Nearly Optimal Algorithms for Linear Contextual Bandits with Adversarial Corruptions

Jiafan He
Department of Computer Science
University of California, Los Angeles
jiafanhe19@ucla.edu

Dongruo Zhou
Department of Computer Science
University of California, Los Angeles
drzhou@cs.ucla.edu

Tong Zhang
Google Research & HKUST
tongzhang@tongzhang-ml.org

Quanquan Gu
Department of Computer Science
University of California, Los Angeles
qgu@cs.ucla.edu

Abstract

We study the linear contextual bandit problem in the presence of adversarial corruption, where the reward at each round is corrupted by an adversary, and the corruption level (i.e., the sum of corruption magnitudes over the horizon) is $C \geq 0$. The best-known algorithms in this setting are limited in that they either are computationally inefficient or require a strong assumption on the corruption, or their regret is at least C times worse than the regret without corruption. In this paper, to overcome these limitations, we propose a new algorithm based on the principle of optimism in the face of uncertainty. At the core of our algorithm is a weighted ridge regression where the weight of each chosen action depends on its confidence up to some threshold. We show that for both known C and unknown C cases, our algorithm with proper choice of hyperparameter achieves a regret that nearly matches the lower bounds. Thus, our algorithm is nearly optimal up to logarithmic factors for both cases. Notably, our algorithm achieves the near-optimal regret for both corrupted and uncorrupted cases ($C = 0$) simultaneously.

1 Introduction

We study linear contextual bandits with adversarial corruptions. At each round, the agent observes a decision set provided by the environment, and selects an action from the decision set. Then an adversary *corrupts* the reward of the action selected by the agent. The agent then receives the corrupted reward of the selected action and proceeds until K rounds. The agent's goal is to minimize the regret $\text{Regret}(K)$, which is the difference between the optimal accumulated reward and the selected accumulated reward. This problem can be regarded as a combination of the two classical bandit problems, *stochastic bandits* and *adversarial bandits* (Lattimore and Szepesvári, 2018). In practice, the contextual bandits with adversarial corruptions can describe many popular decision-making problems such as pay-per-click advertisements with click fraud (Lykouris et al., 2018) and recommendation system with malicious users (Deshpande and Montanari, 2012).

Lykouris et al. (2018) first studied the multi-armed bandit with adversarial corruptions. Specifically, let C denote the *corruption level* which is the sum of the corruption magnitudes at each round. Lykouris et al. (2018) proposed an algorithm with a regret that is C times worse than the regret without corruption. Later, Gupta et al. (2019a) proposed an improved algorithm whose regret consists of two terms: a *corruption-independent* term that matches the optimal regret for multi-armed bandit without corruption, and a *corruption-dependent* term that is linear in C and independent of K, i.e.,

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

$\text{Regret}(K) = o(K) + O(C)$. The lower bound proved in Gupta et al. (2019a) suggests that the linear dependence on C is near-optimal. Such a regret structure reveals an desirable property of corruption-robust bandit algorithms, that is, the algorithm should perform nearly the same as the bandit algorithms without corruption when the corruption level C is small or diminishes.

Based on the above observation, a natural question arises:

Can we design computationally efficient algorithms for linear contextual bandits with corruption that can attain the best possible regret, similar to those in multi-armed bandits?

Some previous works have attempted to answer this question for the simpler stochastic linear bandit setting, where the decision sets at each round are identical and finite. Li et al. (2019) studied the stochastic linear bandits and proposed an instance-dependent regret bound. Later on, Bogunovic et al. (2021) studied the same problem and proposed an algorithm that achieves a regret with the corruption term depending on C linearly and on K logarithmically. However, these algorithms are limited to the stochastic linear bandit setting since their algorithm design highly relies on the experiment design and arm-elimination techniques that require a multiple selection of the same action and can only handle fixed decision set. They are not applicable to contextual bandits, where the decision set is changing over time and can even be infinite. For the more general linear contextual bandit setting, Bogunovic et al. (2021) proved that a simple greedy algorithm based on linear regression can attain an ideal corruption term that has a linear dependence on C and a logarithmic dependence on K, under a stringent diversity assumption on the contexts. Lee et al. (2021) proposed an algorithm and the corruption term in its regret depends on C linearly and on K logarithmically, but only holds for the restricted case when the corruption at each round is a linear function of the action. Without special assumptions on the contexts or corruptions, Zhao et al. (2021); Ding et al. (2021) proposed a variant of the OFUL algorithm (Abbasi-Yadkori et al., 2011) and its regret has a corruption term depending on K polynomially. Recently, Wei et al. (2022) proposed a Robust VOFUL algorithm that achieves a regret with a corruption term linearly dependent on C'[1] and only logarithmically dependent on K. However, Robust VOFUL is computationally inefficient since it needs to solve a maximization problem over a nonconvex confidence set that is defined as the intersection of exponential number of sets, and its regret has a loose dependence on context dimension d. In addition, Wei et al. (2022) also proposed a Robust OFUL algorithm and provided a regret guarantee that has a linear dependence on a different notion of corruption level C_r[2], which is strictly larger than the corruption level C considered in the previous work and the current paper. Thus, the above question remains open.

In this paper, we give an affirmative answer to the above question. We summarize our contributions as follows.

- We propose a computationally efficient algorithm based on the principle of optimism in the face of uncertainty (Abbasi-Yadkori et al., 2011), named Confidence-Weighted OFUL (CW-OFUL). At the core of our algorithm is a weighted ridge regression where the weight of each chosen arm is adaptive to its confidence, which is defined as the truncation of the inverse exploration bonus. Intuitively, such a weighting strategy prevents the algorithm from exploiting the contexts whose rewards are more likely corrupted by a large amount.
- For the case when the corruption level C is known to the agent, we show that the proposed algorithm enjoys a regret $\text{Regret}(K) = \widetilde{O}(d\sqrt{K} + dC)$, where d is the dimension of the contexts, C is the corruption level and K is the number of total iterations. The first term matches the regret lower bound of linear contextual bandits without corruption $\Omega(d\sqrt{K})$ (Lattimore and Szepesvári, 2018). The second term matches the lower bound on the corruption term in regret $\Omega(dC)$ (Bogunovic et al., 2021). They together suggest that our algorithm is not only robust but also near-optimal up to logarithmic factors.
- For the case when the corruption level C is unknown to the agent, we show that CW-OFUL enjoys an $\widetilde{O}(d\sqrt{K})$ regret for the case $C \leq \sqrt{K}$, with proper choice of the hyperparameter. Surprisingly,

[1] In Wei et al. (2022), the adversary adds corruption to all actions in the decision set before observing the agent's action and they define the corruption level C' as the maximum corruption over the decision set. See Remark 2.2 for the formal definition and a more detailed discussion.

[2] The corruption level C_r is defined as $C_r = \sqrt{K \sum_{k=1}^{K} c_k^2}$, where $c_k \geq 0$ is the corruption magnitude at round k. As a comparison, $C = \sum_{k=1}^{K} |c_k|$. In the worst case, $C_r = O(\sqrt{K}C)$ and therefore the corruption term in the regret of Robust OFUL will depend on K polynomially.

by proving a lower bound on the regret, we show that our regret upper bound is already optimal for all algorithms that achieve a near-optimal regret bound for uncorrupted bandits.

We compare our regret bounds with previous ones in Table 1. We can see that our algorithm matches the lower bound up to logarithmic factors in both known C and unknown C cases, and therefore is nearly optimal.

Table 1: Comparisons of regrets for corrupted linear contextual bandits.

Algorithm	Regret	C	Efficiency [3]	Adversary [4]
Robust weighted OFUL (Zhao et al., 2021)	$\widetilde{O}(d\sqrt{K}+dC'\sqrt{K})$	Known	Yes	Weak
Robust OFUL (Wei et al., 2022)	$\widetilde{O}(d\sqrt{K}+C_r)$	Known	Yes	Weak
Robust VOFUL (Wei et al., 2022)	$\widetilde{O}(d^{4.5}\sqrt{K}+d^4C')$	Known	No	Weak
CW-OFUL (Theorem 4.2)	$\widetilde{O}(d\sqrt{K}+dC)$	Known	Yes	Strong
CW-OFUL (Remark 4.4)	$\widetilde{O}(d\sqrt{K}+dC')$	Known	Yes	Weak
Lower bound (Lattimore and Szepesvári, 2018) (Bogunovic et al., 2021)	$\Omega(d\sqrt{K}+dC)$	Known	N/A	Strong
Multi-level weighted OFUL (Zhao et al., 2021)	$\widetilde{O}(dC'^2\sqrt{K}),\ C'=\Omega(1)$	Unknown	Yes	Weak
Greedy (Bogunovic et al., 2021)	$\widetilde{O}\big((\sqrt{dK}+C)/\lambda_0\big)$ [5]	Unknown	Yes	Strong
COBE+OFUL (Wei et al., 2022)	$\widetilde{O}(d\sqrt{K}+C_r)$	Unknown	Yes	Weak
COBE+VOFUL (Wei et al., 2022)	$\widetilde{O}(d^{4.5}\sqrt{K}+d^4C')$	Unknown	No	Weak
CW-OFUL($\bar{C}=\sqrt{K}$) (Theorem 4.9)	$\widetilde{O}(d\sqrt{K}),\ C\le\sqrt{K}$ $O(K),\ C\ge\sqrt{K}$	Unknown	Yes	Strong
COBE + CW-OFUL (Remark 4.10	$\widetilde{O}(d\sqrt{K}+dC')$	Unknown	Yes	Weak
Lower bound[6] (Lattimore and Szepesvári 2018) (Theorem 4.12)	$\Omega(d\sqrt{K}),\ C\le\sqrt{K}$ $\Omega(K),\ C\ge\sqrt{K}$	Unknown	N/A	Strong

1.1 Additional Related Work

Bandits with Misspecification. Bandits with misspecification can be seen as a special case of bandit with adversarial corruption since it is corrupted relative evenly at each round. Let ϵ be the misspecification level. Ghosh et al. (2017) firstly studied the stochastic linear bandits and proved a sublinear regret when ϵ is small. Lattimore and Szepesvari (2019) studied the stochastic linear bandit

[3]The weak adversary must corrupt the rewards before the agent selects its actions, while the powerful adversary (i.e., strong adversary) can corrupt the rewards after seeing the action being selected by the agent.

[4]In this work, we assume there is a computation oracle to solve the linear optimization problems over the decision set $\mathcal{D}_t$ (e.g., Line 3 of Algorithm 1). This is implicitly assumed in almost all existing works for solving contextual linear bandit problems with infinite arms (e.g., OFUL and LinUCB algorithms); otherwise, choosing an arm from the infinite decision set is computationally intractable. In the special case that the decision set is finite or the convex hull of a finite set, such a computation oracle apparently exists.).

[5]Greedy Bogunovic et al. (2021) assumes that each arm in the decision set at each round is sampled from a distribution that satisfies (r,λ_0)-diverse property (Kannan et al., 2018) . A distribution $\mathcal{D}$ is (r,λ_0)-diverse if for any $\mathbf{a}=\boldsymbol{\mu}+\boldsymbol{\xi}$ with $\boldsymbol{\mu}\in\mathbb{R}^d$ and $\boldsymbol{\xi}\sim\mathcal{D}$, $\lambda_{\min}(\mathbb{E}_{\boldsymbol{\xi}\in\mathcal{D}}[\mathbf{a}\mathbf{a}^\top|\boldsymbol{\theta}^\top\boldsymbol{\xi}\ge b])\ge\lambda_0$ holds for all $\boldsymbol{\theta}\in\mathbb{R}^d$ and $b\in\mathbb{R}$ satisfying $b\le r\|\boldsymbol{\theta}\|_2$.

