$ , by*

$$\text{REG}_T \leq \tilde{\mathcal{O}} \left( dT^{\frac{1}{1+\varepsilon}} \right).$$

The proof of Corollary 1 can be found in Appendix C.2. This bound matches the lower bound  $\Omega(dT^{\frac{1}{1+\varepsilon}})$  [[Shao et al., 2018](#)] up to logarithmic factors. Moreover, when  $\varepsilon = 1$ , i.e., the bounded variance setting, our approach recovers the regret bound  $\tilde{\mathcal{O}}(d\sqrt{T})$ , which nearly matches the lower bound  $\Omega(d\sqrt{T})$  [[Dani et al., 2008](#)].

## 4 Analysis Sketch

In this section, we provide a proof sketch for Lemma 1, where we adapt the analysis framework of Online Mirror Descent (OMD) to the SLB setting. To simplify the presentation, we first illustrate this framework using a simpler case where the noise follows a sub-Gaussian distribution in Section 4.1. We then discuss the challenges of extending this framework to heavy-tailed case and present our refined analysis in Section 4.2.



---

**Algorithm 1** Hvt-UCB

---

**Input:** time horizon  $T$ , confidence  $\delta$ , regularizer  $\lambda$ ,  $\sigma_{\min}$ , bounded parameters  $S, L$ , parameter for estimation  $c_0, c_1, \tau_0, \alpha, \varepsilon$ .

1: Set  $\kappa = d \log \left( 1 + \frac{L^2 T}{\sigma_{\min}^2 \lambda \alpha d} \right)$ ,  $V_0 = \lambda I_d$ ,  $\hat{\theta}_1 = \mathbf{0}$  and compute  $\beta_0$  by (10)

2: **for**  $t = 1, 2, \dots, T$  **do**

3:   Select  $X_t = \arg \max_{\mathbf{x} \in \mathcal{X}_t} \left\{ \langle \mathbf{x}, \hat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$

4:   Receive the reward  $r_t$ , and  $\nu_t$

5:   Set  $\sigma_t = \max \left\{ \nu_t, \sigma_{\min}, \sqrt{\frac{2\beta_{t-1}}{\tau_0 \sqrt{\alpha} t^{2(1+\varepsilon)}}} \|X_t\|_{V_{t-1}^{-1}} \right\}$

6:   Set  $\tau_t = \tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}$  with  $w_t = \frac{1}{\sqrt{\alpha}} \left\| \frac{X_t}{\sigma_t} \right\|_{V_{t-1}^{-1}}$

7:   Update  $V_t = V_{t-1} + \alpha^{-1} \sigma_t^{-2} X_t X_t^\top$

8:   Compute  $\hat{\theta}_{t+1}$  by (8) and  $\beta_t$  by (10)

9: **end for**

---

#### 4.1 Case study of sub-Gaussian noise

We first consider that the noise  $\eta_t$  in model (4) follows a  $R$ -sub-Gaussian distribution. Then the robustification parameter  $\tau_t$  should be set as  $+\infty$  and the Huber loss recovers the squared loss by  $\ell_t(\theta) = \frac{1}{2}((r_t - X_t^\top \theta)/\sigma_t)^2$ .

**Estimation error decomposition.** We begin by decomposing the estimation error into the following three components:

**Lemma 2.** Let  $\tilde{\ell}_t(\theta) = \frac{1}{2}(X_t^\top(\theta_* - \theta)/\sigma_t)^2$  be the denoised Huber loss function, then the estimation error of estimator (8) with  $\alpha = 2$  can be decomposed as follows,

$$\left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 \leq \underbrace{2 \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle}_{\text{generalization gap term}} + \underbrace{\sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2}_{\text{stability term}} - \underbrace{\frac{1}{2} \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2}_{\text{negative term}} + 4\lambda S^2,$$

*Proof Sketch.* First based on the standard analysis of OMD [Orabona, 2019, Lemma 6.16], we can have the following inequality,

$$\left\langle \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle \leq \frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{V_t}^2 - \frac{1}{2} \left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 + \frac{1}{2} \left\| \nabla \ell_t(\hat{\theta}_t) \right\|_{V_t^{-1}}^2.$$

Based on the quadratic property of the denoised loss function  $\tilde{\ell}_t(\cdot)$ , we have the following equality,

$$\tilde{\ell}_t(\hat{\theta}_t) - \tilde{\ell}_t(\theta_*) = \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle - \frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{\frac{X_t X_t^\top}{\sigma_t^2}}^2.$$

Since  $\theta_*$  is the minimizer of  $\tilde{\ell}_t(\cdot)$ ,  $\tilde{\ell}_t(\hat{\theta}_t) - \tilde{\ell}_t(\theta_*) \geq 0$ , which implies that  $\frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{X_t X_t^\top / \sigma_t^2}^2 \leq \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle$ , then substituting this into the OMD update rule and we can achieve the iteration inequality for the estimation error, further with a telescoping sum, and we finish the proof.  $\square$

**Stability term analysis.** The first step is to extract stochastic term and deterministic term at each step  $s$  as follows,

$$\left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 = \left( \frac{X_s^\top (\theta_* - \hat{\theta}_s)}{\sigma_s} + \frac{\eta_s}{\sigma_s} \right)^2 \left\| \frac{X_s}{\sigma_s} \right\|_{V_s^{-1}}^2.$$

Setting  $\sigma_t = 1$ , we have  $|X_s^\top (\theta_* - \hat{\theta}_s)| \leq 2LS$ . Then based on the sub-Gaussian property, we have  $|\eta_s|$  bounded by  $\mathcal{O}(\sqrt{\log T})$  with high probability, and  $\sum_{s=1}^t \|X_s\|_{V_s^{-1}}^2$  can be bounded by  $\mathcal{O}(\log T)$  with Lemma 7 (potential lemma), which means  $\sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 \leq \mathcal{O}(\log^2 T)$ .

**Generalization gap term analysis.** By taking the gradient of the loss function, we have

$$\sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle = \sum_{s=1}^t \eta_s \left\langle X_s, \hat{\theta}_s - \theta_* \right\rangle,$$

then we can directly apply the 1-dimension self-normalized concentration (Theorem 2), further based on AM-GM inequality, we can bound the general gap term by a  $\mathcal{O}(\log T)$  term and  $\frac{1}{4} \sum_{s=1}^t \|\hat{\theta}_s - \theta_*\|_{X_s X_s^\top}^2$ , which can be further canceled by the negative term.

By integrating the stability term and generalization gap term into Lemma 2, we can achieve the upper bound for the estimation error as  $\|\hat{\theta}_{t+1} - \theta_*\|_{V_t}^2 \leq \mathcal{O}(\log T)$ .

## 4.2 Analysis for heavy-tailed noise

In this section, we extend our analysis to the heavy-tailed case. The presence of heavy-tailed noise and the partially linear structure of the Huber loss introduces new challenges in the three key analyses mentioned above. We will address these challenges one by one in the following discussion.

**Estimation error decomposition.** With the presence of heavy-tailed noise, the denoised Huber loss function with noise-free data  $\tilde{\ell}_t(\theta) \triangleq \frac{X_t^\top \theta_* - X_t^\top \theta}{\sigma_t}$  is defined as

$$\tilde{\ell}_t(\theta) \triangleq \begin{cases} \frac{1}{2} \left( \frac{X_t^\top (\theta_* - \theta)}{\sigma_t} \right)^2 & \text{if } |\tilde{\ell}_t(\theta)| \leq \tau_t, \\ \tau_t \left| \frac{X_t^\top (\theta_* - \theta)}{\sigma_t} \right| - \frac{\tau_t^2}{2} & \text{if } |\tilde{\ell}_t(\theta)| > \tau_t. \end{cases}$$

Now  $\tilde{\ell}_t(\theta)$  is partially quadratic and partially linear, to further utilize the OMD-based estimation error decomposition, we need to make sure  $\hat{\theta}_t$  and  $\theta_*$  both lie in the quadratic region of the loss function  $\tilde{\ell}_t(\cdot)$ . Assuming the event  $A_t = \{\forall s \in [t], \|\hat{\theta}_s - \theta_*\|_{V_{s-1}} \leq \beta_{s-1}\}$  holds, we set  $\sigma_t^2 \geq (2 \|X_t\|_{V_{t-1}^{-1}}^2 \beta_{t-1}) / (\sqrt{\alpha} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}})$ , we can ensure  $|(X_t^\top \hat{\theta}_t - X_t^\top \theta_*) / \sigma_t| \leq \frac{\tau_t}{2}$ . This guarantees that  $\tilde{\ell}_t(\hat{\theta}_t)$  is always in the quadratic region, allowing us to continue using the previous decomposition. The negative term from the decomposition will then be used for subsequent cancellation.

**Stability term analysis.** Similar to the sub-Gaussian case, we can decompose the stability term into two stochastic and deterministic terms as follows,

$$\sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 \leq \underbrace{2 \sum_{s=1}^t \left( \min \left\{ \left| \frac{\eta_s}{\sigma_s} \right|, \tau_s \right\} \right)^2 \left\| \frac{X_s}{\sigma_s} \right\|_{V_s^{-1}}^2}_{\text{stochastic term}} + \underbrace{2 \sum_{s=1}^t \left( \frac{X_s^\top \theta_* - X_s^\top \hat{\theta}_s}{\sigma_s} \right)^2 \left\| \frac{X_s}{\sigma_s} \right\|_{V_s^{-1}}^2}_{\text{deterministic term}}.$$

Note that the stochastic term is no longer inherently bounded. To address this, we utilize a concentration technique (Lemma 6) to bound it by  $\tilde{\mathcal{O}}(t^{\frac{1-\varepsilon}{1+\varepsilon}})$ . For the deterministic term,  $\frac{X_s^\top \theta_* - X_s^\top \hat{\theta}_s}{\sigma_s}$  is bounded by  $2LS/\sigma_{\min}$ , where  $\sigma_{\min}$  is the minimum of  $\sigma_t$ , which can be as small as  $1/\sqrt{T}$ . If we continue using the analysis based on potential lemma, this term will grow to  $\tilde{\mathcal{O}}(\sqrt{T})$ , which is undesirable. Based on the settings of  $\sigma_t$ , we ensure that  $\|X_s\|_{V_s^{-1}}^2 \leq \frac{1}{8}$ . Therefore, this term can be bounded by  $\frac{1}{4} \sum_{s=1}^t \|\theta_* - \hat{\theta}_s\|_{X_s X_s^\top / \sigma_s^2}^2$ , which can be further canceled out by the negative term.

