Corruption-Robust Variance-aware Algorithms for Generalized Linear Bandits under Heavy-tailed Rewards

Qingyuan Yu[1] **Euijin Baek**[2,4] **Xiang Li**[3] **Qiang Sun**[4,5]

[1]USTC [2]University of Alberta [3]University of Pennsylvania [4]University of Toronto [5] MBZUAI

Abstract

Stochastic linear bandits have recently received significant attention in sequential decision-making. However, real-world challenges such as heavy-tailed noise, reward corruption, and nonlinear reward functions remain difficult to address. To tackle these difficulties, we propose GAdaOFUL, a novel algorithm that leverages adaptive Huber regression to achieve robustness in generalized linear models (GLMs), where rewards can be nonlinear functions of features. GAdaOFUL achieves a state-of-the-art variance-aware regret bound, scaling with the square root of the cumulative reward variance over time, plus an additional term proportional to the level of corruption. The algorithm adapts to problem complexity, yielding improved regret when the cumulative variance is small. Simulation results demonstrate the robustness and effectiveness of GAdaOFUL in practice. The code is available at `https://github.com/NeXAIS/GAdaOFUL`.

1 INTRODUCTION

In online decision-making, stochastic linear bandits have been a powerful framework for balancing exploration and exploitation in sequential processes [Lattimore and Szepesvári, 2020]. However, many real-world applications involve nonlinear relationships between actions and rewards, limiting the effectiveness of standard linear bandit algorithms. To address this limitation, generalized linear models (GLMs) have emerged as a natural extension, allowing expected rewards to be modeled as nonlinear functions of input features via a link function [Filippi et al., 2010]. This flexibility makes GLMs well-suited for applications such as online advertisements and recommendation systems [Li et al., 2010, 2012].

Despite their advantages, designing algorithms that perform well in real-world settings remains challenging. We present some key challenges below.

- **Heavy-tailed rewards:** Many existing methods assume that rewards follow a sub-Gaussian distribution [Filippi et al., 2010, Li et al., 2012, 2017, Zhou et al., 2019, Lu et al., 2021], simplifying analysis but failing to capture the variability commonly observed in practice. In financial markets, for instance, extreme returns occur far more frequently than expected under a normal distribution, a characteristic known as heavy-tailed behavior [Cont and Bouchaud, 2000, Foss et al., 2011]. Real-world reward distributions often exhibit such properties, leading to poor performance for algorithms based on sub-Gaussian assumptions [Bubeck et al., 2013].
- **Adversarial corruptions:** Bandit systems are also susceptible to adversarial manipulations or corrupt reward signals, which can significantly degrade their performance. In recommendation systems, for example, adversaries may inject false feedback or manipulate click-through rates, misleading the algorithm and deteriorating recommendations for genuine users. While recent research has developed corruption-robust algorithms [Lykouris et al., 2018, Kapoor et al., 2019, Bogunovic et al., 2020], many of these approaches do not account for heavy-tailed rewards, where extreme outcomes occur more frequently. Thus, a fundamental challenge remains: designing algorithms that can simultaneously handle corruption and heavy-tailed noise, ensuring robustness in practical settings. Notably, Jun et al. [2018] demonstrate how adversarial manipulations of rewards can dramatically increase regret, underscoring the urgency of robust defenses against adversarial corruptions.
- **Worst-case analysis:** Traditional bandit algorithms often rely on worst-case regret bounds, which tend to be overly conservative. In contrast, variance-aware regret bounds allow regret to scale with the observed

reward variance, rather than an upper bound on reward magnitudes. This provides a more refined measure of problem complexity, adapting to the actual variability in rewards. When cumulative reward variance is small, the regret naturally decreases, indicating that the problem is easier to solve. Consequently, variance-aware regret bounds lead to tighter performance guarantees [Zhang et al., 2021, Dai et al., 2022, Di et al., 2023, Li and Sun, 2024].

This paper investigates whether it is possible to design generalized linear bandit algorithms that simultaneously handle heavy-tailed rewards, protect against adversarial corruption, and incorporate variance-aware regret. Specifically, we consider a setting where the reward follows a generalized linear model (GLM) with heavy-tailed noise. At each round, the algorithm selects an arm ϕ_t from a decision set D_t and observe a reward y_t, which satisfies

$$y_t = f(\langle \phi_t, \theta^* \rangle) + \epsilon_t + c_t,$$

where θ^* is an unknown d-dimensional true parameter, c_t represents adversarial corruption, and ϵ_t denotes the heavy-tailed noise with bounded variance ν_t^2. Our goal is to minimize the cumulative regret over T rounds, defined as the total loss from not selecting the optimal action sequence:

$$\mathrm{Reg}(T) := \sum_{t=1}^{T} \left[\sup_{\phi \in D_t} f(\langle \phi, \theta^* \rangle) - f(\langle \phi_t, \theta^* \rangle) \right]. \quad (1)$$

Our Contributions. While many existing works address some of these challenges individually (see Table 1 for the most relevant studies), we propose the first algorithm that successfully tackles all three key aspects: handling heavy-tailed rewards, resisting adversarial corruption, and adapting to variance-aware regret.

Specifically, we introduce GAdaOFUL (Generalized Adaptive Huber Regression-based OFUL, see Algorithm 1), a novel algorithm designed to address all three challenges simultaneously. At its core, GAdaOFUL leverages adaptive Huber regression [Sun et al., 2020, Sun, 2021, Li and Sun, 2024] to mitigate the impact of heavy-tailed noise. To handle potential adversarial corruption, the algorithm carefully scales each residual error by selecting an appropriate variance parameter σ_t^2. More precisely, σ_t^2 depends on the reward variance ν_t^2, a sample importance weight w_t, and the total corruption level $C = \sum_{t=1}^{T} |c_t|$. Here, w_t quantifies the significance of the t-th sample in improving prediction accuracy.

*Ye et al. [2023] consider a general and abstract class $\mathcal{G}$ where $\mathcal{G}$ has a bounded Eluder dimension. Their general regret is $\widetilde{O}(d_0\sqrt{T} + d_0 \cdot C)$ where d_0 is the Eluder dimension of $\mathcal{G}$. If $\mathcal{G}$ is the GLM family, $d_0 = O(d)$.

†Xue et al. [2024] consider a general heavy-tailed setting where the rewards have only $(1+\epsilon)$-th order moments. For a fair comparison, we translate the results to the finite-variance setting.

Beyond integrating these elements, our work introduces two key technical novelties:

- **Nonlinear Extension Framework**: Unlike most previous works that focus on linear settings, we introduce a novel integral-based loss function (Eq. (2)) that extends naturally to nonlinear GLM cases. Notably, this loss function remains convex, enabling the use of efficient convex optimization techniques, such as (stochastic) gradient descent. This formulation significantly broadens GAdaOFUL's applicability to more complex nonlinear scenarios while maintaining computational efficiency.
- **Corruption-Robust Analysis**: Our proof employs a reduction approach, introducing an auxiliary problem that reformulates the corrupted-reward setting into an equivalent corruption-free counterpart. By carefully selecting hyperparameters, we establish a direct connection between optimality in uncorrupted and corrupted cases. Leveraging constrained convex optimization, we derive state-of-the-art regret bounds that remain valid even under adversarial conditions. See Section 4.3 for a detailed proof overview.

We rigorously prove that GAdaOFUL achieves a state-of-the-art regret bound of

$$\widetilde{\mathcal{O}}\left(d\sqrt{\sum_{t\in[T]} \nu_t^2} + dC + d \right),$$

even in the presence of heavy-tailed noise and adversarial corruption. Here, d represents the feature dimension, T is the total number of rounds, and $\widetilde{\mathcal{O}}(\cdot)$ hides constant factors and logarithmic terms in T. To the best of our knowledge, this work is the first to unify GLMs, heavy-tailed noise, corruption, and variance-awareness into a comprehensive framework. Table 1 provides a comparison with recent state-of-the-art algorithms.

2 PRELIMINARIES

Notation. We denote the ℓ_2-norm in $\mathbb{R}^d$ by $\|\cdot\|$, and $\mathrm{Ball}_d(B)$ represents the ℓ_2-norm ball in $\mathbb{R}^d$ with radius $B > 0$. For a positive definite matrix $H \in \mathbb{R}^{d\times d}$, we define $\|x\|_H = \sqrt{x^T H x}$ for a vector $x \in \mathbb{R}^d$. Additionally, for two positive semidefinite matrices H_1 and H_2, we write $H_1 \succeq H_2$ if $H_2 - H_1$ is positive semidefinite.

Generalized Linear Models (GLMs). Generalized Linear Models (GLMs), first introduced by Nelder and Wedderburn [1972], extend traditional linear regression by allowing more flexible relationships between the response variable y and predictor variables (or feature vector) ϕ. In GLMs, the relationship is modeled through a linear predictor, a linear

Table 1: Theoretical performance comparison of different methods in most related works. For an introduction to earlier studies (which often exhibit poorer performance), see the references in these papers. In this table, ν_t^2 denotes the conditional reward variance for the t-th reward and $C = \sum_{t=1}^{T} |c_t|$ represents the total corruption level with $C \vee 1 = \max\{C, 1\}$. The worst-case lower bound is $\Omega(d\sqrt{T} + dC)$ [Lattimore and Szepesvári, 2020, Bogunovic et al., 2021].

Method	Reward	Noise	Corruption	Variance-aware	Regret
[He et al., 2022b]	linear	Sub-gaussian	✓	✗	$\widetilde{O}(d\sqrt{T} + d \cdot C)$
[Ye et al., 2023]	GLM*	Sub-gaussian	✓	✗	$\widetilde{O}(d\sqrt{T} + d \cdot C)$
[Li and Sun, 2024]	linear	Finite variance	✗	✓	$\widetilde{O}(d\sqrt{\sum_{t=1}^{T} \nu_t^2} + d)$
[Xue et al., 2024]	GLM	Finite variance†	✗	✗	$\widetilde{O}(d\sqrt{T})$
Ours, Theorem 4	GLM	Finite variance	✓	✓	$\widetilde{O}(d\sqrt{\sum_{t=1}^{T} \nu_t^2} + d \cdot C \vee 1)$

combination of the predictors and unknown coefficients θ. Unlike linear regression, where the conditional mean equals the linear predictor $\langle \phi, \theta \rangle$, GLMs link the linear predictor to the conditional mean through a link function f:

$$\mathbb{E}[y|\phi] = f(\langle \phi, \theta \rangle).$$

The link function $f(\cdot)$ is typically an increasing, differentiable function. By choosing different link functions, GLMs can model various types of data. For example, Poisson regression is suitable for count data, where the link function is $f(x) = \exp(x)$ [Coxe et al., 2009]. For binary outcomes, logistic regression is a natural choice, using the logistic function $f(x) = \frac{\exp(x)}{1+\exp(x)}$ [Hilbe, 2011].

