Linear Bandits with Total Variation Contamination

CS 224 Fall 2023 Final Project

Anna L. Trella

Abstract

The linear bandit problem with adversarial corruptions is a variant of the standard stochastic linear bandit problem where the agent observes a reward corrupted by an adversary, rather than one stochastically drawn from the environment. Such an adversary has been proposed with Huber contamination [Chen et al., 2022] and strong contamination with bounded rewards [Kapoor et al., 2019]. In this paper, we specifically focus on total variation (TV) contamination: the formulation of adversarial corruptions where the adversary can change the rewards arbitrarily at every round; however, there is a fixed budget C for the total difference the rewards can change [Zhao et al., 2021]. We offer a deep dive of the analysis presented in [Zhao et al., 2021] for the Robust Weighted OFUL algorithm, an algorithm designed to achieve robust performance in the TV-contamination setting.

1 Introduction

Linear bandits are used in a wide range of real-world applications [Bouneffouf and Rish, 2019], such as online advertising [Li et al., 2010], education [Cai et al., 2021], mobile health [Liao et al., 2019, Trella et al., 2022], etc. The popularity of linear bandits can be attributed to its ability to leverage contextual information for action selection, while also staying tractable and learning well in high-noise or data-sparse regimes. In fact, linear bandits can be seen as a severe form of regularization (i.e., using a lower discount factor than the true discount factor) for Markov Decision Process (MDP) algorithms, which has been shown to lead to selecting more effective actions [Jiang et al., 2015].

Although linear bandits have been studied extensively in non-corrupted settings (i.e., agent observes true rewards stochastically generated by the environment), a body of work closely related to the robust regression literature has emerged to consider settings where the agent receives rewards possibly corrupted by some adversary. These settings are closer to many real world environments. For example, in online advertising, the agent could observe clickfraud from a bot rather than a real user [Lykouris et al., 2018]. It's important to note that not all corruptions need to be malicious or even intentional: Natural outliers (extremely high or low values) can affect the performance of agents trained on with such data. For example, in the online mobile health setting, the agent could receive inaccurate sensory data (used to form the reward) due to real-world challenges (e.g., user did not charge or properly wear their device, user does not have stable WiFi connectivity) [Trella et al., 2022]. Regardless, it is increasing important to consider agents' performance in environment with adversarial corruptions to move towards developing algorithms more robust to real life. Multi-armed bandits with adversarial corruptions have been considered in [Lykouris et al., 2018] and [Gupta et al., 2019], however the stochastic linear bandit setting is more generalizable than the multi-armed bandit setting (See Section 2.2.1).

In this expository work, we first present a generic form for the stochastic linear bandit problem with adversarial corruptions on rewards (Section 2.2) and offer some notable examples. We then provide an in-depth walk-through of an approach and analyses in the TV contamination setting. We formally define the TV contamination setting (Section 3), cover the regret analysis for an algorithm that assumes known corruption level (Section 3.3), and cover a brief overview of an extension when the corruption level is unknown (Section 3.4). Finally, we offer suggestions for further reading (Section 4).

2 Problem Setting

2.1 Notation

We introduce the following notation used throughout the paper. For some vector $v \in \mathbb{R}^d$, we use $\|v\|_2 = \sqrt{\langle v, v \rangle}$ to denote the L2-norm of v and $v^\top$ to denote the transpose of v. For some positive definite matrix $M \in \mathbb{R}^{d \times d}$, M^{-1} denotes the inverse of M, and $\|v\|_M^2 = v^\top M v$ denotes the square of the weighted L2-norm of v. We use $\tilde{O}(\cdot)$ to denote big O notation that ignores logarithm factors. $\mathbb{I}[\cdot]$ denotes the indicator function.

2.2 Stochastic Linear Bandits with Adversarial Corruptions

We set up a general formulation for the linear contextual bandits with adversarial corruptions problem. We first describe standard stochastic linear bandits [Lattimore and Szepesvári, 2020, Abbasi-Yadkori et al., 2011] and then describe adversarial corruptions.

2.2.1 Stochastic Linear Bandits

Let $\mathcal{X}_t \subset \mathbb{R}^d$ be the action space given to the agent at time t. For time-step t, when an agent selects action $x_t \in \mathcal{X}_t$, the environment generates a reward:

$$r_t^* = x_t^\top \theta^* + \epsilon_t \tag{1}$$

where $\theta^* \in \mathbb{R}^{d'}$ is the true reward parameter which is fixed but unknown. ϵ_t is the noise parameter with mean 0. Typically, ϵ_t is also assumed to be conditionally centered and σ-sub-Gaussian given the history $\mathcal{H}_t = \{\mathcal{X}_1, x_1, r_1, ..., \mathcal{X}_{t-1}, x_{t-1}, r_{t-1}, \mathcal{X}_t, x_t\}$. Namely, $\mathbb{E}[\epsilon_t|\mathcal{H}_t] = 0$ and $\mathbb{E}[\exp(\lambda\epsilon_t)|\mathcal{H}_t] \leq \exp(\lambda^2\sigma^2/2)$ for some $\sigma > 0$, $\forall\ \lambda > 0$.