**Generalization gap term analysis.** The non-linear gradient of the Huber loss in the heavy-tailed setting makes it impossible to directly obtain a one-dimensional self-normalized term. We first decompose it into two distinct components as follows,

$$\underbrace{2 \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) + \nabla \ell_s(\theta_*) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle}_{\text{Huber-loss term}} + \underbrace{2 \sum_{s=1}^t \left\langle -\nabla \ell_s(\theta_*), \hat{\theta}_s - \theta_* \right\rangle}_{\text{self-normalized term}},$$

where Huber-loss term represents the influence of the Huber loss structure on the estimation error. Specifically, when all three gradient terms are within the quadratic region of the Huber loss, i.e., when  $|z_t(\theta_*)| \leq \frac{\tau_t}{2}$ , this term becomes zero. Consequently, this term reduces to a sum involving the indicator function  $\sum_{s=1}^t \mathbb{1}\{|z_s(\theta_*)| > \frac{\tau_s}{2}\}$ , which can be bounded by  $\tilde{\mathcal{O}}(t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \beta_t)$  using Eq. (C.12) of Huang et al. [2024]. Then the self-normalized term corresponds to the 1-dimensional version of the self-normalized concentration can be bounded by an  $\tilde{\mathcal{O}}(t^{\frac{1-\varepsilon}{1+\varepsilon}})$  term and  $\frac{1}{2} \sum_{s=1}^t \|\hat{\theta}_s - \theta_*\|_{X_s X_s^\top / \sigma_s^2}^2$  will be canceled by the negative term.

By combining the analysis for stability term and generalization gap term and substituting back to the estimation error decomposition, we obtain:

$$\left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 \lesssim t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \beta_t + t^{\frac{1-\varepsilon}{1+\varepsilon}}.$$

To ensure the event  $A_t$  holds,  $\beta_t$  must satisfy:

$$\beta_t^2 \gtrsim t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \beta_t + t^{\frac{1-\varepsilon}{1+\varepsilon}}.$$

Finally, solving for  $\beta_t$ , we get  $\beta_t = \mathcal{O}\left(t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\right)$ . Thus we finish the proof of Lemma 1.

## 5 Discussion

In this section, we discuss the potential applications of our approach and highlight its advantages compared to previous methods in broader scenarios involving heavy-tailed noise. Specifically, we focus on two settings: heavy-tailed linear MDPs and adaptive control under heavy-tailed noise.

### 5.1 Online Linear MDPs

In the setting of heavy-tailed linear MDPs [Li and Sun, 2023, Huang et al., 2024], the realizable reward for taking action  $a$  in state  $s$  at the  $h$ -th episode is given by:

$$R_h(s, a) = \langle \phi(s, a), \theta_h^* \rangle + \epsilon_h(s, a),$$

where  $\phi(s, a) \in \mathbb{R}^d$  is known feature map,  $\theta_h^* \in \mathbb{R}^d$  is the unknown parameter, and  $\epsilon_h(s, a)$  is the heavy-tailed noise with  $\mathbb{E}_h[\epsilon_h](s, a) = 0$  and  $\mathbb{E}_h[|\epsilon_h|^{1+\varepsilon}](s, a) \leq \nu_R^{1+\varepsilon}$ .

Linear MDPs often adopt techniques from linear bandits, including estimating unknown parameters in a linear model and constructing an upper confidence bound (UCB) based on estimation error analysis. For heavy-tailed linear MDPs, [Huang et al., 2024] employs adaptive Huber regression to estimate the unknown parameter  $\theta_h^*$  in the reward function. This introduces a challenge similar to that in HvtLB: adaptive Huber regression is an offline algorithm, while an online setting requires a one-pass estimator. Existing one-pass estimators for HvtLB, such as the truncation-based and MOM-based algorithms proposed by Xue et al. [2023], are not applicable to heavy-tailed linear MDPs. Truncation-based algorithms rely on assumptions about absolute moments, which are not suitable in this context. MOM-based methods require multiple reward observations for a fixed  $\phi(s, a)$ , but this is infeasible in linear MDPs since state transitions occur probabilistically, preventing repeated observations for the same state-action pair.

In contrast, our proposed Hvt-UCB method can be directly applied to this scenario without requiring additional assumptions. Its estimation error bound analysis naturally extends to linear MDPs, enabling the construction of UCBs for the value function. By leveraging the UCB-to-regret analysis in [Huang et al., 2024], Hvt-UCB has the potential to achieve the same theoretical guarantees as adaptive Huber regression, while significantly reducing computational costs.

## 5.2 Online Adaptive Control

The online adaptive control of Linear Quadratic Systems [Abbasi-Yadkori and Szepesvári, 2011] considers the following state transition system,

$$x_{t+1} = Ax_t + Bu_t + w_{t+1},$$

where  $x_t \in \mathbb{R}^n$  represents the state at time  $t$ ,  $u_t \in \mathbb{R}^d$  is the control input at time  $t$ ,  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{n \times d}$  are unknown system matrices, and  $w_{t+1}$  is the noise. The online adaptive control requires estimating the unknown parameters  $A$  and  $B$  (also known as system identification), and constructing finite-sample guarantees for estimation error. To this end, Abbasi-Yadkori and Szepesvári [2011] transformed the system into the following linear model:

$$x_{t+1} = \Theta_*^\top z_t + w_{t+1},$$

where  $\Theta_* = [A; B]$ ,  $z_t = [x_t; u_t]$  and the noise  $w_t$  in each dimension is assumed to be sub-Gaussian. Then, they used a least squares estimator and leveraged the self-normalized concentration technique from linear bandit analysis [Abbasi-Yadkori et al., 2011] to estimate  $\Theta_*$  and provide finite-sample estimation error guarantees.

As noted by Tsiamis et al. [2023], finite-sample guarantees for system identification under heavy-tailed noise remain an open challenge. Inspired by Abbasi-Yadkori and Szepesvári [2011]’s application of linear bandit techniques to sub-Gaussian cases, we believe methods from HvtLB can similarly benefit adaptive control in heavy-tailed settings. However, MOM methods are infeasible as the feature  $z_t$ , which includes the evolving state  $x_t$ , cannot be repeatedly sampled, and truncation-based methods rely on assumptions of bounded absolute moments for states, which may not hold. While adaptive Huber regression [Li and Sun, 2023, Huang et al., 2024] offers promising theoretical guarantees for heavy-tailed linear system identification, its computational inefficiency makes it unsuitable for adaptive control. In contrast, our proposed Hvt-UCB algorithm provides efficient,

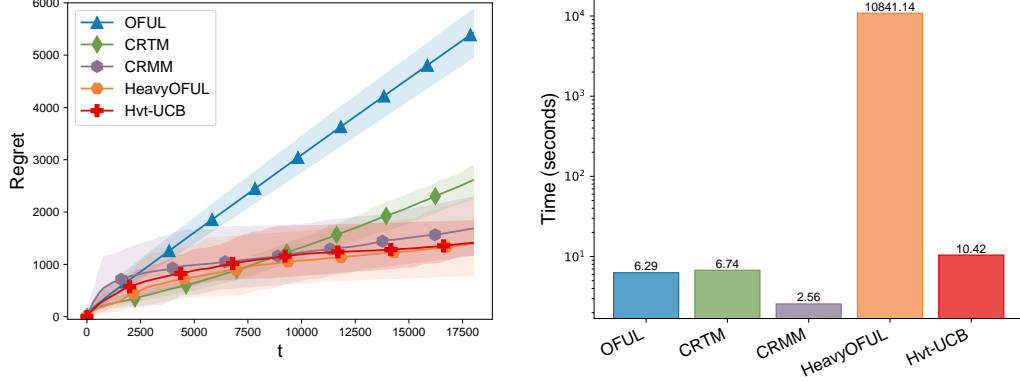


Figure 1: Student  $t$ : regret and time cost.

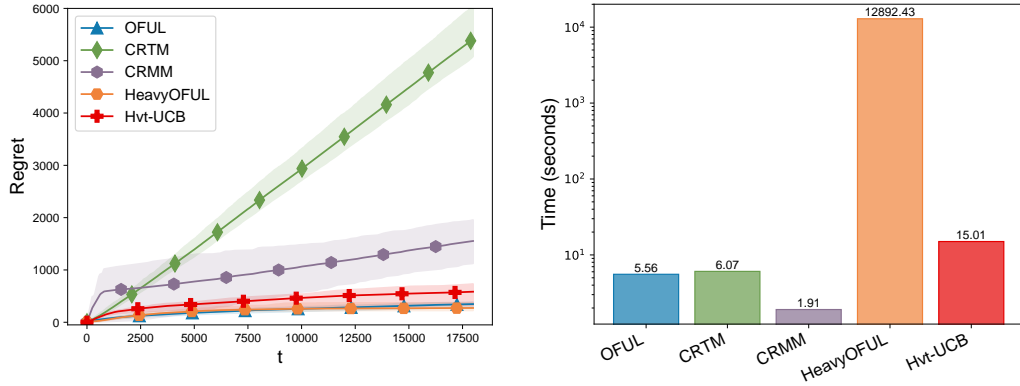


Figure 2: Gaussian: regret and time cost.

one-pass updates without additional assumptions and has the potential to deliver finite-sample guarantees for adaptive control under heavy-tailed noise.