Heavy-Tailed Noise. Unlike standard bandit models that assume sub-Gaussian or bounded noise, we consider a more general and realistic setting where the stochastic noise in rewards may exhibit heavy tails. Formally, we assume that the noise sequence $\{\epsilon_t\}$ forms a martingale difference sequence adapted to the filtration $\{\mathcal{F}_{t-1}\}$, satisfying $\mathbb{E}[\epsilon_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\epsilon_t^2 \mid \mathcal{F}_{t-1}] = \nu_t^2$. This setting accommodates heavy-tailed but variance-bounded noise, which can still exhibit occasional large deviations while being more realistic for applications such as finance, crowdsourcing, and networked systems. This type of noise model has been explored in the context of bandits by Bubeck et al. [2013], and further studied in generalized linear settings by Xue et al. [2024], who developed robust estimation methods for heavy-tailed generalized linear bandits.

Adversarial Corruption. In addition to stochastic noise, we consider adversarial corruption to the reward, modeled by an additive term c_t. We make no assumptions about the distribution or structure of $\{c_t\}$ beyond the total ℓ_1 budget being bounded, i.e., $\sum_{t=1}^{T} |c_t| \leq C$ for some known constant $C > 0$. Importantly, the corruption is allowed to be adaptive: the adversary may observe the realized (possibly noisy) reward before choosing c_t. This corruption model captures various real-world scenarios such as data poisoning, faulty sensor readings, or malicious manipulations. Similar corruption-resilient formulations have been explored in stochastic multi-armed bandits [Lykouris et al., 2018], linear bandits [Bogunovic et al., 2021, He et al., 2022a], though these works typically assume bounded or sub-Gaussian noise.

Problem Setting and Model Assumptions. We study a stochastic generalized linear bandit model with heavy-tailed noise and adversarial corruption. Let $\{D_t\}_{t\geq 1}$ represent a predetermined sequence of decision sets and $\{\mathcal{F}_t\}_{t\geq 1}$ a filtration corresponding to the information available up to time t. At each round t, the agent selects an action $\phi_t \in D_t$ and observes the reward

$$y_t = f(\langle \phi_t, \theta^* \rangle) + \epsilon_t + c_t,$$

where $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector, ϵ_t is a martingale difference noise with $\mathbb{E}[\epsilon_t|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\epsilon_t^2|\mathcal{F}_{t-1}] = \nu_t^2$, and c_t is an adversarial corruption. The cumulative corruption level is $C := \sum_{t=1}^{T} |c_t|$, which is assumed to be known. Additionally, $\|\theta^*\| \leq B$ for some bound B, and both ϕ_t and ν_t are $\mathcal{F}_{t-1}$-measurable with $\|\phi_t\| \leq L$. The function f, referred to as the activation function Zhao et al. [2023], is assumed to be an increasing, differentiable function on $[-BL, BL]$, with constants $k, K \in \mathbb{R}$ such that $0 < k \leq f'(z) \leq K$ for all $z \in [-BL, BL]$.

3 THE GADAOFUL METHOD

In this section, we introduce our algorithm, GAdaOFUL, designed to tackle heavy-tailed noise and adversarial attacks. Heavy-tailed noise refers to rewards with finite variances, while adversarial attacks involve deliberate corruption intended to degrade the reward signals. Our algorithm is applicable to GLMs and achieves state-of-the-art regret.

Adaptive Huber regression modified for GLMs. Our algorithm, GAdaOFUL, is based on adaptive Huber regression [Sun et al., 2020], utilizing the pseudo-Huber loss function

[Sun, 2021] to tackle heavy-tailed issues. Specifically, the Pseudo-Huber loss is defined as $\ell_\tau(x) = \tau(\sqrt{\tau^2 + x^2} - \tau)$. This loss serves as a smooth approximation to the Huber loss [Huber, 1992], transitioning between quadratic penalties for small residuals and linear penalties for larger ones, making it differentiable everywhere.

However, the original adaptive Huber regression is designed for linear models [Sun et al., 2020, Sun, 2021], which limits its theoretical grounding for nonlinear models like GLMs. To address this nonlinearity, we modify the Pseudo-Huber loss to better mitigate its effects. The derivative of $\ell_\tau(x)$ with respect to x is given by $\ell'_\tau(x) = \frac{\tau x}{\sqrt{\tau^2+x^2}}$, and we reformulate the loss in terms of its derivative $\ell'_\tau(x)$. At each round, GAdaOFUL first estimates the ground-truth parameter θ^* by minimizing the following optimization problem:

$$\theta_t := \operatorname{argmin}_{\theta \in \mathrm{Ball}_d(B)} L_t(\theta), \tag{2}$$
$$L_t(\theta) := \frac{\lambda k}{2}\|\theta\|^2 - \sum_{s=1}^{t} \frac{1}{\sigma_s} \int_0^{\langle \phi_s, \theta \rangle} \frac{\tau_s z_s(u)}{\sqrt{\tau_s^2 + z_s^2(u)}}\, du.$$

Here, $k > 0$ is a lower bound for $\min_{|x| \le BL} f'(x)$, $z_s(u) = (y_s - f(u))/\sigma_s$ is the scaled residual error, and σ_t^2 represents surrogate conditional variances.

Remark 1. *The rationale behind using this loss function* (2) *lies in its ability to handle the nonlinearity of GLMs while retaining desirable properties from the linear case. In general, a GLM can be interpreted as a form of weighted linear regression. At a high level, our proposed integral-based loss ensures that the weights used are of the same order, determined by f' , and bounded within the interval $[k, K]$ as per our setup. This point can be verified by computing the derivative and Hessian of the new loss function. Notably, the derivative and Hessian of the proposed loss resemble those of the linear case, as highlighted in [Li and Sun, 2024]. This resemblance enables a natural extension of proof techniques and results from the linear setting to the nonlinear one.*

It is straightforward to verify that

- $L_t(\theta)$ is convex in θ, so the optimization problem in (2) can be efficiently solved by convex solvers.
- $L_t(\theta)$ depends on the adaptive (or varying) values of τ_t, which are essential for achieving optimal regret, as shown by Li and Sun [2024] for non-corrupted cases.
- If $f(\cdot)$ is the identity function, $L_t(\theta)$ reduces to the one used by Li and Sun [2024].

Algorithm description. Next, we outline the steps of the GAdaOFUL algorithm. At round t, construct a confidence ellipsoid:

$$\mathcal{C}_{t-1} := \{\theta \in \mathrm{Ball}_d(B) : \|\theta - \theta_{t-1}\|_{H_{t-1}} \leqslant \beta_{t-1}\},$$

Algorithm 1 Generalized Adaptive Huber regression based OFUL (GAdaOFUL).

Constants: $\lambda = d/B^2, \sigma_{\min} = \frac{1}{\sqrt{T}}, m_0 = \left(6\sqrt{3\log\frac{2T^2}{\delta}}\right)^{-1}$, and $m_1 = \left(42\log\frac{2T^2}{\delta}\right)^{-1}$.
Initialization: $H_0 = \lambda I, \theta_0 = \mathbf{0}, \beta_0 = \sqrt{\lambda}B$.
for $t = 1$ **to** T **do**
  Construct the confidence set C_{t-1}.
  Solve $(\phi_t, \cdot) = \operatorname{argmax}_{\phi \in D_t, \theta \in C_{t-1}} \langle \phi, \theta \rangle$.
  Play ϕ_t and observe (y_t, ν_t).
  Set σ_t, w_t and τ_t according to (19) and record $\{\sigma_s, w_s, \tau_s : 1 \le s \le t\}$.
  Compute θ_t according to (2).
  Define β_t and set $H_t = H_{t-1} + \frac{\phi_t \phi_t^\top}{\sigma_t^2}$.
end for

where H_t is the shape matrix, and β_t is the exploration radius. It can be proven that θ^* lies within $\mathcal{C}_t$ with high probability for all $t \ge 0$. Based on this confidence set, select ϕ_t by maximizing the inner product $\langle \phi, \theta \rangle$:

$$(\phi_t, \cdot) = \operatorname{argmax}_{\phi \in D_t, \theta \in C_{t-1}} \langle \phi, \theta \rangle \tag{3}$$

Play the chosen arm, then observe the reward y_t and its conditional variance v_t^2. Next, compute (σ_t, w_t, τ_t) according to (19), and solve the optimization problem in (2) to update θ. Finally, update the shape matrix H_t and the exploration radius β_t, then proceed to the next round.

Parameter selection. In the following, we specify the parameters used in Algorithm 1. The parameter σ_t represents the surrogate conditional variance, w_t quantifies the importance of the t-th sample (y_t, ϕ_t, σ_t), and τ_t is a robustification parameter used in Pseudo-Huber regression. Their expressions are given as follows:

$$\begin{aligned} \sigma_t &= \max\left\{\nu_t, \sigma_{\min}, \|\phi_t\|_{H_{t-1}^{-1}}/m_0, \alpha\|\phi_t\|_{H_{t-1}^{-1}}^{1/2}\right\}, \\ w_t &= \left\|\frac{\phi_t}{\sigma_t}\right\|_{H_{t-1}^{-1}}, \text{ and } \tau_t = \tau_0 \frac{\sqrt{1+w_t^2}}{w_t}, \end{aligned} \tag{19}$$

where

$$\alpha = \max\left\{\frac{\sqrt{LBK}}{m_1^{1/4} d^{1/4}}, C^{\frac{1}{2}} \kappa^{-\frac{1}{4}}\right\},$$

$C = \sum_{t=1}^{T} |c_t|$ is the corruption level, and $\kappa = d \cdot \log\left(1 + TL^2/(d\lambda\sigma_{\min}^2)\right)$ is a constant.