Different choices of $\mathcal{X}_t$ define other bandit frameworks. For example if $\mathcal{X}_t = \{e_1, .., e_d\}$ where $(e_i)_i$ are standard basis vectors, then $\theta^* = [\mu_1^*, ..., \mu_d^*] \in \mathbb{R}^d$ becomes a vector of the true mean reward of each arm, and the stochastic linear bandit problem becomes the finite multi-armed bandit setting. Additionally, let $\mathcal{X}_t = \{\psi(c_t, i) : i \in [k]\}$, for $c_t \in \mathbb{R}^d$, where $\psi(c_t, i)$ denotes a vector in $\mathbb{R}^{k \cdot d}$ that has c_t in the $d \cdot (i-1)$th entry and 0s elsewhere. Then $\theta^* = [\theta_1^*, ..., \theta_k^*] \in \mathbb{R}^{k \cdot d}$ becomes a vector of the true reward parameters θ_a^* for each action a, and the stochastic linear bandit problem becomes the linear contextual bandit setting.

2.2.2 Adversarial Corruptions

At every time-step t, instead of observing the true reward r_t^*, the adversary may or may not corrupt the reward following some corruption procedure (with limitations). The agent then observes r_t. We offer a brief overview of two notable examples of formulating adversarial corruptions.

Huber Contamination [Chen et al., 2022] formulates corrupted rewards under the Huber contamination model. Under the Huber contamination model, every time-step, the reward has probability $\eta \in (0, \frac{1}{2})$ (corruption parameter) of being corrupted.

Definition 1. *(Huber-Contaminated Linear Contextual Bandits) Let $\mathcal{S}$ denote the arbitrary state space and let $\mathcal{A}$ be the action space of size K where every action in $\mathcal{A}$ is available at every time-step t. Before the interaction between the environment and the agent, an oblivious adversary selects distributions $p_{r_t^*}[\cdot|s_t]$ over reward functions $r_t^* : \mathcal{A} \to [0, R]$ for all possible contexts s_t and time-steps $t \in [T]$. Let noise value $\epsilon_t(s, a) = \xi_t$ (i.e., noise is drawn independent of action and state) be bounded by σ^2 (i.e., $\sigma^2 := \sup_{s,t} \mathbb{E}[\xi_t^2|s_{ta} = s]$).*

For each time-step $t \in [T]$:

1. *Nature chooses contexts $S_t = \{s_{t1}, s_{t2}, ..., s_{t|\mathcal{A}|}\}$, potentially adversarially based on history $\mathcal{H}_t$.*
2. *Agent selects action $a_t \in \mathcal{A}$.*
3. *A Bern(η) coin is flipped giving sample O_t that decides whether this time-step is corrupted.*
4. *If $O_t = 1$ (time-step is not corrupted), then agent obtains reward $r_t = r_t^*(a)$, specified by Equation 1 with noise value $\epsilon_t(s, a) = \xi_t$.*
5. *If $O_t = 0$ (time-step is corrupted), then the agent obtains an arbitrary reward $r_t = r_t(a_t)$ chosen by the adversary based on $s_t, a_t, \mathcal{H}_t$.*

The goal of the agent is to minimize clean regret against the best policy π^* as measured by the true uncorrupted rewards.

$$Regret_{HCB}(T) := \mathbb{E}\left[\sum_{t=1}^{T} r_t^*(a_t) - r_t^*(\pi^*(z_t))\right] \tag{2}$$

where $\pi^*(s) := \arg\max_a f(s, a)$, the supremum ranges over all (non-adaptive) policies and the expectation is over the randomness of the Bernoulli draw, the rewards, the choice of contexts, and the policy of the agent.

[Chen et al., 2022] proves a high-probability regret bound for their method of learning in the Huber-Contaminated Linear Contextual Bandit setting. They first offer an algorithm for solving the offline case of Huber-contaminated linear regression. They then convert the offline approach into achieving clean square loss for online linear regression. Finally, they use a reduction from online regression to contextual bandits [Foster and Rakhlin, 2020] to obtain an algorithm for contextual bandits.

Strong Contamination with Bounded Rewards [Kapoor et al., 2019] formulates an adaptive adversary that has access to the true environment process, the history $\mathcal{H}_t$ (which includes the current action set and the current action chosen by the agent), and the uncorrupted reward value. At every time-step t, the adversary can only add a corruption value b_t to the environment-generated rewards and rewards (corrupted or not) are bounded (i.e., $r_t \in [-B, B]$ for some $B > 0$).

Namely, the final reward given to the agent is:

$$r_t = x_t^\top \theta^* + \epsilon_t + b_t = r_t^* + b_t \tag{3}$$

The only restriction for the adversary is that, at every point in the online process, the adversary can only corrupt up to an η fraction of the rewards. Formally, let $G_t = \{\tau < t \mid b_\tau = 0\}$ and $B_t = \{\tau < t \mid b_\tau \neq = 0\}$ denote the set of corrupted and uncorrupted time-steps, respectively. Then, $|B_t| \leq \eta \cdot t\ \ \forall t$.