## 6 Experiments

In this section, we empirically evaluate the performance and time efficiency of our proposed algorithm under two different noise distributions: Student's  $t$  and Gaussian.

**Settings.** We consider the linear model  $r_t = X_t^\top \theta_* + \eta_t$ , where the dimension  $d = 2$ , the number of rounds  $T = 18000$ , and the number of arms  $n = 50$ . Each dimension of the feature vectors for the arms is uniformly sampled from  $[-1, 1]$  and subsequently rescaled to satisfy  $L = 1$ . Similarly,  $\theta_*$  is sampled in the same way and rescaled to satisfy  $S = 1$ . We conduct two synthetic experiments with different distributions for noise  $\eta_t$ : (a) Student's  $t$ -distribution with degree of freedom  $df = 2.1$  to represent heavy-tailed noise; and (b) Gaussian noise sampled from  $\mathcal{N}(0, 1)$  to represent light-tailed noise. For the heavy-tailed experiment, we set  $\varepsilon = 0.99$  and  $\nu = 1.31$ , while for the light-tailed experiment, we set  $\varepsilon = 1$ . We compare the performance of our proposed Hvt-UCB algorithm with the following baselines: (a) the OFUL algorithm [Abbasi-Yadkori et al., 2011]; (b) the one-pass truncation-based algorithm CRTM [Xue et al., 2023]; (c) the one-pass MOM-based algorithm CRMM [Xue et al., 2023]; and (d) the Huber-based algorithm HEAVY-OFUL [Huang et al., 2024].

**Results.** We conducted 10 independent trials and averaged the results. Figure 1 shows the cumulative regret and computational time under Student’s  $t$  noise, while Figure 2 presents the same metrics under Gaussian noise, with shaded regions indicating the variance across trials. Under both noise settings, our algorithm demonstrates comparable regret performance to Heavy-OFUL while achieving remarkable computational efficiency, with a speedup factor exceeding  $800\times$ . This efficiency advantage becomes more important as the number of rounds  $T$  increases. Notably, among the one-pass algorithms, CRMM exhibits the fastest runtime because it updates only every  $\mathcal{O}(\log T)$  rounds rather than at each round. In the heavy-tailed noise scenario (Figure 1), the performance of OFUL drops significantly, demonstrating its sensitivity to heavy-tailed distributions. In the Gaussian noise setting (Figure 2), our algorithm maintains competitive performance alongside OFUL. However, the truncation-based CRTM algorithm shows suboptimal performance due to excessive reward truncation in light-tailed noise scenarios, which leads to the loss of valuable data. These experimental results validate that our Hvt-UCB algorithm achieves robust performance across different noise environments while maintaining low computational cost.

## 7 Conclusion

In this paper, we investigate the heavy-tailed linear bandits problem. We identify the advantages of Huber loss-based methods over the truncation and median-of-means strategies, while also pointing out the inefficiencies of previous Huber loss-based methods. We then propose the Huber loss-based one-pass algorithm Hvt-UCB based on the Online Mirror Descent (OMD) framework. Hvt-UCB achieves the optimal and instance-dependent regret bound  $\tilde{\mathcal{O}}(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}})$  while only requiring  $\mathcal{O}(1)$  per-round computational cost. Furthermore, it can be extended to online linear MDPs and online adaptive control, broadening its applicability. The key contribution is our adaptation of the OMD framework to stochastic linear bandits, which addresses the challenges introduced by heavy-tailed noise and the structure of the Huber loss.

Both Huang et al. [2024] and our work rely on knowledge of the moment  $\nu_t$  to achieve instance-dependent regret guarantees. Extending these Huber loss-based methods to handle unknown  $\nu_t$  while maintaining instance-dependent guarantees is an interesting problem for future investigation.

## References

- Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference Computational Learning Theory (COLT)*, volume 19, pages 1–26, 2011.
- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2312–2320, 2011.
- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–9, 2012.
- Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm



- for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 263–274, 2008.
- Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.
- Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l’IHP Probabilités et statistiques*, 48(4):1148–1185, 2012.
- Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.
- Hana Choi, Carl F. Mela, Santiago R. Balseiro, and Adam Leary. Online display advertising markets: A literature review and future directions. *Information Systems Research*, 31(2):556–575, 2020.
- Rama Cont and Jean-Philippe Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(2):170–196, 2000.
- Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference Computational Learning Theory (COLT)*, pages 355–366, 2008.
- Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 586–594, 2010.
- Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, page to appear, 2024.
- Peter J Huber. Robust estimation of a location parameter. In *The Annals of Mathematical Statistics*, pages 73–101. 1964.
- John Hull. *Risk management and financial institutions, + Web Site*, volume 733. John Wiley & Sons, 2012.
- Andrew Jacobsen and Ashok Cutkosky. Parameter-free mirror descent. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178, pages 4160–4211, 2022.
- Charles Jebarajakirthy, Haroon Iqbal Maseeh, Zakir Morshed, Amit Shankar, Denni Arli, and Robin Pentecost. Mobile advertising: A systematic literature review and future research agenda. *International Journal of Consumer Studies*, 45(6):1258–1291, 2021.
- Kwang-Sung Jun, Aniruddha Bhargava, Robert D. Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 99–109, 2017.
- Minhyun Kang and Gi-Soo Kim. Heavy-tailed linear bandit with Huber regression. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1027–1036, 2023.

- Joongkyu Lee and Min-hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, page to appear, 2024.
- Long-Fei Li, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Provably efficient reinforcement learning with multinomial logit function approximation. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, page to appear, 2024.
- Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. Provably efficient RLHF pipeline: A unified view from contextual bandits. *ArXiv preprint*, 2502.07193, 2025.
- Xiang Li and Qiang Sun. Variance-aware robust reinforcement learning with linear function approximation with heavy-tailed rewards. *ArXiv preprint*, arXiv:2303.05606, 2023.
- Yingkai Li, Yining Wang, Xi Chen, and Yuan Zhou. Tight regret bounds for infinite-armed linear contextual bandits. In *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 370–378, 2021.
- Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.
- Andres Muñoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1642–1650, 2016.
- Francesco Orabona. A modern introduction to online learning. *ArXiv preprint*, arXiv:1912.13213, 2019.
- Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 3066–3074, 2013.
- Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 8430–8439, 2018.
- Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.
- Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 2989–2997, 2015.
- Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.
- Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 1263–1291, 2018.

- Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2936–2942, 2020.
- Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, page to appear, 2023.
- Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 392–401, 2016.
- Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 29741–29782, 2023.
- Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. A simple approach for non-stationary linear bandits. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 746–755, 2020.
- Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1 – 52, 2024.

## A Properties of Huber Loss

**Property 1** (Property 1 of [Huang et al. \[2024\]](#)). Let  $f_\tau(x)$  be the Huber loss defined in Definition 1, then the followings are true:

1.  $|f'_\tau(x)| = \min\{|x|, \tau\}$ ,
2.  $f'_\tau(x) = \tau f'_1\left(\frac{x}{\tau}\right)$ ,
3.  $-\log(1 - x + |x|^{1+\varepsilon}) \leq f'_1(x) \leq \log(1 + x + |x|^{1+\varepsilon})$  for any  $\varepsilon \in (0, 1]$ ,
4.  $f''_\tau(x) = \mathbf{1}\{|x| \leq \tau\}$ .

Furthermore, the gradient of the loss function (7) is,

$$\nabla \ell_t(\theta) = \begin{cases} -z_t(\theta) \frac{X_t}{\sigma_t} & \text{if } |z_t(\theta)| \leq \tau_t, \\ -\tau_t \frac{X_t}{\sigma_t} & \text{if } z_t(\theta) > \tau_t, \\ \tau_t \frac{X_t}{\sigma_t} & \text{if } z_t(\theta) < -\tau_t. \end{cases} \quad (12)$$

Based on Property 1, we have

$$\|\nabla \ell_t(\theta)\| = \left\| \min\{|z_t(\theta)|, \tau_t\} \frac{X_t}{\sigma_t} \right\|, \quad \nabla^2 \ell_t(\theta) = \mathbf{1}\{|z_t(\theta)| \leq \tau_t\} \frac{X_t X_t^\top}{\sigma_t^2}. \quad (13)$$

## B Estimation Error Analysis

For the analysis of estimation error, we begin by defining the following denoised loss function based on noise-free data  $\tilde{z}_t(\theta) \triangleq \frac{X_t^\top \theta_* - X_t^\top \theta}{\sigma_t}$ ,

$$\tilde{\ell}_t(\theta) \triangleq \begin{cases} \frac{1}{2} \left( \frac{X_t^\top (\theta_* - \theta)}{\sigma_t} \right)^2 & \text{if } |\tilde{z}_t(\theta)| \leq \tau_t, \\ \tau_t \left| \frac{X_t^\top (\theta_* - \theta)}{\sigma_t} \right| - \frac{\tau_t^2}{2} & \text{if } |\tilde{z}_t(\theta)| > \tau_t. \end{cases}$$

The gradient and Hessian of the denoised loss function  $\tilde{\ell}_t(\theta)$  can be written as

$$\nabla \tilde{\ell}_t(\theta) = \begin{cases} -\tilde{z}_t(\theta) \frac{X_t}{\sigma_t} & \text{if } |\tilde{z}_t(\theta)| \leq \tau_t, \\ -\tau_t \frac{X_t}{\sigma_t} & \text{if } \tilde{z}_t(\theta) > \tau_t, \\ \tau_t \frac{X_t}{\sigma_t} & \text{if } \tilde{z}_t(\theta) < -\tau_t. \end{cases} \quad \nabla^2 \tilde{\ell}_t(\theta) = \mathbf{1}\{|\tilde{z}_t(\theta)| \leq \tau_t\} \frac{X_t X_t^\top}{\sigma_t^2}. \quad (14)$$

We further set the robustification parameter  $\tau_t$  as following,

$$\tau_t = \tau_0 \frac{\sqrt{1 + w_t^2}}{w_t} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}, \quad \tau_0 = \frac{\sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}{\left(\log \frac{2T^2}{\delta}\right)^{\frac{1}{1+\varepsilon}}}, \quad \text{where } w_t \triangleq \frac{1}{\sqrt{\alpha}} \left\| \frac{X_t}{\sigma_t} \right\|_{V_{t-1}^{-1}}, \quad (15)$$

and we denote event  $A_t = \{\forall s \in [t], \|\hat{\theta}_s - \theta_*\|_{V_{s-1}} \leq \beta_{s-1}\}$ . Based on the parameter setting and the event  $A_t$ , we derive three useful lemmas for the analysis of estimation error in the following section.

