Best-of-Three-Worlds Linear Bandit Algorithm
with Variance-Adaptive Regret Bounds

Shinji Ito   I-SHINJI@NEC.COM

Kei Takemura   KEI_TAKEMURA@NEC.COM

NEC Corporation

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

This paper proposes a linear bandit algorithm that is adaptive to environments at two different levels of hierarchy. At the higher level, the proposed algorithm adapts to a variety of types of environments. More precisely, it achieves best-of-three-worlds regret bounds, i.e., of $O(\sqrt{T \log T})$ for adversarial environments and of $O(\frac{\log T}{\Delta_{\min}} + \sqrt{\frac{C \log T}{\Delta_{\min}}})$ for stochastic environments with adversarial corruptions, where T, $\Delta_{\min}$, and C denote, respectively, the time horizon, the minimum sub-optimality gap, and the total amount of the corruption. Note that polynomial factors in the dimensionality are omitted here. At the lower level, in each of the adversarial and stochastic regimes, the proposed algorithm adapts to certain environmental characteristics, thereby performing better. The proposed algorithm has data-dependent regret bounds that depend on all of the cumulative loss for the optimal action, the total quadratic variation, and the path-length of the loss vector sequence. In addition, for stochastic environments, the proposed algorithm has a variance-adaptive regret bound of $O(\frac{\sigma^2 \log T}{\Delta_{\min}})$ as well, where σ^2 denotes the maximum variance of the feedback loss. The proposed algorithm is based on the SCRiBLe algorithm (Abernethy et al., 2012). By incorporating into this a new technique we call *scaled-up sampling*, we obtain high-level adaptability, and by incorporating the technique of optimistic online learning, we obtain low-level adaptability.

1. Introduction

This paper considers linear bandit problems. In this class of problems, a player chooses, in each round t, an action a_t from a given *action set* $\mathcal{A}$, which is a subset of a d-dimensional linear space. The player then observes the incurred loss $f_t(a_t) \in [-1, 1]$, where the (conditional) expectation of f_t is assumed to be a linear function, i.e., f_t is expressed as $f_t(a) = \langle \ell_t, a \rangle + \varepsilon_t(a)$ with some vector $\ell_t \in \mathbb{R}^d$ and some noise ε_t. The performance of the player is evaluated in terms of *regret* R_T defined as $R_T(a^*) = \mathbf{E}\left[\sum_{t=1}^T f_t(a_t) - \sum_{t=1}^T f_t(a^*)\right]$ and $R_T = \max_{a^* \in \mathcal{A}} R_T(a^*)$.

Algorithms for linear bandit problems have been proposed mainly for two different types of environments: *stochastic* and *adversarial*. In stochastic environments, $\{f_t\}$ are assumed to follow an unknown distribution $\mathcal{D}$ independently for all t. Consequently, we may assume that there exists $\ell^* \in \mathbb{R}^d$ such that $\ell_t = \ell^*$ and $\varepsilon_t(a)$ follows an identical distribution for all t.[1] In adversarial environments, the distributions of f_t (and thus also ℓ_t) are decided arbitrarily depending on the action sequence $(a_s)_{s=1}^{t-1}$ that the player has chosen so far.

What we can do in linear bandit problems varies greatly depending on the type of environment. For stochastic environments, it is known that the optimal regret is of $\Theta(\log T)$ (Lattimore and

---

1. In standard stochastic settings, it is assumed that $\varepsilon_t(a)$ follows a zero-mean distribution for all $a \in \mathcal{A}$. Our proposed algorithm, however, works well under milder assumptions, details of which are given in Section 2.1 and Remark 12.

Szepesvari, 2017), ignoring the factor dependent on d, $\mathcal{A}$ and ℓ^*. For adversarial environments, the minimax optimal regret is $\tilde{\Theta}(d\sqrt{T})$ (Bubeck et al., 2012), where we ignore poly-logarithmic factors with respect to d and T in the notation of $\tilde{O}, \tilde{\Omega}$ and $\tilde{\Theta}$. A class of intermediate settings between these two types of environments are called stochastic environments with adversarial corruption (Lykouris et al., 2018), or corrupted stochastic environments. Environments in this regime are parametrized by corruption level C, which measures the amount of the adversarial component. For this setting, an algorithm achieving $O((\log T)^2 + C)$-regret has been proposed (Lee et al., 2021).