We briefly explain the selection of σ_t, the most important parameter. First, we set $\sigma_t \ge \nu_t$ to ensure it is larger than the true conditional variance, which keeps the conditional

variance of y_t/σ_t always less than 1. Note that we do not require ν_t itself to be known; any valid upper bound suffices to ensure the theoretical properties of the adaptive Huber loss. We also impose $\sigma_t \geq \sigma_{\min}$ to avoid numerical instability. Additionally, we require $\sigma_t \geq \|\phi_t\|_{H_{t-1}^{-1}}/m_0$ to ensure that the importance measure w_t remains bounded by a constant m_0. Finally, we set $\sigma_t \geq \alpha\|\phi_t\|_{H_{t-1}^{-1}}^{1/2}$ with a carefully chosen α to mitigate potential corruptions. This α depends on $C^{\frac{1}{2}}\kappa^{-\frac{1}{4}}$, which accounts for the effects of corruption—a factor not considered by Li and Sun [2024].

4 REGRET ANALYSIS

In this section, we present the theoretical analysis for Algorithm 1.

4.1 NON-CORRUPTED CASE

To begin, we consider the non-corrupted case where $C = 0$ and show the results in Theorem 2.

Theorem 2 (Uncorrupted Case)**.** *Assume $C = 0$. Let $\lambda = d/B^2$ and $\kappa = d \cdot \log\left(1 + \frac{TL^2}{d\lambda\sigma_{\min}^2}\right)$. If $\tau_0\sqrt{\log\left(\frac{2T^2}{\delta}\right)} \geq \max\{\sqrt{2\kappa}, 2\sqrt{d}\}$, then with probability $1-3\delta$, it holds that for all $0 \leq t \leq T$,*

$$\|\theta_t - \theta^*\|_{H_t} \leq \beta_t,$$

where

$$\beta_t = \frac{32}{k}\left[\frac{\kappa}{\tau_0} + \sqrt{\kappa \log\left(\frac{2t^2}{\delta}\right)} + \tau_0 \log\left(\frac{2t^2}{\delta}\right)\right] + 5\sqrt{\lambda}B. \tag{4}$$

Then, with probability at least $1 - 3\delta$, we have

$$\mathrm{Reg}(T) \leq 4K\beta_T\left[\sqrt{\kappa} \cdot \sqrt{\sum_{t\in[T]} \nu_t^2 + 1} + \frac{L\kappa}{m_0^2\sqrt{\lambda}} + \frac{LBK\kappa}{\sqrt{m_1 d}}\right]. \tag{5}$$

Similar to previous work [Li and Sun, 2024], in Theorem 2, we demonstrate that (i) the true parameter θ^* falls within the constructed confidence intervals $\mathcal{C}_t$ with high probability, and (ii) the regret scales as $\mathrm{Reg}(T) = \widetilde{O}(d\sqrt{\sum_{t=1}^T \nu_t^2} + d)$, which matches the results for linear bandits with heavy-tailed rewards [Li and Sun, 2024]. The key takeaway is that even when considering a nonlinear GLM model for rewards, using our modified loss in (2) allows us to maintain the same level of regret performance. The proof of Theorem 2 largely follows the approach in [Li and Sun, 2024], utilizing a quadratic approximation of the nonlinear loss $L_t(\theta)$; see the appendix for the details.

4.2 CORRUPTED CASE

Next, we consider the case with corruption, where $C > 0$ is assumed to be known.[1] Theorem 3 guarantees the high probability coverage, while Theorem 4 upper bounds the regret.

Theorem 3. *Let $\kappa = d \cdot \log\left(1 + TL^2/(d\lambda\sigma_{min}^2)\right)$. If $\tau_0\sqrt{\log(2T^2/\delta)} \geq \max\{\sqrt{2\kappa}, 2\sqrt{d}\}$,then with probability $1 - 3\delta$, it holds that, for all $0 \leq t \leq T$,*

$$\|\theta_t - \theta^*\|_{H_t} \leq \beta_t,$$

where

$$\beta_t = \frac{4\sqrt{\kappa}}{k} + \frac{32}{k}\left[\frac{\kappa}{\tau_0} + \sqrt{\kappa \log\frac{2t^2}{\delta}} + \tau_0 \log\frac{2t^2}{\delta}\right] + 5\sqrt{\lambda}B. \tag{6}$$

Theorem 3 demonstrates that θ^* falls within the set $\mathcal{C}_t := \{\theta \in \mathrm{Ball}_d(B) : \|\theta - \theta_t\|_{H_t} \leqslant \beta_t\}$ for any $t \geq 1$ with high probability. In contrast to the non-corrupted case, the presence of corruption introduces an additional constant term of $\frac{4\sqrt{\kappa}}{k}$ to β_t. However, this term is negligible when $\tau_0 = \widetilde{O}(\sqrt{d})$, which also leads to $\beta_t = \widetilde{O}(\sqrt{d})$.

With the above confidence region, the regret bound of Algorithm 1 for corrupted cases is explicitly given as follows.

Theorem 4. *Then with probability at least $1 - 3\delta$,*

$$\mathrm{Reg}(T) \leq 4K\beta_T\left[\sqrt{\kappa} \cdot \sqrt{\sum_{t\in[T]} \nu_t^2 + 1} + \frac{L\kappa}{m_0^2\sqrt{\lambda}} + \frac{LBK\kappa}{\sqrt{m_1 d}} + 2C\sqrt{\kappa}\right].$$

where β_T is defined in (6).

By setting λ and τ_0 carefully, $\mathrm{Reg}(T)$ is simplified to $\widetilde{O}(d\sqrt{\sum_{t=1}^n \nu_t^2} + d \cdot C \vee 1)$ where $C \vee 1 = \max\{C, 1\}$.

Corollary 5. *Let $\lambda = d/B^2$ and $\tau_0 = \max\left\{\sqrt{2\kappa}, 2\sqrt{d}\right\}/\sqrt{\log(2T^2/\delta)}$. The regret bound in Theorem 4 becomes*

$$\mathrm{Reg}(T) = \widetilde{\mathcal{O}}\left(\frac{Kd}{k}\sqrt{\sum_{t\in[T]} \nu_t^2} + \frac{Kd}{k} \cdot \max\{LBK, C\}\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ hides constant factors and logarithmic dependence on T.

[1] In fact, it is sufficient to have a valid upper bound for $\sum_{t=1}^T |c_t|$. If C is unknown, one can employ the doubling trick [Besson and Kaufmann, 2018] to estimate a valid upper bound. After a logarithmic number of guesses, we can reliably determine a true upper bound.

Comparison with previous works. We emphasize that the regret bound in Theorem 4 and Corollary 5 is the first variance-aware regret to simultaneously address heavy-tailed rewards, adversarial corruption, and nonlinear settings. While numerous studies have explored bandits under the GLM framework and light-tailed noise scenarios, to our knowledge, none have accomplished all three aspects. For a comparison among the most related and competitive results, see Table 1. We discuss their differences below.

In the absence of corruptions (i.e., $C = 0$), the regret bound simplifies to $\widetilde{\mathcal{O}}\left(d\sqrt{\sum_{t\in[T]}\nu_t^2} + d\right)$, aligning with the results obtained by AdaOFUL [Li and Sun, 2024]. However, AdaOFUL is restricted to linear bandit problems, while GAdaOFUL is applicable to the more general GLM setting. Recently, Xue et al. [2024] introduced an algorithm that addresses heavy-tailed noise within the GLM context; however, their results do not account for adversarial corruption, and their regret fails to be variance-aware, thus lacking adaptability to the problem's difficulty.

In cases where corruptions are present (i.e, $C > 0$), He et al. [2022b] proposed the CW-OFUL algorithm, which achieves a minimax optimal regret bound of $\widetilde{\mathcal{O}}\left(d\sqrt{T} + d \cdot C\right)$. This bound is also obtained by Ye et al. [2023] for GLM rewards setting. However, both of them pertain to the worst-case scenario. We argue that our bound, $\widetilde{\mathcal{O}}\left(d\sqrt{\sum_{t\in[T]}\nu_t^2} + d \cdot (C \vee 1)\right)$, is significantly more adaptive than theirs. Specifically, if we assume that the variance ν_t remains constant and significant (i.e., $\nu_t = \Theta(1)$ for all $t \geq 1$), our regret reduces to theirs, implying that our approach is also minimax optimal in the worst case.

4.3 PROOF SKETCH

At the end of this section, we provide a proof sketch of Theorem 3 and 4.

Proof of Theorem 3. The proof of Theorem 3 consists of two key steps. In the first step, we focus on bounding $\|\nabla L_T(\theta^*)\|_{H_t^{-1}}$. Thanks to the loss function (2), the gradient estimator can be written as

$$\nabla L_T(\theta) = \lambda k\theta - \sum_{t=1}^{T} \frac{\tau_t z_t(\theta)}{\sqrt{\tau_t^2 + z_t^2(\theta)}} \frac{\phi_t}{\sigma_t}$$

where $z_t(\theta) = \dfrac{y_t - f(\langle \phi_t, \theta\rangle)}{\sigma_t}$ is the standardized residual error. With this expression, we show that, with high probability,

$$\|\nabla L_T(\theta^*)\|_{H_t^{-1}} = \widetilde{O}\left(\frac{\kappa}{\tau_0} + \tau_0 + \sqrt{\kappa} + B\sqrt{\lambda}\right) \quad (7)$$

holds for any $T \geq 1$. In other words, $\|\nabla L_T(\theta^*)\|_{H_t^{-1}}$ is uniformly bounded in terms of τ_0. To achieve a smaller bound, we should tune $\tau_0 = \widetilde{O}(\sqrt{\kappa})$, which is precisely what we set in Theorem 3.

In the second step, we aim to control $\nabla^2 L_T(\theta)$. More specifically, we demonstrate that, with high probability, for all $T \geq 0$ and any $\|\theta\| \leq B$,

$$\nabla^2 L_T(\theta) \succeq \frac{k}{4} H_T. \quad (8)$$

A similar lower bound to (8) appears in previous work [Li and Sun, 2024]; however, our loss formulation in (2) enables its extension to GLMs. Combining these two steps in (7) and (8), we apply the mean value theorem and obtain that

$$\nabla L_T(\theta_T) - \nabla L_T(\theta^*) = \nabla^2 L_T(\theta_T^*)(\theta_T - \theta^*)$$

for some vector θ_T^* satisfying $\|\theta_T^*\| \leq B$. The first-order stationary condition of the constrained convex optimization in (2) implies that $\langle \nabla L_T(\theta_T), \theta_T - \theta^*\rangle \leq 0$.