[6]The lower bound under a large corruption level $C\ge\sqrt{K}$ only holds for algorithms that can achieve near-optimal regret for uncorrupted bandits. It is possible for an algorithm that does not achieve the optimal regret for uncorrupted bandits (e.g., $R_K=O(K^{0.75})$) to achieve a sub-linear regret in the presence of corruptions.

setting under milder assumptions. With the knowledge of ϵ, they proposed an algorithm with an $\widetilde{O}(\sqrt{dK\log(N)}+\epsilon\sqrt{d}K)$ regret, where d is the dimension of the contextual vector, N is the number of arms. Their regret bound matches their proved lower bound up to logarithmic factors. Foster et al. (2020) further considered the more general linear contextual bandits with misspecification when ϵ is unknown to the agent, and proposed an algorithm equipped with a CORRAL meta algorithm (Agarwal et al., 2017) to deal with the unknown ϵ. Their algorithm enjoys an $\widetilde{O}(d\sqrt{K}+\epsilon\sqrt{d}K)$ regret. Krishnamurthy et al. (2021) proposed an algorithm without using a meta algorithm which has the same order of regret as Foster et al. (2020). Our algorithm can be directly applied to the misspecification setting by choosing the corruption level C to be $K\epsilon$, which immediately gives us an $\widetilde{O}(d\sqrt{K}+dK\epsilon)$ regret upper bound.

Bandits with Adversarial Rewards. There exists a large body of literature on the problems of adversarial multi-armed bandits (Auer et al., 2002; Bubeck and Cesa-Bianchi, 2012). There is also a line of works trying to design algorithms that can achieve near-optimal regret bounds for both stochastic bandits and adversarial bandits simultaneously (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Zimmert and Seldin, 2019; Lee et al., 2021). However, most of these algorithms focus on the general adversarial reward setting without specifying the total amount of corruption. One of the notable exceptions is Lee et al. (2021), which assumed that the adversarial corruptions are generated through the inner product of an adversarial vectors and the contextual vector. As a comparison, our algorithm and result do not need such additional assumption on the structure of the corruption. Our algorithm can be applied to both corrupted and uncorrupted settings with different choices of hyperparameters, and achieves a near-optimal regret for both cases.

Notation We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. We denote by $[n]$ the set $\{1,\ldots,n\}$. For a vector $\mathbf{x}\in\mathbb{R}^d$ and a positive semi-definite matrix $\mathbf{\Sigma}\in\mathbb{R}^{d\times d}$, we denote by $\|\mathbf{x}\|_2$ the vector's ℓ_2 norm and by $\|\mathbf{x}\|_{\mathbf{\Sigma}}=\sqrt{\mathbf{x}^\top\mathbf{\Sigma}\mathbf{x}}$ the Mahalanobis norm. For two positive sequences $\{a_n\}$ and $\{b_n\}$ with $n=1,2,\ldots$, we write $a_n=O(b_n)$ if there exists an absolute constant $C>0$ such that $a_n\le Cb_n$ holds for all $n\ge 1$ and write $a_n=\Omega(b_n)$ if there exists an absolute constant $C>0$ such that $a_n\ge Cb_n$ holds for all $n\ge 1$. We use $\widetilde{O}(\cdot)$ to further hide the polylogarithmic factors. We use $\mathbb{1}\{\cdot\}$ to denote the indicator function.

2 Preliminaries

In this section, we introduce the setting of linear contextual bandit with adversarial corruption.

Linear contextual bandit with corruption. We define linear contextual bandits with corruption as follows: at the beginning of each round $k\in[K]$, the agent receives a decision set $\mathcal{D}_k\subseteq\mathbb{R}^d$ from the environment and it chooses an action (i.e., arm, contextual vector) $\mathbf{x}\in\mathcal{D}_k$. After choosing the action $\mathbf{x}_k$ at round k, the environment generates the corresponding r_k based on the stochastic linear model $r_k=\langle\boldsymbol{\theta}^*,\mathbf{x}\rangle+\eta_k$, where $\boldsymbol{\theta}^*\in\mathbb{R}^d$ is an unknown environment parameter and η_k is the stochastic noise. After seeing the stochastic reward r_k, the adversary (i.e., attacker) introduces an adversarial corruption c_k onto the reward, which may depend on the decision set $\mathcal{D}_k$, action $\mathbf{x}_k$, stochastic reward r_k. Finally, the agent observes the corrupted reward $\widehat{r}_k=\langle\boldsymbol{\theta}^*,\mathbf{x}\rangle+\eta_k+c_k$ at round k. Following Abbasi-Yadkori et al. (2011), we make the following assumptions on the bandit model.

Assumption 2.1. The linear contextual bandit satisfies the following conditions:

- At each round k and any action $\mathbf{x}\in\mathcal{D}_k$, we have $\|\mathbf{x}\|_2\le L$.
- For the unknown environment parameter $\boldsymbol{\theta}^*$, it satisfies $\|\boldsymbol{\theta}^*\|_2\le S$.
- At each round k, the corresponding stochastic noise η_k is conditional R-sub-Gaussian, i.e.,

$$\forall\lambda\in\mathbb{R},\ \mathbb{E}\big[e^{\lambda\eta_k}|\mathbf{x}_{1:k},\eta_{1:k-1},c_{1:k-1}\big]\le\exp(R^2\lambda^2/2).$$

Regret. The goal of the agent is to minimize the pseudo-regret in the first K rounds, which is defined as follows:

$$\text{Regret}(K)=\textstyle\sum_{k=1}^K\max_{\mathbf{x}\in\mathcal{D}_k}\langle\boldsymbol{\theta}^*,\mathbf{x}\rangle-\langle\boldsymbol{\theta}^*,\mathbf{x}_k\rangle.$$

Corruption level. To measure the level of adversarial corruptions, we define the *corruption level* as $C := \sum_{k=1}^{K} |c_k|$. With this definition, we say a linear contextual bandit problem is C-corrupted if and only if the corruption level is no larger than C.

Remark 2.2. The adversary in our setting and the corresponding definition of corruption level is the same as that in Bogunovic et al. (2021) and slightly different from that in prior works such as Lykouris et al. (2018); Gupta et al. (2019b); Zhao et al. (2021). More specifically, in these works, the adversarial corruption c_k is chosen before the choice of action $\mathbf{x}_k \in \mathcal{D}_k$. Since the actions selected by the agent may not be deterministic, the adversary chooses different corruption $c_{k,\mathbf{x}}$ for different action $\mathbf{x} \in \mathcal{D}_k$. With this notion of corruption, the corresponding corruption level is defined as $C' = \sum_{k=1}^{K} \max_{\mathbf{x}\in\mathcal{D}_k} |c_{k,\mathbf{x}}|$. As a comparison, our adversary chooses the corruption after observing the action x_k and for the corruption level. We have

$$C = \sum_{k=1}^{K} |c_{k,\mathbf{x}_k}| \leq \sum_{k=1}^{K} \max_{\mathbf{x}\in\mathcal{D}_k} |c_{k,\mathbf{x}}| = C',$$

which implies that our corruption level C is always no larger than the corruption level C' in Lykouris et al. (2018); Gupta et al. (2019b); Zhao et al. (2021).

3 Algorithms

In this section, we review existing algorithms for linear contextual bandits (and stochastic linear bandits) and discuss their limitations when they are applied to the adversarial corruption setting. Then we present our algorithm CW-OFUL and illustrate how our algorithm design can overcome the these limitations.

3.1 Existing Algorithms

We begin with reviewing the classical OFUL algorithm (Abbasi-Yadkori et al., 2011). Under Assumption 2.1, at round k, OFUL estimates $\boldsymbol{\theta}^*$ by online ridge regression over all the past observed actions and rewards, i.e.,

$$\boldsymbol{\theta}_k \leftarrow \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \lambda\|\boldsymbol{\theta}\|_2^2 + \textstyle\sum_{i=1}^{k-1} \big(\boldsymbol{\theta}^\top \mathbf{x}_i - r_i\big)^2. \tag{3.1}$$

With $\boldsymbol{\theta}_k$ in hand, OFUL constructs a confidence set for $\boldsymbol{\theta}^*$ as follows $\mathcal{C}_k = \Big\{\boldsymbol{\theta} : \|\boldsymbol{\theta}_k - \boldsymbol{\theta}\|_{\boldsymbol{\Sigma}_k} \leq \beta\Big\}$, where β is the confidence radius and $\boldsymbol{\Sigma}_k = \lambda\mathbf{I} + \sum_{i=1}^{k-1} \mathbf{x}_i\mathbf{x}_i^\top$ is the covariance matrix of contexts $\mathbf{x}_i, i = 1, \ldots, k$. Without corruption, it is known that setting $\beta = \widetilde{O}(R\sqrt{d})$ guarantees that $\boldsymbol{\theta}^* \in \mathcal{C}_k$ with high probability, which further leads to a sublinear regret $\widetilde{O}(d\sqrt{K})$. However, with corruption, such a choice of β is not sufficient. To see why, we take a closer look at the closed-form solution $\boldsymbol{\theta}_k$ to (3.1):

$$\boldsymbol{\theta}_k = \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{k-1} \mathbf{x}_i r_i = \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{k-1} \mathbf{x}_i(\mathbf{x}_i^\top \boldsymbol{\theta}^* + \eta_i) + \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{k-1} \mathbf{x}_i c_i.$$

By simple calculation and assuming λ to be a constant, we can show that $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_k}$ can be upper bounded by

$$\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_k} \leq O\Bigg(\underbrace{\Bigg\| \sum_{i=1}^{k-1} \mathbf{x}_i \eta_i \Bigg\|_{\boldsymbol{\Sigma}_k^{-1}}}_{I_1} + \underbrace{\Bigg\| \sum_{i=1}^{k-1} \mathbf{x}_i c_i \Bigg\|_{\boldsymbol{\Sigma}_k^{-1}}}_{I_2} \Bigg).$$

The first term I_1 is corruption-independent and bounded by $\widetilde{O}(R\sqrt{d})$ according to Abbasi-Yadkori et al. (2011). The challenge is to bound the second term I_2, which depends on the corruption. Existing approaches (Zhao et al., 2021; Ding et al., 2021) bound I_2 by triangle inequality and Cauchy-Schwarz inequality,

$$I_2 \leq \sum_{i=1}^{k-1} \|\mathbf{x}_i c_i\|_{\boldsymbol{\Sigma}_k^{-1}} \leq \sum_{i=1}^{k-1} |c_i| \max_{1\leq j\leq k-1} \|\mathbf{x}_j\|_{\boldsymbol{\Sigma}_k^{-1}} \leq \sum_{i=1}^{k-1} |c_i| L/\sqrt{\lambda} = O(C), \tag{3.2}$$

Algorithm 1 CW-OFUL

Require: Regularization parameter λ, confidence radius β and threshold parameter α

1: **for** round $k = 1, 2, ..$ **do**
2: Set $\mathbf{\Sigma}_k = \lambda \mathbf{I} + \sum_{i=1}^{k-1} w_i \mathbf{x}_i \mathbf{x}_i^\top$
3: Set $\mathbf{b}_k = \sum_{i=1}^{k-1} w_i \mathbf{x}_i r_i$ and $\boldsymbol{\theta}_k = \mathbf{\Sigma}_k^{-1} \mathbf{b}_k$
4: Receive the decision set $\mathcal{D}_k$
5: Choose action $\mathbf{x}_k \leftarrow \operatorname{argmax}_{\mathbf{x} \in \mathcal{D}_k} \boldsymbol{\theta}_k^\top \mathbf{x} + \beta \sqrt{\mathbf{x}^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}}$
6: Set $w_k = \min\{1, \alpha / \|\mathbf{x}_k\|_{\mathbf{\Sigma}_k^{-1}}\}$
7: **end for**

where C is the corruption level and $\max_{1 \le j \le k-1} \|\mathbf{x}_j\|_{\mathbf{\Sigma}_k^{-1}}$ is bounded by the crude upper bound $L/\sqrt{\lambda}$. Unfortunately, such a bound makes the confidence radius be the order of $O(R\sqrt{d} + C)$, which eventually leads to an term $O(C\sqrt{K})$ in the regret, which is C times worse than the regret without corruption.

