Achieving Near Instance-Optimality and Minimax-Optimality in Stochastic and Adversarial Linear Bandits Simultaneously

Chung-Wei Lee [1] **Haipeng Luo** [1] **Chen-Yu Wei** [1] **Mengxiao Zhang** [1] **Xiaojin Zhang** [2]

Abstract

In this work, we develop linear bandit algorithms that automatically adapt to different environments. By plugging a novel loss estimator into the optimization problem that characterizes the instance-optimal strategy, our first algorithm not only achieves nearly instance-optimal regret in stochastic environments, but also works in corrupted environments with additional regret being the amount of corruption, while the state-of-the-art (Li et al., 2019) achieves neither instance-optimality nor the optimal dependence on the corruption amount. Moreover, by equipping this algorithm with an adversarial component and carefully-designed testings, our second algorithm *additionally* enjoys minimax-optimal regret in completely adversarial environments, which is the first of this kind to our knowledge. Finally, all our guarantees hold with high probability, while existing instance-optimal guarantees only hold in expectation.

1. Introduction

We consider the linear bandit problem with a finite and fixed action set. In this problem, the learner repeatedly selects an action from the action set and observes her loss whose mean is the inner product between the chosen action and an unknown loss vector determined by the environment. The goal is to minimize the *regret*, which is the difference between the learner's total loss and the total loss of the best action in hindsight. Two standard environments are heavily-studied in the literature: the stochastic environment and the adversarial environment. In the stochastic environment, the loss vector is fixed over time, and we are interested in instance-optimal regret bounds of order $o(T^\epsilon)$ for any $\epsilon > 0$, where T is the number of rounds and $o(\cdot)$ hides some instance-dependent constants. On the other hand, in the adversarial environment, the loss vector can be arbitrary in each round, and we are interested in minimax-optimal regret bound $\widetilde{\mathcal{O}}(\sqrt{T})$, where $\widetilde{\mathcal{O}}(\cdot)$ hides the problem dimension and logarithmic factors in T.

While there are many algorithms obtaining such optimal bounds in either environment (e.g., (Lattimore & Szepesvari, 2017) in the stochastic setting and (Bubeck et al., 2012) in the adversarial setting), a natural question is whether there exists an algorithm achieving both guarantees simultaneously without knowing the type of the environment. Indeed, the same question has been studied extensively in recent years for the special case of multi-armed bandits where the action set is the standard basis (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Wei & Luo, 2018; Zimmert & Seldin, 2019). Notably, Zimmert & Seldin (2019) developed an algorithm that is optimal up to universal constants for both stochastic and adversarial environments, and the techniques have been extended to combinatorial semi-bandits (Zimmert et al., 2019) and finite-horizon tabular Markov decision processes (Jin & Luo, 2020). Despite all these advances, however, it is still open whether similar results can be achieved for general linear bandits.

On the other hand, another line of recent works study the robustness of stochastic linear bandit algorithms from a different perspective and consider a corrupted setting where an adversary can corrupt the stochastic losses up to some limited amount C. This was first considered in multi-armed bandits (Lykouris et al., 2018; Gupta et al., 2019; Zimmert & Seldin, 2019; 2021) and later extended to linear bandits (Li et al., 2019) and Markov decision processes (Lykouris et al., 2019; Jin & Luo, 2020). Ideally, the regret of a robust stochastic algorithm should degrade with an additive term $\mathcal{O}(C)$ in this setting, which is indeed the case in (Gupta et al., 2019; Zimmert & Seldin, 2019; 2021; Jin & Luo, 2020) for multi-armed bandits or Markov decision processes, but is not achieved yet for general linear bandits.

In this paper, we make significant progress in this direction and develop algorithms with near-optimal regret simultaneously for different environments. Our main contributions are as follows.

* Authors are listed in alphabetical order. [1]University of Southern California [2]The Chinese University of Hong Kong. Correspondence to: Mengxiao Zhang <mengxiao.zhang@usc.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

- In Section 4, we first introduce Algorithm 1, a simple algorithm that achieves $\mathcal{O}(c(\mathcal{X},\theta)\log^2 T + C)$ regret with high probability in the corrupted setting,[1] where $c(\mathcal{X},\theta)$ is an instance-dependent quantity such that the instance-optimal bound for the stochastic setting (i.e. $C = 0$) is $\Theta(c(\mathcal{X},\theta)\log T)$. This result significantly improves (Li et al., 2019) which only achieves $\mathcal{O}\big(\frac{d^6\log^2 T}{\Delta_{\min}^2} + \frac{d^{2.5}C\log T}{\Delta_{\min}}\big)$ where d is the dimension of the actions and $\Delta_{\min}$ is the minimum sub-optimality gap satisfying $c(\mathcal{X},\theta) \le \mathcal{O}\big(\frac{d}{\Delta_{\min}}\big)$. Moreover, Algorithm 1 also ensures an instance-independent bound $\widetilde{\mathcal{O}}(d\sqrt{T} + C)$ that some existing instance-optimal algorithms fail to achieve even when $C = 0$ (e.g., (Jun & Zhang, 2020)).

- In Section 5, based on Algorithm 1, we further propose Algorithm 2 which not only achieves nearly instance-optimal regret $\mathcal{O}(c(\mathcal{X},\theta)\log^2 T)$ in the stochastic setting, but also achieves the minimax optimal regret $\widetilde{\mathcal{O}}(\sqrt{T})$ in the adversarial setting (both with high probability). To the best of our knowledge, this is the first algorithm that enjoys the best of both worlds for linear bandits. Additionally, the same algorithm also guarantees $\widetilde{\mathcal{O}}\big(\frac{d\log^2 T}{\Delta_{\min}} + C\big)$ in the corrupted setting, which is slightly worse than Algorithm 1 but still significantly better than (Li et al., 2019).

- Finally, noticing the extra $\log T$ factor in our bound for the stochastic setting, in Appendix D we also prove that this is in fact inevitable if the same algorithm simultaneously achieves sublinear regret in the adversarial setting with high probability (which is the case for Algorithm 2). This generalizes the result of (Auer & Chiang, 2016) for two-armed bandits.

At a high level, Algorithm 1 utilizes a well-known optimization problem (that characterizes the lower bound in the stochastic setting) along with a robust estimator to determine a randomized strategy for each round. This ensures the near instance-optimality of the algorithm in the stochastic setting, and also the robustness to corruption when combined with a doubling trick. To handle the adversarial setting as well, Algorithm 2 switches between an adversarial linear bandit algorithm with high-probability regret guarantees and a variant of Algorithm 1, depending on the results of some carefully-designed statistical tests on the stochasticity of the environment.

[1] In the texts, $\mathcal{O}(\cdot)$ often hides lower-order terms (in terms of T dependence) for simplicity. However, in all formal theorem/lemma statements, we use $\mathcal{O}(\cdot)$ to hide universal constants only.