Definition 2. *(Linear Contextual Bandits with Strong Contamination on Bounded Rewards) .*

For each time-step $t \in [T]$:

1. *Nature chooses contexts $S_t = \{s_{t1}, s_{t2}, ..., s_{t|\mathcal{A}|}\}$*
2. *Agent selects action $a_t \in \mathcal{A}$.*
3. *Clean reward r_t^* is generated according to Equation 1 conditioned on history $\mathcal{H}_t$.*
4. *Adversary selects a corruption amount of b_t after viewing agent selected action a_t, clean reward r_t^*, and history $\mathcal{H}_t$, while satisfying the constraint that $|B_t| \leq \eta \cdot t$*
5. *Agent observes reward $r_t = r_t^* + b_t$*

The goal of the agent is to minimize cumulative pseudo regret against an oracle measured by the true uncorrupted rewards. Let $a_t^* = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta^* \rangle$ be the best action that yields the highest expected, uncorrupted reward. The cumulative pseudo regret is:

$$Regret_{strong}(T) := \sum_{t=1}^{T} \langle a_t^*, \theta^* \rangle - \mathbb{E}[r_t] \tag{4}$$

[Kapoor et al., 2019] presents RUCB-LIN, which preforms an estimation of θ_t of the true model parameter θ^*, maintains a confidence set to model the region of uncertainty, and selects actions in a UCB-based way, like the OFUL algorithm [Abbasi-Yadkori et al., 2011]. However, unlike OFUL which uses regularized least squares to update θ_t, RUCB-LIN uses the TORRENT algorithm [Bhatia et al., 2017], which is a simple and easily, implementable approach for robust regression against an adaptive adversary. At update time, RUCB-LIN first obtains a model estimate from running TORRENT, and then uses the model estimate to perform a pruning step (i.e., constructs an estimate of the set of uncorrupted points $\tilde{G}_t$) and constructs a confidence set whose center is the OLS estimate from the subset of data indexed by time-steps in $\tilde{G}_t$ (the datapoints that the algorithm believes are uncorrupted). [Kapoor et al., 2019] also prove a regret bound for their algorithm that scales sub-linearly with T and linearly with ηT.

3 Total Variation (TV) Contamination

We now offer an in-depth walk-through of the algorithm and analysis for the formulation presented in [Zhao et al., 2021]. In this formulation, the adversary is able to alter the reward at every time-step, however, the adversary has a limited corruption budget on how much they can alter the true rewards. Namely, there is a bound on the total variation distance between true rewards and corrupted rewards (See Equation 5).

Definition 3. *(Total Variation (TV) Contamination Linear Bandits) The interaction between the agent and the environment contaminated by an adversary is as follows:*

For each time-step $t \in [T]$:

1. *Nature chooses action set $\mathcal{A}_t = \{a_{t1}, a_{t2}, ..., a_{t|\mathcal{A}_t|}\} \subseteq \mathbb{R}^d$, where each element represents a feasible action that can be selected by the agent.*
2. *Nature generates stochastic reward function $r_t^*(a) = \langle a, \theta^* \rangle + \epsilon_t(a)$ and variance bound $\sigma_t(a)$ on $\epsilon_t(a)$, for each $a \in \mathcal{A}_t$.*
3. *The adversary observes $\mathcal{A}_t, r_t^*(a), \sigma_t(a)$ and decides a corrupted reward function r_t.*
4. *Agent observes $\mathcal{A}_t$ and selects action $a_t \in \mathcal{A}_t$.*
5. *Adversary returns $r_t(a_t)$ and $\sigma_t(a)$ to the agent.*

Such an environment is called ***C-corrupted*** *if the corruption level is:*

$$C = \frac{1}{R+1} \sum_{t=1}^{T} \sup_{a \in \mathcal{A}_t} |r_t^*(a) - r_t(a)| \tag{5}$$

where R is such that $|\epsilon_t(a)| \leq R$ (range of the noise process).

3.1 Assumptions

Let the filtration $\mathcal{F}_t$ be $\mathcal{F}_t = \sigma(\mathcal{A}_{1:t}, a_{1:t-1}, \epsilon_{1:t-1}, r_{1:t-1}, \sigma_{1:t-1})$. For all time-steps $t \in [T]$ and actions $a \in \bigcup_{t=1}^{T} \mathcal{A}_t$, we have the following assumptions:

Assumption 1. *(Bounded Action) $\|a\|_2 \leq A$*

Assumption 2. *(Bounded Mean Reward) $|\langle a, \theta^* \rangle| \leq 1$*

Assumption 3. *Noise process $\epsilon_t(a)$ satisfies:*

- *(Boundedness) $|\epsilon_t(a)| \leq R$*
- *(Martingale Difference Sequence) $\mathbb{E}[\epsilon_t(a)|\mathcal{F}_t] = 0$*
- *(Bounded Variance) $\mathbb{E}[\epsilon_t(a)^2|\mathcal{F}_t] \leq \sigma_t^2(a)$*

Assumption 4. *(Bounded True Parameter) $\|\theta^*\|_2 \leq B$*

Notice that because of Assumption 2 and 3, the drawn stochastic reward from the environment r_t is bounded $|r_t| \leq R + 1$.