Combining all the above results, we have:

$$\begin{aligned}
\frac{k}{4}\|\theta_T - \theta^*\|_{H_T}^2 \leq& \langle \nabla^2 L_T(\theta_T^*)(\theta_T - \theta^*), \theta_T - \theta^*\rangle \\
=& \langle \nabla L_T(\theta_T) - \nabla L_T(\theta^*), \theta_T - \theta^*\rangle \\
\leq& \langle -\nabla L_T(\theta^*), \theta_T - \theta^*\rangle \\
\leq& \|\nabla L_T(\theta^*)\|_{H_t^{-1}} \cdot \|\theta_T - \theta^*\|_{H_T}
\end{aligned}$$

This leads to the implication:

$$\begin{aligned}
\|\theta_T - \theta^*\|_{H_T} \leq& \frac{4}{k}\|\nabla L_T(\theta^*)\|_{H_t^{-1}} \\
=& \widetilde{O}\left(\frac{\kappa}{\tau_0} + \tau_0 + \sqrt{\kappa} + B\sqrt{\lambda}\right). \quad (9)
\end{aligned}$$

Proof of Theorem 4. By the Lipschitz continuity of the (nonlinear) link function f (i.e., $\sup_{|x|\leq BL} f'(x) \leq K$), we bound the regret in (1) by

$$\mathrm{Reg}(T) \leq K \sum_{t=1}^{T} \left[\sup_{\phi\in\mathcal{D}_t} \langle \phi, \theta^*\rangle - \langle \phi_t, \theta^*\rangle\right].$$

To bound the right-hand side, we apply a standard argument:

$$\begin{aligned}
&\sum_{t=1}^{T} \left[\sup_{\phi\in\mathcal{D}_t} \langle \phi, \theta^*\rangle - \langle \phi_t, \theta^*\rangle\right] \\
\leqslant& \sum_{t=1}^{T} \left[\sup_{\phi\in\mathcal{D}_t, \theta\in\mathcal{C}_{t-1}} \langle \phi, \theta\rangle - \langle \phi_t, \theta^*\rangle\right] \\
\overset{(a)}{=}& \sum_{t=1}^{T} \left[\sup_{\theta\in\mathcal{C}_{t-1}} \langle \phi_t, \theta\rangle - \langle \phi_t, \theta^*\rangle\right] \\
\leqslant& \sum_{t=1}^{T} \|\phi_t\|_{H_{t-1}^{-1}} \cdot \sup_{\theta\in\mathcal{C}_{t-1}} \|\theta - \theta^*\|_{H_{t-1}} \\
\overset{(b)}{\leq}& 2\beta_T \cdot \sum_{t=1}^{T} \|\phi_t\|_{H_{t-1}^{-1}} \overset{(c)}{=} 2\beta_T \cdot \sum_{t=1}^{T} \sigma_t w_t,
\end{aligned}$$

where (a) uses the selection rule for ϕ_t in (3) and (b) follows from the implications of $\mathcal{C}_{t-1}$. Specifically, for any $\theta \in \mathcal{C}_{t-1}$,

$$\begin{aligned}\|\theta - \theta^*\|_{H_{t-1}} &\leq \|\theta - \theta_{t-1}\|_{H_{t-1}} + \|\theta_{t-1} - \theta^*\|_{H_{t-1}} \\ &\leq 2\beta_{t-1} \leq 2\beta_T.\end{aligned}$$

The inequality (c) uses definition of w_t from (19) where $w_t = \|\phi_t\|_{H_{t-1}^{-1}} / \sigma_t$.

Since σ_t is defined as the maximum of several expressions in (19), specifically, $\sigma_t = \max\{\nu_t, \sigma_{1,t}, \sigma_{2,t}, \sigma_{3,t}\}$ for certain $\sigma_{i,t}$ $(i = 1, 2, 3)$, we can use the bound $\sigma_t \leq \nu_t + \sigma_{1,t} + \sigma_{2,t} + \sigma_{3,t}$. We then bound the remaining sums, either $\sum_{t=1}^T \nu_t w_t$ or $\sum_{t=1}^T \sigma_{i,t} w_t$. A key reason our algorithm achieves a finer variance-aware bound is the careful selection of the parameter σ_g. In particular, the dominant term is bounded as

$$\sum_{t=1}^T \nu_t w_t \leq \sqrt{\sum_{t=1}^T \nu_t^2} \cdot \sqrt{\sum_{t=1}^T w_t^2} = \widetilde{O}\left(d \cdot \sqrt{\sum_{t=1}^T \nu_t^2}\right).$$

We then make efforts to show that the remaining term, $\sum_{t=1}^T \sigma_{i,t} w_t$, is at most $\widetilde{O}(d) \cdot C$. The specific techniques used to establish these bounds are detailed in Appendix C.

5 NUMERICAL STUDIES

5.1 EXPERIMENTAL SETUP

We conduct numerical experiments to compare different online bandit algorithms. We consider a 10-dimensional space ($d = 10$), where the vector dimensions are specified with $B = 1$ and $L = 1$, and conduct the following experimental setup. The target vector θ^* is randomly chosen from the unit sphere. The experiment is repeated ten times to reduce the effect of random results.

- **Decision set:** The decision set D_t comprises 20 random unit vectors in $\mathbb{R}^d$ ($|D_t| = 20$). Each vector is independently generated in the same manner as θ^*. Notably, D_t is not a fixed set; instead, it is dynamically and randomly generated for each trial.
- **Noise distribution:** We use noise from a t-distribution with 3 degrees of freedom (denoted as t_3). The probability density function of t_3 is given by:

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}},$$

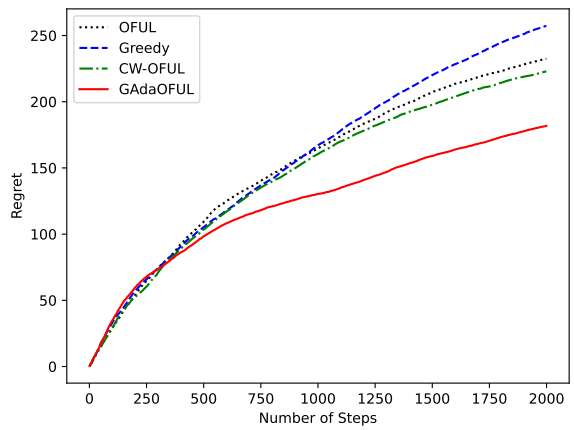
  where $v = 3$, Γ is the gamma function, and t is the random variable. The second moment (variance) of t_3 is 3, while higher moments do not exist (i.e., $E[X^k]$ is undefined for $k \geq 3$). Compared to Gaussian noise, the t_3-distribution better simulates real-world stochastic disturbances with occasional extreme values, challenging algorithms to be robust under non-sub-Gaussian conditions. This choice allows us to test the heavy-tail robustness of the algorithms.
- **Nonlinear function:** We select the exponential function $y = \exp(x)$ for the mapping f. This function is commonly used in various exponential models and is monotonically increasing. Within the interval $x \in [-1, 1]$, the derivative values range from $\exp(-1)$ to $\exp(1)$, which we denote as k and K, respectively. Additional experimental results using other three nonlinear link functions—yielding similar findings—are provided in Appendix E.
- **Corruption:** To simulate corruption, we employ the flipping technique [Bogunovic et al., 2021] during the first n steps. Specifically, for each reward y computed as $y = f(\langle\theta, \phi\rangle) + \epsilon$, where ϵ is noise from the t_3 distribution, we flip the reward to $y' = -f(\langle\theta, \phi\rangle) + \epsilon$. This manipulation misleads the bandit into making completely opposite decisions regarding the position of θ. This simulates a corruption level of $C = 2Kn$, where K is the maximum value of f'. The inequality $|y - y'| = 2f(\langle\theta, \phi\rangle) \leq 2K$ ensures the bound on the corruption. For linear function, $K = 1$, while for the exponential function, $K = e$. In this experiment, we choose $n = 50$.

5.2 EXPERIMENTAL RESULTS

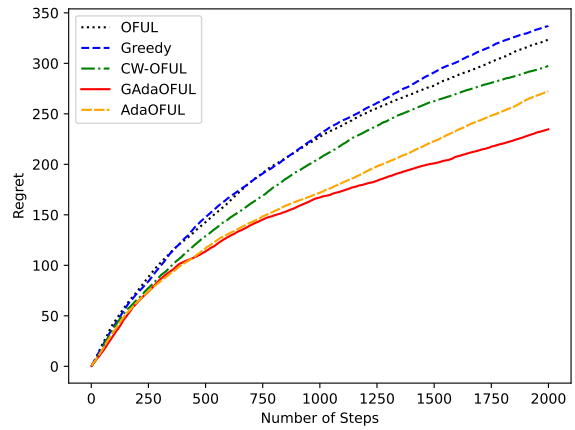
Alternative algorithms. We conducted experiments to compare the performance of the GAdaOFUL algorithm with several competing algorithms, including Greedy [Kannan et al., 2018], OFUL [Abbasi-Yadkori et al., 2011], CW-OFUL [He et al., 2022a], and AdaOFUL [Li and Sun, 2024], across diverse conditions. The Greedy algorithm and OFUL are classic bandit learning methods that provide baselines for our comparisons. CW-OFUL extends OFUL to enhance robustness against corrupted rewards, making it a suitable benchmark for comparison with GAdaOFUL in corrupted environments. While AdaOFUL performs well under heavy-tailed noise, it does not explicitly address corruption, enabling a clear comparison of its performance against GAdaOFUL in such settings.

Experimental results. The results are presented in Figure 1. The x-axis represents the number of steps, while the y-axis indicates the averaged regret over 10 repeated trials. The regret-iteration plot illustrates how regret accumulates as the number of steps increases.