2. Related Work

Linear Bandits. Linear bandits is a classic model to study sequential decision problems. The stochastic setting dates back to (Abe & Long, 1999). Auer (2002) first used the optimism principle to solve this problem. Later, several algorithms were proposed based on confidence ellipsoids, further improving the regret bounds (Dani et al., 2008a; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Chu et al., 2011).

On the other hand, the adversarial setting was introduced by Awerbuch & Kleinberg (2004). Dani et al. (2008b) achieved the first $\mathcal{O}(\sqrt{T})$ expected regret bound using the Geometric Hedge algorithm (also called Exp2) with uniform exploration over a barycentric spanner. Abernethy et al. (2008) proposed the first computational efficient algorithm that achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret using the Following-the-Regularized-Leader framework. Bubeck et al. (2012) further tightened the bound by improving Exp2 with John's exploration. Our Algorithm 2 makes use of any adversarial linear bandit algorithm with high-probability guarantees (e.g., (Bartlett et al., 2008; Lee et al., 2020)) in a black-box manner.

Instance Optimality for Bandit Problems. In the stochastic setting, Lattimore & Szepesvari (2017) showed that, unlike multi-armed bandits, optimism-based algorithms or Thompson sampling can be arbitrarily far from optimal in some simple instances. They proposed an algorithm also based on the lower bound optimization problem to achieve instance-optimality, but their algorithm is deterministic and cannot be robust to an adversary. Instance-optimality was also considered in other related problems lately such as linear contextual bandits (Hao et al., 2020; Tirinzoni et al., 2020), partial monitoring (Komiyama et al., 2015), and structured bandits (Combes et al., 2017; Jun & Zhang, 2020). Most of these works only consider expected regret, while our guarantees all hold with high probability.

Best-of-Both-Worlds. Algorithms that are optimal for both stochastic and adversarial settings were studied in multi-armed bandits (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Wei & Luo, 2018; Zimmert & Seldin, 2019), semi-bandits (Zimmert et al., 2019), and Markov Decision Processes (Jin & Luo, 2020). On the other hand, linear bandits, a generalization of multi-armed bandits and semi-bandits, is much more challenging and currently underexplored in this direction. To the best of our knowledge, our algorithm is the first that guarantees near-optimal regret bounds in both stochastic and adversarial settings simultaneously.

Stochastic Bandits with Corruption. Lykouris et al. (2018) first considered the corrupted setting for multi-armed

bandits. Their results were improved by (Gupta et al., 2019; Zimmert & Seldin, 2019; 2021) and extended to linear bandits (Li et al., 2019; Bogunovic et al., 2020) and reinforcement learning (Lykouris et al., 2019). As mentioned, our results significantly improve those of (Li et al., 2019) (although their corruption model is slightly more general than ours; see Section 3). On the other hand, the results of (Bogunovic et al., 2020) are incomparable to ours, because they consider a setting where the adversary has even more power and can decide the corruption after seeing the chosen action. Finally, we note that (Lykouris et al., 2019, Theorem 3.2) considers episodic linear Markov decision processes in the corrupted setting, which can be seen as a generalization of linear bandits. However, this result is highly suboptimal when specified to linear bandits ($\Omega(C^2\sqrt{T})$ ignoring other parameters).

3. Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^d$ be a finite set that spans $\mathbb{R}^d$. Each element in $\mathcal{X}$ is called an *arm* or an *action*. We assume that $\|x\|_2 \le 1$ for all $x \in \mathcal{X}$. A linear bandit problem proceeds in T rounds. In each round $t = 1, \ldots, T$, the learner selects an action $x_t \in \mathcal{X}$. Simultaneously, the environment decides a hidden loss vector $\ell_t \in \mathbb{R}^d$ and generates some independent zero-mean noise $\epsilon_t(x)$ for each action x. Afterwards, the learner observes her loss $y_t = \langle x_t, \ell_t \rangle + \epsilon_t(x_t)$. We consider three different types of settings: stochastic, corrupted, and adversarial, explained in detail below.

In the stochastic setting, ℓ_t is fixed to some unknown vector $\theta \in \mathbb{R}^d$. We assume that there exists a unique optimal arm $x^* \in \mathcal{X}$ such that $\langle x^*, \theta \rangle < \min_{x^* \neq x \in \mathcal{X}} \langle x, \theta \rangle$, and define for each $x \in \mathcal{X}$, its *sub-optimality gap* as $\Delta_x = \langle x - x^*, \theta \rangle$. Also denote the minimum gap $\min_{x \neq x^*} \Delta_x$ by $\Delta_{\min}$.

The corrupted setting is a generalization of the stochastic setting, where in addition to a fixed vector θ, the environment also decides a corruption vector $c_t \in \mathbb{R}^d$ for each round (before seeing x_t) so that $\ell_t = \theta + c_t$.[2] We define the total amount of corruption as $C = \sum_t \max_{x \in \mathcal{X}} |\langle x, c_t \rangle|$. The stochastic setting is clearly a special case with $C = 0$. In both of these settings, we define the regret as $\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \sum_{t=1}^T \langle x_t - x, \theta \rangle = \sum_{t=1}^T \Delta_{x_t}$.

Finally, in the adversarial setting, ℓ_t can be chosen arbitrarily (possibly dependent on the learner's algorithm and her previously chosen actions). The difference compared to the corrupted setting (which also has potentially arbitrary loss vectors) is that the regret is now defined in terms of ℓ_t: $\mathrm{Reg}(T) = \max_{x \in \mathcal{X}} \sum_{t=1}^T \langle x_t - x, \ell_t \rangle$.

[2] In other words, the environment corrupts the observation y_t by adding $\langle x_t, c_t \rangle$. The setting of (Li et al., 2019) is slightly more general with the corruption on y_t being $c_t(x_t)$ for some function c_t that is not necessarily linear.

In all settings, we assume $\langle x, \theta \rangle, \langle x, c_t \rangle, \langle x, \ell_t \rangle$ and y_t are all in $[-1, 1]$ for all t and $x \in \mathcal{X}$. We also denote $\langle x, \ell_t \rangle$ by $\ell_{t,x}$ and similarly $\langle x, c_t \rangle$ by $c_{t,x}$.