3.2 Regret

Let the regret for an agent interacting in this environment be:

$$\text{Regret}_{\text{TV}}(T) = \sum_{t=1}^{T} \langle a_t^*, \theta^* \rangle - \mathbb{E}\Big[\sum_{t=1}^{T} \langle a_t, \theta^* \rangle\Big] \tag{6}$$

where $a_t^* = \arg\max_{a \in \mathcal{A}_t} \langle a, \theta^* \rangle$ is the optimal action for that time-step, and the expectation is with respect to the randomness in the agent's policy.

3.3 Algorithm for Known Corruption Level

If the corruption level C in Equation 5 is known to the agent, then a robust version of weighted OFUL [Zhou et al., 2021, Abbasi-Yadkori et al., 2011] can achieve a regret upper bound of $\tilde{O}(CRd\sqrt{T})$.

3.3.1 Concentration Inequality with Enlarged Ellipsoid

We first present and prove Lemma 1 which is needed to prove Lemma 2 which states that the confidence set $\mathcal{C}_t$ we choose contains the true parameter θ^* at all time-steps t.

Let the confidence set $\mathcal{C}_t$ be:

$$C_t := \{\theta \mid \|\theta - \theta_t\|_{\Sigma_t} \leq \alpha_t\} \tag{7}$$

where $\Sigma_t = \lambda I + \sum_{i=1}^{t} \frac{1}{\bar{\sigma}_t^2} a_i a_i^\top$ (see Algorithm 1 for definition of $\bar{\sigma}_t$) and

$$\alpha_t = 8\sqrt{d \log \frac{(R+1)^2\lambda + tA^2}{(R+1)^2\lambda} \log(4t^2/\delta)} + 4\sqrt{d}\log(4t^2/\delta) + C\sqrt{d} + \sqrt{\lambda}B \tag{8}$$

Notice that $(\alpha_t)_t$ is monotonically increasing with t.

Lemma 1. *(Bernstein Inequality for Vector-Valued Martingales with Corruptions) Let $\{\mathcal{F}_t\}_t$ be a filtration, $\{x_t, \eta_t\}_{t \geq 1}$ be a stochastic process such that $x_t \in \mathbb{R}^d$ is $\mathcal{F}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{F}_{t+1}$-measurable. Fix constants $R, L, \sigma, \lambda > 0, \theta^* \in \mathbb{R}^d$. For $t \geq 1$, let $y_t^{stoch} = \langle \theta^*, x_t \rangle + \eta_t$ and suppose η_t, x_t also satisfy:*

$$|\eta_t| \leq R, \;\; \mathbb{E}[\eta_t|\mathcal{F}_t] = 0, \;\; \mathbb{E}[\eta_t^2|\mathcal{F}_t] \leq \sigma^2, \;\; \|x_t\|_2 \leq L$$

Suppose $\{y_t\}$ is a sequence such that $\sum_{i=1}^{t} |y_i - y_i^{stoch}| = C(t) \quad \forall t \geq 1$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have for all $t \geq 1$:

$$\|\theta_t - \theta^*\|_{V_t} \leq \beta_t + C(t) + \sqrt{\lambda}\|\theta^*\|_2$$

where $\theta_t = V_t^{-1} b_t, V_t = \lambda I + \sum_{i=1}^{t} x_i x_i^\top, b_t = \sum_{i=1}^{t} y_i x_i$, and

$$\beta_t = 8\sigma\sqrt{d \log \frac{d\lambda + tL^2}{d\lambda} \log(4t^2/\delta)} + 4R\log(4t^2/\delta)$$

Proof. Let $S(t) = \{1 \leq i \leq t | y_i \neq y_i^{\text{stoch}}\}$, $b_t^{\text{stoch}} = \sum_{i=1}^{t} y_i^{\text{stoch}} x_i$ and $\theta_t^{\text{stoch}} = V_t^{-1} b_t^{\text{stoch}}$ Using Theorem 4.1 from [Zhou et al., 2021] (Bernstein inequality for vector-valued martingales without corruptions), we have that with probability at least $1 - \delta$, for all $t \geq 1$:

$$\|\theta_t^{\text{stoch}} - \theta^*\|_{V_t} \leq \beta_t + \sqrt{\lambda}\|\theta^*\|_2$$

We also know that:

$$\|\theta_t - \theta_t^{\text{stoch}}\|_{V_t} = \|V_t^{-1}(b_t - b_t^{\text{stoch}})\|_{V_t}$$

$$= \left\| V_t^{-1}\Big(\sum_{i=1}^{t} y_i x_i - \sum_{i=1}^{t} y_i^{\text{stoch}} x_i\Big) \right\|_{V_t} = \left\| \Big(\sum_{i=1}^{t} V_t^{-1}(y_i - y_i^{\text{stoch}}) x_i\Big) \right\|_{V_t}$$

$$\leq \sum_{i=1}^{t} \left\| V_t^{-1}(y_i - y_i^{\text{stoch}}) x_i \right\|_{V_t} \quad \text{(extension of triangle-inequality)}$$

$$= \sum_{i=1}^{t} |y_i - y_i^{\text{stoch}}| \cdot \left\| V_t^{-1} x_i \right\|_{V_t}$$

$$= \sum_{i=1}^{t} |y_i - y_i^{\text{stoch}}| \cdot (V_t^{-1} x_i)^\top V_t V_t^{-1} x_i$$