In order to obtain a tighter bound of I_2, for stochastic linear bandits, Bogunovic et al. (2021) proposed a Robust Phase Elimination (RPE) algorithm, which employs *optimal design* (Lattimore and Szepesvári, 2018) to select the arms. In this setting, the decision set is finite and fixed over time, i.e., $\mathcal{D}_k = \mathcal{D}$ for all $k \in [K]$ and $|\mathcal{D}| \le \infty$. More specifically, RPE divides the time horizon into several phases. Within each phase, RPE performs linear regression on a multiset $\mathcal{A} \subset \mathcal{D}$, which is the G-optimal design of $\mathcal{D}$. Here the multiset means $\mathcal{A}$ has duplicate elements. Let $\mathbf{\Sigma}$ be the covariance matrix defined over $\mathcal{A}$, then the following upper bound holds (Lattimore and Szepesvári, 2018):

$$\forall \mathbf{x} \in \mathcal{D},\ \|\mathbf{x}\|_{\mathbf{\Sigma}^{-1}} = O\big(|\mathcal{A}|^{-1/2}\big). \tag{3.3}$$

By choosing a large enough $|\mathcal{A}|$, (3.3) provides a *uniformly small* upper bound for $\max_{1 \le j \le k-1} \|\mathbf{x}_j\|_{\mathbf{\Sigma}_k^{-1}}$ for any k. Substituting (3.3) back into (3.2) with $|\mathcal{A}| = O(C^2)$, we can show that I_2 is bounded by some small constant, which therefore eliminates the $O(C\sqrt{K})$ term in the final regret. Although the optimal design-based approach RPE (Bogunovic et al., 2021) successfully eliminates the multiplicative term $C\sqrt{K}$, it is not applicable to our linear contextual bandit setting: (1) it needs to select a *multiset* from the decision set, which is impossible for the general contextual bandit setting; (2) the complexity of optimal design introduces some additional quadratic term C^2 in their final regret, which makes their algorithm non-optimal (See Bogunovic et al. (2021) for more details).

3.2 Our Algorithm

As we have seen before, it is pivotal to bound the corruption-dependent term I_2 tightly. To overcome the limitations of existing approaches, we propose a fundamentally new approach and present our CW-OFUL in Algorithm 1. At a high level, Algorithm 1 is an extension of the OFUL algorithm (Abbasi-Yadkori et al., 2011), which is also based on the principle of optimism in the face of uncertainty.

Our algorithm assigns a weight w_k to each selected action $\mathbf{x}_k$. More specifically, at round k, we use the following weighted ridge regression to estimate the unknown vector $\boldsymbol{\theta}^*$:

$$\boldsymbol{\theta}_k \leftarrow \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \lambda \|\boldsymbol{\theta}\|_2^2 + \textstyle\sum_{i=1}^{k-1} w_i \big(\boldsymbol{\theta}^\top \mathbf{x}_i - r_i\big)^2. \tag{3.4}$$

The closed-form solution to the above optimization problem is displayed in Line 3 of Algorithm 1. While weighted ridge regression is not new and has been used in prior work on bandits (Kirschner and Krause, 2018; Zhou et al., 2021; Russac et al., 2019), the setting, motivation and the choice of weight are fundamentally different. More specifically, we choose the weight as the *truncation* of the inverse exploration bonus, which is $w_k = \min\left\{1, \alpha / \|\mathbf{x}_k\|_{\mathbf{\Sigma}_k^{-1}}\right\}$. Here $\alpha > 0$ is a threshold parameter. We can see that for action $\mathbf{x}_k$ with a large exploration bonus $\|\mathbf{x}_k\|_{\mathbf{\Sigma}_k^{-1}}$ (low confidence), CW-OFUL will assign a small weight to it to avoid the potentially large regret caused by both the stochastic noise and the adversarial corruption. On the other hand, for the action with a small exploration bonus (high

confidence), CW-OFUL will assign a large weight to it (it can be as large as 1). Another interesting observation is that by setting α to be sufficiently large, the weight will become 1 for every action, and CW-OFUL will degenerate to OFUL (Abbasi-Yadkori et al., 2011).

As a comparison, Kirschner and Krause (2018); Zhou et al. (2021) used the inverse of the noise variance as the weight to normalize the noise and derived tight variance-dependent regret guarantees. Russac et al. (2019) set the weight as a geometric sequence to perform moving average to deal with the non-stationary environment.

To see how our choice of weight can lead to tighter regret, we first write down the closed-form solution to (3.4)

$$\boldsymbol{\theta}_k = \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^{k-1} w_i \mathbf{x}_i (\mathbf{x}_i^\top \boldsymbol{\theta}^* + \eta_i) + \sum_{i=1}^{k-1} \boldsymbol{\Sigma}_k^{-1} w_i \mathbf{x}_i c_i,$$

where the covariance matrix $\boldsymbol{\Sigma}_k = \lambda \mathbf{I} + \sum_{i=1}^{k-1} w_i \mathbf{x}_i \mathbf{x}_i^\top$. With some calculation and assuming λ to be a constant, we can obtain

$$\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_k} \leq O\Bigg(\underbrace{\Bigg\| \sum_{i=1}^{k-1} w_i \mathbf{x}_i \eta_i \Bigg\|_{\boldsymbol{\Sigma}_k^{-1}}}_{I_1} + \underbrace{\Bigg\| \sum_{i=1}^{k-1} w_i \mathbf{x}_i c_i \Bigg\|_{\boldsymbol{\Sigma}_k^{-1}}}_{I_2} \Bigg).$$

I_1 is the corruption-independent term and can still be bounded by $\widetilde{O}(R\sqrt{d})$ according to Abbasi-Yadkori et al. (2011). For I_2, we have

$$\Bigg\| \sum_{i=1}^{k-1} w_i \mathbf{x}_i c_i \Bigg\|_{\boldsymbol{\Sigma}_k^{-1}} \leq \sum_{i=1}^{k-1} |c_i| w_i \big\|\mathbf{x}_i\big\|_{\boldsymbol{\Sigma}_k^{-1}} \leq \sum_{i=1}^{k-1} |c_i| \alpha = C\alpha,$$

It is evident that with our carefully designed weight, the corruption-dependent term I_2 can be uniformly bounded by some constant $C\alpha$, the same as that in Bogunovic et al. (2021). Therefore, by setting α to be sufficiently small, our CW-OFUL can get rid of the $C\sqrt{K}$ term in the final regret.

4 Main Results

In this section, we present the main theoretical guarantees of CW-OFUL.

4.1 Known Corruption Level C: Upper Bound

We first consider the case when C is known to the agent. In this case, we choose $\alpha = R\sqrt{d}/C$. The following lemma characterizes the estimation error of $\boldsymbol{\theta}_k$ with respect to $\boldsymbol{\theta}^*$, which is a formal summary of our discussion in Section 3.

Lemma 4.1. Suppose that Assumption 2.1 holds. For any $0 < \delta < 1$ and corruption budget $C \geq 0$, set the confidence radius $\beta = R\sqrt{d \log\big((1 + KL^2/\lambda)/\delta\big)} + \sqrt{\lambda} S + \alpha C$ in Algorithm 1, then with probability at least $1 - \delta$, for every round k, the estimator $\boldsymbol{\theta}_k$ satisfies that $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_k} \leq \beta$.

The following theorem provides the regret bound of Algorithm 1.

Theorem 4.2. Suppose that Assumption 2.1 holds. For any $0 < \delta < 1$ and corruption budget $C \geq 0$, set the confidence radius β in Algorithm 1 as follows:

$$\beta = R\sqrt{d \log\big((1 + KL^2/\lambda)/\delta\big)} + \alpha C + \sqrt{\lambda} S.$$

Then with probability at least $1 - \delta$, its regret in the first K rounds is upper bounded by

$$\begin{aligned}\text{Regret}(K) = O\Bigg(& dR\sqrt{K \log^2\big((1 + KL^2/\lambda)/\delta\big)} + \alpha C \sqrt{dK \log^2\big((1 + KL^2/\lambda)/\delta\big)} \\ & + S\sqrt{d\lambda K \log(1 + KL^2/\lambda)} + \frac{Rd^{1.5}}{\alpha} \times \sqrt{\log^3\big((1 + KL^2/\lambda)/\delta\big)}\end{aligned}$$

$$+ \frac{dS\sqrt{\lambda}}{\alpha} \times \sqrt{\log^2\left((1+KL^2/\lambda)/\delta\right)} + dC\sqrt{\log^2\left((1+KL^2/\lambda)/\delta\right)}\Bigg).$$

In addition, if choosing $\alpha = (R\sqrt{d} + \sqrt{\lambda}S)/C$ and $\lambda = R^2/S^2$, its regret can be upper bounded by

$$\text{Regret}(K) = \widetilde{O}(d\sqrt{K} + dC).$$

A few remarks about Theorem 4.2 are in order.

Remark 4.3. Compared with the $\widetilde{O}(d\sqrt{K} + dC\sqrt{K})$ regret proved in Zhao et al. (2021); Ding et al. (2021), our algorithm improves the multiplicative dependence on corruption level C to additive dependence. In particular, CW-OFUL achieves the same order of regret as the uncorrupted setting when $C = O(\sqrt{K})$, and it attains a sublinear regret as long as $C = o(K)$. In sharp contrast, the algorithm proposed in Zhao et al. (2021) achieves the same order of regret as the uncorrupted setting only when $C = O(1)$, and has a sublinear regret only when $C = o(\sqrt{K})$.

Remark 4.4. We also compare our result with that in Wei et al. (2022). The Robust+OFUL algorithm in Wei et al. (2022) achieves an $\widetilde{O}(d\sqrt{K} + C_r)$ regret with $C_r = \sqrt{T\sum_{k=1}^{K} c_k^2}$, which will degenerate to $\widetilde{O}(d\sqrt{K} + d\sqrt{K}C)$ in the worst case. Their regret guarantee is always worse than ours when $C < \sqrt{K}$. In addition, according to the discussion in Remark 2.2, Theorem 4.2 also implies an $\widetilde{O}\big(d\sqrt{K} + dC'\big)$ regret under the notion of the corruption level C'. In contrast, the Robust VOFUL algorithm in Wei et al. (2022) has an $\widetilde{O}(d^{4.5}\sqrt{K} + d^4C')$ regret, which is also inferior to our regret. Furthermore, Robust VOFUL is computationally inefficient.

Remark 4.5. We further compare our result with previous additive regrets derived for stochastic linear bandits. Let $\mathcal{D}_k = \mathcal{D}$ be the decision set. Compared with the $O(\sqrt{dK\log|\mathcal{D}|} + Cd^{3/2})$ regret for stochastic linear bandit with corruption derived in Bogunovic et al. (2021), our regret improves the corruption term by a factor of $\sqrt{d}$. Note that the $\sqrt{d}$ difference in the leading $\sqrt{K}$ term between our regret and theirs is caused by the fact that Bogunovic et al. (2021) considered the finite-arm setting, while we consider the infinite-arm setting. Our algorithm will have the same regret as theirs when $|\mathcal{D}| = O(\exp(d))$.