It is known that the minimax optimal regret in the adversarial setting is $\Theta(d\sqrt{T})$ (Dani et al., 2008b; Bubeck et al., 2012). The instance-optimality in the stochastic case, on the other hand, is slightly more complicated. Specifically, an algorithm is called *consistent* if it guarantees $\mathbb{E}[\mathrm{Reg}(T)] = o(T^\epsilon)$ for any θ, $\mathcal{X}$, and $\epsilon > 0$. Then, a classic lower bound result (see e.g., (Lattimore & Szepesvari, 2017)) states that: for a particular instance $(\mathcal{X}, \theta)$, all consistent algorithms satisfy:[3]

$$\liminf_{T \to \infty} \frac{\mathbb{E}[\mathrm{Reg}(T)]}{\log T} \ge \Omega(c(\mathcal{X}, \theta)),$$

where $c(\mathcal{X}, \theta)$ is the objective value of the following optimization problem:

$$\inf_{N \in [0,\infty)^{\mathcal{X}}} \sum_{x \in \mathcal{X} \setminus \{x^*\}} N_x \Delta_x \tag{1}$$

$$\text{subject to } \|x\|^2_{H^{-1}(N)} \le \frac{\Delta_x^2}{2}, \quad \forall x \in \mathcal{X} \setminus \{x^*\} \tag{2}$$

and $H(N) = \sum_{x \in \mathcal{X}} N_x x x^\top$ (the notation $\|x\|_M$ denotes the quadratic norm $\sqrt{x^\top M x}$ with respect to a matrix M). This implies that the best instance-dependent bound for $\mathrm{Reg}(T)$ one can hope for is $\mathcal{O}(c(\mathcal{X}, \theta) \log T)$ (and more generally $\mathcal{O}(c(\mathcal{X}, \theta) \log T + C)$ for the corrupted setting). It can be shown that $c(\mathcal{X}, \theta) \le \mathcal{O}\left(\frac{d}{\Delta_{\min}}\right)$ (see Lemma 16), but this upper bound can be arbitrarily loose as shown in (Lattimore & Szepesvari, 2017).

The solution N_x in the optimization problem above specifies the least number of times action x should be drawn in order to distinguish between the present environment and any other alternative environment with a different optimal action. Many previous instance-optimal algorithms try to match their number of pulls for x to the solution N_x under some estimated gap $\widehat{\Delta}_x$ (Lattimore & Szepesvari, 2017; Hao et al., 2020; Jun & Zhang, 2020). While these algorithms are asymptotically optimal, their regret usually grows linearly when T is small (Jun & Zhang, 2020). Furthermore, they are all deterministic algorithms and by design cannot tolerate corruptions. We will show how these issues can be addressed in the next section.

Notations. We use $\mathcal{P}_{\mathcal{S}}$ to denote the probability simplex over $\mathcal{S}$: $\left\{p \in \mathbb{R}^{|\mathcal{S}|}_{\ge 0} : \sum_{s \in \mathcal{S}} p_s = 1\right\}$, and define the clipping operator $\mathrm{Clip}_{[a,b]}(v)$ as $\min(\max(v, a), b)$ for $a \le b$.

[3] The original proof is under the Gaussian noise assumption. To meet our boundedness assumption on y_t, it suffices to consider the case when y_t is a Bernoulli random variable, which only affects the constant of the lower bound.

4. A New Algorithm for the Corrupted Setting

In this section, we focus on the corrupted setting (hence covering the stochastic setting as well). We introduce a new algorithm that achieves with high probability an instance-dependent regret bound of $\mathcal{O}(c(\mathcal{X},\theta)\log^2 T + C)$ for large T and also an instance-independent regret bound of $\widetilde{\mathcal{O}}(d\sqrt{T}+C)$ for any T. This improves over previous instance-optimal algorithms (Lattimore & Szepesvari, 2017; Hao et al., 2020; Jun & Zhang, 2020) from several aspects: 1) first and foremost, our algorithm handles corruption optimally with extra $\mathcal{O}(C)$ regret, while previous algorithms can fail completely due to their deterministic nature; 2) previous bounds only hold in expectation; 3) previous algorithms might suffer linear regret when T is small, while ours is always $\widetilde{\mathcal{O}}(d\sqrt{T}+C)$ for any T. The price we pay is an additional $\log T$ factor in the instance-dependent bound. On the other hand, compared to the work of (Li et al., 2019) that also covers the same corrupted setting and achieves $\mathcal{O}\left(\frac{d^6\log^2 T}{\Delta_{\min}^2}+\frac{d^{2.5}C\log T}{\Delta_{\min}}\right)$, our results are also significantly better (recall $c(\mathcal{X},\theta)\le\mathcal{O}(d/\Delta_{\min})$), although as mentioned in Footnote 2, their results hold for an even more general setting with non-linear corruption.

Our algorithm is presented in Algorithm 1, which proceeds in blocks of rounds whose length grows in a doubling manner $(2^0, 2^1, \ldots)$. At the beginning of block m (denoted as $\mathcal{B}_m$), we compute a distribution p_m over actions by solving an optimization problem **OP** (Figure 1) using the empirical gap $\widehat{\Delta}_{m,x}$ estimated in the previous block (Line 5). Then we use p_m to sample actions for the entire block m, and construct an unbiased loss estimator $\widehat{\ell}_{t,x}$ in every round for every action x (Line 9). At the end of each block m, we use $\{\widehat{\ell}_{\tau,x}\}_{\tau\in\mathcal{B}_m}$ to construct a *robust loss estimator* $\mathrm{Rob}_{m,x}$ for each action (Line 12), which will then be used to construct $\widehat{\Delta}_{m+1,x}$ for the next block. We next explain the optimization problem **OP** and the estimators in detail.

OP is inspired by the lower bound optimization (Eq. (1) and Eq. (2)), where we normalize the pull counts N as a distribution p over the arms such that for a large m, $p_{m,x}\approx\frac{N_x}{2^m}$ holds for $x\neq x^*$. One key difference between our algorithm and previous ones (Lattimore & Szepesvari, 2017; Hao et al., 2020; Jun & Zhang, 2020) is exactly that we select actions randomly according to these distributions, while they try to deterministically match the pull count of each arm to N. Our randomized strategy not only prevents the environment from exploiting the knowledge on the learner's choices, but also allows us to construct unbiased estimator $\widehat{\ell}_{t,x}$ (Line 9) following standard adversarial linear bandit algorithms (Dani et al., 2008b; Bubeck et al., 2012). In fact, as shown in our analysis, the variance of the estimator $\widehat{\ell}_{t,x}$ is exactly bounded by $\|x\|^2_{S_m^{-1}}$ (for $t\in\mathcal{B}_m$), which is in turn bounded in terms of the sub-optimality gap of x in light of the constraint Eq. (4). The similar idea of imposing explicit constraints on the variance of loss estimators appears before in for example (Dudik et al., 2011; Agarwal et al., 2014) for contextual bandits. Finally, we point out that **OP** always has a solution due to the additive term $4d$ in Eq. (4) (see Lemma 11), and it can be solved efficiently by standard methods since Eq. (4) is a convex constraint.