$$= \sum_{i=1}^{t} |y_i - y_i^{\text{stoch}}| \cdot \|x_i\|_{V_t^{-1}} \quad \text{(since } V_t^{-1} \text{ is symmetric)}$$

$$\leq C(t) \quad \text{(since } \|x_i\|_{V_t^{-1}} \leq 1)$$

Therefore:

$$\|\theta_t - \theta^*\|_{V_t} = \|\theta_t - \theta_t^{\text{stoch}} + \theta_t^{\text{stoch}} - \theta^*\|_{V_t}$$

$$\leq \|\theta_t - \theta_t^{\text{stoch}}\|_{V_t} + \|\theta_t^{\text{stoch}} - \theta^*\|_{V_t} \quad \text{(by triangle-inequality)]}$$

$$\leq \beta_t + C(t) + \lambda\|\theta^*\|_2$$

□

Algorithm 1: Robust Weighted OFUL, an extension of Weighted OFUL [Zhou et al., 2021]

```
Input:
$V_1 = \lambda I, \theta_1 = 0, b_1 = 0$
for $t = 1, 2, ..., T$ do
    Observed action set $\mathcal{A}_t$
    Set $\mathcal{C}_t$ defined in Equation 7
    Select action $a_t = \underset{a \in \mathcal{A}_t}{\arg\max} \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle$.
    Execute action $a_t$ and observe $r_t, \sigma_t$.
    Set $\bar{\sigma}_t = \max\{\sigma_t, (R+1)/\sqrt{d}\}$.
    Update weighted RLS estimator: $\theta_{t+1} = \Sigma_{t+1}^{-1} b_{t+1}$, $\Sigma_{t+1} = \Sigma_t + \frac{1}{\bar{\sigma}_t^2} a_t a_t^\top$, $b_{t+1} = b_t + \frac{1}{\bar{\sigma}_t^2} r_t a_t$
```

Lemma 2. *(True Parameter Contained in Enlarged Confidence Ellipsoid) Suppose the assumptions of Lemma 1 hold. With probability at least* $1 - \delta$*, we have* $\theta^* \in \mathcal{C}_t$ *for all* $t \geq 1$*.*

Proof. Using Lemma 1, we have with probability at least $1 - \delta$:

$$\|\theta_t - \theta^*\|_{\Sigma_t} \leq \beta_t + C(t) + \sqrt{\lambda}\|\theta^*\|_2$$

$$= 8\sigma \sqrt{d \log \frac{d\lambda + tA^2}{d\lambda} \log(4t^2/\delta)} + 4R \log(4t^2/\delta) + C(t) + \sqrt{\lambda}\|\theta^*\|_2$$

$$\leq 8\sqrt{d \log \frac{(R+1)^2\lambda + tA^2}{(R+1)^2\lambda} \log(4t^2/\delta)} + 4\sqrt{d}\log(4t^2/\delta) + C(R+1) + \sqrt{\lambda}\|\theta^*\|_2 \quad \text{(since authors assume } R+1 = O(\sqrt{d})\text{)}$$

$$\leq 8\sqrt{d \log \frac{(R+1)^2\lambda + tA^2}{(R+1)^2\lambda} \log(4t^2/\delta)} + 4\sqrt{d}\log(4t^2/\delta) + C\sqrt{d} + \sqrt{\lambda}\|\theta^*\|_2$$

$$\leq 8\sqrt{d \log \frac{(R+1)^2\lambda + tA^2}{(R+1)^2\lambda} \log(4t^2/\delta)} + 4\sqrt{d}\log(4t^2/\delta) + C\sqrt{d} + \sqrt{\lambda}B$$

$$= \alpha_t$$

Therefore, $\theta^* \in \mathcal{C}_t$ for all $t \geq 1$ □

3.3.2 Regret Bound

Theorem 3. *(Regret Bound for Known Corruption Level) Set* $\lambda = 1/B^2$*. Suppose the corruption level* C *is a known constant and* $R \in \mathbb{R}$ *is an upper bound for the noise process. Assume Assumptions 1, 2, 3 all hold. Then with probability at least* $1 - \delta$*, Algorithm 1 achieves regret bounded by:*

$$\textit{Regret}(T) \leq \tilde{O}\left(Cd\sqrt{\sum_{t=1}^{T} \sigma_t^2} + C(R+1)\sqrt{dT}\right) \tag{9}$$

Proof. From Lemma 2, we know that $\theta^* \in \mathcal{C}_t$ for all $t \in [T]$. Then the total regret (w.r.t. a standard oracle) is bounded as follows:

$$\text{Regret}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \langle a_t^*, \theta^* \rangle - \langle a_t, \theta^* \rangle\right]$$

Let $\delta_t := \langle a_t^*, \theta^* \rangle - \langle a_t, \theta^* \rangle$ and $\tilde{\theta}_t \in \mathcal{C}_t$ be the parameter in the confidence set for which $\langle a_t, \tilde{\theta}_t \rangle = \max_{\theta \in \mathcal{C}_t} \langle a_t, \theta \rangle$. Notice that

$$\langle a_t^*, \theta^* \rangle \leq \max_{\theta \in \mathcal{C}_t} \langle a_t^*, \theta \rangle \quad (\text{since } \theta^* \in \mathcal{C}_t)$$

$$\leq \langle a_t, \tilde{\theta}_t \rangle \quad (\text{by def. of algorithm which selects } a_t = \underset{a \in \mathcal{A}_t}{\arg\max} \ \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle)$$