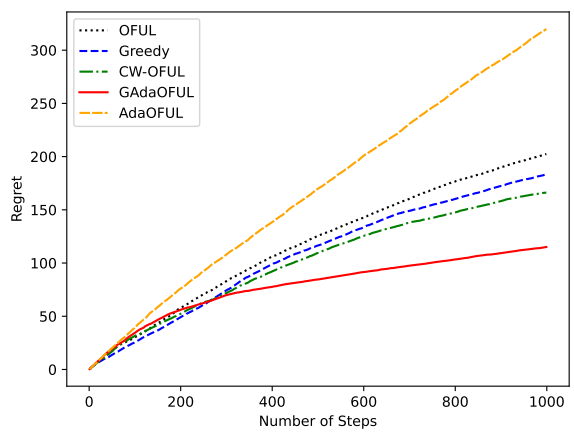
When there is no corruption (i.e., $C = 0$), the results in subfigures (a) and (c) demonstrate that GAdaOFUL achieves the smallest regret among all considered baselines. Notably, in scenarios with linear rewards and no corruption, GAdaOFUL coincides with the previous AdaOFUL, resulting in



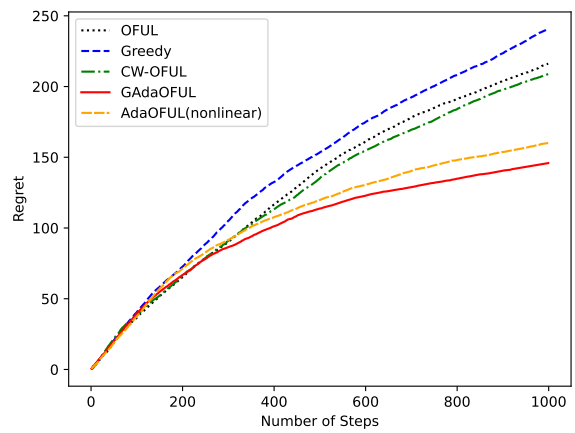
(a) Linear reward with $C = 0$.



(b) Linear reward with $C = 100$.



(c) Nonlinear reward with $C = 0$.



(d) Nonlinear reward with $C = 100e$.

Figure 1: Comparison results of various online bandit algorithms are presented under four scenarios: with linear or nonlinear rewards and in the presence or absence of corruption. The total number of iterations is set to $T = 2000$.

overlapping curves in subfigure (a). Furthermore, because AdaOFUL is specifically designed for linear rewards, its performance significantly degrades when the underlying reward deviates from a linear model, as shown in subfigure (c).

Next, we consider the case where corruption exists (i.e., $C = 1$). Again, the results in subfigures (b) and (d) reveal that GAdaOFUL achieves the smallest regret among all baselines. The original AdaOFUL does not account for nonlinear rewards. To further substantiate the superiority of our method, we also analyze a stronger competitor, AdaOFUL(nonlinear), which employs the same loss function (2) as GAdaOFUL for computing θ and is designed for nonlinear rewards. The only difference is that GAdaOFUL considers potential corruption in the rewards and modifies the selection of σ in (19). Interestingly, even with corrupted and nonlinear rewards, GAdaOFUL maintains its superiority, demonstrating its effectiveness in managing both nonlinearities and robustness against corruption.

Additionally, to assess the robustness of our method under a broader range of nonlinear reward structures, we conducted further experiments with alternative nonlinear functions. The results, consistent with our main findings, are presented in Appendix E.

In summary, the experimental results consistently show that GAdaOFUL outperforms other algorithms across various conditions. GAdaOFUL exhibits remarkable robustness, particularly in the presence of corruption and heavy-tailed noise, while also remaining effective under nonlinear conditions. These results validate the theoretical foundations of GAdaOFUL and suggest its practical utility in real-world ap-

plications where the reliability of feedback may be uncertain or compromised.

6 CONCLUSIONS AND DISCUSSIONS

In this paper, we introduced GAdaOFUL, a novel online bandit algorithm that achieves a state-of-the-art regret bound of $\widetilde{\mathcal{O}}\left(d\sqrt{\sum_{t\in[T]}\nu_t^2} + d\cdot(C\vee 1)\right)$. This bound highlights the algorithm's efficiency in low-variance environments and its resilience against corrupted, nonlinear, and heavy-tailed rewards. Specifically, our results demonstrate that sublinear regret is attainable even in the presence of highly non-standard reward characteristics commonly observed in real-world scenarios, such as adversarial corruption, nonlinearity, and heavy-tailed noise. Empirical evaluations show that GAdaOFUL outperforms existing methods.

There are several promising directions for future research building on our work.

- First, our algorithm is primarily based on adaptive Huber regression [Sun et al., 2020], originally designed for data with only $1+\delta$ ($\delta \leq 1$) moments. This paper focuses on rewards with bounded variance, aligning with the assumptions of most variance-aware algorithms. A natural direction would be to generalize our results to settings where rewards possess only $1+\delta$ moments, in the spirit of Huang et al. [2024], which builds upon Li and Sun [2024].
- Second, a key limitation of this work is the dependence on the assumption that the reward model admits a generalized linear form. However, real-world reward structures often deviate from the GLM framework, introducing significant model mismatch. One possible approach is to consider a broader class of reward models, such as nonparametric and neural-network-based reward models.
- Third, a promising direction is to leverage our algorithm as a modular component for tackling more complex tasks, such as linear Markov decision processes (MDPs) [He et al., 2023].
- Fourth, it is practically valuable to develop parameter-free algorithms, along the lines of Sun [2021], that do not require prior knowledge of problem- or data-dependent constants.
- Finally, it is of interest to design more comprehensive evaluations that stress-test the algorithm under diverse and potentially misspecified environments. Investigating the performance of GAdaOFUL under model misspecification could inspire more robust algorithms and provide deeper insights into the limitations of reward-model-based approaches.

Acknowledgements

Qiang Sun's research is partially supported by the Natural Sciences and Engineering Research Council of Canada (Grant RGPIN-2018-06484), computing resources provided by the Digital Research Alliance of Canada, and MBZUAI.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Ilija Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1071–1081. PMLR, 2020.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021.

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

Rama Cont and Jean-Philipe Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic dynamics*, 4(2):170–196, 2000.

Stefany Coxe, Stephen G West, and Leona S Aiken. The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of personality assessment*, 91(2):121–136, 2009.

Yan Dai, Ruosong Wang, and Simon S Du. Variance-aware sparse linear bandits. *arXiv preprint arXiv:2205.13450*, 2022.

Qiwei Di, Tao Jin, Yue Wu, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. *arXiv preprint arXiv:2310.00968*, 2023.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in neural information processing systems*, volume 23, 2010.

Sergey Foss, Dmitry Korshunov, Stan Zachary, et al. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34614–34625. Curran Associates, Inc., 2022a.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions, 2022b.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

Joseph M Hilbe. Logistic regression. *International encyclopedia of statistical science*, 1:15–32, 2011.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. In *Advances in neural information processing systems*, volume 31, 2018.

Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 19–36. JMLR Workshop and Conference Proceedings, 2012.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

Xiang Li and Qiang Sun. Variance-aware decision making with linear function approximation under heavy-tailed rewards. *Transactions on Machine Learning Research*, 2024.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pages 460–468. PMLR, 2021.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

Qiang Sun. Self-tuned robust mean estimators. *arXiv preprint arXiv:2107.00118*, 2021.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 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*, volume 36, 2024.

Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*, pages 39834–39863. PMLR, 2023.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. In *Advances in Neural Information Processing Systems*, volume 34, pages 4342–4355, 2021.

Heyang Zhao, Dongruo Zhou, Jiafan He, and Quanquan Gu. Optimal online generalized linear regression with stochastic noise and its application to heteroscedastic bandits. In *International Conference on Machine Learning*, pages 42259–42279. PMLR, 2023.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32, 2019.

Appendix

Qingyuan Yu[1] **Euijin Baek**[2,4] **Xiang Li**[3] **Qiang Sun**[4,5]

[1]USTC [2]University of Alberta [3]University of Pennsylvania [4]University of Toronto [5] MBZUAI

In the appendix, we provide the proofs of the main results and the supporting lemmas.

A PROOF OF THEOREM 1

Initially, we introduce two lemmas to help our discussion. The first lemma establishes a bound of $\|\nabla L_T(\theta^*)\|_{H_t^{-1}}$. The second lemma asserts that $\nabla^2 L_T(\theta)$ is positive with high probability.

Lemma 6. *Assume* $\mathbb{E}\left[z_t^2\left(\theta^*\right) \mid \mathcal{F}_{t-1}\right] \leqslant b^2$ *for all* $t \geqslant 1$*, where* $z_t(\theta) = \dfrac{y_t - f(\langle \phi_t, \theta\rangle)}{\sigma_t}$*. With probability at least* $1-\delta$*, for all* $T \geqslant 1$*, it follows that*

$$\|\nabla L_T\left(\theta^*\right)\|_{H_T^{-1}} \leqslant 8\left[\frac{\kappa b^2}{\tau_0} + b\sqrt{\kappa \log \frac{2T^2}{\delta}} + \tau_0 \log \frac{2T^2}{\delta}\right] + kB\sqrt{\lambda}, \tag{10}$$

where $\kappa = d \cdot \log\left(1 + TL^2/(d\lambda\sigma_{\min}^2)\right)$ *is a constant.*

Lemma 7. *Assume* $\mathbb{E}\left[z_t^2\left(\theta^*\right) \mid \mathcal{F}_{t-1}\right] \leqslant b^2$ *for all* $t \geqslant 1$*, where* $z_t(\theta) = \dfrac{y_t - f(\langle \phi_t, \theta\rangle)}{\sigma_t}$*. If we set*

$$\tau_0\sqrt{\log \frac{2T^2}{\delta}} \geqslant \max\{\sqrt{2\kappa}b, 2\sqrt{d}\},$$

with probability at least $1-2\delta$*, we have that for all* $T \geqslant 0$*,*

$$\nabla^2 L_T(\theta) \succeq \frac{k}{4} H_T \quad \textit{for any} \quad \|\theta\| \leqslant B. \tag{11}$$