Remark 4.6. For the uncorrupted setting where $C = 0$, Theorem 4.2 suggests that the threshold parameter α should be set to infinity. Then by Line 6 in Algorithm 1, each weight w_k becomes 1, and CW-OFUL degenerates to OFUL. Meanwhile, the regret in Theorem 1 also becomes $\widetilde{O}(d\sqrt{K})$ that matches the regret of OFUL (Abbasi-Yadkori et al., 2011).

4.2 Known Corruption Level C: Lower Bound

In this subsection, we refer to two existing lower bound results to show that when C is known, our $\widetilde{O}(d\sqrt{K} + dC)$ regret is optimal up to logarithmic factors. The first proposition shows that the $\widetilde{O}(d\sqrt{K})$ corruption-independent term in our regret is near-optimal.

Proposition 4.7 (Theorem 24.2, Lattimore and Szepesvári 2018)**.** Assume $d \le 2K$, $R = 1$ and $\mathcal{D}_k = \{\|\mathbf{x}\|_2 \le 1\}$ for all $k \ge 1$. Then for any algorithm, there exists an environment parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ satisfying $\|\boldsymbol{\theta}^*\|_2^2 = d^2/(48K)$ such that $\mathbb{E}(\text{Regret}(K)) \ge d\sqrt{K}/(16\sqrt{3})$.

The second proposition suggests that the $O(dC)$ corruption term in our regret is optimal.

Proposition 4.8 (Theorem 3, Bogunovic et al. 2021)**.** For any dimension d, for any algorithm that has the knowledge of C, there exists an instance satisfying with probability at least 0.5, $\text{Regret}(K) = \Omega(dC)$.

Combining Propositions 4.7 and 4.8, we can conclude that for any algorithm, there exists a corrupted bandit instance such that the algorithm suffers at least $\Omega(\max\{d\sqrt{K}, dC\})$ regret. Such a lower bound matches our upper bound up to logarithmic factors. Therefore, our algorithm is nearly optimal.

4.3 Unknown Corruption Level C: Upper Bound

Now we consider the case when C is unknown. Our solution is quite simple for this case: we introduce a tuning parameter $\bar{C}$, which can be viewed as an estimate of C, and select the threshold

parameter α as Theorem 4.2 suggests. The following theorem gives the regret upper bound of CW-OFUL for the unknown C case.

Theorem 4.9. Under the same conditions of Theorem 4.2 except that we set $\alpha = (R\sqrt{d} + \sqrt{\lambda}S)/\bar{C}$ with $\bar{C}$ being an estimated corruption level, $\lambda = R^2/S^2$ and $\beta = 2R\sqrt{d\log\big((1+KL^2/\lambda)/\delta\big)} + 2\sqrt{\lambda}S$ in Algorithm 1. The regret of CW-OFUL can be upper bounded in the following two cases:

- If the corruption level C satisfies that $0 \le C \le \bar{C}$, then with probability at least $1-\delta$, the regret is upper bounded by $\text{Regret}(K) = \widetilde{O}(dR\sqrt{K} + d\bar{C})$.
- If the corruption level C satisfies that $C > \bar{C}$, the regret is upper bounded by $\text{Regret}(K) = O(K)$.

In addition, if we set the estimation $\bar{C} = \sqrt{K}$, then when $0 \le C \le \sqrt{K}$, the regret is upper bounded by $\widetilde{O}(d\sqrt{K})$.

Remark 4.10. Zhao et al. (2021) proposed an $\widetilde{O}(C^2 d\sqrt{K})$ regret with unknown $C = \Omega(1)$. Compared with their result, our regret (with $\bar{C} = \sqrt{K}$) is strictly better in the corruption term. Bogunovic et al. (2021) proposed an $\widetilde{O}(\sqrt{dK\log|\mathcal{D}|} + Cd^{1.5} + C^2)$ regret for the stochastic linear bandit with unknown C, in the regime $C = \widetilde{O}(\sqrt{K}/d)$, where $\mathcal{D}$ is the finite decision set. Such a regret becomes $\widetilde{O}(d\sqrt{K} + Cd^{1.5} + C^2)$ when the size of $\mathcal{D}$ becomes exponentially large in d or even infinite. Compared with their regret, our regret is not only smaller, but also holds for a wider regime (i.e., $C = O(\sqrt{K})$). Compared with the greedy algorithm in Bogunovic et al. (2021), our result does not rely on the stringent (r, λ_0)-diverse property assumption on the contexts.

Remark 4.11. We also compare our result (choosing $\bar{C} = \sqrt{K}$) with those in Wei et al. (2022) for the unknown C case. Wei et al. (2022) proposed a COBE+OFUL algorithm with an $\widetilde{O}(d\sqrt{K}+C_r)$ regret, and a COBE+VOFUL algorithm with an $\widetilde{O}(d^{4.5}\sqrt{K} + d^4 C')$ regret, analogous to their results for the known C case discussed in Remark 4.4. Our CW-OFUL enjoys a better regret than COBE+OFUL for all C, and it is better than COBE+VOFUL for $C < \sqrt{K}$. In addition, for the modified notion of corruption level C', if we choose the basic algorithm in COBE (Wei et al., 2022) as our CW-OFUL algorithm, then Theorem 3 in Wei et al. (2022) suggests that COBE+CW-OFUL can deal with unknown corruption level C' and obtained an $\widetilde{O}(d\sqrt{K} + dC')$ regret guarantee, which matches the regret of CW-OFUL algorithm with known corruption level C'. Note that COBE+VOFUL is also computationally inefficient.

4.4 Unknown Corruption Level C: Lower Bound

With $\bar{C} = \sqrt{K}$, for the case when $0 \le C \le \sqrt{K}$, our regret result is already near-optimal, due to the lower bound for the uncorrupted bandit in Proposition 4.7. Now we show that our $O(K)$ bound, seemingly trivial, is actually optimal for a large class of bandit algorithms. In detail, the following theorem provides a lower bound result for any algorithm for the unknown C case. This is an extension of the lower bound result in Bogunovic et al. (2021) from $d = 2$ to general d.

Theorem 4.12. For any algorithm **Alg**, let R_K be an upper bound of $\text{Regret}(K)$ such that for any bandit instance satisfying Assumption 2.1 with $C = 0$, it satisfies the $\mathbb{E}\big[\text{Regret}(K)\big] \le R_K \le O(K)$, where the expectation is with respect to the randomness of the algorithm and the stochastic noise. Then for the general case with $C = \Omega(R_K/d)$, such an algorithm will have $\mathbb{E}\big[\text{Regret}(K)\big] = \Omega(K)$.

Remark 4.13. If we selects the estimated corruption $\bar{C} = \Omega(R_K/d)$, Theorem 4.9 immediately implies that CW-OFUL enjoys a $O(R_K)$ regret when corruption level $C < \Omega(R_K/d)$ and $O(K)$ regret when corruption level $C \ge \Omega(R_K/d)$. Compared with the algorithm **Alg**, Theorem 4.12 suggests that CW-OFUL is no worse than the algorithm **Alg** no matter whether the corruption level $C < \Omega(R_K/d)$. More discussion can be found in Appendix A.1.

5 Conclusion and Future Work

In this work, we study corrupted linear contextual bandits. We propose a CW-OFUL algorithm based on a weighted ridge regression with truncated inverse exploration bonus weights. We show that for

both cases when the corruption level C is known or unknown to the agent, CW-OFUL achieves a regret that matches the lower bound up to logarithmic factors.

We are also interested in achieving the optimal regret when specializing our algorithm to the misspecified linear contextual bandits.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and area chair for their helpful comments. JH, DZ and QG are supported in part by the National Science Foundation CAREER Award 1906169 and the Sloan Research Fellowship. TZ is supported in part by the GRF 16201320. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*.

AGARWAL, A., LUO, H., NEYSHABUR, B. and SCHAPIRE, R. E. (2017). Corralling a band of bandit algorithms. In *Conference on Learning Theory*. PMLR.

AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing* **32** 48–77.

AUER, P. and CHIANG, C.-K. (2016). An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*. PMLR.

BOGUNOVIC, I., LOSALKA, A., KRAUSE, A. and SCARLETT, J. (2021). Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

BUBECK, S. and CESA-BIANCHI, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721* .

BUBECK, S. and SLIVKINS, A. (2012). The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings.

CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, learning, and games*. Cambridge university press.

DESHPANDE, Y. and MONTANARI, A. (2012). Linear bandits in high dimension and recommendation systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE.

DING, Q., HSIEH, C.-J. and SHARPNACK, J. (2021). Robust stochastic linear contextual bandits under adversarial attacks. *arXiv preprint arXiv:2106.02978* .

FOSTER, D. J., GENTILE, C., MOHRI, M. and ZIMMERT, J. (2020). Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems* **33** 11478–11489.

GHOSH, A., CHOWDHURY, S. R. and GOPALAN, A. (2017). Misspecified linear bandits. In *Thirty-First AAAI Conference on Artificial Intelligence*.

GUPTA, A., KOREN, T. and TALWAR, K. (2019a). Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*. PMLR.

GUPTA, A., KOREN, T. and TALWAR, K. (2019b). Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*. PMLR.

HORN, R. A. and JOHNSON, C. R. (2012). *Matrix analysis*. Cambridge university press.

KANNAN, S., MORGENSTERN, J. H., ROTH, A., WAGGONER, B. and WU, Z. S. (2018). A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems* **31**.

KIRSCHNER, J. and KRAUSE, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*. PMLR.

KRISHNAMURTHY, S. K., HADAD, V. and ATHEY, S. (2021). Adapting to misspecification in contextual bandits with offline regression oracles. In *International Conference on Machine Learning*. PMLR.

LATTIMORE, T. and SZEPESVÁRI, C. (2018). Bandit algorithms. *preprint* 28.

LATTIMORE, T. and SZEPESVARI, C. (2019). Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676* .

LEE, C.-W., LUO, H., WEI, C.-Y., ZHANG, M. and ZHANG, X. (2021). Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*. PMLR.

LI, Y., LOU, E. Y. and SHAN, L. (2019). Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109* .

LYKOURIS, T., MIRROKNI, V. and PAES LEME, R. (2018). Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*.

RUSSAC, Y., VERNADE, C. and CAPPÉ, O. (2019). Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems* **32**.

SELDIN, Y. and LUGOSI, G. (2017). An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*. PMLR.

SELDIN, Y. and SLIVKINS, A. (2014). One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*. PMLR.

WEI, C.-Y., DANN, C. and ZIMMERT, J. (2022). A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*. PMLR.

ZHAO, H., ZHOU, D. and GU, Q. (2021). Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615* .

ZHOU, D., GU, Q. and SZEPESVARI, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

ZIMMERT, J. and SELDIN, Y. (2019). An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We study corrupted linear contextual bandits and propose a computation efficient algorithm with a near-optimal regret guarantee.
   (b) Did you describe the limitations of your work? [Yes] See the discussion about Misspecified linear bandits in subsection 4.1.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work focuses on a purely theoretical problem and does not have any negative social impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumption 2.1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional Results

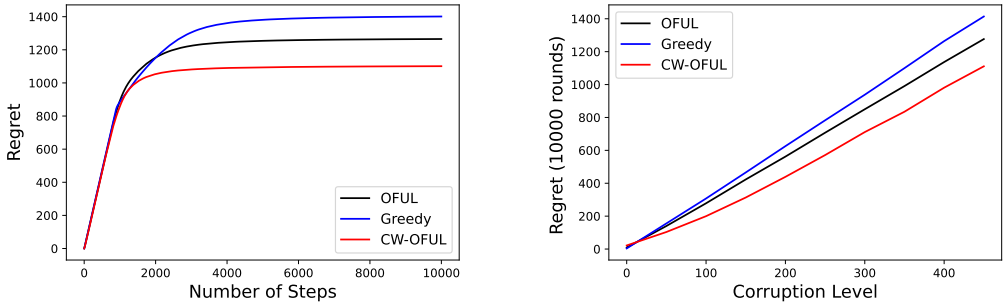
A.1 Discussion on the Lower Bound for Unknown Corruption Level C

Consider a class of algorithms $\mathcal{A}$ whose worst-case regret is R_K in the uncorrupted case. Here we only need to consider $\Omega(d\sqrt{K}) \leq R_K \leq O(K)$, since for any algorithm, $\Omega(d\sqrt{K})$ is the lowest possible worst-case regret (Lattimore and Szepesvári, 2018) and $O(K)$ is the highest possible regret. We first show that CW-OFUL belongs to $\mathcal{A}$. Choosing $\bar{C} = R_K/d$, Theorem 4.9 immediately suggests that CW-OFUL enjoys a R_K regret in the uncorrupted case (i.e., $C = 0$). Thus CW-OFUL belongs to $\mathcal{A}$. Then we will show that CW-OFUL is the best possible one in $\mathcal{A}$. On the one hand, Theorem 4.9 suggests that CW-OFUL suffers a linear regret when $C > \bar{C} = R_K/d$. On the other hand, Theorem 4.12 shows that any algorithm with R_K regret in the uncorrupted case should have a linear regret when $C = \Omega(R_K/d)$. These together imply that CW-OFUL is optimal within $\mathcal{A}$.