Therefore:

$$\delta_t \leq \langle a_t, \tilde{\theta}_t \rangle - \langle a_t, \theta^* \rangle = \langle a_t, \tilde{\theta}_t - \theta^* \rangle$$

$$\leq \|a_t\|_{\Sigma_t^{-1}} \|\tilde{\theta}_t - \theta^*\|_{\Sigma_t} \quad (\text{Cauchy-Schwartz and bounding } \|\cdot\|_{\Sigma_t^{-1}} \leq \|\cdot\|_{\Sigma_t} \text{ Lemma 5})$$

$$\leq \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T \quad (\text{bounding by diameter of ellipsoid})$$

$$\leq \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T \quad (\text{by monotonicity of } (\alpha_t)_t)$$

Notice also that by Assumption 2, $\delta_t \leq |\langle a_t^*, \theta^* \rangle - \langle a_t, \theta^* \rangle| \leq |\langle a_t^*, \theta^* \rangle| + |\langle a_t, \theta^* \rangle| \ \leq 2$. So:

$$\text{Regret}(T) \leq \mathbb{E}\left[\sum_{t=1}^{T} \min(2, \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T)\right]$$

We now break up the indices t into two cases and bound each case. Define $\mathcal{I}_1 := \{t \in [T] : \|a_t/\bar{\sigma}_t\|_{\Sigma_t^{-1}} > 1\}$ and $\mathcal{I}_2 := \{t \in [T] : \|a_t/\bar{\sigma}_t\|_{\Sigma_t^{-1}} \leq 1\}$. Notice that by construction, indices in $\mathcal{I}_1$ are when the summand is 2, and indices in $\mathcal{I}_2$ are when the summand is $2\|a_t\|_{\Sigma_t^{-1}} \cdot \alpha_T \leq 2$.

Focusing on $\mathcal{I}_1$:

$$|\mathcal{I}_1| \leq \sum_{t=1}^{T} \min(1, \|a_t/\bar{\sigma}_t\|^2_{\Sigma_t^{-1}})$$

$$\leq 2d \log \frac{(R+1)^2\lambda + TA^2}{(R+1)^2\lambda} \quad \text{(using Lemma 4)}$$

Therefore:

$$\sum_{t \in \mathcal{I}_1} \min(2, \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T) = 2|\mathcal{I}_1| \leq 4d \log \frac{(R+1)^2\lambda + TA^2}{(R+1)^2\lambda}$$

Now considering $\mathcal{I}_2$:

$$\sum_{t \in \mathcal{I}_2} \min(2, \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T)$$

$$= 2\alpha_T \sum_{t \in \mathcal{I}_2} \|a_t\|_{\Sigma_t^{-1}} \quad \text{(by construction of } \mathcal{I}_2\text{)}$$

$$= 2\alpha_T \sum_{t \in \mathcal{I}_2} \bar{\sigma}_t \|a_t/\bar{\sigma}_t\|_{\Sigma_t^{-1}}$$

$$\leq 2\alpha_T \sqrt{\sum_{t \in \mathcal{I}_2} \bar{\sigma}_t^2} \cdot \sqrt{\sum_{t \in \mathcal{I}_2} \|a_t/\bar{\sigma}_t\|^2_{\Sigma_t^{-1}}} \quad \text{(Cauchy-Schwartz)}$$

Now notice the first term:

$$\sqrt{\sum_{t \in \mathcal{I}_2} \bar{\sigma}_t^2} \leq \sqrt{\sum_{t \in \mathcal{I}_2} \sigma_t^2 + (R+1)^2/d} \quad \text{(by definition of } \bar{\sigma}_t = \max\{\sigma_t, (R+1)/\sqrt{d}\}\text{)}$$

$$\leq \sqrt{\sum_{t \in \mathcal{I}_2} \sigma_t^2} + \sqrt{\sum_{t \in \mathcal{I}_2} (R+1)^2/d} \quad \text{(since } \sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \text{ for } a, b > 0\text{)}$$

$$\leq \sqrt{\sum_{t=1}^{T} \sigma_t^2} + \sqrt{(R+1)^2 T/d}$$

The second term:

$$\sqrt{\sum_{t \in \mathcal{I}_2} \|a_t/\bar{\sigma}_t\|^2_{\Sigma_t^{-1}}} \leq \sqrt{\sum_{t=1}^{T} \min(1, \|a_t/\bar{\sigma}_t\|^2_{\Sigma_t^{-1}})} \quad \text{(by construction of } \mathcal{I}_2\text{)}$$

$$\leq \sqrt{\log \frac{(R+1)^2\lambda + TA^2}{(R+1)^2\lambda}} \quad \text{(using Lemma 4)}$$

Therefore:

$$\sum_{t \in \mathcal{I}_2} \min(2, \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T) \leq 2\alpha_T \cdot \left(\sqrt{\sum_{t=1}^{T} \sigma_t^2} + \sqrt{(R+1)^2 T/d} \right) \cdot \sqrt{2d \log \frac{(R+1)^2\lambda + TA^2}{(R+1)^2\lambda}}$$