## B.1 Useful Lemmas

In this section, we provide some useful lemmas for the estimation error analysis. We provide the estimation error decomposition in Lemma 3, the stability term analysis in Lemma 4, and the generalization gap term analysis in Lemma 5.

**Lemma 3** (Estimation error decomposition). *When event  $A_t$  holds, by setting  $\tau_t$  as (15), and  $\sigma_t$  as*

$$\sigma_t^2 \geq \frac{2 \|X_t\|_{V_{t-1}^{-1}}^2 \beta_{t-1}}{\sqrt{\alpha} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}, \quad (16)$$

*then the estimation error can be decomposed as following three terms,*

$$\begin{aligned} & \left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 \\ & \leq 4\lambda S^2 + \sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 + 2 \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle + \left( \frac{1}{\alpha} - 1 \right) \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2. \end{aligned}$$

**Lemma 4.** *By setting  $\tau_t$  as (15) and  $\sigma_t = \max \left\{ \nu_t, \sigma_{\min}, 2\sqrt{2} \|X_t\|_{V_{t-1}^{-1}} \right\}$ , with probability at least  $1 - \delta$ , we have  $\forall t \geq 1$ ,*

$$\sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 \leq 6\alpha \left( t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa} (\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \log \frac{2T^2}{\delta} \right)^{\frac{\varepsilon}{1+\varepsilon}} \right)^2 + \frac{1}{4} \sum_{s=1}^t \left\| \theta_* - \hat{\theta}_s \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2.$$

*where  $\alpha$  is the learning rate need to be tuned and  $\kappa \triangleq d \log \left( 1 + \frac{L^2 T}{\sigma_{\min}^2 \lambda \alpha d} \right)$ .*

**Lemma 5.** *By setting  $\tau_t$  as (15) and  $\sigma_t$  as following,*

$$\sigma_t = \max \left\{ \nu_t, \sigma_{\min}, 2\sqrt{2} \|X_t\|_{V_{t-1}^{-1}}, \sqrt{\frac{2 \|X_t\|_{V_{t-1}^{-1}}^2 \beta_{t-1}}{\sqrt{\alpha} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}} \right\}, \quad (17)$$

*with probability at least  $1 - 3\delta$ , we have  $\forall t \geq 1$ ,*

$$\begin{aligned} \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle \mathbb{1}_{A_s} & \leq 23 \sqrt{\alpha + \frac{1}{8} \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}} \max_{s \in [t+1]} \beta_{s-1} \\ & + \frac{1}{4} \left( \lambda \alpha + \sum_{s=1}^t \left\langle \frac{X_s}{\sigma_s}, \hat{\theta}_s - \theta_* \right\rangle^2 \right) + \left( 8 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa} (\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \log \left( \frac{2T^2}{\delta} \right) \right)^{\frac{\varepsilon}{1+\varepsilon}} \right)^2. \end{aligned}$$

## B.2 Proof of Lemma 3

*Proof.* Since  $A_t = \{ \forall s \in [t], \|\hat{\theta}_s - \theta_*\|_{V_{s-1}} \leq \beta_{s-1} \}$  holds,  $\tau_t$  and  $\sigma_t$  satisfies (15) and (16), we have

$$\left| \frac{X_t^\top \hat{\theta}_t - X_t^\top \theta_*}{\sigma_t} \right| \leq \left\| \frac{X_t}{\sigma_t} \right\|_{V_{t-1}^{-1}} \left\| \theta_* - \hat{\theta}_t \right\|_{V_{t-1}} \leq \frac{\tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}{2w_t \beta_{t-1}} \beta_{t-1} \leq \frac{1}{2} \tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} = \frac{\tau_t}{2}, \quad (18)$$

where  $w_t = \frac{1}{\sqrt{\alpha}} \left\| \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}}$ . Then, based on definition of  $\tilde{z}_t(\cdot)$ , we have  $\tilde{z}_t(\theta_*) = 0 \leq \tau_t$ , and (18) shows that  $\tilde{z}_t(\hat{\theta}_t) \leq \frac{\tau_t}{2} \leq \tau_t$ , which means both points  $\theta_*$  and  $\hat{\theta}_t$  lie on the quadratic side of the denoised loss function  $\tilde{\ell}_t(\cdot)$ . This allows us to apply Taylor's Formula with lagrange remainder to obtain

$$\tilde{\ell}_t(\theta_*) = \tilde{\ell}_t(\hat{\theta}_t) + \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t), \theta_* - \hat{\theta}_t \right\rangle + \frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{\nabla^2 \tilde{\ell}_t(\xi_t)}^2, \quad (19)$$

where  $\xi_t = \gamma \hat{\theta}_t + (1 - \gamma) \theta_*$  for some  $\gamma \in (0, 1)$ , which means  $\xi_t$  also lie on the quadratic side of the denoised loss function:

$$|\tilde{z}_t(\xi_t)| = \left| \frac{X_t^\top \xi_t - X_t^\top \theta_*}{\sigma_t} \right| = \left| \frac{X_t^\top (\gamma \hat{\theta}_t + (1 - \gamma) \theta_*) - X_t^\top \theta_*}{\sigma_t} \right| = \gamma \left| \frac{X_t^\top \hat{\theta}_t - X_t^\top \theta_*}{\sigma_t} \right| \leq \gamma \frac{\tau_t}{2} \leq \tau_t,$$

where the first inequality comes from (18). Then we have  $\nabla^2 \tilde{\ell}_t(\xi_t) = \mathbf{1}_{\{\tilde{z}_t(\xi_t) \leq \tau_t\}} \frac{X_t X_t^\top}{\sigma_t^2} = \frac{X_t X_t^\top}{\sigma_t^2}$ . At the same time, since  $\theta_* = \arg \min_{\theta \in \mathbb{R}^d} \tilde{\ell}_t(\theta)$ , we have  $0 \leq \tilde{\ell}_t(\hat{\theta}_t) - \tilde{\ell}_t(\theta_*)$ . Substituting these into (19), we have

$$0 \leq \tilde{\ell}_t(\hat{\theta}_t) - \tilde{\ell}_t(\theta_*) = \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle - \frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{\frac{X_t X_t^\top}{\sigma_t^2}}^2, \quad (20)$$

which means  $\frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{\frac{X_t X_t^\top}{\sigma_t^2}}^2 \leq \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle$ , then we have

$$\begin{aligned} & \frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{\frac{X_t X_t^\top}{\sigma_t^2}}^2 \\ & \leq \left\langle \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_{t+1} - \theta_* \right\rangle + \left\langle \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \hat{\theta}_{t+1} \right\rangle + \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle, \end{aligned} \quad (21)$$

where the first term can be bounded by the Bregman proximal inequality (9), we have

$$\begin{aligned} \left\langle \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_{t+1} - \theta_* \right\rangle & \leq \mathcal{D}_{\psi_t}(\theta_*, \hat{\theta}_t) - \mathcal{D}_{\psi_t}(\theta_*, \hat{\theta}_{t+1}) - \mathcal{D}_{\psi_t}(\hat{\theta}_{t+1}, \hat{\theta}_t) \\ & = \frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{V_t}^2 - \frac{1}{2} \left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 - \frac{1}{2} \left\| \hat{\theta}_{t+1} - \theta_t \right\|_{V_t}^2, \end{aligned}$$

substituting the above inequality into (21), we have

$$\begin{aligned} \frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{\frac{X_t X_t^\top}{\sigma_t^2}}^2 & \leq \frac{1}{2} \left\| \hat{\theta}_t - \theta_* \right\|_{V_t}^2 - \frac{1}{2} \left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 - \frac{1}{2} \left\| \hat{\theta}_{t+1} - \theta_t \right\|_{V_t}^2 \\ & \quad + \left\langle \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \hat{\theta}_{t+1} \right\rangle + \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle. \end{aligned} \quad (22)$$

Rearranging the above inequality and using AM-GM inequality, we have

$$\begin{aligned} \left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 & \leq \left\| \hat{\theta}_t - \theta_* \right\|_{V_t}^2 + \left\| \nabla \ell_t(\hat{\theta}_t) \right\|_{V_t^{-1}}^2 + 2 \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle - \left\| \hat{\theta}_t - \theta_* \right\|_{\frac{X_t X_t^\top}{\sigma_t^2}}^2 \\ & \leq \left\| \hat{\theta}_t - \theta_* \right\|_{V_{t-1}}^2 + \left\| \nabla \ell_t(\hat{\theta}_t) \right\|_{V_t^{-1}}^2 + 2 \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle + \left( \frac{1}{\alpha} - 1 \right) \left\| \hat{\theta}_t - \theta_* \right\|_{\frac{X_t X_t^\top}{\sigma_t^2}}^2, \end{aligned}$$



where the last equality comes from that  $V_t = V_{t-1} + \frac{1}{\alpha} \frac{X_t X_t^\top}{\sigma_t^2}$ . Taking the summation of the above inequality over  $t$  rounds and we have

$$\begin{aligned}
& \left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 \\
& \leq \left\| \hat{\theta}_1 - \theta_* \right\|_{V_0}^2 + \sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 + 2 \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle + \left( \frac{1}{\alpha} - 1 \right) \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2 \\
& \leq 4\lambda S^2 + \sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 + 2 \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle + \left( \frac{1}{\alpha} - 1 \right) \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2,
\end{aligned}$$

where the last inequality comes from  $V_0 = \lambda I_d$  and Assumption 1, thus we complete the proof.  $\square$