Algorithm 1 Randomized Instance-optimal Algorithm

1 **Input**: $\delta<0.1$
2 $t\leftarrow 1$.
3 **for** $m=0,1,2\ldots$ **do**
4 Define block $\mathcal{B}_m=\{t,\ t+1,\ldots,\ t+2^m-1\}$.
5 Find a randomized strategy $p_m=\textbf{OP}(2^m,\widehat{\Delta}_m)$ with $\widehat{\Delta}_{m,x}=\begin{cases}0 & \text{if } m=0,\\ \mathrm{Rob}_{m-1,x}-\min_{x'\in\mathcal{X}}\mathrm{Rob}_{m-1,x'} & \text{else.}\end{cases}$
6 Compute second moment $S_m=\sum_{x\in\mathcal{X}}p_{m,x}xx^\top$.
7 **while** $t\in\mathcal{B}_m$ **do**
8 Sample $x_t\sim p_m$ and observe y_t.
9 Compute for all $x\in\mathcal{X}$, $\widehat{\ell}_{t,x}=x^\top S_m^{-1}x_t y_t$.
10 $t\leftarrow t+1$.
11 **for** $x\in\mathcal{X}$ **do**
12 Construct robust loss estimators $\mathrm{Rob}_{m,x}=\mathrm{Clip}_{[-1,1]}\left(\textbf{Catoni}_{\alpha_x}\left(\{\widehat{\ell}_{\tau,x}\}_{\tau\in\mathcal{B}_m}\right)\right)$ with $\alpha_x=\sqrt{\frac{4\log(2^m|\mathcal{X}|/\delta)}{2^m\cdot\|x\|^2_{S_m^{-1}}+2^m}}$.

OP$(t,\widehat{\Delta})$: return any minimizer p^* of the following:

$$\min_{p\in\mathcal{P}_\mathcal{X}}\sum_x p_x\widehat{\Delta}_x, \tag{3}$$

$$\text{s.t. } \|x\|^2_{S(p)^{-1}}\le\frac{t\widehat{\Delta}_x^2}{\beta_t}+4d,\ \ \forall x\in\mathcal{X}, \tag{4}$$

where $S(p)=\sum_{x\in\mathcal{X}}p_xxx^\top$ and $\beta_t=2^{15}\log\frac{t|\mathcal{X}|}{\delta}$.

Figure 1. Optimization Problem (OP)

Catoni$_\alpha$ $(\{X_1,X_2,\ldots,X_n\})$: return $\widehat{X}$, the unique root of the function $f(z)=\sum_{i=1}^n\psi(\alpha(X_i-z))$ where $\psi(y)=\begin{cases}\ln(1+y+y^2/2), & \text{if } y\ge 0,\\ -\ln(1-y+y^2/2), & \text{else.}\end{cases}$

Figure 2. Catoni's Estimator

Another important ingredient of our algorithm is the robust estimator $\text{Rob}_{m,x}$, which is a clipped version of the Catoni's esimator (Catoni, 2012) constructed using all the unbiased estimators $\{\widehat{\ell}_{\tau,x}\}_{\tau\in\mathcal{B}_m}$ from this block for action x (Figure 2). From a technical perspective, this avoids a lower-order term in Bernstein-style concentration bounds and is critical for our analysis. We in fact also believe that this is necessary since there is no explicit regularization on the magnitude of $\widehat{\ell}_{t,x}$, and it can indeed have a heavy-tailed distribution. While other robust estimators are possible, we use the Catoni's estimator which was analyzed in (Wei et al., 2020) for non-i.i.d. random variables (again important for our analysis).

The following theorem summarizes the nearly instance-optimal regret bound of Algorithm 1.

Theorem 1. *In the corrupted setting, Algorithm 1 guarantees that with probability at least* $1-\delta$,

$$\text{Reg}(T) = \mathcal{O}\left(c(\mathcal{X},\theta)\log T\log\frac{T|\mathcal{X}|}{\delta} + M^*\log^{\frac{3}{2}}\frac{1}{\delta} + C + d\sqrt{\frac{C}{\Delta_{\min}}}\log\frac{C|\mathcal{X}|}{\Delta_{\min}\delta}\right),$$

where M^* is some constant that depends on $\mathcal{X}$ and θ only.

The dominating term of this regret bound is thus $\mathcal{O}(c(\mathcal{X},\theta)\log^2 T + C)$ as claimed. The definition of M^* can be found in the proof (Appendix B) and is importantly independent of T. In fact, in Theorem 19, we also provide an alternative (albeit weaker) bound $\mathcal{O}(\frac{d^2(\log T)^2}{\Delta_{\min}} + C)$ for Algorithm 1 without the dependence on M^*.

The next theorem shows an instance-independent bound of order $\widetilde{\mathcal{O}}(d\sqrt{T}+C)$ for Algorithm 1, which previous instance-optimal algorithms fail to achieve as mentioned.

Theorem 2. *In the corrupted setting, Algorithm 1 guarantees that with probability at least* $1-\delta$, $\text{Reg}(T) \le \mathcal{O}(d\sqrt{T}\log(T|\mathcal{X}|/\delta)+C)$.

We emphasize that Algorithm 1 is parameter-free and does not need to know C to achieve these bounds. In the rest of the section, we provide a proof sketch for Theorem 1 and Theorem 2. First, we show that the estimated gap $\widehat{\Delta}_{m,x}$ is close to the true gap Δ_x with a constant multiplicative factor and some additive terms that go down at the rate of roughly $1/\sqrt{t}$ up to the some average amount of corruption.

Lemma 3. *With probability at least* $1-\delta$, *Algorithm 1 ensures for all m and all x,*

$$\Delta_x \le 2\widehat{\Delta}_{m,x} + \sqrt{\frac{d\gamma_m}{4\cdot 2^m}} + 2\rho_{m-1}, \tag{5}$$

$$\widehat{\Delta}_{m,x} \le 2\Delta_x + \sqrt{\frac{d\gamma_m}{4\cdot 2^m}} + 2\rho_{m-1}, \tag{6}$$

where $\rho_m = \sum_{k=0}^{m}\frac{2^k C_k}{4^{m-1}}$ (ρ_{-1} is defined as 0), $C_k = \sum_{\tau\in\mathcal{B}_k}\max_{x\in\mathcal{X}}|c_{\tau,x}|$ is the amount of corruption within block k, and $\gamma_m = 2^{15}\log(2^m|\mathcal{X}|/\delta)$.

As mentioned, the proof of Lemma 3 heavily relies on the robust estimators we use as well as the variance constraint Eq. (4). Next, we have the following lemma which bounds the objective value of **OP**.

Lemma 4. *Let p be the solution of **OP**$(t,\widehat{\Delta})$, where $\widehat{\Delta}\in\mathbb{R}^{|\mathcal{X}|}_{\ge 0}$. Then we have $\sum_{x\in\mathcal{X}}p_x\widehat{\Delta}_x = \mathcal{O}\left(\frac{d\log(t|\mathcal{X}|/\delta)}{\sqrt{t}}\right)$.*

Combining Lemma 3 and Lemma 4, we see that in block m, the regret of Algorithm 1 can be upper bounded by

$$\begin{aligned}&\mathcal{O}\left(2^m\sum_x p_{m,x}\Delta_x\right)\\ &= \widetilde{\mathcal{O}}\left(2^m\sum_x p_{m,x}\left(\widehat{\Delta}_{m,x}+\sqrt{\frac{d}{2^m}}+\rho_{m-1}\right)\right)\\ &= \widetilde{\mathcal{O}}\left(d\sqrt{2^m}+2^m\rho_{m-1}\right),\end{aligned}$$

where in the first equality we use Lemma 3 and in the second equality we use Lemma 4 with the fact that $p_m = \textbf{OP}(2^m,\widehat{\Delta}_m)$. Further summing this over m and relating $\sum_m 2^m\rho_{m-1}$ to C proves Theorem 2.