A.2 Discussion on the Misspecified Linear Bandits

We consider the misspecified linear bandit setting which assumes that the corruption at each round is uniformly bounded by ϵ. Clearly, the misspecified linear bandit is a special case of corrupted linear contextual bandit with $C = K\epsilon$. Theorem 4.2 suggests that a direct application of our algorithm to this special setting incurs an $\widetilde{O}(d\sqrt{K} + dK\epsilon)$ regret, which differs from the near-optimal regret $\widetilde{O}(d\sqrt{K} + \sqrt{d}K\epsilon)$ (Lattimore and Szepesvari, 2019; Foster et al., 2020) by a $\sqrt{d}$ factor on the corruption term. Whether our algorithm is able to achieve the near-optimal regret for both misspecified linear bandit and corrupted linear contextual bandit simultaneously remains an open question.

B Experiments



(a) Corruption level $C = 450$. Cumulative regret versus Round

(b) Cumulative regret versus Corruption level

Figure 1: Comparison of CW-OFUL(ours), Greedy (Bogunovic et al., 2021) and OFUL(Abbasi-Yadkori et al., 2011). Experiments are run for unknown corruption levels C from $\{0, 50, 100, ..., 450\}$ (10 different corruption levels) , and results are averaged over 100 runs. Figure 1(a) presents the cumulative regrets with unknown corruption $C = 450$; Figure 1(b) shows the cumulative regret versus unknown corruption level C.

In this section, we run experiments and evaluate the performance of our algorithm CW-OFUL with an unknown corruption level C, which corroborates our theory. The code and data for our experiments can be found on Github [3].

Model Parameters We construct a linear bandit instance with dimension $d = 5$ and the true model parameter $\boldsymbol{\theta}^*$ is denoted by

$$\boldsymbol{\theta}^* = \left[\frac{1}{\sqrt{d}}, ..., \frac{1}{\sqrt{d}}\right]^\top \in \mathbb{R}^d.$$

[3] https://github.com/uclaml/CW-OFUL

During each round $k \in [K]$, the decision set $\mathcal{D}_k$ consists of 20 different actions and each action is uniformly sampled from the space $[+1/\sqrt{d}, -1/\sqrt{d}]^d$, which satisfies the positive minimum eigenvalue assumption for greedy algorithm (Bogunovic et al., 2021). In addition, after choosing the action, $\mathbf{x}_k$ in round $k \in [K]$, an 0.1-Gaussian noise η_k will be added to the reward.

Attack method For the attack method, we choose the flip-$\boldsymbol{\theta}$ attack. More specifically, with a corruption level of C, the adversary tricks the learner by flipping the value, i.e., $r_t(\mathbf{x}_t) = -\langle \mathbf{x}_t, \theta^* \rangle + \eta_k$ in the first C rounds. In the remaining rounds, the adversary does not corrupt the reward.

Results and discussions In our experiments, we make a simulation with the total number of rounds $K = 10000$ (repeating 100 times and taking the average) and corruption levels from $\{0, 50, 100, ..., 450\}$ (10 different corruption levels). We applied our CW-OFUL algorithm and compared its performance with greedy (Bogunovic et al., 2021) and OFUL (Abbasi-Yadkori et al., 2011).

The experimental results are shown in Figure 1 These simulation results suggest that our CW-OFUL algorithm outperforms both the Greedy and OFUL algorithms. Specifically, Figure 1(a) displays the cumulative regret of our algorithm, OFUL, and Greedy algorithm under the same number of rounds. They show that our algorithm slightly outperforms them in terms of regret.

Figure 1(b) plots the cumulative regret versus the unknown corruption level. We can see that all of the three algorithms demonstrate an additive linear dependence on the unknown corruption level C, which corroborate our theoretical guarantee.

C Instance-dependent Regrets

Prior works (Lykouris et al., 2018; Li et al., 2019; Zhao et al., 2021) have proved instance-dependent regret bounds for corruption-robust linear bandits. We show that CW-OFUL also enjoys an instance-dependent regret bound. Following Abbasi-Yadkori et al. (2011), we define the minimal sub-optimality gap as follows.

Definition C.1 (Minimal sub-optimality gap)**.** For each round $k \in [K]$ and any action $\mathbf{x} \in \mathcal{D}_k$, the sub-optimality gap $\Delta_{\mathbf{x},k}$ is defined as

$$\Delta_{\mathbf{x},k} = \max_{\mathbf{x}^* \in \mathcal{D}_k} \langle \boldsymbol{\theta}^*, \mathbf{x}^* \rangle - \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle,$$

and the minimal sub-optimality gap is defined as

$$\Delta = \min_{k \in [K], \mathbf{x} \in \mathcal{D}_k} \big\{ \Delta_{\mathbf{x},k} : \Delta_{\mathbf{x},k} \neq 0 \big\}. \tag{C.1}$$

We assume that the minimal sub-optimality gap is strictly positive.

Assumption C.2. The minimal sub-optimality gap is strictly positive, i.e., $\Delta > 0$.

Under the assumption of positive minimal sub-optimality, the following theorem provides an instance-dependent regret guarantee for CW-OFUL.

Theorem C.3. Under the same conditions of Theorem 4.2, with high probability at least $1 - \delta$, the regret of Algorithm 1 in the first K rounds is upper bounded by

$$\begin{aligned}\text{Regret}(K) \leq O\Bigg(& R^2 d^2 \log^2\big((1 + KL^2/\lambda)/\delta\big)/\Delta + \frac{\alpha^2 d C^2}{\Delta} \times \sqrt{\log\big(3 + C^2 L^2 K/(R^2 \lambda \delta)\big)} \\ & + S^2 d \lambda \log(1 + KL^2/\lambda)/\Delta + \frac{R d^{1.5}}{\alpha} \times \sqrt{\log^3\big((1 + KL^2/\lambda)/\delta\big)} \\ & + \frac{dS\sqrt{\lambda}}{\alpha} \times \sqrt{\log^2\big((1 + KL^2/\lambda)/\delta\big)} + dC\sqrt{\log^2\big((1 + KL^2/\lambda)/\delta\big)}\Bigg)\end{aligned}$$

In addition, if choosing $\alpha = (R\sqrt{d} + \sqrt{\lambda} S)/C$ and $\lambda = R^2/S^2$, the regret can be upper bounded by

$$\text{Regret}(K) \leq \widetilde{O}(d^2/\Delta + dC).$$

Remark C.4. Our regret is strictly better than the $\widetilde{O}(d^{2.5}C/\Delta + d^6/\Delta^2)$ regret proved by Li et al. (2019) under a stronger assumption. Meanwhile, Zhao et al. (2021) implies an $\widetilde{O}(d^2 C/\Delta)$ regret for their algorithm under the known C case, which is also worse than our result.

D Overview of Key Proof Techniques

In this section, we give an overview of the main technical difficulty and our proof technique to derive Theorem 4.2.

By the standard regret decomposition technique from Abbasi-Yadkori et al. (2011), we upper bound the regret by the sum of the exploration bonuses times the confidence radius:

$$\text{Regret}(K) = O\bigg(\beta \cdot \sum_{k=1}^{K} \sqrt{\mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_k}\bigg). \tag{D.1}$$

Lemma 4.1 suggests $\beta \sim R\sqrt{d} + \alpha C$. Therefore, we only need to bound the summation of the exploration bonuses. For the basic case when $w_k = 1$, we bound it using the elliptical potential lemma (Abbasi-Yadkori et al., 2011) as follows

$$\sum_{w_k=1} \sqrt{\mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_k} \leq \sum_{k=1}^{K} \sqrt{\mathbf{x}_k^\top \bigg(\lambda \mathbf{I} + \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{x}_i^\top\bigg)^{-1} \mathbf{x}_k} \sim \widetilde{O}(\sqrt{dK}), \tag{D.2}$$

which contributes to the corruption-independent term $dR\sqrt{K}$ in our regret. For the case when $w_k < 1$, however, we are facing the *weighted* covariance matrix and cannot directly use the elliptical potential lemma. A trivial approach is to lower bound the weights by their *uniform* lower bound, i.e.,

$$\lambda \mathbf{I} + \sum_{i=1}^{k-1} w_i \mathbf{x}_i \mathbf{x}_i^\top \succeq \min_{1 \leq i \leq k-1} w_i \cdot \bigg(\lambda \mathbf{I} + \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{x}_i^\top\bigg). \tag{D.3}$$

By the definition of the weight w_i in Algorithm 1 and a crude upper bound for the exploration bonus, we conclude from the definition of w_k that $w_k = \Omega(\alpha)$. Substituting it into (D.3), we only obtain a regret $\widetilde{O}(\sqrt{dK}/\alpha)$, which is not satisfying.

To overcome this issue, we recall the definition for weight $w_k < 1$ in Algorithm 1: $w_k = \alpha / \|\mathbf{x}_k\|_{\boldsymbol{\Sigma}_k^{-1}}$ and we can bound the summation of the exploration bonuses as

$$\sum_{w_k<1} \sqrt{\mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_k} = \sum_{k=1}^{K} w_k \mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_k / \alpha \sim \widetilde{O}(d/\alpha). \tag{D.4}$$

Combining the results in (D.2) and (D.4) into (D.1), we can prove the final regret.

E Proof of Theorem 4.2

In this section, we provide the proof of Theorem 4.2. For simplicity, we use $\mathcal{E}$ to denote the following event:

$$\mathcal{E} = \bigg\{ \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_k} \leq \beta, \forall k \in [K] \bigg\}.$$

Lemma 4.1 shows that $\Pr(\mathcal{E}) \geq 1 - \delta$.

Lemma E.1. If setting the confidence radius $\beta = R\sqrt{d \log\big((1 + KL^2/\lambda)/\delta\big)} + \alpha C + \sqrt{\lambda} S$ in Algorithm 1, then on the event $\mathcal{E}$, for each round $k \in [K]$, the regret at round k is upper bounded by

$$\Delta_k = \max_{\mathbf{x} \in \mathcal{D}_k} \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle - \langle \boldsymbol{\theta}^*, \mathbf{x}_k \rangle \leq 2\beta \sqrt{\mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_k}.$$

Proof of Theorem 4.2. Based on the event $\mathcal{E}$, the regret in the first K round can be decomposed into two parts based on the weight w_k:

$$\text{Regret}(K) = \sum_{k=1}^{K} \max_{\mathbf{x} \in \mathcal{D}_k} \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle - \langle \boldsymbol{\theta}^*, \mathbf{x}_k \rangle$$

$$
\begin{aligned}
&\leq \min\left(2, \sum_{k=1}^{K} 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right) \\
&= \underbrace{\sum_{k:w_k=1} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right)}_{I_1} + \underbrace{\sum_{k:w_k<1} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right)}_{I_2}, \qquad \text{(E.1)}
\end{aligned}
$$

where the inequality holds due to the Lemma E.1 with the fact that the suboptimality in each round k is no more than 2.