Since $\sigma_t \geq v_t$, where v_t is the variance of $\epsilon_t = y_t - f(\langle \phi_t, \theta\rangle)$, here we set $b = 1$.
Let $\theta(\eta) = (1-\eta)\theta^* + \eta\theta_T$. Using the mean value theorem for vector-valued functions, we have

$$\nabla L_T(\theta_T) - \nabla L_T(\theta^*) = \int_0^1 \nabla^2 L_T(\theta(\eta))\, d\eta \cdot (\theta_T - \theta^*). \tag{12}$$

Combining (11) and $\|\theta(\eta)\| \leq B$ for all $\eta \in [0,1]$, it follows that

$$\frac{k}{4}\|\theta_T - \theta^*\|_{H_T}^2 \leqslant \left\langle \theta_T - \theta^*, \nabla L_T\left(\theta_T\right) - \nabla L_T\left(\theta^*\right)\right\rangle. \tag{13}$$

The first-order stationary condition of the constrained convex optimization that $\theta_T := \operatorname{argmin}_{\theta \in \text{Ball}_d(B)} L_T(\theta)$ implies

$$\langle \nabla L_T(\theta_T), \theta_T - \theta^* \rangle \leq 0.$$

Consequently,

$$\begin{aligned} & \langle \theta_T - \theta^*, \nabla L_T(\theta_T) - \nabla L_T(\theta^*) \rangle . \\ \leqslant & \langle \theta_T - \theta^*, -\nabla L_T(\theta^*) \rangle \\ \leqslant & \|\theta_T - \theta^*\|_{H_T} \|\nabla L_T(\theta^*)\|_{H_T^{-1}} . \end{aligned} \tag{14}$$

By (13) and (14), we have

$$\begin{aligned} \|\theta_T - \theta^*\|_{H_T} &\leq \frac{4}{k} \|\nabla L_T(\theta^*)\|_{H_t^{-1}} \\ &\leq \frac{32}{k} \left[\frac{\kappa}{\tau_0} + \sqrt{\kappa \log \frac{2T^2}{\delta}} + \tau_0 \log \frac{2T^2}{\delta} \right] + 5B\sqrt{\lambda}. \end{aligned} \tag{15}$$

Then the regret can be bounded as

$$\text{Reg}(T) \leq 2K\beta_T \left[\sqrt{2\kappa} \cdot \sqrt{\sum_{t \in [T]} \nu_t^2 + 1} + \frac{2L\kappa}{m_0^2 \sqrt{\lambda}} + \frac{2LBK\kappa}{\sqrt{m_1 d}} \right],$$

The proof of this can refer to the proof of Theorem 3, with the only difference being that C is set to 0.

A.1 PROOF OF LEMMA 1

Let $z_t(\theta) = \dfrac{y_t - f(\langle \phi_t, \theta \rangle)}{\sigma_t}$. The gradient is given by

$$\nabla L_T(\theta) = \lambda k \theta - \sum_{t=1}^{T} \frac{\tau_t z_t(\theta)}{\sqrt{\tau_t^2 + z_t^2(\theta)}} \frac{\phi_t}{\sigma_t}.$$

By triangle inequality, we have

$$\|\nabla L_T(\theta^*)\|_{H_T^{-1}} \leq \|\lambda k \theta^*\|_{H_T^{-1}} + \underbrace{\left\| \sum_{t=1}^{T} \frac{\tau_t z_t(\theta^*)}{\sqrt{\tau_t^2 + z_t^2(\theta^*)}} \frac{\phi_t}{\sigma_t} \right\|_{H_T^{-1}}}_{d_T}.$$

For the first term $\|\lambda k \theta^*\|_{H_T^{-1}}$, we have $H_T^{-1} \preceq \lambda^{-1} I$, due to $H_T \succeq \lambda I$. Thus, $\|\lambda k \theta^*\|_{H_T^{-1}} \leq kB\sqrt{\lambda}$.

For the second term $\|d_T\|_{H_T^{-1}}$, we have $\|d_T\|_{H_T^{-1}} \leq \alpha_T$, where

$$\alpha_T = 8 \left[\frac{\kappa b^2}{\tau_0} + b\sqrt{\kappa \log \frac{2T^2}{\delta}} + \tau_0 \log \frac{2T^2}{\delta} \right].$$

The proof of this inequality follows exactly the same steps as the proof of Lemma B.2 in [Li and Sun, 2024], and is therefore omitted here for brevity. The reader is referred to [Li and Sun, 2024] for a detailed proof.

A.2 PROOF OF LEMMA 2

The second-order gradient is given by

$$\begin{aligned}\nabla^2 L_T(\theta) &= \lambda k I + \sum_{t=1}^{T} \left(\frac{\tau_t}{\sqrt{\tau_t^2 + z_t^2}} \right)^3 f'\left(\langle \phi_t, \theta^* \rangle\right) \frac{\phi_t \phi_t^\top}{\sigma_t^2} \\ &\succeq k \left(\lambda I + \sum_{t=1}^{T} \left(\frac{\tau_t}{\sqrt{\tau_t^2 + z_t^2}} \right)^3 \frac{\phi_t \phi_t^\top}{\sigma_t^2} \right).\end{aligned}$$

where the last inequality holds due to $f'(x) \geq k$, when $x \in [-BL, BL]$. In this case, we can treat it in the same way as the linear case by decomposing the equation into three parts for processing.

$$\begin{aligned}&\lambda I + \sum_{t=1}^{T} \left(\frac{\tau_t}{\sqrt{\tau_t^2 + z_t^2}} \right)^3 \frac{\phi_t \phi_t^\top}{\sigma_t^2} \\ =& H_T - \underbrace{\sum_{t=1}^{T} \left[1 - \left(\frac{\tau_t}{\sqrt{\tau_t^2 + z_t^2\left(\theta^*\right)}} \right)^3 \right] \frac{\phi_t \phi_t^\top}{\sigma_t^2}}_{H_{1,T}} - \underbrace{\sum_{t=1}^{T} \left[\left(\frac{\tau_t}{\sqrt{\tau_t^2 + z_t^2\left(\theta^*\right)}} \right)^3 - \left(\frac{\tau_t}{\sqrt{\tau_t^2 + z_t^2(\theta)}} \right)^3 \right] \frac{\phi_t \phi_t^\top}{\sigma_t^2}}_{H_{2,T}}.\end{aligned}$$

Then we can prove that $H_{1,T} \preceq \frac{1}{4} H_T$, and that $H_{2,T} \preceq \frac{1}{2} H_T$. For detailed proves, we refer the reader to Lemma B.1 in [Li and Sun, 2024]. Thus,

$$\nabla^2 L_T(\theta) \succeq k \left(\lambda I + \sum_{t=1}^{T} \left(\frac{\tau_t}{\sqrt{\tau_t^2 + z_t^2}} \right)^3 \frac{\phi_t \phi_t^\top}{\sigma_t^2} \right) \succeq \frac{k}{4} H_T.$$

B PROOF OF THEOREM 2

To prove the main theorems, we introduce an auxiliary problem:

$$\begin{aligned}z_t(s) &= \frac{y_t - f(s)}{\sigma_t},\ \tilde{z}_t(s) := \frac{y_t - f(s) - c_t}{\sigma_t}. \\ \tilde{L}_T(\theta) &:= \frac{\lambda k}{2} \|\theta\|^2 + \sum_{t=1}^{T} \frac{1}{\sigma_t} \int_0^{\langle \phi_t, \theta \rangle} \frac{\tau_t \tilde{z}_t(s)}{\sqrt{\tau_t^2 + \tilde{z}_t^2(s)}} ds. \\ \theta_T &= \arg \min_{\|\theta\| \leq B} L_T(\theta),\ \tilde{\theta}_T := \arg \min_{\|\theta\| \leq B} \tilde{L}_T(\theta).\end{aligned}$$

The symbol $\tilde{z}_t(s)$ is defined as the standardized difference between the observed value y_t and the prediction $f(s)$ adjusted for the corruption c_t, all scaled by the noise level σ_t. A crucial property of $\tilde{z}_t(s)$ is that $\mathbb{E}[\tilde{z}_t(s) | \mathcal{F}_{t-1}] = 0$. It means $\tilde{L}_T$ is essentially the non-corrupted objective, which we studied in Theorem 1 and we can use (15) to bound $\left\| \tilde{\theta}_T - \theta_T^* \right\|_{H_T}$.

Therefore, the only thing we need to do is to find the relationship between θ_T and $\tilde{\theta}_T$.

$$
\begin{aligned}
\frac{k}{4}\|\theta_T - \tilde{\theta}_T\|_{H_T}^2 &\overset{(a)}{\leq} \langle \nabla \tilde{L}_T(\tilde{\theta}_T) - \nabla \tilde{L}_T(\theta_T), \tilde{\theta}_T - \theta_T \rangle \\
&= \langle \nabla \tilde{L}_T(\tilde{\theta}_T) - \nabla L_T(\theta_T), \tilde{\theta}_T - \theta_T \rangle + \langle \nabla L_T(\theta_T) - \nabla \tilde{L}_T(\theta_T), \tilde{\theta}_T - \theta_T \rangle \\
&\overset{(b)}{\leq} \langle \nabla L_T(\theta_T) - \nabla \tilde{L}_T(\theta_T), \tilde{\theta}_T - \theta_T \rangle \\
&= \sum_{t=1}^{T} \left(\frac{\tau_t z_t(\langle \phi_t, \theta \rangle)}{\sqrt{\tau_t^2 + z_t^2(\langle \phi_t, \theta \rangle)}} - \frac{\tau_t \tilde{z}_t(\langle \phi_t, \theta \rangle)}{\sqrt{\tau_t^2 + \tilde{z}_t{}^2(\langle \phi_t, \theta \rangle)}} \right) \left\langle \frac{\phi_t}{\sigma_t}, \tilde{\theta}_T - \theta_T \right\rangle \\
&\overset{(c)}{\leq} \sum_{t=1}^{T} |z_t(\langle \phi_t, \theta \rangle) - \tilde{z}_t(\langle \phi_t, \theta \rangle)| \left| \left\langle \frac{\phi_t}{\sigma_t}, \tilde{\theta}_T - \theta_T \right\rangle \right| \\
&= \sum_{t=1}^{T} \frac{|c_t|}{\sigma_t} \left| \left\langle \frac{\phi_t}{\sigma_t}, \tilde{\theta}_T - \theta_T \right\rangle \right| \\
&\overset{(d)}{\leq} \sum_{t=1}^{T} \frac{|c_t|\, w_t}{\sigma_t} \|\theta_T - \tilde{\theta}_T\|_{H_T} \\
&\overset{(e)}{\leq} \sqrt{\kappa} \|\theta_T - \tilde{\theta}_T\|_{H_T}.
\end{aligned}
$$