In addition, based on Lemma 3, we show that when $t\in\mathcal{B}_m$ is larger than $\Omega(C/\Delta_{\min})$ plus some problem-dependent constant, the estimated gap $\widehat{\Delta}_{m,x}$ becomes $\Theta(\Delta_x)$. Therefore, the solution $\{p_{m,x}\}_{x\in\mathcal{X}\setminus\{x^*\}}$ from **OP** is very close to $\{\frac{N_x}{2^m}\}_{x\in\mathcal{X}\setminus\{x^*\}}$, where N_x is the optimal solution of Eq. (1) and Eq. (2), except that we have an additional $\log(2^m|\mathcal{X}|/\delta)$ factor in the constraint (coming from β_{2^m}). Therefore, the regret is bounded by $\mathcal{O}(c(\mathcal{X},\theta)\log(T)\log(T|\mathcal{X}|/\delta))$ for large enough T. Formally, we have the following lemma.

Lemma 5. *Algorithm 1 guarantees with probability at least* $1-\delta$, *for some constant T^* depending on $\mathcal{X},\theta$, and C:*

$$\sum_{t=T^*+1}^{T}\sum_x p_{t,x}\Delta_x \le \mathcal{O}\left(c(\mathcal{X},\theta)\log(T)\log(T|\mathcal{X}|/\delta)\right).$$

Finally, to obtain Theorem 1, it suffices to apply Theorem 2 for the regret before round T^* and Lemma 5 for the regret after.

5. Best of Three Worlds

In this section, building on top of Algorithm 1, we develop another algorithm that enjoys similar regret guarantees in the stochastic or corrupted setting, and additionally guarantees $\widetilde{\mathcal{O}}(\sqrt{T})$ regret in the adversarial setting, without having any prior knowledge on which environment it is facing. To the

best of our knowledge, this kind of *best-of-three-worlds* guarantee only appears before for multi-armed bandits (Wei & Luo, 2018; Zimmert & Seldin, 2019) and Markov decision processes (Jin & Luo, 2020), but not for linear bandits.

Our algorithm requires a block-box access to an adversarial linear bandit algorithm $\mathcal{A}$ that satisfies the following:

Assumption 1. *$\mathcal{A}$ is a linear bandit algorithm that outputs a loss estimator $\widehat{\ell}_{t,x}$ for each action x after each time t. There exist L_0, $C_1 \geq 2^{15} d \log(T|\mathcal{X}|/\delta)$, and universal constant $C_2 \geq 20$, such that for all $t \geq L_0$, $\mathcal{A}$ guarantees the following with probability at least $1 - \frac{\delta}{T}$: $\forall x \in \mathcal{X}$,*

$$\sum_{s=1}^{t}(\ell_{s,x_s} - \ell_{s,x}) \leq \sqrt{C_1 t} - C_2 \left|\sum_{s=1}^{t}(\ell_{s,x} - \widehat{\ell}_{s,x})\right|. \quad (7)$$

Eq. (7) states that the regret of $\mathcal{A}$ against action x is bounded by a $\sqrt{t}$-order term minus the deviation between the loss of x and its estimator. While this might not seem intuitive, in fact, *all* existing linear bandit algorithms with a near-optimal high-probability bound satisfy Assumption 1, even though this may not have been stated explicitly (and one may need to slightly change the constant parameters in these algorithms to satisfy the conditions on C_1 and C_2). Below, we give two examples of such $\mathcal{A}$ and justify them in Appendix E.

- A variant of GeometricHedge.P (Bartlett et al., 2008) with an improved exploration scheme satisfies Assumption 1 with ($\delta' = \delta/(|\mathcal{X}| \log_2 T)$)

  $$C_1 = \Theta\left(d \log(T/\delta')\right), \quad L_0 = \Theta\left(d \log^2(T/\delta')\right).$$

- The algorithm of (Lee et al., 2020) satisfies Assumption 1 with ($\lg = \log(dT)$, $\delta'' = \delta/(|\mathcal{X}|T)$)

  $$C_1 = \Theta\left(d^6 \lg^8 \log(\lg/\delta'')\right), \quad L_0 = \Theta\left(\log(\lg/\delta'')\right).$$

With such a black-box at hand, our algorithm BOTW is shown in Algorithm 2. We first present its formal guarantees in different settings.

Theorem 6. *Algorithm 2 guarantees that with probability at least $1 - \delta$, in the stochastic setting ($C = 0$), $\mathrm{Reg}(T)$ is at most*

$$\mathcal{O}\left(c(\mathcal{X},\theta) \log T \log \frac{T|\mathcal{X}|}{\delta} + \frac{C_1 \sqrt{\log T}}{\Delta_{\min}} + M^* \log^{\frac{3}{2}} \frac{1}{\delta} + \sqrt{C_1 L_0}\right),$$

where M^* is the same problem-dependent constant as in Theorem 1; and in the corrupted setting ($C > 0$), $\mathrm{Reg}(T)$ is at most

$$\mathcal{O}\left(\frac{C_1 \log T}{\Delta_{\min}} + C + \sqrt{C_1 L_0}\right).$$

Algorithm 2 BOTW (Best of Three Worlds)

Input: an algorithm $\mathcal{A}$ satisfying Assumption 1.
Initialize: $L \leftarrow L_0$ (L_0 defined in Assumption 1).
while *true* **do**
  Run BOTW-SE with input L, and receive output t_0.
  $L \leftarrow 2t_0$.

In the case when $\mathcal{A}$ is the variant of GeometricHedge.P, the last bound is

$$\mathcal{O}\left(\frac{d \log(T|\mathcal{X}|/\delta) \log T}{\Delta_{\min}} + C\right).$$

Therefore, Algorithm 2 enjoys the nearly instance-optimal regret $\mathcal{O}(c(\mathcal{X},\theta) \log^2 T)$ in the stochastic setting as Algorithm 1[4], but slightly worse regret $\mathcal{O}(\frac{d \log^2 T}{\Delta_{\min}} + C)$ in the corrupted setting (recall again $c(\mathcal{X},\theta) \leq d/\Delta_{\min}$). In exchange, however, Algorithm 2 enjoys the following worst-case robustness in the adversarial setting.

Theorem 7. *In the adversarial setting, Algorithm 2 guarantees that with probability at least $1 - \delta$, $\mathrm{Reg}(T)$ is at most $\mathcal{O}\left(\sqrt{C_1 T \log T} + \sqrt{C_1 L_0}\right)$.*

The dependence on T in this bound is minimax-optimal as mentioned, while the dependence on d depends on the coefficient C_1 of the black-box. Note that because of this adversarial robustness, the $\log^2 T$ dependence in Theorem 6 turns out to be unavoidable, as we show in Theorem 27. In addition, Theorem 7 also works for the stochastic setting, which implies a regret bound of $\mathcal{O}(\sqrt{dT} \log(T|\mathcal{X}| \log_2 T/\delta))$. This is a factor of $\sqrt{d}$ better than the guarantee of Algorithm 1 shown in Theorem 2.