For the term I_1, we consider for all rounds $k \in [K]$ with $w_k = 1$ and we assume these rounds can be listed as $\{k_1, .., k_m\}$ for simplicity. With this notation, for each $i \leq m$, we can construct the auxiliary covariance matrix $\mathbf{A}_i = \lambda\mathbf{I} + \sum_{j=1}^{i-1} \mathbf{x}_{k_j}\mathbf{x}_{k_j}^\top$. Due to the definition of original covariance matrix $\mathbf{\Sigma}_k$ in Algorithm (Line 2), we have

$$
\mathbf{\Sigma}_{k_i} \geq \lambda\mathbf{I} + \sum_{j=1}^{i-1} w_{k_j}\mathbf{x}_{k_j}\mathbf{x}_{k_j}^\top = \mathbf{A}_i.
$$

According to Lemma J.4, it further implies that for vector $\mathbf{x}_{k_i}$, we have

$$
\mathbf{x}_{k_i}^\top \mathbf{\Sigma}_{k_i}^{-1}\mathbf{x}_{k_i} \leq \mathbf{x}_{k_i}^\top (\mathbf{A}_i)^{-1}\mathbf{x}_{k_i}. \qquad \text{(E.2)}
$$

Therefore, the term I_1 can be bounded by

$$
\begin{aligned}
I_1 &= \sum_{k:w_k=1} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right) \\
&\leq \sum_{i=1}^{m} 2\beta \min\left(1, \sqrt{\mathbf{x}_{k_i}^\top \mathbf{\Sigma}_{k_i}^{-1}\mathbf{x}_{k_i}}\right) \\
&\leq 2\beta \sum_{i=1}^{m} \min\left(1, \sqrt{\mathbf{x}_{k_i}^\top (\mathbf{A}_i)^{-1}\mathbf{x}_{k_i}}\right) \\
&\leq 2\beta \sqrt{\sum_{i=1}^{m} 1 \times \sum_{i=1}^{m} \min\left(1, \mathbf{x}_{k_i}^\top (\mathbf{A}_i)^{-1}\mathbf{x}_{k_i}\right)} \\
&\leq 2\beta\sqrt{2dK\log(1 + KL^2/\lambda)}, \qquad \text{(E.3)}
\end{aligned}
$$

where the first inequality holds since $\beta \geq 1$, the second inequality holds due to (E.2), the third inequality holds due to Cauchy-Schwarz inequality, the last inequality holds due to Lemma J.3 with the facts that $m \leq K$ and $\|\mathbf{x}_{k_i}\|_2 \leq L$.

For the second term I_2, according to the definition for weight $w_k < 1$ in Algorithm 1, we have $w_k = \alpha/\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}$, which implies that

$$
\begin{aligned}
I_2 &= \sum_{k:w_k<1} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right) \\
&= \sum_{k:w_k<1} \min\left(2, 2\beta w_k \mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k/\alpha\right) \\
&\leq \sum_{k:w_k<1} \min\left((2 + 2\beta/\alpha), (2 + 2\beta/\alpha) w_k \mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k\right) \\
&= \sum_{k:w_k<1} (2 + 2\beta/\alpha) \min\left(1, w_k \mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k\right), \qquad \text{(E.4)}
\end{aligned}
$$

where the second equation holds due to the definition of weight w_k. Now, we assume the rounds with weight $w_k < 1$ can be listed as $\{k_1, .., k_m\}$ for simplicity. In addition, we introduce the auxiliary vector $\mathbf{x}_i'$ as $\mathbf{x}_i' = \sqrt{w_{k_i}}\mathbf{x}_{k_i}$ and matrix $\mathbf{\Sigma}_i'$ as

$$
\mathbf{\Sigma}_i' = \lambda\mathbf{I} + \sum_{j=1}^{i-1} w_{k_j}\mathbf{x}_{k_j}\mathbf{x}_{k_j}^\top = \lambda\mathbf{I} + \sum_{j=1}^{i-1} \mathbf{x}_j'(\mathbf{x}_j')^\top.
$$

According to Lemma J.4, we have $(\mathbf{\Sigma}_i')^{-1} \succeq \mathbf{\Sigma}_{k_i}^{-1}$. Therefore, for each $i \in [m]$, we have

$$\mathbf{x}_{k_i}^\top (\mathbf{\Sigma}_i')^{-1} \mathbf{x}_{k_i} \geq \mathbf{x}_{k_i}^\top \mathbf{\Sigma}_{k_i}^{-1} \mathbf{x}_{k_i}, \tag{E.5}$$

where the inequality holds due to $(\mathbf{\Sigma}_i')^{-1} \succeq \mathbf{\Sigma}_{k_i}^{-1}$. Now, taking a summation of (E.5) over all rounds k_i, we have

$$\begin{aligned}
\sum_{i=1}^{m} \min\left(1, w_{k_i} \mathbf{x}_{k_i}^\top \mathbf{\Sigma}_{k_i}^{-1} \mathbf{x}_{k_i}\right) &\leq \sum_{i=1}^{m} \min\left(1, w_{k_i} \mathbf{x}_{k_i}^\top (\mathbf{\Sigma}_i')^{-1} \mathbf{x}_{k_i}\right) \\
&= \sum_{i=1}^{m} \min\left(1, (\mathbf{x}_i')^\top (\mathbf{\Sigma}_i')^{-1} \mathbf{x}_i'\right) \\
&\leq 2d \log(1 + KL^2/\lambda),
\end{aligned} \tag{E.6}$$

where the first inequality holds due to (E.5), the second inequality holds due to Lemma J.3 with the facts that $m \leq K$. Substituting the result in (E.6) into (E.4), the term I_2 can be upper bounded by

$$\begin{aligned}
I_2 &\leq \sum_{k: w_k < 1} (2 + 2\beta/\alpha) \min\left(1, w_k \mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}_k\right) \\
&\leq (2 + 2\beta/\alpha) \times 2d \log(1 + KL^2/\lambda).
\end{aligned} \tag{E.7}$$

Finally, substituting the results in (E.3) and (E.7) into (E.1), the regret can be upper bounded by

$$\begin{aligned}
\text{Regret}(K) &\leq 2\beta \sqrt{2dK \log(1 + KL^2/\lambda)} + (2 + 2\beta/\alpha) \times 2d \log(1 + KL^2/\lambda) \\
&= O\Bigg(dR \sqrt{K \log^2\left((1 + KL^2/\lambda)/\delta\right)} + \alpha C \sqrt{dK \log^2\left((1 + KL^2/\lambda)/\delta\right)} \\
&\quad + S \sqrt{d\lambda K \log(1 + KL^2/\lambda)} + \frac{Rd^{1.5}}{\alpha} \times \sqrt{\log^3\left((1 + KL^2/\lambda)/\delta\right)} \\
&\quad + \frac{dS\sqrt{\lambda}}{\alpha} \times \sqrt{\log^2\left((1 + KL^2/\lambda)/\delta\right)} + dC \sqrt{\log^2\left((1 + KL^2/\lambda)/\delta\right)} \Bigg).
\end{aligned}$$

Therefore, we complete the proof of Theorem 4.2. □

F Proof of Theorem C.3

In this section, we present the detailed proof of Theorem 4.2.

Proof of Theorem C.3. Based on the event $\mathcal{E}$, the regret in round $k \in [K]$ is upper bounded by

$$\Delta_k = \max_{\mathbf{x} \in \mathcal{D}_k} \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle - \langle \boldsymbol{\theta}^*, \mathbf{x}_k \rangle \leq 2\beta \sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}_k}.$$

On the other hand, according to Assumption C.2, the regret in round $k \in [K]$ satisfies that $\Delta_k = 0$ or $\Delta_k \geq \Delta$. Combining these two results, for round $k \in [K]$ with uncertainty $2\beta \sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}_k} < \Delta$, the regret must satisfy $\Delta_k = 0$. Therefore, the regret in the first K rounds can be decomposed to two part based on the weight w_k and exploration bonus $\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}_k}$:

$$\begin{aligned}
\text{Regret}(K) &= \sum_{k=1}^{K} \max_{\mathbf{x} \in \mathcal{D}_k} \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle - \langle \boldsymbol{\theta}^*, \mathbf{x}_k \rangle \\
&= \sum_{k: 2\beta \sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}_k} \geq \Delta} \max_{\mathbf{x} \in \mathcal{D}_k} \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle - \langle \boldsymbol{\theta}^*, \mathbf{x}_k \rangle \\
&\leq \min\left(2, \sum_{k: 2\beta \sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}_k} \geq \Delta} 2\beta \sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1} \mathbf{x}_k}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k:w_k=1, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\geq \Delta} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right) \\
&\quad + \sum_{k:w_k<1, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\geq \Delta} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right) \\
&\leq \underbrace{\sum_{k:w_k=1, k:2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\geq \Delta} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right)}_{J_1} \\
&\quad + \underbrace{\sum_{k:w_k<1} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right)}_{J_2}, \qquad \text{(F.1)}
\end{aligned}
$$

where the inequality holds due to Lemma E.1 with the fact that the suboptimality in each round is no more than 2. Notice that the term J_2 is equal to the term I_2 in the proof of Theorem 4.2 (See (E.1)) and with the same argument, it can be upper bounded by

$$
\begin{aligned}
J_2 \leq O\Bigg(&\frac{Rd^{1.5}}{\alpha} \times \sqrt{\log^3\left((1+KL^2/\lambda)/\delta\right)} + \frac{dS\sqrt{\lambda}}{\alpha} \times \sqrt{\log^2\left((1+KL^2/\lambda)/\delta\right)} \\
&+ dC\sqrt{\log^2\left((1+KL^2/\lambda)/\delta\right)}\Bigg), \qquad \text{(F.2)}
\end{aligned}
$$

where the inequality comes from (E.7). For the term J_1, we consider for all rounds $k \in [K]$ with $w_k = 1$ and exploration bonus $2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k} \geq \Delta$. For simplicity, we assume these rounds can be listed as $\{k_1, .., k_m\}$. With this notation, for each $i \leq m$, we can construct the auxiliary covariance matrix $\mathbf{A}_i = \lambda\mathbf{I} + \sum_{j=1}^{i-1} \mathbf{x}_{k_j}\mathbf{x}_{k_j}^\top$. Due to the definition of original covariance matrix $\mathbf{\Sigma}_k$ in Algorithm (Line 2), we have

$$
\mathbf{\Sigma}_{k_i} \geq \lambda\mathbf{I} + \sum_{j=1}^{i-1} w_{k_j}\mathbf{x}_{k_j}\mathbf{x}_{k_j}^\top = \mathbf{A}_i.
$$

According to Lemma J.4, it further implies that for vector $\mathbf{x}_{k_i}$, we have

$$
\mathbf{x}_{k_i}^\top \mathbf{\Sigma}_{k_i}^{-1}\mathbf{x}_{k_i} \leq \mathbf{x}_{k_i}^\top (\mathbf{A}_i)^{-1}\mathbf{x}_{k_i}. \qquad \text{(F.3)}
$$