Inequality (a) uses mean value theorem and $\nabla^2 L_T(\theta) \succeq \frac{k}{4} H_T$, the same as (13). The first-order stationary condition of the constrained convex optimization implies that $\langle \nabla \tilde{L}_T(\tilde{\theta}_T), \tilde{\theta}_T - \theta_T \rangle \leq 0$ and $\langle \nabla L_T(\theta_T), \theta_T - \tilde{\theta}_T \rangle \leq 0$, thus proving inequality (b). Inequality (c) comes from the fact that $0 \leq \frac{d}{dx} \frac{\tau x}{\sqrt{\tau^2 + x^2}} \leq 1$. Inequality (d) uses $\left| \left\langle \frac{\phi_t}{\sigma_t}, \tilde{\theta}_T - \theta_T \right\rangle \right| \leq \left\| \frac{\phi_t}{\sigma_t} \right\|_{H_T^{-1}} \left\| \tilde{\theta}_T - \theta_T \right\|_{H_T}$ and $\left\| \frac{\phi_t}{\sigma_t} \right\|_{H_T^{-1}} \leq \left\| \frac{\phi_t}{\sigma_t} \right\|_{H_{t-1}^{-1}} = w_t$. Inequality (e) comes from $\sigma_t = \sigma_t^2 / \sigma_t \geq C \|\phi_t\|_{H_{t-1}^{-1}} / \sqrt{\kappa} \sigma_t = C w_t / \sqrt{\kappa}$. Thus, we have

$$
\left\| \tilde{\theta}_T - \theta_T \right\|_{H_T} \leq \frac{4\sqrt{\kappa}}{k}.
$$

Combining the upper bound of $\left\| \tilde{\theta}_T - \theta_T^* \right\|_{H_T}$ in (15),

$$
\begin{aligned}
\|\theta_T - \theta_T^*\|_{H_T} &\leq \left\| \tilde{\theta}_T - \theta_T \right\|_{H_T} + \left\| \tilde{\theta}_T - \theta_T^* \right\|_{H_T} \\
&\leq \frac{4\sqrt{\kappa}}{k} + \frac{32}{k} \left[\frac{\kappa}{\tau_0} + \sqrt{\kappa \log \frac{2T^2}{\delta}} + \tau_0 \log \frac{2T^2}{\delta} \right] + 5\sqrt{\lambda} B.
\end{aligned}
$$

C PROOF OF THEOREM 3

In this proof, we will bound the regret in the event that high probability coverage holds.

By the Lipschitz continuity of the (nonlinear) link function f (i.e., $\sup_{|x| \leq BL} f'(x) \leq K$), we bound the regret from its definition by

$$
\mathrm{Reg}(T) := \sum_{t=1}^{T} \left[\sup_{\phi \in D_t} f(\langle \phi, \theta^* \rangle) - f(\langle \phi_t, \theta^* \rangle) \right] \leq K \sum_{t=1}^{T} \left[\sup_{\phi \in \mathcal{D}_t} \langle \phi, \theta^* \rangle - \langle \phi_t, \theta^* \rangle \right]. \tag{16}
$$

To bound the right-hand side, we apply a standard argument:

$$
\begin{aligned}
&\sum_{t=1}^{T}\left[\sup_{\phi\in\mathcal{D}_t}\langle\phi,\theta^*\rangle-\langle\phi_t,\theta^*\rangle\right]\\
\leqslant&\sum_{t=1}^{T}\left[\sup_{\phi\in\mathcal{D}_t,\theta\in\mathcal{C}_{t-1}}\langle\phi,\theta\rangle-\langle\phi_t,\theta^*\rangle\right]\\
\overset{(a)}{=}&\sum_{t=1}^{T}\left[\sup_{\theta\in\mathcal{C}_{t-1}}\langle\phi_t,\theta\rangle-\langle\phi_t,\theta^*\rangle\right]\\
\leqslant&\sum_{t=1}^{T}\|\phi_t\|_{H_{t-1}^{-1}}\cdot\sup_{\theta\in\mathcal{C}_{t-1}}\|\theta-\theta^*\|_{H_{t-1}}\\
\overset{(b)}{\leq}&2\beta_T\cdot\sum_{t=1}^{T}\|\phi_t\|_{H_{t-1}^{-1}}\overset{(c)}{=}2\beta_T\cdot\sum_{t=1}^{T}\sigma_t w_t,
\end{aligned}
\tag{17}
$$

where (a) uses the selection rule for ϕ_t that $(\phi_t,\cdot)=\mathrm{argmax}_{\phi\in D_t,\theta\in C_{t-1}}\langle\phi,\theta\rangle$, and (b) follows from the implications of $\mathcal{C}_{t-1}$. Specifically, for any $\theta\in\mathcal{C}_{t-1}$,

$$
\|\theta-\theta^*\|_{H_{t-1}}\leq\|\theta-\theta_{t-1}\|_{H_{t-1}}+\|\theta_{t-1}-\theta^*\|_{H_{t-1}}\leq 2\beta_{t-1}\leq 2\beta_T.
$$

The inequality (c) uses definition of w_t that $w_t=\|\phi_t\|_{H_{t-1}^{-1}}/\sigma_t$.

Here, the problem is converted to how to bound $\sum_{t=1}^{T}\sigma_t w_t$.

Since $\sigma_t\geq\|\phi_t\|_{H_{t-1}^{-1}}/m_0$ and $m_0\leq 1$, we have $w_t\leqslant 1$. Notice that $\frac{\|\phi_t\|}{\sigma_t}\leqslant\frac{\|\phi_t\|}{\sigma_{\min}}\leqslant\frac{L}{\sigma_{\min}}$. Then by Lemma 8,

$$
\sum_{t=1}^{T}w_t^2=\sum_{t=1}^{T}\min\left\{1,w_t^2\right\}=\sum_{t=1}^{T}\min\left\{1,\left\|\frac{\phi_t}{\sigma_t}\right\|_{H_{t-1}^{-1}}^2\right\}\leqslant 2d\log\left(1+\frac{TL^2}{d\lambda\sigma_{\min}^2}\right)=2\kappa,
\tag{18}
$$

where $\kappa=d\cdot\log\left(1+TL^2/(d\lambda\sigma_{\min}^2)\right)$ is a constant.

Recall the definition of σ_t:

$$
\sigma_t=\max\left\{\nu_t,\sigma_{\min},\|\phi_t\|_{H_{t-1}^{-1}}/m_0,\alpha\|\phi_t\|_{H_{t-1}^{-1}}^{1/2}\right\},
\tag{19}
$$

where $\alpha=\max\left\{\frac{\sqrt{LBK}}{m_1^{1/4}d^{1/4}},C^{\frac{1}{2}}\kappa^{-\frac{1}{4}}\right\}$, $C=\sum_{t=1}^{T}|c_t|$ is the corruption level.

According to what value σ_t takes, we decompose $[T]$ into three sets $[T]\subseteq\cup_{i=1}^{3}\mathcal{J}_i$ where

$$
\begin{aligned}
\mathcal{J}_1&=\left\{t\in[T]:\sigma_t\in\{\nu_t,\sigma_{\min}\}\right\},\\
\mathcal{J}_2&=\left\{t\in[T]:\sigma_t=\frac{\|\phi_t\|_{H_{t-1}^{-1}}}{m_0}\right\},\\
\mathcal{J}_3&=\left\{t\in[T]:\sigma_t=\alpha\|\phi_t\|_{H_{t-1}^{-1}}^{1/2}\right\}.
\end{aligned}
$$

First, for any $t\in\mathcal{J}_1$,

$$
\begin{aligned}
\sum_{t\in\mathcal{J}_1}\sigma_t w_t &\leqslant \sum_{t\in\mathcal{J}_1}\max\left\{\nu_t,\sigma_{\min}\right\}w_t\\
&\leqslant \sum_{t\in[T]}\max\left\{\nu_t,\sigma_{\min}\right\}w_t\\
&\overset{(a)}{\leqslant} \sqrt{\sum_{t\in[T]}(\nu_t^2+\sigma_{\min}^2)}\sqrt{\sum_{t\in[T]}w_t^2}\\
&\overset{(b)}{\leqslant} \sqrt{2\kappa}\cdot\sqrt{\sum_{t\in[T]}\nu_t^2+1}.
\end{aligned}
\tag{20}
$$

Here (a) holds due to Cauchy-Schwarz inequality and (b) uses (18) and $\sigma_{\min}=\frac{1}{\sqrt{T}}$.

Second, for any $t\in\mathcal{J}_2$, we have

$$
\begin{aligned}
\sum_{t\in\mathcal{J}_2}\sigma_t w_t &= \frac{1}{m_0}\sum_{t\in\mathcal{J}_2}\sigma_t w_t^2 \leqslant \frac{\sup_{t\in\mathcal{J}_2}\sigma_t}{m_0}\sum_{t\in\mathcal{J}_2}w_t^2\\
&\leqslant \frac{\sup_{t\in[T]}\left\|\phi_t\right\|_{H_{t-1}^{-1}}}{m_0^2}\cdot\sum_{t\in\mathcal{J}_2}w_t^2\\
&\leqslant \frac{\sup_{t\in[T]}\left\|\phi_t\right\|_{H_{t-1}^{-1}}}{m_0^2}\cdot\sum_{t\in[T]}w_t^2 \leqslant \frac{2L\kappa}{m_0^2\sqrt{\lambda}}
\end{aligned}
\tag{21}
$$

where the last inequality holds due to $\left\|\phi_t\right\|_{H_{t-1}^{-1}}\leqslant\frac{1}{\sqrt{\lambda}}\left\|\phi_t\right\|\leqslant\frac{L}{\sqrt{\lambda}}$ for all $t\geqslant 1$ and (18).

Finally, for any $t\in\mathcal{J}_3$, we have $\sigma_t^2=\alpha^2\left\|\phi_t\right\|_{H_{t-1}^{-1}}$, which implies $\sigma_t=\alpha^2 w_t$ due to $w_t=\left\|\frac{\phi_t}{\sigma_t}\right\|_{H_{t-1}^{-1}}$.