Next, in Section 5.1, we describe our algorithm in detail. Then in Section 5.2 and Section 5.3, we provide proof sketches for Theorem 7 and Theorem 6 respectively.

5.1. The algorithm

Algorithm 2 BOTW takes a black-box $\mathcal{A}$ satisfying Assumption 1 (with parameter L_0) as input, and then proceeds in epochs until the game ends. In each epoch, it runs its single-epoch version BOTW-SE (Algorithm 3) with a minimum duration L (initialized as L_0). Based on the results of some statistical tests, at some point BOTW-SE will terminate with an output $t_0 \geq L$. Then BOTW enters into the next epoch with L updated to $2t_0$, so that the number of epochs is always $\mathcal{O}(\log T)$.

BOTW-SE has two phases. In Phase 1, the learner executes the adversarial linear bandit algorithm $\mathcal{A}$. Starting from

[4]Note that when we choose $\mathcal{A}$ as the variant of GeometricHedge.P, $\frac{C_1\sqrt{\log T}}{\Delta_{\min}} = \mathcal{O}\left(\frac{d}{\Delta_{\min}} \log^{\frac{3}{2}} T\right)$ which is dominated by the term $\mathcal{O}(c(\mathcal{X},\theta) \log^2 T)$ when T is sufficiently large.

$t = L$ (i.e. after the minimum duration specified by the input), the algorithm checks in every round whether Eq. (9) and Eq. (10) hold for some action $\widehat{x}$ (Line 3). If there exists such an $\widehat{x}$, Phase 1 terminates and the algorithm proceeds to Phase 2. This test is to detect whether the environment is likely stochastic. Indeed, Eq. (9) and Eq. (10) imply that the performance of the learner is significantly better than all but one action (i.e., $\widehat{x}$). In the stochastic environment, this event happens at roughly $t \approx \Theta\left(\frac{d}{\Delta_{\min}^2}\right)$ with $\widehat{x} = x^*$. This is exactly the timing when the learner should stop using $\mathcal{A}$ whose regret grows as $\widetilde{\mathcal{O}}(\sqrt{t})$ and start doing more exploitation on the better actions, in order to keep the regret logarithmic in time for the stochastic environment. We define t_0 to be the time when Phase 1 ends, and $\widehat{\Delta}_x$ be the empirical gap for action x with respect to the estimators obtained from $\mathcal{A}$ so far (Line 4). In the stochastic setting, we can show that $\widehat{\Delta}_x = \Theta(\Delta_x)$ holds with high probability.

In the second phase, we calculate the action distribution using **OP** with the estimated gap $\{\widehat{\Delta}_x\}_{x\in\mathcal{X}}$. Indeed, if $\widehat{\Delta}_x$'s are accurate, the distribution returned by **OP** is close to the optimal way of allocating arm pulls, leading to near-optimal regret.[5] For technical reasons, there are some differences between Phase 2 and Algorithm 1. First, instead of using p_t, the distribution returned by **OP**, to draw actions, we mix it with $e_{\widehat{x}}$ (the distribution that concentrates on $\widehat{x}$), and draw actions using $\widetilde{p}_t = \frac{1}{2}e_{\widehat{x}} + \frac{1}{2}p_t$. This way, $\widehat{x}$ is drawn with probability at least $\frac{1}{2}$. Moreover, the loss estimator $\widehat{\ell}_{t,x}$ is now defined as the following:

$$\widehat{\ell}_{t,x} = \begin{cases} x^\top \widetilde{S}_t^{-1} x_t y_t, & x \neq \widehat{x} \\ \frac{y_t}{\widetilde{p}_{t,\widehat{x}}}\mathbb{I}\{x_t = \widehat{x}\}, & x = \widehat{x} \end{cases} \tag{8}$$

where $\widetilde{S}_t = \sum_{x\in\mathcal{X}} \widetilde{p}_{t,x} x x^\top$. While the construction of $\widehat{\ell}_{t,x}$ for $x \neq \widehat{x}$ is the same as Algorithm 1, we see that the construction of $\widehat{\ell}_{t,\widehat{x}}$ is different and is based on standard inverse probability weighting. These differences are mainly because we later use the average estimator instead of the robust mean estimator for $\widehat{x}$ (the latter produces a slightly looser concentration bound in our analysis). Therefore, we must ensure that $\widehat{x}$ is drawn with enough probability, and that the magnitude of $\widehat{\ell}_{t,\widehat{x}}$ is well-controlled.

Then, we define the average empirical gap in $[1, t]$ for $x \neq \widehat{x}$ and t in Phase 2 as the following:

$$\widehat{\Delta}_{t,x} = \frac{1}{t}\left(\sum_{s=1}^{t_0} \widehat{\ell}_{s,x} + (t - t_0)\mathrm{Rob}_{t,x} - \sum_{s=1}^{t} \widehat{\ell}_{s,\widehat{x}}\right) \tag{13}$$

where

$$\mathrm{Rob}_{t,x} = \mathrm{Clip}_{[-1,1]}\left(\textbf{Catoni}_{\alpha_x}\left(\left\{\widehat{\ell}_{\tau,x}\right\}_{\tau=t_0+1}^{t}\right)\right)$$

[5] Here, we solve **OP** at every iteration for simplicity. It can in fact be done only when time doubles, just like Algorithm 1.

Algorithm 3 BOTW-SE (BOTW – Single Epoch)

Input: L (minimum duration)
Define: $f_T = \log T$
Initialize: a new instance of $\mathcal{A}$.
// Phase 1
1 **for** $t = 1, 2, \ldots$ **do**
2 Execute and update $\mathcal{A}$. Receive estimators $\{\widehat{\ell}_{t,x}\}_{x\in\mathcal{X}}$.
3 **if** $t \geq L$ and there exists an action $\widehat{x}$ such that

$$\sum_{s=1}^{t} y_s - \sum_{s=1}^{t} \widehat{\ell}_{s,\widehat{x}} \geq -5\sqrt{f_T C_1 t}, \tag{9}$$

$$\sum_{s=1}^{t} y_s - \sum_{s=1}^{t} \widehat{\ell}_{s,x} \leq -25\sqrt{f_T C_1 t}, \quad \forall x \neq \widehat{x}, \tag{10}$$