### B.3 Proof of Lemma 4

*Proof.* We first analyze the upper bound of single term  $\left\| \nabla \ell_t(\hat{\theta}_t) \right\|_{V_t^{-1}}^2$ , based on the definition of loss function (13), we have

$$\begin{aligned}
\left\| \nabla \ell_t(\hat{\theta}_t) \right\|_{V_t^{-1}}^2 &= \left\| \min \left\{ \left| \frac{r_t - X_t^\top \hat{\theta}_t}{\sigma_t} \right|, \tau_t \right\} \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}}^2 \\
&= \left\| \min \left\{ \left| \frac{\eta_t}{\sigma_t} + \frac{X_t^\top \theta_* - X_t^\top \hat{\theta}_t}{\sigma_t} \right|, \tau_t \right\} \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}}^2 \\
&\leq \left( \min \left\{ \left| \frac{\eta_t}{\sigma_t} \right| + \left| \frac{X_t^\top \theta_* - X_t^\top \hat{\theta}_t}{\sigma_t} \right|, \tau_t \right\} \right)^2 \left\| \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}}^2 \\
&\leq \left( \min \left\{ \left| \frac{\eta_t}{\sigma_t} \right|, \tau_t \right\} + \min \left\{ \left| \frac{X_t^\top \theta_* - X_t^\top \hat{\theta}_t}{\sigma_t} \right|, \tau_t \right\} \right)^2 \left\| \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}}^2 \\
&\leq 2 \left( \min \left\{ \left| \frac{\eta_t}{\sigma_t} \right|, \tau_t \right\} \right)^2 \left\| \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}}^2 + 2 \left( \frac{X_t^\top \theta_* - X_t^\top \hat{\theta}_t}{\sigma_t} \right)^2 \left\| \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}}^2. \quad (23)
\end{aligned}$$

We define  $\tilde{V}_t \triangleq \lambda \alpha I + \sum_{s=1}^t \frac{X_s X_s^\top}{\sigma_s^2} = \alpha V_t$ , then we have  $w_t = \frac{1}{\sqrt{\alpha}} \left\| \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}} = \left\| \frac{X_t}{\sigma_t} \right\|_{\tilde{V}_t^{-1}}$ . Based on Sherman-Morrison-Woodbury formula of  $\tilde{V}_t^{-1}$  we have

$$\left\| \frac{X_t}{\sigma_t} \right\|_{\tilde{V}_t^{-1}}^2 = \frac{X_t^\top \tilde{V}_t^{-1} X_t}{\sigma_t^2} = \frac{1}{\sigma_t^2} X_t^\top \left( \tilde{V}_{t-1}^{-1} - \frac{\tilde{V}_{t-1}^{-1} X_t X_t^\top \tilde{V}_{t-1}^{-1}}{\sigma_t^2 (1 + w_t^2)} \right) X_t = w_t^2 - \frac{w_t^4}{1 + w_t^2} = \frac{w_t^2}{1 + w_t^2},$$

thus  $\left\| \frac{X_t}{\sigma_t} \right\|_{V_t^{-1}}^2 = \alpha \frac{w_t^2}{1 + w_t^2}$ , then taking the summation over inequality (23) over  $t$  rounds and we have

$$\sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 \leq \underbrace{2\alpha \sum_{s=1}^t \left( \min \left\{ \left| \frac{\eta_s}{\sigma_s} \right|, \tau_s \right\} \right)^2 \frac{w_s^2}{1 + w_s^2}}_{\text{TERM (A.1)}} + \underbrace{2\alpha \sum_{s=1}^t \left( \frac{X_s^\top \theta_* - X_s^\top \hat{\theta}_s}{\sigma_s} \right)^2 \frac{w_s^2}{1 + w_s^2}}_{\text{TERM (A.2)}}.$$

**For term (A.1).** By applying Lemma C.5 of Huang et al. [2024], restated in Lemma 6, with  $\tau_t = \tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}$  and  $b = \max_{t \in [T]} \frac{\nu_t}{\sigma_t} \leq 1$ , we have with probability at least  $1 - \delta, \forall t \geq 1$ ,

$$\sum_{s=1}^t \left( \min \left\{ \left| \frac{\eta_s}{\sigma_s} \right|, \tau_s \right\} \right)^2 \frac{w_s^2}{1+w_s^2} \leq \left[ t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \sqrt{\tau_0^{1-\varepsilon} (\sqrt{2\kappa})^{1+\varepsilon} (\log 3T)^{\frac{1-\varepsilon}{2}}} + \tau_0 \sqrt{2 \log \frac{2T^2}{\delta}} \right) \right]^2,$$

by choosing  $\tau_0 = \frac{\sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}{\left(\log \frac{2T^2}{\delta}\right)^{\frac{1}{1+\varepsilon}}}$ , we have  $\sqrt{\tau_0^{1-\varepsilon} (\sqrt{2\kappa})^{1+\varepsilon} (\log 3T)^{\frac{1-\varepsilon}{2}}} = \tau_0 \sqrt{\log \frac{2T^2}{\delta}}$ , which means

$$\sum_{s=1}^t \left( \min \left\{ \left| \frac{\eta_s}{\sigma_s} \right|, \tau_s \right\} \right)^2 \frac{w_s^2}{1+w_s^2} \leq 6 \left( t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa} (\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \log \frac{2T^2}{\delta} \right)^{\frac{\varepsilon}{1+\varepsilon}} \right)^2. \quad (24)$$

**For term (A.2).** Since  $\sigma_t \geq 2\sqrt{2} \|X_t\|_{V_{t-1}^{-1}}$ , which means  $\alpha w_t^2 = \left\| \frac{X_t}{\sigma_t} \right\|_{V_{t-1}^{-1}}^2 \leq 1/8$ , then we have

$$2\alpha \sum_{s=1}^t \left( \frac{X_s^\top \theta_* - X_s^\top \hat{\theta}_s}{\sigma_s} \right)^2 \frac{w_s^2}{1+w_s^2} \leq \frac{1}{4} \sum_{s=1}^t \left( \frac{X_s^\top \theta_* - X_s^\top \hat{\theta}_s}{\sigma_s} \right)^2 = \frac{1}{4} \sum_{s=1}^t \left\| \theta_* - \hat{\theta}_s \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2, \quad (25)$$

Combining (24) and (25), we have with probability at least  $1 - \delta, \forall t \geq 1$

$$\sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 \leq 6\alpha \left( t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa} (\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \log \frac{2T^2}{\delta} \right)^{\frac{\varepsilon}{1+\varepsilon}} \right)^2 + \frac{1}{4} \sum_{s=1}^t \left\| \theta_* - \hat{\theta}_s \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2.$$

Thus we complete the proof.  $\square$

## B.4 Proof of Lemma 5

*Proof.* We first analyze the upper bound of single term,

$$\begin{aligned} & \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle \mathbf{1}_{A_t} \\ &= \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) - \nabla \ell_t(\theta_*) + \nabla \ell_t(\theta_*) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle \mathbf{1}_{A_t} \\ &= \underbrace{\left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) + \nabla \ell_t(\theta_*) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle \mathbf{1}_{A_t}}_{\text{TERM (B.1)}} + \underbrace{\left\langle -\nabla \ell_t(\theta_*), \hat{\theta}_t - \theta_* \right\rangle \mathbf{1}_{A_t}}_{\text{TERM (B.2)}}. \end{aligned}$$

**For term (B.1).** We define  $\psi_t(z)$  as the gradient of Huber loss function (5), then we have

$$\psi_t(z) = \begin{cases} z & \text{if } |z| \leq \tau_t, \\ \tau_t & \text{if } z > \tau_t, \\ -\tau_t & \text{if } z < -\tau_t. \end{cases} \quad (26)$$

Based on (12) and (14),  $\nabla \ell_t(\theta) = -\psi_t(z_t(\theta)) \frac{X_t}{\sigma_t}$  and  $\nabla \tilde{\ell}_t(\theta) = -\psi_t(z_t(\theta) - z_t(\theta_*)) \frac{X_t}{\sigma_t}$ , which means

$$\begin{aligned} & \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) + \nabla \ell_t(\theta_*) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle \mathbf{1}_{A_t} \\ &= \left( \psi_t(z_t(\hat{\theta}_t)) - \psi_t(z_t(\theta_*)) - \psi_t(z_t(\hat{\theta}_t) - z_t(\theta_*)) \right) \frac{\left\langle X_t, \hat{\theta}_t - \theta_* \right\rangle}{\sigma_t} \mathbf{1}_{A_t}, \end{aligned}$$

we first analyze term  $\psi_t(z_t(\hat{\theta}_t)) - \psi_t(z_t(\theta_*)) - \psi_t(z_t(\hat{\theta}_t) - z_t(\theta_*))$ . When event  $A_t$  holds, similar to (18), we have

$$\left| z_t(\hat{\theta}_t) - z_t(\theta_*) \right| = \left| \frac{X_t^\top \hat{\theta}_t - X_t^\top \theta_*}{\sigma_t} \right| \leq \frac{\tau_t}{2},$$

then based on (26), we have  $\psi_t(z_t(\hat{\theta}_t) - z_t(\theta_*)) = z_t(\hat{\theta}_t) - z_t(\theta_*)$ . Next, we analyze the different situation of  $z_t(\theta_*)$ . When  $|z_t(\theta_*)| \leq \frac{\tau_t}{2}$ , we have

$$\left| z_t(\hat{\theta}_t) \right| = \left| z_t(\hat{\theta}_t) - z_t(\theta_*) + z_t(\theta_*) \right| \leq \left| z_t(\hat{\theta}_t) - z_t(\theta_*) \right| + |z_t(\theta_*)| \leq \frac{\tau_t}{2} + \frac{\tau_t}{2} = \tau_t, \quad (27)$$

based on (26) we have

$$\psi_t(z_t(\hat{\theta}_t)) - \psi_t(z_t(\theta_*)) - \psi_t(z_t(\hat{\theta}_t) - z_t(\theta_*)) = z_t(\hat{\theta}_t) - z_t(\theta_*) - z_t(\hat{\theta}_t) + z_t(\theta_*) = 0.$$