Therefore, the term J_1 can be bounded by

$$
\begin{aligned}
J_1 &= \sum_{k:w_k=1, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\geq \Delta} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right) \\
&\leq \sum_{i=1}^{m} 2\beta \min\left(1, \sqrt{\mathbf{x}_{k_i}^\top \mathbf{\Sigma}_{k_i}^{-1}\mathbf{x}_{k_i}}\right) \\
&\leq 2\beta \sum_{i=1}^{m} \min\left(1, \sqrt{\mathbf{x}_{k_i}^\top (\mathbf{A}_i)^{-1}\mathbf{x}_{k_i}}\right) \\
&\leq 2\beta \sqrt{\sum_{i=1}^{m} 1 \times \sum_{i=1}^{m} \min\left(1, \mathbf{x}_{k_i}^\top (\mathbf{A}_i)^{-1}\mathbf{x}_{k_i}\right)} \\
&\leq 2\beta\sqrt{2dm\log(1+KL^2/\lambda)}, \qquad \text{(F.4)}
\end{aligned}
$$

where the first inequality holds since $\beta \geq 1$, the second inequality holds due to (F.3), the third inequality holds due to Cauchy-Schwarz inequality, the fourth inequality holds due to Lemma J.3 with the facts that $m \leq K$ and $\|\mathbf{x}_{k_i}\|_2 \leq L$. On the other hand, the term J_1 is lower bounded by

$$
J_1 = \sum_{k:w_k=1, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\geq \Delta} \min\left(2, 2\beta\sqrt{\mathbf{x}_k^\top \mathbf{\Sigma}_k^{-1}\mathbf{x}_k}\right) \geq m \times \Delta, \qquad \text{(F.5)}
$$

where the inequality holds due to the definition of k_i with the fact that $\Delta \leq 2$. Combining the upper and lower bound for term J_1, we have

$$m \times \Delta \leq 2\beta\sqrt{2dm\log(1+KL^2/\lambda)},$$

which further implies that

$$m \leq O\big(\beta^2 d\log(1+KL^2/\lambda)/\text{gap}_{\min}^2\big). \tag{F.6}$$

Substituting the upper bound of m in (F.6) into (F.4), the term J_1 can be upper bounded by

$$J_1 \leq O\big(\beta^2 d\log(1+KL^2/\lambda)/\Delta\big). \tag{F.7}$$

Finally, substituting the upper bounds of term J_2 in (F.2) and term J_1 in (F.7) into (F.1), the regret can be upper bounded by

$$\begin{aligned}
\text{Regret}(K) &\leq O\big(\beta^2 d\log(1+KL^2/\lambda)/\Delta\big) + O\bigg(\frac{Rd^{1.5}}{\alpha} \times \sqrt{\log^3\big((1+KL^2/\lambda)/\delta\big)} \\
&\quad + \frac{dS\sqrt{\lambda}}{\alpha} \times \sqrt{\log^2\big((1+KL^2/\lambda)/\delta\big)} + dC\sqrt{\log^2\big((1+KL^2/\lambda)/\delta\big)}\bigg) \\
&= O\bigg(R^2d^2\log^2\big((1+KL^2/\lambda)/\delta\big)/\Delta + \frac{\alpha^2 dC^2}{\Delta} \times \sqrt{\log\big(3+C^2L^2K/(R^2\lambda\delta)\big)} \\
&\quad + S^2 d\lambda\log(1+KL^2/\lambda)/\Delta + \frac{Rd^{1.5}}{\alpha} \times \sqrt{\log^3\big((1+KL^2/\lambda)/\delta\big)} \\
&\quad + \frac{dS\sqrt{\lambda}}{\alpha} \times \sqrt{\log^2\big((1+KL^2/\lambda)/\delta\big)} + dC\sqrt{\log^2\big((1+KL^2/\lambda)/\delta\big)}\bigg).
\end{aligned}$$

Therefore, we complete the proof of Theorem C.3. □

G Proof of Theorem 4.9

Proof of Theorem 4.9. We discuss two cases here.

- For the case $C \leq \bar{C}$, we know that $\bar{C}$ is still a valid upper bound of the corruption level. Thus, CW-OFUL with a $\bar{C}$ corruption level runs successfully, and its regret is upper bounded by $\widetilde{O}(dR\sqrt{K} + d\bar{C}) = \widetilde{O}(dR\sqrt{K} + d\bar{C})$ as Theorem 4.2 suggests.
- For the case $C = \Omega(\bar{C})$, CW-OFUL can not guarantee a sublinear regret. Thus a trivial regret bound (i.e., regret at each round is bounded by 2) applies.

□

H Proof of Theorem 4.12

We introduce our proof of Theorem 4.12, which is adapted from Bogunovic et al. (2021).

Proof of Theorem 4.12. In this proof, we consider an arbitrary algorithm satisfying the conditions in the statement of Theorem 4.12, which will run K rounds for any bandit instance. We consider an uncorrupted bandit instance A_0 defined as follows. A_0 has the decision sets $\mathcal{D}_k = \mathcal{D}$. Here $\mathcal{D} = \{\mathbf{a}_i\}_{1\leq i\leq d}$, where $\mathbf{a}_i = \mathbf{e}_i$ is the basis in the d-dimensional space. Let $\boldsymbol{\theta}_0^* = (1/4, \underbrace{1/8, \ldots, 1/8}_{(d-1)-\text{times}}) \in \mathbb{R}^d$ and $\epsilon_i = 0$. It is easy to see that the optimal policy is to select $\mathbf{a}_1$ at each round, and the regret to select a sub-optimal arm is $1/8$. Since the regret of the algorithm without corruption satisfies $\mathbb{E}\big[\text{Regret}(K)\big] < R_K$, and all the regret comes from selecting $\mathbf{a}_2, \ldots, \mathbf{a}_d$, we have the expected number of rounds to select $\mathbf{a}_2, \ldots, \mathbf{a}_d$ is at most $R_K/(1/8) = 8R_K$. Then by the pigeonhole principle, there exists some $2 \leq i \leq d$ such that the expected number of times to select $\mathbf{a}_i$ is less than $8R_K/(d-1)$. Without loss of generality, we suppose $i = 2$. Then by Markov inequality, with probability at least $1/2$, the number of times to select $\mathbf{a}_2$ is less than $16R_K/(d-1)$.

Next, we consider a corrupted bandit instance A_1 defined as follows. A_1 has the same decision set $\mathcal{D} = \{\mathbf{e}_i\}$ as A_0, while it has a different $\boldsymbol{\theta}_1^* = (1/4, 3/8, \underbrace{1/8, \ldots, 1/8}_{(d-2)-\text{times}})$. A_1 is also noiseless, i.e., $\epsilon_i = 0$. Unlike A_0, we have an adversary to attack A_1 as follows: whenever $\mathbf{a}_2$ is selected and the total corruption level up to the previous step is no more than $4R_K/(d-1) - 1/4$, the adversary corrupts the reward from $3/8$ to $1/8$. Otherwise, the adversary stops to corrupt the reward. With this adversary, the corruption level C is upper bounded by $4R_K/(d-1) - 1/4 + 1/4 = \Omega(R_K/d)$.

For this adversary, since for A_1, each selection of $\mathbf{a}_2$ returns a reward $1/8$, then the agent can not tell the difference between A_0 and A_1 until the total corruption level reaches the threshold $4R_k/(d-1)$ and the adversary stops to corrupt the reward. Therefore, the sequence of rounds for the agent to select $\mathbf{a}_2$ with A_1 instance is the same as the sequence for the agent to select $\mathbf{a}_2$ with A_0, until the number of rounds to select action $\mathbf{a}_2$ reaches $4R_k/(d-1)/(1/4) = 16R_k/(d-1)$. However, when the total number of times to select $\mathbf{a}_2$ is less than $16R_K/(d-1)$, the agent cannot differentiate A_0 and A_1 and will follow the same action sequence as A_0. In this case, since for A_1, $\mathbf{a}_2$ is the optimal action, and all the other actions suffer a $1/8$ regret, then the regret on A_1 is at least $1/8 \cdot (K - 16R_K/(d-1)) = \Omega(K)$, where we use the fact that $R_K \leq O(K)$. Therefore, with probability at least $1/2$, the regret is at least $\Omega(K)$, which further implies that the expected regret is lower bounded by $\mathbb{E}[\text{Regret(K)}] \geq 1/2 \times \Omega(K) = \Omega(K)$. Thus, we finish the proof of Theorem 4.12. □

I Proof of Lemmas in Sections 4, D and Appendix E

I.1 Proof of Lemma 4.1

Proof of Lemma 4.1. According to the definition of estimated vector $\boldsymbol{\theta}_k$ in Algorithm 1 (Line 3), we have

$$\boldsymbol{\theta}_k = \boldsymbol{\Sigma}_k^{-1}\mathbf{b}_k = \boldsymbol{\Sigma}_k^{-1}\sum_{i=1}^{k-1} w_i\mathbf{x}_i r_i = \boldsymbol{\Sigma}_k^{-1}\sum_{i=1}^{k-1} w_i\mathbf{x}_i(\mathbf{x}_i^\top\boldsymbol{\theta} + \eta_i + c_i).$$

This equation further implies that the difference between estimated vector $\boldsymbol{\theta}_k$ and the unknown vector $\boldsymbol{\theta}^*$ can be decomposed as:

$$\begin{aligned}
\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_k} &= \Big\|\boldsymbol{\Sigma}_k^{-1}\sum_{i=1}^{k-1} w_i\mathbf{x}_i(\mathbf{x}_i^\top\boldsymbol{\theta}^* + \eta_i + c_i) - \boldsymbol{\theta}^*\Big\|_{\boldsymbol{\Sigma}_k} \\
&= \Big\|\boldsymbol{\Sigma}_k^{-1}\sum_{i=1}^{k-1} w_i\mathbf{x}_i(\mathbf{x}_i^\top\boldsymbol{\theta} + \eta_i + c_i) - \boldsymbol{\Sigma}_k^{-1}\Big(\sum_{i=1}^{k-1} w_i\mathbf{x}_i\mathbf{x}_i^\top + \lambda\mathbf{I}\Big)\boldsymbol{\theta}^*\Big\|_{\boldsymbol{\Sigma}_k} \\
&= \Big\|\boldsymbol{\Sigma}_k^{-1}\sum_{i=1}^{k-1} w_i\mathbf{x}_i\eta_i + \boldsymbol{\Sigma}_k^{-1}\sum_{i=1}^{k-1} w_i\mathbf{x}_i c_i - \lambda\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\theta}^*\Big\|_{\boldsymbol{\Sigma}_k} \\
&\leq \underbrace{\Big\|\boldsymbol{\Sigma}_k^{-1}\sum_{i=1}^{k-1} w_i\mathbf{x}_i\eta_i\Big\|_{\boldsymbol{\Sigma}_k}}_{\text{Stochastic error}: I_1} + \underbrace{\Big\|\boldsymbol{\Sigma}_k^{-1}\sum_{i=1}^{k-1} w_i\mathbf{x}_i c_i\Big\|_{\boldsymbol{\Sigma}_k}}_{\text{Corruption error}: I_2} + \underbrace{\Big\|\lambda\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\theta}^*\Big\|_{\boldsymbol{\Sigma}_k}}_{\text{Regularization error}: I_3},
\end{aligned} \tag{I.1}$$

where the inequality holds due to the fact that $\|\mathbf{a} + \boldsymbol{b} + \boldsymbol{c}\|_{\boldsymbol{\Sigma}_k} \leq \|\mathbf{a}\|_{\boldsymbol{\Sigma}_k} + \|\boldsymbol{b}\|_{\boldsymbol{\Sigma}_k} + \|\boldsymbol{c}\|_{\boldsymbol{\Sigma}_k}$.