4 **then** $t_0 \leftarrow t$, $\widehat{\Delta}_x \leftarrow \frac{1}{t_0}\left(\sum_{s=1}^{t_0} \widehat{\ell}_{s,x} - \widehat{\ell}_{s,\widehat{x}}\right)$, **break**.
// Phase 2
5 **for** $t = t_0 + 1, \ldots$ **do**
6 Let $p_t = \textbf{OP}(t, \widehat{\Delta})$ and $\widetilde{p}_t = \frac{1}{2}e_{\widehat{x}} + \frac{1}{2}p_t$.
7 Sample $x_t \sim \widetilde{p}_t$ and observe y_t.
8 Calculate $\widehat{\ell}_{t,x}$ and $\widehat{\Delta}_{t,x}$ based on Eq. (8) and Eq. (13).
9 **if**

$$\exists x \neq \widehat{x}, \ \widehat{\Delta}_{t,x} \notin \left[0.39\widehat{\Delta}_x, 1.81\widehat{\Delta}_x\right] \text{ or} \tag{11}$$

$$\sum_{s=t_0+1}^{t}\left(y_s - \widehat{\ell}_{s,\widehat{x}}\right) \geq 20\sqrt{f_T C_1 t_0}. \tag{12}$$

10 **then break**.
11 **Return** t_0.

with $\alpha_x = \left(\frac{4\log(t|\mathcal{X}|/\delta)}{t - t_0 + \sum_{\tau=t_0+1}^{t} 2\|x\|^2_{S_\tau^{-1}}}\right)^{\frac{1}{2}}$ (*c.f.* Figure 2). Note that we use a simple average estimator for $\widehat{x}$, but a hybrid of average estimator of Phase 1 and robust estimator of Phase 2 for other actions. These gap estimators are useful in monitoring the non-stochasticity of the environment, which is done via the tests Eq. (11) and Eq. (12). The first condition (Eq. (11)) checks whether the average empirical gap $\widehat{\Delta}_{t,x}$ is still close to the estimated gap $\widehat{\Delta}_x$ at the end of Phase 1. The second condition (Eq. (12)) checks whether the regret against $\widehat{x}$ incurred in Phase 2 is still tolerable. It can be shown that (see Lemma 10), with high probability Eq. (11) and Eq. (12) do not hold in a stochastic environment. Therefore, when either event is detected, BOTW-SE terminates and returns the value of t_0 to BOTW, which will then run BOTW-SE again from scratch with $L = 2t_0$.

In the following subsections, we provide a sketch of analysis for BOTW, further revealing the ideas behind our design.

5.2. Analysis for the Adversarial Setting (Theorem 7)

We first show that at any time t in Phase 2, with high probability, $\widehat{x}$ is always the best action so far.

Lemma 8. *With probability at least $1-\delta$, for at any t in Phase 2, we have $\widehat{x} \in \operatorname{argmin}_{x\in\mathcal{X}} \sum_{s=1}^{t} \ell_{s,x}$.*

Proof sketch. The idea is to prove that for any $x \neq \widehat{x}$, the deviation between the actual gap $\sum_{s=1}^{t}(\ell_{s,x} - \ell_{s,\widehat{x}})$ and the estimated gap $t\widehat{\Delta}_{t,x}$ is no larger than $\mathcal{O}(t\widehat{\Delta}_x)$. This is enough to prove the statement since $t\widehat{\Delta}_{t,x}$ is of order $\Omega(t\widehat{\Delta}_x)$ in light of the test in Eq. (11).

Bounding the derivation for Phase 2 is somewhat similar to the analysis of Algorithm 1, and here we only show how to bound the derivation for Phase 1: $\sum_{s=1}^{t_0}(\ell_{s,x} - \widehat{\ell}_{s,x})$. We start by rearranging Eq. (7) to get: $(C_2 - 1)\left|\sum_{s=1}^{t_0}(\ell_{s,x} - \widehat{\ell}_{s,x})\right| \le \sqrt{C_1 t_0} - \sum_{s=1}^{t_0}(\ell_{s,x_s} - \widehat{\ell}_{s,x}) = \sqrt{C_1 t_0} - \sum_{s=1}^{t_0}(\ell_{s,x_s} - \widehat{\ell}_{s,\widehat{x}}) + t_0\widehat{\Delta}_x$. By the termination conditions of Phase 1, we have $\sum_{s=1}^{t_0}(\ell_{s,x_s} - \widehat{\ell}_{s,\widehat{x}}) \ge -5\sqrt{f_T C_1 t_0}$ and $\widehat{\Delta}_x \ge 20\sqrt{f_T C_1/t_0}$, which then shows $\left|\sum_{s=1}^{t_0}(\ell_{s,x} - \widehat{\ell}_{s,x})\right| \le \frac{6\sqrt{f_T C_1 t_0} + t_0\widehat{\Delta}_x}{C_2 - 1} = \mathcal{O}(t_0\widehat{\Delta}_x)$ as desired. (See Appendix C for the full proof.) □

We then prove that, importantly, the regret in each epoch is bounded by $\widetilde{\mathcal{O}}(\sqrt{t_0})$ (not square root of the epoch length):

Lemma 9. *With probability at least $1-\delta$, for any time t in Phase 2, we have for any $x \in \mathcal{X}$,*

$$\sum_{s=1}^{t}(\ell_{s,x_s} - \ell_{s,x}) = \mathcal{O}\left(\sqrt{C_1 t_0 f_T}\right).$$

Proof sketch. By Lemma 8, it suffices to consider $x = \widehat{x}$. By Eq. (7), we know that the regret for the first t_0 rounds is directly bounded by $\mathcal{O}\left(\sqrt{C_1 t_0}\right)$. For the regret incurred in Phase 2, we decompose it as the sum of $\sum_{s=t_0+1}^{t}(y_s - \widehat{\ell}_{s,\widehat{x}})$, $\sum_{s=t_0+1}^{t}(\widehat{\ell}_{s,\widehat{x}} - \ell_{s,\widehat{x}} - \epsilon_s(\widehat{x}))$, and $\sum_{s=t_0+1}^{t}(\epsilon_s(\widehat{x}) - \epsilon_s(x_s))$. The first term is controlled by the test in Eq. (12). The second and third terms are martingale difference sequences with variance bounded by $\mathcal{O}(1-\widetilde{p}_{s,\widehat{x}})$, which as we further show is at most $1/s\widehat{\Delta}^2_{\min}$ with $\widehat{\Delta}_{\min} = \min_{x\neq\widehat{x}}\widehat{\Delta}_x$. By combining Eq. (9) and Eq. (10), it is clear that $\widehat{\Delta}_{\min} \ge 20\sqrt{f_T C_1/t_0}$ and thus the variance is in the order of t_0/s. Applying Freedman's inequality, the last two terms are thus bounded by $\widetilde{\mathcal{O}}(\sqrt{t_0})$ as well, proving the claimed result (see Appendix C for the full proof). □

Finally, to obtain Theorem 7, it suffices to apply Lemma 9 and the fact that the number of epochs is $\mathcal{O}(\log T)$.

5.3. Analysis for the Corrupted Setting (Theorem 6)

The key for this analysis is the following lemma.