For another situation such that  $|z_t(\theta_*)| > \frac{\tau_t}{2}$ , based on (26) we have

$$\begin{aligned} & \psi_t(z_t(\hat{\theta}_t)) - \psi_t(z_t(\theta_*)) - \psi_t(z_t(\hat{\theta}_t) - z_t(\theta_*)) \\ & \leq \left| \psi_t(z_t(\hat{\theta}_t)) \right| + \left| \psi_t(z_t(\theta_*)) \right| + \left| \psi_t(z_t(\hat{\theta}_t) - z_t(\theta_*)) \right| \leq 3\tau_t. \end{aligned} \quad (28)$$

Combine this two situations (27) and (28), we have for TERM (B.1),

$$\begin{aligned} & \left\langle \nabla \tilde{\ell}_t(\hat{\theta}_t) + \nabla \ell_t(\theta_*) - \nabla \ell_t(\hat{\theta}_t), \hat{\theta}_t - \theta_* \right\rangle \mathbf{1}_{A_t} \\ &= \left( \psi_t(z_t(\hat{\theta}_t)) - \psi_t(z_t(\theta_*)) - \psi_t(z_t(\hat{\theta}_t) - z_t(\theta_*)) \right) \frac{\left\langle X_t, \hat{\theta}_t - \theta_* \right\rangle}{\sigma_t} \mathbf{1}_{A_t} \\ &\leq \left| \psi_t(z_t(\hat{\theta}_t)) - \psi_t(z_t(\theta_*)) - \psi_t(z_t(\hat{\theta}_t) - z_t(\theta_*)) \right| \left\| \frac{X_t}{\sigma_t} \right\|_{V_{t-1}^{-1}} \left\| \hat{\theta}_t - \theta_* \right\|_{V_{t-1}} \mathbf{1}_{A_t} \\ &\leq \mathbf{1} \left\{ |z_t(\theta_*)| \leq \frac{\tau_t}{2} \right\} 0 + \mathbf{1} \left\{ |z_t(\theta_*)| > \frac{\tau_t}{2} \right\} 3\tau_t \left\| \frac{X_t}{\sigma_t} \right\|_{V_{t-1}^{-1}} \left\| \hat{\theta}_t - \theta_* \right\|_{V_{t-1}} \mathbf{1}_{A_t} \\ &\leq \mathbf{1} \left\{ |z_t(\theta_*)| > \frac{\tau_t}{2} \right\} 3\tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\alpha} w_t \beta_{t-1} \\ &= \mathbf{1} \left\{ |z_t(\theta_*)| > \frac{\tau_t}{2} \right\} 3\tau_0 \sqrt{\alpha + \alpha w_t^2 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}} \beta_{t-1} \\ &\leq \mathbf{1} \left\{ |z_t(\theta_*)| > \frac{\tau_t}{2} \right\} 3\sqrt{\alpha + \frac{1}{8} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}} \beta_{t-1}, \end{aligned}$$

where the third inequality comes from the definition of event  $A_t$  and the last third inequality comes from  $\sigma_t \geq 2\sqrt{2}\|X_t\|_{V_{t-1}^{-1}}$ , such that  $\alpha w_t^2 = \left\|\frac{X_t}{\sigma_t}\right\|_{V_{t-1}^{-1}}^2 \leq 1/8$ . Then sum up for  $t$  round, and we have

$$\begin{aligned} \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) + \nabla \ell_s(\theta_*) - \nabla \ell_t(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle \mathbb{1}_{A_s} &\leq 3\sqrt{\alpha + \frac{1}{8}}\tau_0 \sum_{s=1}^t s^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \beta_{s-1} \mathbb{1} \left\{ |z_s(\theta_*)| > \frac{\tau_s}{2} \right\} \\ &\leq 3\sqrt{\alpha + \frac{1}{8}}\tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \max_{s \in [t+1]} \beta_s \right) \sum_{s=1}^t \mathbb{1} \left\{ |z_s(\theta_*)| > \frac{\tau_s}{2} \right\}. \end{aligned}$$

where the last inequality comes from that  $\max_{s \in [t]} \beta_{s-1} \leq \max_{s \in [t+1]} \beta_{s-1}$ . Same as Eq. (C.12) of [Huang et al. \[2024\]](#), by setting  $\tau_0$  as (17), with probability at least  $1 - \delta$ , for all  $t \geq 1$ , we have

$$\sum_{s=1}^t \mathbb{1} \left\{ |z_s(\theta_*)| > \frac{\tau_s}{2} \right\} \leq \frac{23}{3} \log \frac{2T^2}{\delta},$$

then we have for TERM (B.1), with probability at least  $1 - \delta$ , for all  $t \geq 1$ ,

$$\sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) + \nabla \ell_s(\theta_*) - \nabla \ell_t(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle \mathbb{1}_{A_t} \leq 23\sqrt{\alpha + \frac{1}{8}} \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \max_{s \in [t+1]} \beta_{s-1} \right). \quad (29)$$

**For term (B.2).** We first have

$$\begin{aligned} \sum_{s=1}^t \left\langle -\nabla \ell_s(\theta_*), \hat{\theta}_s - \theta_* \right\rangle \mathbb{1}_{A_s} &= \sum_{s=1}^t \psi_s(z_s(\theta_*)) \left\langle \frac{X_s}{\sigma_s}, \hat{\theta}_s - \theta_* \right\rangle \mathbb{1}_{A_s} \\ &\leq \left| \sum_{s=1}^t \psi_s(z_s(\theta_*)) \left\langle \frac{X_s}{\sigma_s}, \hat{\theta}_s - \theta_* \right\rangle \right|, \end{aligned} \quad (30)$$

where the last inequality comes from that  $\mathbb{1}_{A_s} \leq 1$ . Then, Lemma C.2 of [Huang et al. \[2024\]](#) (Self-normalized concentration) shows that by setting  $\tau_0$  as (17) and  $b = \max_{t \in [T]} \frac{\nu_t}{\sigma_t} \leq 1$ , we have with probability at least  $1 - 2\delta$ ,  $\forall t \geq 1$ ,

$$\left\| \sum_{s=1}^t \psi_s(z_s(\theta_*)) \frac{X_s}{\sigma_s} \right\|_{\tilde{V}_t^{-1}} \leq 8t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa} (\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \log \left( \frac{2T^2}{\delta} \right) \right)^{\frac{\varepsilon}{1+\varepsilon}}. \quad (31)$$

where  $\tilde{V}_t = \lambda\alpha I + \sum_{s=1}^t \frac{X_s X_s^\top}{\sigma_s}$  and  $\kappa = d \log \left( 1 + \frac{L^2 T}{\sigma_{\min}^2 \lambda \alpha d} \right)$ . Now we need to convert it into a 1-dimensional version, if  $Z_s$  is scalar, we have:

$$\left\| \sum_{s=1}^t \psi_s(z_s(\theta_*)) \frac{Z_s}{\sigma_s} \right\|_{\tilde{V}_t^{-1}}^2 = \frac{\left( \sum_{s=1}^t \psi_s(z_s(\theta_*)) \frac{Z_s^2}{\sigma_s^2} \right)}{\lambda\alpha + \sum_{s=1}^t \frac{Z_s^2}{\sigma_s^2}},$$

Based on inequality (31), we have

$$\begin{aligned} \frac{\left(\sum_{s=1}^t \psi_s(z_s(\theta_*)) \frac{Z_s}{\sigma_s}\right)^2}{\lambda\alpha + \sum_{s=1}^t \frac{Z_s^2}{\sigma_s^2}} &\leq \left(8t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left(\log\left(\frac{2T^2}{\delta}\right)\right)^{\frac{\varepsilon}{1+\varepsilon}}\right)^2 \\ \left|\sum_{s=1}^t \psi_s(z_s(\theta_*)) \frac{Z_s}{\sigma_s}\right| &\leq \sqrt{\left(\lambda\alpha + \sum_{s=1}^t \frac{Z_s^2}{\sigma_s^2}\right) \left(8t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left(\log\left(\frac{2T^2}{\delta}\right)\right)^{\frac{\varepsilon}{1+\varepsilon}}\right)^2}. \end{aligned}$$

Then based on AM-GM, we have

$$\left|\sum_{s=1}^t \psi_s(z_s(\theta_*)) \frac{Z_s}{\sigma_s}\right| \leq \frac{1}{4} \left(\lambda\alpha + \sum_{s=1}^t \frac{Z_s^2}{\sigma_s^2}\right) + \left(8t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left(\log\left(\frac{2T^2}{\delta}\right)\right)^{\frac{\varepsilon}{1+\varepsilon}}\right)^2,$$

where we use  $\sqrt{ab} \leq \frac{1}{4}a + b$ . Let  $Z_s = \langle X_s, \hat{\theta}_s - \theta_* \rangle$ , then put it back to (30), for TERM (2), we have with probability at least  $1 - 2\delta$ ,  $\forall t \geq 1$ ,

$$\begin{aligned} \sum_{s=1}^t \langle -\nabla \ell_s(\theta_*), \hat{\theta}_s - \theta_* \rangle &\leq \frac{1}{4} \left(\lambda\alpha + \sum_{s=1}^t \left\langle \frac{X_s}{\sigma_s}, \hat{\theta}_s - \theta_* \right\rangle^2\right) \\ &\quad + \left(8t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left(\log \frac{2T^2}{\delta}\right)^{\frac{\varepsilon}{1+\varepsilon}}\right)^2. \end{aligned} \quad (32)$$