Finally:

$$\mathbb{E}\left[\sum_{t=1}^{T} \min(2, \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T) \right] = \mathbb{E}\left[\sum_{t \in \mathcal{I}_1} \min(2, \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T) + \sum_{t \in \mathcal{I}_2} \min(2, \|a_t\|_{\Sigma_t^{-1}} 2\alpha_T) \right]$$

$$\leq 4d \log \frac{(R+1)^2\lambda + TA^2}{(R+1)^2\lambda} + 2\alpha_T \cdot \left(\sqrt{\sum_{t=1}^{T} \sigma_t^2} + \sqrt{(R+1)^2 T/d} \right) \cdot \sqrt{2d \log \frac{(R+1)^2\lambda + TA^2}{(R+1)^2\lambda}}$$

$$\leq \tilde{O}\left(Cd \sqrt{\sum_{t=1}^{T} \sigma_t^2} + C(R+1)\sqrt{dT} \right) \quad \text{(since } \alpha_T \leq \tilde{O}(C\sqrt{d}) \text{ when } \lambda = 1/B^2\text{)}$$

□

Remarks on Regret Bound. Theorem 3 shows that for fixed and known corruption level C, Algorithm 1 achieves sub-linear regret in terms of T, but linear regret in terms of C. Next notice the $\sqrt{\sum_{t=1}^{T} \sigma_t^2}$ term. If we trivially upper bound each σ_t^2 by R^2, then the regret becomes $\tilde{O}(CdR\sqrt{T})$. This indicates that we may be able to improve the regret bound if we incorporate some information about the variance.

3.4 Extension: Algorithm for Unknown Corruption Level

If the corruption level C is unknown to the agent, then [Zhao et al., 2021] proposed an algorithm, Multi-level weighted OFUL. The algorithm implements an action partition scheme to group historical data and maintains several additional estimators besides the original estimator θ_t. During action-selection, the algorithm randomly selects one of the learners with different probabilities at each time-step. More specifically, the algorithm partitions the historical data into $l_{\max}$ levels and maintain $l_{\max}$ sub-sampled estimators $\theta_{t,1}, ..., \theta_{t,l_{\max}}$ at time-step t. The observed data obtained in time-step t goes into level l with probability 2^{-l} if $1 < l \leq l_{\max}$ and it goes to level 1 with probability $1 - \sum_{l=2}^{l_{\max}} 2^{-l} = 1/2 + 2^{-l_{\max}}$.

Let the corruption experienced at level l, up to time-step t be:

$$\text{Corruption}_{t,l} = \sum_{i=1}^{t} \frac{\mathbb{I}[f(i) = l]}{R+1} \sup_{a \in \mathcal{A}_i} |r_i(a) - r'_i(a)| \tag{10}$$

where $f(i) \in \{1, ..., l_{\max}\}$ denotes the level that was chosen at time-step i.

The intuition is that if the corruption level C is less than or equal to 2^l, then $C_{t,l}$ can be upper-bounded by some quantity that is *independent* of C. In addition, the learners whose level if greater than $\log C$ can still learn θ^* even with the corruption. On the other hand, for learners whose level if less than $\log C$, the algorithm can control the error by controlling the probability of selecting them. [Zhao et al., 2021] is able to prove a $\tilde{O}(C^2 d\sqrt{\sum_{t=1}^{T} \sigma_t^2} + C^2 R\sqrt{dT})$ regret bound for Multi-level weighted OFUL, which has an extra factor of C as compared to the regret bound for known corruption level (Equation 9) due to the multi-level structure maintained to deal with the unknown corruption level.

4 Further Reading

4.1 Adversarial Bandits

A closely-related, yet subtly different, bandit problem is the adversarial bandit problem [Auer et al., 1995] where the environment is dynamic and can strategically adapt based on the agent's past actions and observations. Some formulations of adversarial bandits even assume that the environment has access to the policy or strategy of the proposed algorithms. For example, the environment could strategically change the reward distributions over time or even adversarially generate the arm set or observed context observed by the agent [Neu and Olkhovskaya, 2020]. This is in contrast to bandits with adversarial corruptions which assumes the true environment is stationary, but an adversary corrupts the information on the rewards provided to the agent. In adversarial bandit environments, algorithms must implement randomized policies to gather diverse information and make it difficult for the adversary to predict and counter the agent's decisions. One of the most simple yet foundational algorithms in this environment is Exp3 [Auer et al., 2002].

4.2 Stochastic Bandits with Adversarial Corruptions

4.3 Robust Regression

Bandits with adversarial corruptions is closely tied to the long-studied problem of robust statistics [Li, 2018] and more specifically robust regression [Hopkins and Li, 2018, Klivans et al., 2018]. In these problems, the environment operates in a stochastic setting where the covariates are drawn i.i.d. from some data distribution, but the adversary can arbitrarily alter any η fraction of the responses or labels and the corresponding covariates. Algorithms in these settings are evaluated by their ability to minimize the clean mean squared error (MSE), The mean squared error over the subset of data points that are uncorrupted. Notice that η must be less than 1/2 (optimal breakdown point) or no estimator can accurately nor tractably achieve low non-trivial clean MSE. A majority of the work assume that the uncorrupted data is evenly spread out: such as assuming the generative model is Gaussian [Diakonikolas et al., 2019], or atleast hypercontractive [Klivans et al., 2018], or certifiably sub-Gaussian [Hopkins and Li, 2018].