For the stochastic error term I_1, it can be bounded by the concentration Lemma J.2 in Abbasi-Yadkori et al. (2011). More specifically, we introduce the auxiliary vector $\mathbf{x}_i'$ and noise η_i' such that $\mathbf{x}_i' = \sqrt{w_i}\mathbf{x}_i$ and $\eta_i' = \sqrt{w_i}\eta_i$. According to the definition of weight $\boldsymbol{\theta}_i$ in Algorithm (Line 6), both of these two situations satisfies that the weight $\boldsymbol{\theta}_i$ is bounded by $w_i \leq 1$. Since the original vector $\mathbf{x}_i$ satisfies that $\|\mathbf{x}_i\|_2 \leq L$ and the original stochastic noise η_i is R-sub Gaussian, these results further imply that

$$\|\mathbf{x}_i'\|_2 = \|\sqrt{w_i}\mathbf{x}_i\|_2 \leq L, \eta_i' = \sqrt{w_i}\eta_i \text{ is } R\text{-sub Gaussian.}$$

With this notation, the covariance matrix $\mathbf{\Sigma}_k$ and the stochastic error term I_1 can be rewritten and bounded as:

$$\mathbf{\Sigma}_k = \lambda \mathbf{I} + \sum_{i=1}^{k-1} w_i \mathbf{x}_i \mathbf{x}_i^\top = \lambda \mathbf{I} + \sum_{i=1}^{k-1} \mathbf{x}_i' (\mathbf{x}_i')^\top$$

$$\begin{aligned}
I_1 &= \left\| \mathbf{\Sigma}_k^{-1} \sum_{i=1}^{k-1} w_i \mathbf{x}_i \eta_i \right\|_{\mathbf{\Sigma}_k} \\
&= \left\| \sum_{i=1}^{k-1} w_i \mathbf{x}_i \eta_i \right\|_{\mathbf{\Sigma}_k^{-1}} \\
&= \left\| \sum_{i=1}^{k-1} \mathbf{x}_i' \eta_i' \right\|_{\mathbf{\Sigma}_k^{-1}} \\
&\leq \sqrt{2R^2 \log \left(\frac{\det(\mathbf{\Sigma}_k)^{1/2} \det(\mathbf{\Sigma}_1)^{-1/2}}{\delta} \right)} \\
&\leq R \sqrt{d \log \big((1 + KL^2/\lambda)/\delta \big)}, \qquad \text{(I.2)}
\end{aligned}$$

where the first inequality holds due to Lemma J.2 and the second inequality holds due to the facts that $\mathbf{\Sigma}_k = \lambda \mathbf{I}_+ \sum_{i=1}^{k-1} \mathbf{x}_i' (\mathbf{x}_i')^\top$ and $\|\mathbf{x}'\|_2 \leq L$.

For the corruption error term I_2, it can be bounded by

$$\begin{aligned}
I_2 &= \left\| \mathbf{\Sigma}_k^{-1} \sum_{i=1}^{k-1} w_i \mathbf{x}_i c_i \right\|_{\mathbf{\Sigma}_k} \\
&= \left\| \mathbf{\Sigma}_k^{-1/2} \sum_{i=1}^{k-1} w_i \mathbf{x}_i c_i \right\|_2 \\
&\leq \sum_{i=1}^{k-1} \left\| \mathbf{\Sigma}_k^{-1/2} w_i \mathbf{x}_i c_i \right\|_2 \\
&= \sum_{i=1}^{k-1} |c_i| \times w_i \|\mathbf{\Sigma}_k^{-1/2} \mathbf{x}_i\| \\
&\leq \sum_{i=1}^{k-1} |c_i| \alpha \\
&\leq \alpha C, \qquad \text{(I.3)}
\end{aligned}$$

where the first inequality holds due to the fact that $\|\mathbf{a} + \boldsymbol{b}\|_2 \leq \|\mathbf{a}\|_2 + \|\boldsymbol{b}\|_2$, the second inequality holds due to the definition of weight w_i in Algorithm (Line 6) with the fact that $\mathbf{\Sigma}_k \succeq \mathbf{\Sigma}_i$ and the last inequality holds due to the definition of corruption level C.

For the regularization error term I_3, we have

$$I_3 = \left\| \lambda \mathbf{\Sigma}_k^{-1} \boldsymbol{\theta}^* \right\|_{\mathbf{\Sigma}_k} = \lambda \left\| \boldsymbol{\theta}^* \right\|_{\mathbf{\Sigma}_k^{-1}} \leq \sqrt{\lambda} \|\boldsymbol{\theta}^*\|_2 \leq \sqrt{\lambda} S, \qquad \text{(I.4)}$$

where the first inequality holds due to $\left\| \boldsymbol{\theta}^* \right\|_{\mathbf{\Sigma}_k} \leq \|\boldsymbol{\theta}^*\|_2 / \sqrt{\lambda_{\min}(\mathbf{\Sigma}_k)}$ with the fact that $\mathbf{\Sigma}_k = \lambda \mathbf{I} + \sum_{i=1}^{k-1} w_i \mathbf{x}_i \mathbf{x}_i^\top \succeq \lambda \mathbf{I}$ and the last inequality holds due to the assumption that $\|\boldsymbol{\theta}^*\|_2 \leq S$.

Finally, substituting the results in (I.2), (I.3) and (I.4) into (I.1), we have

$$\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\mathbf{\Sigma}_k} \leq I_1 + I_2 + I_3 \leq R \sqrt{d \log \big((1 + KL^2/\lambda)/\delta \big)} + \alpha C + \sqrt{\lambda} S.$$

Therefore, we finish the proof of Lemma 4.1. □

I.2 Proof of Lemma E.1

Proof of Lemma E.1. Firstly, on the event $\mathcal{E}$, for each round $k \in [K]$ and each action $\mathbf{x} \in \mathcal{D}_k$, we have

$$\begin{aligned}
\boldsymbol{\theta}_k^\top \mathbf{x} + \beta\sqrt{\mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}} - (\boldsymbol{\theta}^*)^\top \mathbf{x} &= (\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)^\top \mathbf{x} + \beta\sqrt{\mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}} \\
&\geq -\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_k} \times \|\mathbf{x}\|_{\boldsymbol{\Sigma}_k^{-1}} + \beta\sqrt{\mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}} \\
&\geq -\beta\|\mathbf{x}\|_{\boldsymbol{\Sigma}_k^{-1}} + \beta\sqrt{\mathbf{x}^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}} \\
&= 0, \qquad \text{(I.5)}
\end{aligned}$$

where the first inequality holds due to the Cauchy-Schwarz inequality and the last inequality holds due to the definition of $\mathcal{E}$ in Lemma 4.1. (I.5) shows that our estimator in Algorithm 1 is optimistic for each action $\mathbf{x} \in \mathcal{D}_k$. For simplicity, we denote the optimal action at round k as $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{D}_k} (\boldsymbol{\theta}^*)^\top \mathbf{x}$ and (I.5) further implies that the regret at round k can be upper bounded by

$$\begin{aligned}
\Delta_k &= (\boldsymbol{\theta}^*)^\top \mathbf{x}^* - (\boldsymbol{\theta}^*)^\top \mathbf{x}_k \\
&\leq \boldsymbol{\theta}_k^\top \mathbf{x}^* + \beta\sqrt{(\mathbf{x}^*)^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}^*} - (\boldsymbol{\theta}^*)^\top \mathbf{x}_k \\
&\leq \boldsymbol{\theta}_k^\top \mathbf{x}_k + \beta\sqrt{\mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}_k} - (\boldsymbol{\theta}^*)^\top \mathbf{x}_k \\
&= (\boldsymbol{\theta}_k - \boldsymbol{\theta}^*)^\top \mathbf{x}_k + \beta\sqrt{\mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}_k} \\
&\leq \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_k} \times \|\mathbf{x}_k\|_{\boldsymbol{\Sigma}_k^{-1}} + \beta\sqrt{\mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}_k} \\
&\leq 2\beta\sqrt{\mathbf{x}_k^\top \boldsymbol{\Sigma}_k^{-1}\mathbf{x}_k},
\end{aligned}$$

where the first inequality holds due to (I.5), the second inequality holds due to the selection rule in Algorithm (Line 5), the third inequality holds due to the Cauchy-Schwarz inequality and the last inequality holds due to the definition of $\mathcal{E}$ in Lemma 4.1. Thus, we finish the proof of Lemma E.1. □

J Auxiliary Lemmas

Lemma J.1 (Azuma–Hoeffding inequality, Cesa-Bianchi and Lugosi 2006)**.** Let $\{\eta_k\}_{k=1}^K$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_k\}$ satisfying $|\eta_k| \leq R$ for some constant R, η_k is $\mathcal{G}_{k+1}$-measurable, $\mathbb{E}\big[\eta_k|\mathcal{G}_k\big] = 0$. Then for any $0 < \delta < 1$, with high probability at least $1 - \delta$, we have

$$\sum_{k=1}^K \eta_k \leq R\sqrt{2K\log(1/\delta)}.$$

Lemma J.2 (Lemma 9 in Abbasi-Yadkori et al. 2011)**.** Let $\{\epsilon_k\}_{k=1}^K$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_k\}_{k=0}^K$ such that ϵ_k is $\mathcal{F}_k$-measure and ϵ_k is conditionally R-sub-Gaussian, $i.e.$

$$\forall\lambda \in \mathbb{R}, \mathbb{E}\big[e^{\lambda\epsilon_k}|\mathcal{F}_{k-1}\big] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{\mathbf{x}_k\}_{k=1}^K$ be an $\mathbb{R}^d$-valued stochastic process where $\mathbf{x}_k$ is $\mathcal{F}_{k-1}$-measurable and for any $k \in [K]$, we further define $\boldsymbol{\Sigma}_k = \lambda\mathbf{I} + \sum_{i=1}^k \mathbf{x}_i\mathbf{x}_i^\top$. Then with probability at least $1 - \delta$, for all $k \in [K]$, we have

$$\left\|\sum_{i=1}^k \mathbf{x}_i\eta_i\right\|_{\boldsymbol{\Sigma}_k^{-1}}^2 \leq 2R^2\log\left(\frac{\det(\boldsymbol{\Sigma}_k)^{1/2}\det(\boldsymbol{\Sigma}_0)^{-1/2}}{\delta}\right).$$

Lemma J.3 (Lemma 11 in Abbasi-Yadkori et al. 2011)**.** Let $\{\mathbf{x}_k\}_{k=1}^K$ be a sequence of vectors in $\mathbb{R}^d$, matrix $\mathbf{\Sigma}_0$ a $d \times d$ positive definite matrix and define $\mathbf{\Sigma}_k = \mathbf{\Sigma}_0 + \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^\top$, then we have

$$\sum_{i=1}^k \min\left\{1, \mathbf{x}_i^\top \mathbf{\Sigma}_{i-1}^{-1} \mathbf{x}_i\right\} \leq 2 \log\left(\frac{\det \mathbf{\Sigma}_k}{\det \mathbf{\Sigma}_0}\right).$$

In addition, if $\|\mathbf{x}_i\|_2 \leq L$ holds for all $i \in [K]$, then

$$\sum_{i=1}^k \min\left\{1, \mathbf{x}_i^\top \mathbf{\Sigma}_{i-1}^{-1} \mathbf{x}_i\right\} \leq 2 \log\left(\frac{\det \mathbf{\Sigma}_k}{\det \mathbf{\Sigma}_0}\right) \leq 2\Big(d \log\big((\text{trace}(\mathbf{\Sigma}_0) + kL^2)/d\big) - \log\det \mathbf{\Sigma}_0\Big).$$

Lemma J.4 (Corollary 7.7.4. (a) in Horn and Johnson 2012)**.** Let $\mathbf{A}, \mathbf{B}$ be a Hermitian matrix in $\mathbb{R}^{d \times d}$ and suppose $\mathbf{A}, \mathbf{B} \succ \mathbf{0}$, then $\mathbf{A} \succeq \mathbf{B}$ if and only if $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$.