Combine (29) and (32) together, with union bound we have with probability at least  $1 - 3\delta$ ,  $\forall t \geq 1$

$$\begin{aligned} \sum_{s=1}^t \langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \rangle \mathbf{1}_{A_s} &\leq 23\sqrt{\alpha + \frac{1}{8} \log \frac{2T^2}{\delta}} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left(\max_{s \in [t+1]} \beta_{s-1}\right) \\ &\quad + \frac{1}{4} \left(\lambda\alpha + \sum_{s=1}^t \left\langle \frac{X_s}{\sigma_s}, \hat{\theta}_s - \theta_* \right\rangle^2\right) + \left(8t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left(\log \frac{2T^2}{\delta}\right)^{\frac{\varepsilon}{1+\varepsilon}}\right)^2. \end{aligned}$$

□

## B.5 Proof of Lemma 1

*Proof.* Combining the results of Lemma 4 and Lemma 5, with applying the union bound, we obtain that, with probability at least  $1 - 4\delta$ , the following holds for all  $t \geq 1$ ,

$$\begin{aligned} &4\lambda S^2 + \sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 + 2 \sum_{s=1}^t \langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \rangle \mathbf{1}_{A_s} + \left(\frac{1}{\alpha} - 1\right) \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2 \\ &\leq 4\lambda S^2 + 6\alpha \left(8t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left(\log \frac{2T^2}{\delta}\right)^{\frac{\varepsilon}{1+\varepsilon}}\right)^2 + \frac{1}{4} \sum_{s=1}^t \left\| \theta_* - \hat{\theta}_s \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2 \\ &\quad + 46\sqrt{\alpha + \frac{1}{8} \log \frac{2T^2}{\delta}} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left(\max_{s \in [t+1]} \beta_{s-1}\right) + \frac{1}{2} \left(\lambda\alpha + \sum_{s=1}^t \left\langle \frac{X_s}{\sigma_s}, \hat{\theta}_s - \theta_* \right\rangle^2\right) \end{aligned}$$

$$\begin{aligned}
& + 2 \left( 8t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa} (\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \log \frac{2T^2}{\delta} \right)^{\frac{\varepsilon}{1+\varepsilon}} \right)^2 + \left( \frac{1}{\alpha} - 1 \right) \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2 \\
& \leq 94 \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \max_{s \in [t+1]} \beta_{s-1} \right) + \lambda(2 + 4S^2) + 152 \left( \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \right)^2,
\end{aligned}$$

where the least inequality comes from that  $\alpha = 4$ , and

$$\log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} = t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{2\kappa} (\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \left( \log \frac{2T^2}{\delta} \right)^{\frac{\varepsilon}{1+\varepsilon}}.$$

Then we choose  $\forall t \geq 1$ ,

$$\beta_t = 107 \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} + \sqrt{\lambda(2 + 4S^2)},$$

and we can verify that with probability at least  $1 - 4\delta$ , the following holds for all  $t \geq 1$ ,

$$\begin{aligned}
\beta_t^2 & \geq 4\lambda S^2 + \sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 + 2 \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle \mathbb{1}_{A_s} \\
& \quad + \left( \frac{1}{\alpha} - 1 \right) \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2.
\end{aligned} \tag{33}$$

Let  $B$  denote the event that the conditions in (33) hold  $\forall t \geq 1$ , then  $\mathbb{P}(B) \geq 1 - 4\delta$ . We now introduce a new event  $C$  that is define as

$$C \triangleq \left\{ \forall t \geq 1, \left\| \hat{\theta}_t - \theta_* \right\|_{V_{t-1}} \leq \beta_{t-1} \right\} = \bigcap_{t=0}^{\infty} A_t.$$

In the following we will show that if  $B$  is true,  $C$  must be true, which means  $\mathbb{P}(C) \geq \mathbb{P}(B) \geq 1 - 4\delta$  by mathematical induction. When  $t = 1$ ,  $A_1$  is true by definition such that  $\left\| \hat{\theta}_1 - \theta_* \right\|_{V_0} \leq \sqrt{4\lambda S^2} \leq \sqrt{\lambda(2 + 4S^2)} = \beta_0$ . Suppose that at iteration  $t$ , for all  $s \in [t]$ ,  $A_s$  is true, then we are going to show that  $A_{t+1}$  is also true.

$$\begin{aligned}
& \left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t}^2 \\
& \leq 4\lambda S^2 + \sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 + 2 \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle + \left( \frac{1}{\alpha} - 1 \right) \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2 \\
& = 4\lambda S^2 + \sum_{s=1}^t \left\| \nabla \ell_s(\hat{\theta}_s) \right\|_{V_s^{-1}}^2 + 2 \sum_{s=1}^t \left\langle \nabla \tilde{\ell}_s(\hat{\theta}_s) - \nabla \ell_s(\hat{\theta}_s), \hat{\theta}_s - \theta_* \right\rangle \mathbb{1}_{A_s} + \left( \frac{1}{\alpha} - 1 \right) \sum_{s=1}^t \left\| \hat{\theta}_s - \theta_* \right\|_{\frac{X_s X_s^\top}{\sigma_s^2}}^2 \\
& \leq \beta_t^2,
\end{aligned}$$

where the first inequality comes from Lemma 3, the second equality comes from that  $\forall s \in [t]$ ,  $A_s$  holds, and the last inequality comes from condition (33). As a result, we can conclude that all  $\{A_t\}_{t \geq 1}$  is true and thus we have  $\mathbb{P}(C) \geq 1 - 4\delta$ . And further we can find that

$$\sqrt{\frac{2 \left\| X_t \right\|_{V_{t-1}^{-1}}^2 \beta_{t-1}}{\sqrt{\alpha} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}} = \sqrt{\frac{107 \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} + \sqrt{\lambda(2 + 4S^2)}}{\tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}} \left\| X_t \right\|_{V_{t-1}^{-1}} \geq \sqrt{107} \left\| X_t \right\|_{V_{t-1}^{-1}} \geq 2\sqrt{2} \left\| X_t \right\|_{V_{t-1}^{-1}},$$



thus by setting

$$\sigma_t = \max \left\{ \nu_t, \sigma_{\min}, \sqrt{\frac{\|X_t\|_{V_{t-1}^{-1}}^2 \beta_{t-1}}{\tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}} \right\}, \quad \tau_t = \tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}, \quad \tau_0 = \frac{\sqrt{2\kappa}(\log 3T)^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}{\left(\log \frac{2T^2}{\delta}\right)^{\frac{1}{1+\varepsilon}}},$$

we obtain for any  $\delta \in (0, 1)$ , with probability at least  $1 - 4\delta$ , the following holds for all  $t \geq 1$ ,

$$\left\| \hat{\theta}_{t+1} - \theta_* \right\|_{V_t} \leq 107 \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} + \sqrt{\lambda(2 + 4S^2)}.$$

□

## C Regret Analysis

### C.1 Proof of Theorem 1

*Proof.* Let  $X_t^* = \arg \max_{\mathbf{x} \in \mathcal{X}_t} \mathbf{x}^\top \theta_*$ . Due to Lemma 1 and the fact that  $X_t^*, X_t \in \mathcal{X}_t$ , each of the following holds with probability at least  $1 - 4\delta$

$$\begin{aligned} \forall t \in [T], \quad X_t^{*\top} \theta_* &\leq X_t^{*\top} \hat{\theta}_t + \beta_{t-1} \|X_t^*\|_{V_{t-1}^{-1}} \\ \forall t \in [T], \quad X_t^\top \theta_* &\geq X_t^\top \hat{\theta}_t - \beta_{t-1} \|X_t\|_{V_{t-1}^{-1}}. \end{aligned}$$

By the union bound, the following holds with probability at least  $1 - 8\delta$ ,

$$\forall t \in [T], \quad X_t^{*\top} \theta_* - X_t^\top \theta_* \leq X_t^{*\top} \hat{\theta}_t - X_t^\top \hat{\theta}_t + \beta_{t-1} (\|X_t^*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}}) \leq 2\beta_{t-1} \|X_t\|_{V_{t-1}^{-1}},$$

where the last inequality comes from the arm selection criteria (11) such that

$$X_t^{*\top} \hat{\theta}_t + \beta_{t-1} \|X_t^*\|_{V_{t-1}^{-1}} \leq X_t^\top \hat{\theta}_t + \beta_{t-1} \|X_t\|_{V_{t-1}^{-1}}.$$

Hence the following regret bound holds with probability at least  $1 - 8\delta$ ,

$$\text{REG}_T = \sum_{t=1}^T X_t^{*\top} \theta_* - \sum_{t=1}^T X_t^\top \theta_* \leq 2\beta_T \sum_{t=1}^T \|X_t\|_{V_{t-1}^{-1}} = 4\beta_T \sum_{t=1}^T \sigma_t w_t,$$

where  $\beta_T = 107T^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \tau_0 \log \frac{2T^2}{\delta} + \sqrt{\lambda(2 + 4S^2)}$ .

Next we bound the sum of bonus  $\sum_{t=1}^T \sigma_t w_t$  separately by the value of  $\sigma_t$ . Recall the definition of  $\sigma_t$  in Algorithm 2, we decompose  $[T]$  as the union of three disjoint sets  $\mathcal{J}_1, \mathcal{J}_2$ ,

$$\mathcal{J}_1 = \{t \in [T] \mid \sigma_t \in \{\nu_t, \sigma_{\min}\}\}, \quad \mathcal{J}_2 = \left\{ t \in [T] \mid \sigma_t = \sqrt{\frac{\|X_t\|_{V_{t-1}^{-1}}^2 \beta_{t-1}}{\tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}} \right\}. \quad (34)$$

For the summation over  $\mathcal{J}_1$ , since  $\sigma_{\min} = \frac{1}{\sqrt{T}}$ ,

$$\sum_{t \in \mathcal{J}_1} \sigma_t w_t = \sum_{t \in \mathcal{J}_1} \max\{\nu_t, \sigma_{\min}\} w_t \leq \sqrt{\sum_{t=1}^T (\nu_t^2 + \sigma_{\min}^2)} \sqrt{\sum_{t=1}^T w_t^2} \leq \sqrt{2\kappa} \sqrt{\sum_{t=1}^T \nu_t^2 + 1},$$

where the last inequality comes from Lemma 7.