Lemma 10. *In the corrupted setting,* BOTW-SE *ensures with probability at least $1-15\delta$:*

- $t_0 \le \max\left\{\frac{900 f_T C_1}{\Delta^2_{\min}}, \frac{900C^2}{f_T C_1}, L\right\}$.
- *If $C \le \frac{1}{30}\sqrt{f_T C_1 L}$, then 1) $\widehat{x} = x^*$; 2) $\widehat{\Delta}_x \in [0.7\Delta_x, 1.3\Delta_x]$ for all x; and 3) Phase 2 never ends.*

Using this lemma, we show a proof sketch of Theorem 6 for the stochastic case (i.e. $C = 0$). The full proof is deferred to Appendix C.2.

Proof sketch for Theorem 6 with $C = 0$. By Lemma 10, we know that after roughly $\Theta\left(\frac{f_T C_1}{\Delta^2_{\min}}\right)$ rounds in Phase 1, the algorithm finds $\widehat{x} = x^*$, estimates $\widehat{\Delta}_x$ up to a constant factor of Δ_x, and enters Phase 2 without ever going back to Phase 1. By Eq. (7), the regret in Phase 1 can be upper bounded by $\mathcal{O}\left(\sqrt{C_1 \cdot \frac{f_T C_1}{\Delta^2_{\min}}}\right) = \mathcal{O}\left(\frac{C_1}{\Delta_{\min}}\sqrt{f_T}\right)$.

To bound the regret in Phase 2, we show that as long as t is larger than a problem-dependent constant T^*, there exist $\{N_x\}_{x\in\mathcal{X}}$ satisfying $\sum_{x\in\mathcal{X}} N_x\Delta_x \le 2c(\mathcal{X},\theta)$ such that $\{p^*_{t,x}\}_{x\in\mathcal{X}\setminus\{x^*\}} = \left\{\frac{\beta_t N_x}{2t}\right\}_{x\in\mathcal{X}\setminus\{x^*\}}$ is a feasible solution of Eq. (4). Therefore, we can bound the regret in this regime as follows:

$$\begin{aligned}
&\sum_{s=T^*+1}^{t}\sum_{x\in\mathcal{X}} \widetilde{p}_{s,x}\Delta_x \\
&= \sum_{s=T^*+1}^{t}\sum_{x\in\mathcal{X}} \frac{1}{2}p_{s,x}\Delta_x && (x^* = \widehat{x}) \\
&\le \sum_{s=T^*+1}^{t}\sum_{x\in\mathcal{X}} \frac{1}{1.4}p_{s,x}\widehat{\Delta}_x && (\widehat{\Delta}_x \in [0.7\Delta_x, 1.3\Delta_x]) \\
&\le \sum_{s=T^*+1}^{t}\sum_{x\in\mathcal{X}} \frac{1}{1.4}p^*_{s,x}\widehat{\Delta}_x && (\text{optimality of } p_s) \\
&\le \sum_{s=T^*+1}^{t}\sum_{x\in\mathcal{X}} p^*_{s,x}\Delta_x && (\widehat{\Delta}_x \in [0.7\Delta_x, 1.3\Delta_x]) \\
&\le \mathcal{O}\left(c(\mathcal{X},\theta)\log^2 T\right). && (\text{definition of } p^*_s)
\end{aligned}$$

Combining the regret bounds in Phase 1 and Phase 2, we prove the results for the stochastic setting. □

6. Conclusion

In this work, we make significant progress on improving the robustness and adaptivity of linear bandit algorithms. Our

algorithms are the first to achieve near-optimal regret in various different settings, without having any prior knowledge on the environment. Our techniques might also be useful for more general problems such as linear contextual bandits.

In light of the work (Zimmert & Seldin, 2019) for multi-armed bandits that shows a simple Follow-the-Regularized-Leader algorithm achieves optimal regret in different settings, one interesting open question is whether there also exists such a simple Follow-the-Regularized-Leader algorithm for linear bandit with the same adaptivity to different settings. In fact, it can be shown that their algorithm has a deep connection with **OP** in the special case of multi-armed bandits, but we are unable to extend the connection to general linear bandits.

Acknowledgements

We thank Tor Lattimore and Julian Zimmert for helpful discussions. HL thanks Ilias Diakonikolas and Anastasia Voloshinov for initial discussions in this direction. The first four authors are supported by NSF Awards IIS-1755781 and IIS-1943607.

References

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

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, 1999.

Abernethy, J., Hazan, E., and Rakhlin, A. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory, COLT 2008*, pp. 263–273, 2008.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pp. 1638–1646. PMLR, 2014.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Auer, P. and Chiang, C.-K. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 116–120, 2016.

Awerbuch, B. and Kleinberg, R. D. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 45–53, 2004.

Bartlett, P., Dani, V., Hayes, T., Kakade, S., Rakhlin, A., and Tewari, A. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory-COLT 2008*, pp. 335–342. Omnipress, 2008.

Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. Stochastic linear bandits robust to adversarial attacks. *arXiv:2007.03285*, 2020.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, 2012.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, 2012.

Catoni, O. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pp. 1148–1185, 2012.

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

Combes, R., Magureanu, S., and Proutiere, A. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, 2017.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. 2008a.

Dani, V., Kakade, S. M., and Hayes, T. P. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, 2008b.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*, pp. 169, 2011.

Gerchinovitz, S. and Lattimore, T. Refined lower bounds for adversarial bandits. In *NeurIPS*, 2016.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, 2019.

Hao, B., Lattimore, T., and Szepesvari, C. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in Neural Information Processing Systems*, 2020.

Jun, K.-S. and Zhang, C. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. *Advances in Neural Information Processing Systems*, 2020.

Komiyama, J., Honda, J., and Nakagawa, H. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. *Advances in Neural Information Processing Systems*, 2015.

Lattimore, T. and Szepesvari, C. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pp. 728–737. PMLR, 2017.

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in neural information processing systems*, 2020.

Li, Y., Lou, E. Y., and Shan, L. Stochastic linear optimization with adversarial corruption. *arXiv:1909.02109*, 2019.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 2018.

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. Corruption robust exploration in episodic reinforcement learning. *arXiv:1911.08689*, 2019.

Mond, B. and Pecaric, J. A mixed arithmetic-mean-harmonic-mean matrix inequality. *Linear algebra and its applications*, 237:449–454, 1996.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

Seldin, Y. and Lugosi, G. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pp. 1743–1759. PMLR, 2017.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pp. 1287–1295. PMLR, 2014.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Tirinzoni, A., Pirotta, M., Restelli, M., and Lazaric, A. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. In *Advances in Neural Information Processing Systems*, 2020.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pp. 1263–1291. PMLR, 2018.

Wei, C.-Y., Luo, H., and Agarwal, A. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pp. 3583–3634. PMLR, 2020.

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

Zimmert, J. and Seldin, Y. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *J. Mach. Learn. Res.*, 22:28–1, 2021.

Zimmert, J., Luo, H., and Wei, C.-Y. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pp. 7683–7692. PMLR, 2019.