References

[Abbasi-Yadkori et al., 2011] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

[Auer et al., 1995] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE.

[Auer et al., 2002] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

[Bhatia et al., 2017] Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2017). Consistent robust regression. *Advances in Neural Information Processing Systems*, 30.

[Bouneffouf and Rish, 2019] Bouneffouf, D. and Rish, I. (2019). A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*.

[Cai et al., 2021] Cai, W., Grossman, J., Lin, Z. J., Sheng, H., Wei, J. T.-Z., Williams, J. J., and Goel, S. (2021). Bandit algorithms to personalize educational chatbots. *Machine Learning*, pages 1–30.

[Chen et al., 2022] Chen, S., Koehler, F., Moitra, A., and Yau, M. (2022). Online and distribution-free robustness: Regression and contextual bandits with huber contamination. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 684–695. IEEE.

[Diakonikolas et al., 2019] Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864.

[Foster and Rakhlin, 2020] Foster, D. and Rakhlin, A. (2020). Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR.

[Gupta et al., 2019] Gupta, A., Koren, T., and Talwar, K. (2019). Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR.

[Hopkins and Li, 2018] Hopkins, S. B. and Li, J. (2018). Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034.

[Jiang et al., 2015] Jiang, N., Kulesza, A., Singh, S., and Lewis, R. (2015). The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189. Citeseer.

[Kapoor et al., 2019] Kapoor, S., Patel, K. K., and Kar, P. (2019). Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715.

[Klivans et al., 2018] Klivans, A., Kothari, P. K., and Meka, R. (2018). Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR.

[Lattimore and Szepesvári, 2020] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms.* Cambridge University Press.

[Li, 2018] Li, J. Z. (2018). *Principled approaches to robust machine learning and beyond.* PhD thesis, Massachusetts Institute of Technology.

[Li et al., 2010] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

[Liao et al., 2019] Liao, P., Greenewald, K. H., Klasnja, P. V., and Murphy, S. A. (2019). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *CoRR*, abs/1909.03539.

[Lykouris et al., 2018] 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*, pages 114–122.

[Neu and Olkhovskaya, 2020] Neu, G. and Olkhovskaya, J. (2020). Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, pages 3049–3068. PMLR.

[Trella et al., 2022] Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., and Murphy, S. A. (2022). Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8):255.

[Zhao et al., 2021] Zhao, H., Zhou, D., and Gu, Q. (2021). Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615.*

[Zhou et al., 2021] Zhou, D., Gu, Q., and Szepesvari, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR.

A Auxillary Lemmas

Lemma 4. *(Lemma 11 in [Abbasi-Yadkori et al., 2011]) For any $\lambda > 0$ and sequence $\{x_t\}_{t=1}^T \subset \mathbb{R}^d$ for $t \in 0 \cup [T]$, let $V_t = \lambda I + \sum_{i=1}^t x_i x_i^\top$. Then provided that $\|x_t\|_2 \leq L$ holds for all $t \in [T]$, we have:*

$$\sum_{t=1}^T \min\{1, \|x_t\|^2_{V_{t-1}^{-1}}\} \leq 2d \log \frac{d\lambda + TL^2}{d\lambda}$$

Lemma 5. *Let $V_t = \lambda I + \sum_{j=1}^t a_t a_t^\top$ for $\lambda > 0 \in \mathbb{R}^{d \times d}$ be the matrix defined in regularized least squares. Then for any vector $x \in \mathbb{R}^d$,*

$$\|x\|_{V_t^{-1}} \leq \|x\|_{V_t}$$

Proof. First confirm that V_t is a positive definite (PD) matrix. This is because each $a_t a_t^\top$ and λI are PD matrices and the sum of PD matrices is also a PD matrix. By definition of V_t being a PD matrix, all eigenvalues $\lambda_1, ..., \lambda_d$ of V_t are positive with corresponding eigenvectors $v_1, ..., v_d$. By definition, V_t^{-1} has the same eigenvectors but with reciprocal eigenvalues $\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_d}$ also positive. Then:

$$\|x\|_{V_t} = x^\top V_t x = x^\top PDP^{-1}x = x^\top \sum_{i=1}^{d} \lambda_i v_i v_i^\top x \quad \text{(eigendecomposition of } V_t\text{)}$$

Similarly,

$$\|x\|_{V_t^{-1}} = x^\top V_t^{-1} x = x^\top PD^{-1}P^{-1}x = x^\top \sum_{i=1}^{d} \frac{1}{\lambda_i} v_i v_i^\top x \quad \text{(eigendecomposition of } V_t^{-1}\text{)}$$

$$\begin{aligned} \implies x^\top V_t x - x^\top V_t^{-1} x &= x^\top \sum_{i=1}^{d} \left(\lambda_i - \frac{1}{\lambda_i}\right) v_i v_i^\top x \\ &= \sum_{i=1}^{d} \left(\lambda_i - \frac{1}{\lambda_i}\right) \langle x, v_i \rangle^2 \geq 0 \end{aligned}$$

□