For the summation over  $\mathcal{J}_2$ , we first have

$$w_t^{-2} = \frac{\beta_{t-1}}{\tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}} = \frac{107 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \tau_0 \log \frac{2T^2}{\delta} + \sqrt{\lambda(2+4S^2)}}{\tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}} \leq \frac{\sqrt{\lambda(2+4S^2)}}{\tau_0} + 107 \log \frac{2T^2}{\delta},$$

we denote  $c = \frac{\sqrt{\lambda(2+4S^2)}}{\tau_0} + 107 \log \frac{2T^2}{\delta}$  and have

$$\sum_{t \in \mathcal{J}_2} \sigma_t w_t = \sum_{t \in \mathcal{J}_2} \frac{1}{w_t^2} \|X_t\|_{\tilde{V}_{t-1}^{-1}} w_t^2 \leq \frac{cL}{2\sqrt{\lambda}} \sum_{t \in \mathcal{J}_2} w_t^2 \leq \frac{cL\kappa}{\sqrt{\lambda}}. \quad (35)$$

where the last inequality comes from Lemma 7. Combine these two cases together, by choosing  $\delta = \frac{1}{8T}$ ,  $\lambda = d$ , we have

$$\begin{aligned} \text{REG}_T &\leq 4\beta_T \left( \sqrt{2\kappa} \sqrt{\sum_{t=1}^T \nu_t^2 + 1} + \frac{cL\kappa}{\sqrt{\lambda}} \right) \leq \tilde{\mathcal{O}} \left( \sqrt{dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}} \left( \sqrt{d \sum_{t=1}^T \nu_t^2 + \sqrt{d}} \right) \right) \\ &= \tilde{\mathcal{O}} \left( dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}} \sqrt{\sum_{t=1}^T \nu_t^2 + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}} \right). \end{aligned}$$

Thus we complete the proof.  $\square$

## C.2 Proof of Corollary 1

*Proof.* Proof of Corollary 1 is similar to Theorem 1, only need to change the condition  $\mathcal{J}_1$  in (34) to  $\mathcal{J}_1 = \{t \in [T] \mid \sigma_t \in \{\nu, \sigma_{\min}\}\}$  and for the summation over  $\mathcal{J}_1$ , since  $\sigma_{\min} = \frac{1}{\sqrt{T}}$ ,

$$\sum_{t \in \mathcal{J}_1} \sigma_t w_t = \sum_{t \in \mathcal{J}_1} \max\{\nu, \sigma_{\min}\} w_t \leq \sqrt{\sum_{t=1}^T (\nu^2 + \sigma_{\min}^2)} \sqrt{\sum_{t=1}^T w_t^2} \leq \sqrt{2\kappa\nu\sqrt{T+1}}, \quad (36)$$

where the last inequality comes from Lemma 7. Then combine condition (36) with condition (35), by choosing  $\delta = \frac{1}{8T}$ ,  $\lambda = d$ , we have

$$\begin{aligned} \text{REG}_T &\leq 4\beta_T \sum_{t=1}^T \sigma_t w_t \leq 4\beta_T \left( \sqrt{2\kappa\nu\sqrt{T+1}} + \frac{cL\kappa}{\sqrt{\lambda}} \right) \\ &\leq \tilde{\mathcal{O}} \left( \sqrt{dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}} \left( \sqrt{d\nu\sqrt{T}} + \sqrt{d} \right) \right) \\ &= \tilde{\mathcal{O}} \left( d\nu T^{\frac{1}{1+\varepsilon}} \right). \end{aligned}$$

Thus we complete the proof.  $\square$

## D Technical Lemmas

This section contains some useful technical lemmas that are used in the proofs.

## D.1 Concentrations

**Theorem 2** (Self-normalized concentration for scalar [Abbasi-Yadkori et al., 2012, Lemma 7]). *Let  $\{F_t\}_{t=0}^\infty$  be a filtration. Let  $\{\eta_t\}_{t=0}^\infty$  be a real-valued stochastic process such that  $\eta_t$  is  $F_t$ -measurable and  $R$ -sub-Gaussian. Let  $\{Z_t\}_{t=1}^\infty$  be a sequence of real-valued variables such that  $Z_t$  is  $F_{t-1}$ -measurable. Assume that  $V > 0$  be deterministic. For any  $\delta > 0$ , with probability at least  $1 - \delta$ :*

$$\forall t \geq 0, \quad \left| \sum_{s=1}^t \eta_s Z_s \right| \leq R \sqrt{2 \left( V + \sum_{s=1}^t Z_s^2 \right) \ln \left( \frac{\sqrt{V + \sum_{s=1}^t Z_s^2}}{\delta \sqrt{V}} \right)}.$$

**Lemma 6** (Lemma C.5 of Huang et al. [2024]). *By setting  $\tau_t = \tau_0 \frac{\sqrt{1+w_t^2}}{w_t} t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}$ , assume  $\mathbb{E} \left\{ \frac{\eta_t}{\sigma_t} \mid \mathcal{F}_{t-1} \right\} = 0$  and  $\mathbb{E} \left\{ \left| \frac{\eta_t}{\sigma_t} \right|^{1+\varepsilon} \mid \mathcal{F}_{t-1} \right\} \leq b^{1+\varepsilon}$ . Then with probability at least  $1 - \delta$ , we have  $\forall t \geq 1$ ,*

$$\sum_{s=1}^t \left( \min \left\{ \left| \frac{\eta_s}{\sigma_s} \right|, \tau_s \right\} \right)^2 \frac{w_s^2}{1+w_s^2} \leq t^{\frac{1-\varepsilon}{1+\varepsilon}} \left( \sqrt{\tau_0^{1-\varepsilon} (\sqrt{2\kappa}b)^{1+\varepsilon} (\log 3t)^{\frac{1-\varepsilon}{2}}} + \tau_0 \sqrt{2 \log \frac{2t^2}{\delta}} \right)^2.$$

## D.2 Potential Lemma

**Lemma 7** (Elliptical Potential Lemma). *Suppose  $U_0 = \lambda I$ ,  $U_t = U_{t-1} + X_t X_t^\top$ , and  $\|X_t\|_2 \leq L$ , denote  $\forall t \geq 1$ ,  $\|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2 \leq c_{\max}$ , then*

$$\sum_{t=1}^T \|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2 \leq 2 \max\{1, c_{\max}\} d \log \left( 1 + \frac{L^2 T}{\lambda d} \right). \quad (37)$$

*Proof.* First, we have the following decomposition,

$$U_t = U_{t-1} + X_t X_t^\top = U_{t-1}^{\frac{1}{2}} (I + U_{t-1}^{-\frac{1}{2}} X_t X_t^\top U_{t-1}^{-\frac{1}{2}}) U_{t-1}^{\frac{1}{2}}.$$

Taking the determinant on both sides, we get

$$\det(U_t) = \det(U_{t-1}) \det(I + U_{t-1}^{-\frac{1}{2}} X_t X_t^\top U_{t-1}^{-\frac{1}{2}}),$$

which in conjunction with Lemma 8 yields

$$\begin{aligned} \det(U_t) &= \det(U_{t-1}) (1 + \|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2) \geq \det(U_{t-1}) \left( 1 + \frac{1}{\max\{1, c_{\max}\}} \|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2 \right) \\ &\geq \det(U_{t-1}) \exp \left( \frac{1}{2 \max\{1, c_{\max}\}} \|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2 \right). \end{aligned}$$

Note that in the last inequality, we utilize the fact that  $\|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2 \leq c_{\max}$  and  $1 + x \geq \exp(x/2)$  holds for any  $x \in [0, 1]$ . By taking advantage of the telescope structure, we have

$$\sum_{t=1}^T \|U_{t-1}^{-\frac{1}{2}} X_t\|_2^2 \leq 2 \max\{1, c_{\max}\} \log \frac{\det(U_T)}{\det(U_0)} \leq 2 \max\{1, c_{\max}\} d \log \left( 1 + \frac{L^2 T}{\lambda d} \right),$$

where the last inequality follows from the fact that  $\text{Tr}(U_T) \leq \text{Tr}(U_0) + L^2 T = \lambda d + L^2 T$ , and thus  $\det(U_T) \leq (\lambda + L^2 T/d)^d$ .  $\square$

**Lemma 8** (Lemma 5 of [Zhao et al. \[2020\]](#)). *For any  $\mathbf{v} \in \mathbb{R}^d$ , we have*

$$\det(I + \mathbf{v}\mathbf{v}^\top) = 1 + \|\mathbf{v}\|_2^2.$$

### D.3 Useful Lemma for OMD

**Lemma 9** (Bregman proximal inequality [[Chen and Teboulle, 1993](#), Lemma 3.2]). *Let  $\mathcal{X}$  be a convex set in a Banach space. Let  $f : \mathcal{X} \mapsto \mathbb{R}$  be a closed proper convex function on  $\mathcal{X}$ . Given a convex regularizer  $\psi : \mathcal{X} \mapsto \mathbb{R}$ , we denote its induced Bregman divergence by  $\mathcal{D}_\psi(\cdot, \cdot)$ . Then, any update of the form*

$$\mathbf{x}_k = \arg \min_{\mathbf{x} \in \mathcal{X}} \{f(\mathbf{x}) + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_{k-1})\},$$

*satisfies the following inequality for any  $\mathbf{u} \in \mathcal{X}$ ,*

$$f(\mathbf{x}_k) - f(\mathbf{u}) \leq \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_{k-1}) - \mathcal{D}_\psi(\mathbf{u}, \mathbf{x}_k) - \mathcal{D}_\psi(\mathbf{x}_k, \mathbf{x}_{k-1}).$$