Therefore,

$$
\begin{aligned}
\sum_{t\in\mathcal{J}_3}\sigma_t w_t &= \sum_{t\in\mathcal{J}_3}\alpha^2 w_t^2 \leq \alpha^2\sum_{t\in[T]}w_t^2\\
&\leq \left(\frac{LBK}{\sqrt{m_1 d}}+\frac{C}{\sqrt{\kappa}}\right)\sum_{t\in[T]}w_t^2\\
&\leq \left(\frac{LBK}{\sqrt{m_1 d}}+\frac{C}{\sqrt{\kappa}}\right)\cdot 2\kappa\\
&= \frac{2LBK\kappa}{\sqrt{m_1 d}}+2C\sqrt{\kappa}.
\end{aligned}
\tag{22}
$$

Plugging (20), (21) and (22) into (16) and (17), we have

$$
\mathrm{Reg}(T)\leqslant 2K\beta_T\left[\sqrt{2\kappa}\cdot\sqrt{\sum_{t\in[T]}\nu_t^2+1}+\frac{2L\kappa}{m_0^2\sqrt{\lambda}}+\frac{2LBK\kappa}{\sqrt{m_1 d}}+2C\sqrt{\kappa}\right].
$$

D AUXILIARY LEMMAS

Lemma 8. *(Lemma 11 in [Abbasi-Yadkori et al., 2011]). Let $\{x_t\}_{t\geqslant 1}\subset\mathbb{R}^d$ and assume $\|x_t\|\leqslant L$ for all $t\geqslant 1$. Set $Z_t=\sum_{s=1}^t x_s x_s^\top+\lambda I$. Then it follows that*

$$\sum_{t=1}^{T} \min\left\{1, \|x_t\|^2_{Z_{t-1}^{-1}}\right\} \leqslant 2d \log\left(\frac{d\lambda + TL^2}{d\lambda}\right).$$

Lemma 9. *(Lemma B.1 in [Li and Sun, 2024]). Assume* $z_t(\theta) = \dfrac{y_t - \langle \phi_t, \theta \rangle}{\sigma_t}$, $\mathbb{E}[z_t|\mathcal{F}_{t-1}] = 0$ *and* $\mathbb{E}\left[z_t^2\left(\theta^*\right) \mid \mathcal{F}_{t-1}\right] \leqslant b^2$ *for all* $t \geqslant 1$. *If we set*

$$\tau_0 \sqrt{\log \frac{2T^2}{\delta}} \geqslant \max\{\sqrt{2\kappa} b, 2\sqrt{d}\},$$

with probability at least $1 - 2\delta$, *we have that for all* $T \geqslant 0$,

$$\frac{1}{4} H_T \leq \lambda I + \sum_{t=1}^{T} \left(\frac{\tau_T}{\sqrt{\tau_T^2 + z_T\left(\theta^2\right)}}\right)^3 \frac{\phi_T \phi_T^\top}{\sigma_T^2} \leq H_T \text{ for any } \|\theta\| \leqslant B.$$

Lemma 10. *(Lemma B.2 in [Li and Sun, 2024]). Assume* $z_t(\theta) = \dfrac{y_t - \langle \phi_t, \theta \rangle}{\sigma_t}$, $\mathbb{E}[z_t|\mathcal{F}_{t-1}] = 0$ *and* $\mathbb{E}\left[z_t^2\left(\theta^*\right) \mid \mathcal{F}_{t-1}\right] \leqslant b^2$ *for all* $t \geqslant 1$. *With probability at least* $1 - \delta$, *for all* $T \geqslant 1$, *it follows that*

$$\left\| \lambda\theta^* - \sum_{t=1}^{T} \frac{\tau_T z_T(\theta^*)}{\sqrt{\tau_T^2 + z_T^2(\theta^*)}} \frac{\phi_T}{\sigma_T} \right\|_{H_T^{-1}} \leqslant 8\left[\frac{\kappa b^2}{\tau_0} + b\sqrt{\kappa \log \frac{2T^2}{\delta}} + \tau_0 \log \frac{2T^2}{\delta}\right] + \sqrt{\lambda} B$$

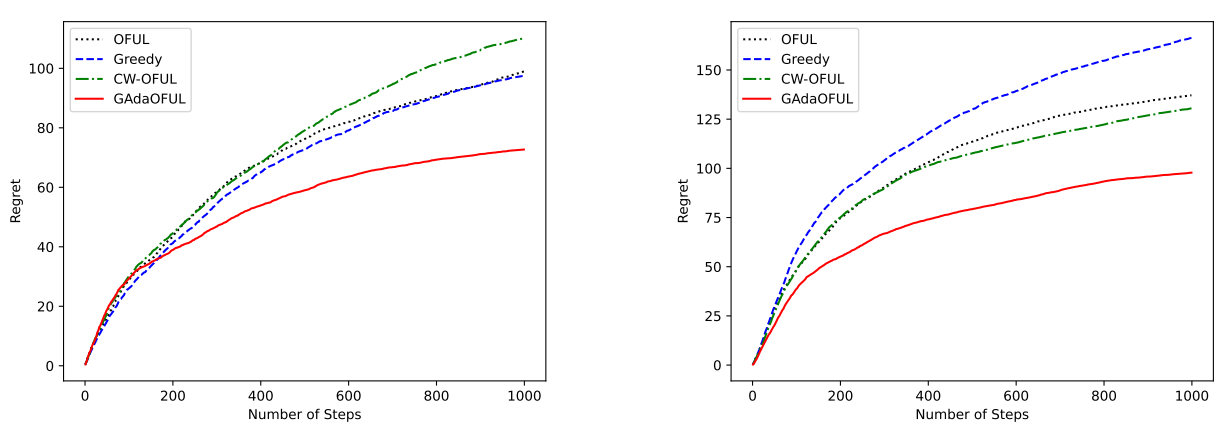
where $\kappa = d \cdot \log\left(1 + TL^2/(d\lambda\sigma_{min}^2)\right)$.

E ADDITIONAL EXPERIMENTS ON NONLINEAR REWARD FUNCTIONS

In this section, we present additional experimental results to further show the effectiveness of GAdaOFUL across other types of nonlinear reward functions. Specifically, we consider the following mappings f: (1) logistic function: $f(x) = \dfrac{5}{1 + \exp(-x)}$, (2) quadratic function: $f(x) = (x + 1.5)^2$, and (3) logarithmic function: $f(x) = 3\log(x + 2)$.

Each function is monotonically increasing and has been appropriately translated and scaled to meet our modeling assumptions. The corresponding results are presented in the following subsections. As shown, our method consistently achieves the lowest regret, regardless of whether reward corruption is present.

E.1 LOGISTIC LINK FUNCTION

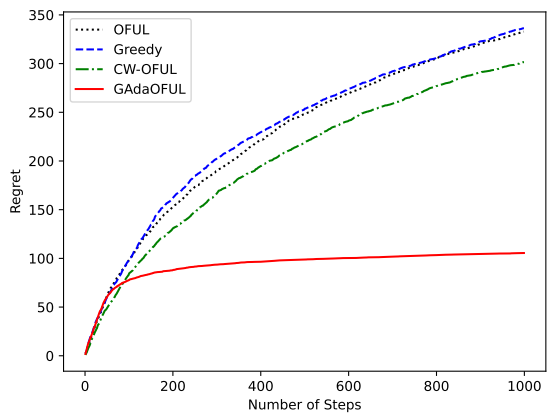


(a) No corruption

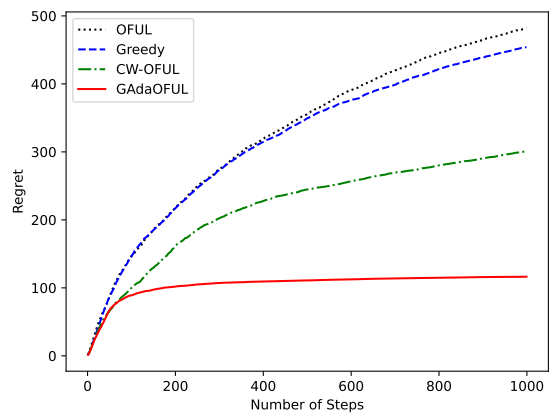
(b) Corruption level = 100

Figure 2: Performance under clean and corrupted settings with the logistic link function.

E.2 QUADRATIC LINK FUNCTION



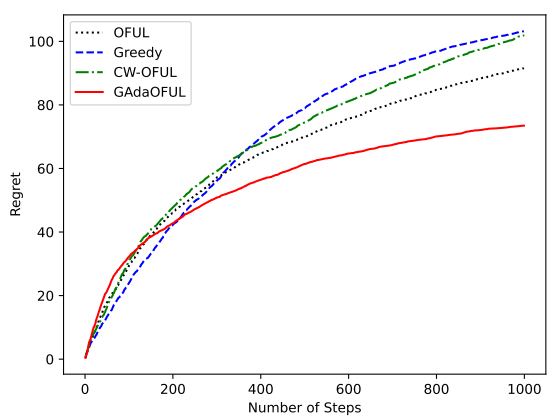
(a) No corruption



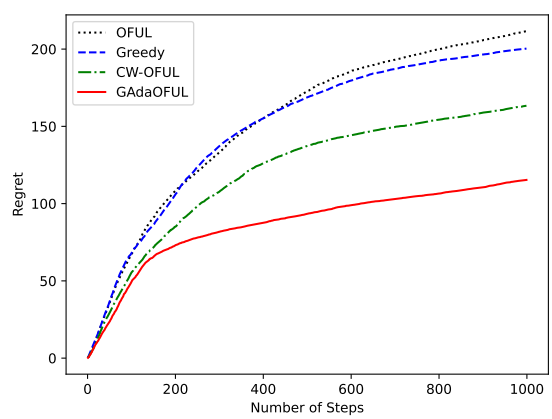
(b) Corruption level = 100

Figure 3: Performance under clean and corrupted settings with the quadratic link function.

E.3 LOGARITHMIC LINK FUNCTION



(a) No corruption



(b) Corruption level = 100

Figure 4: Performance under clean and corrupted settings with the logarithmic link function.