The aim of this paper is to make possible the construction of *adaptive* algorithms that automatically exploit certain specific characteristics of environments. In existing studies of bandit algorithms, the concept of adaptability has been considered at two different levels, regarding which we here refer to high-level adaptability and low-level adaptability. Algorithms with high-level adaptability are designed to work well for different types of environments, e.g., stochastic and adversarial types. Algorithms with low-level adaptability perform better in specific individual environments by exploiting certain favorable characteristics that they possess, e.g., small cumulative loss or small variance in loss sequences.

High-level-adaptive bandit algorithms that perform (nearly) optimally for both stochastic and adversarial environments are called best-of-both-worlds (BOBW) algorithms (Bubeck and Slivkins, 2012). Among such algorithms, those that can adapt to corrupted stochastic environments are referred to as best-of-all-worlds (Erez and Koren, 2021) or best-of-three-worlds (BOTW) algorithms (Lee et al., 2021). For linear bandit problems, Lee et al. (2021) provide a best-of-three-worlds algorithm that achieves regret bounds of $O((\log T)^2)$ for stochastic environments, of $\tilde{O}(\sqrt{T})$ for adversarial environments, and of $O((\log T)^2 + C)$ for corrupted stochastic environments.

Various types of low-level-adaptive algorithms have been considered for adversarial bandit problems. Representative examples are algorithms with $\tilde{O}(\sqrt{L^*})$-regret, where L^* represents the cumulative loss for the optimal action; such examples are said to have *first-order* regret bounds. In addition to such algorithms, Hazan and Kale (2011) proposed an algorithm with a *second-order* regret bound of $\tilde{O}(\sqrt{Q})$ that depends on the total quadratic variation Q of loss vectors. Algorithms of Bubeck et al. (2019) have regret bounds depending on *path-length* P of the loss sequence. An algorithm for combinatorial semi-bandits proposed by Ito (2021a) achieves $\tilde{O}(\sqrt{\min\{L^*, Q, P\}})$-regret, which means that the algorithm simultaneously has first-order and second-order bounds as well as a bound depending on the path-length of the loss sequence. These regret bounds, which are referred to as *data-dependent regret bounds*, imply that algorithm performance can be improved by exploiting certain environmental characteristics that are common in applications, such as small variations in loss sequences or sparsity of loss. For the stochastic multi-armed bandit problem, Audibert et al. (2007) proposed an algorithm with an $O(\sum_i (\frac{\sigma_i^2}{\Delta_i} + 1) \log T)$-regret bound that depended not only on the sub-optimality gap Δ_i but also on the variance σ_i^2 of the loss. We refer to such bounds as *variance-adaptive* bounds, and they can be considered to represent low-level adaptability in stochastic regimes.

1.1. Contribution of this work

The main contribution of this paper is the proposal of a linear bandit algorithm that combines high-level adaptability and low-level adaptability. It is a BOTW algorithm that achieves regret bounds of $O(\log T)$ in stochastic environments, $\tilde{O}(\sqrt{T})$ in adversarial environments, and $O(\log T + \sqrt{C \log T})$ in corrupted stochastic environments, ignoring factors depending on d, $\mathcal{A}$ and ℓ^*. Fur-

Table 1: List of parameters in regret bounds.

Parameter	Description
$T \in \mathbb{N}$	time horizon
$d \in \mathbb{N}$	dimensionality of action set
$\mathcal{A} \subseteq \mathbb{R}^d$	action set (One may assume $\log(|\mathcal{A}|) = O(d \log T)$ w.l.o.g.)
$\vartheta \geq 1$	parameter of a self-concordant barrier ψ used in the algorithm (One can choose ψ so that $\vartheta = O(d)$)
$\Delta_{\min} > 0$	minimum suboptimality gap: $\Delta_{\min} = \min_{a \in \mathcal{A} \backslash \{a^*\}} \langle \ell^*, a - a^* \rangle$
$c^* > 0$	asymptotic lower bound parameter: $c^* = c(\mathcal{A}, \ell^*) = O(d/\Delta_{\min})$
$\sigma^2 \geq 0$	maximum variance of loss: $\sigma^2 = \max_{a \in \mathcal{A}, t} \mathbf{E}\left[(f_t(a) - \langle \ell^*, a \rangle)^2 \right]$
$L^* \geq 0$	minimum cumulative loss: $L^* = \min_{a^* \in \mathcal{A}} \mathbf{E}\left[\sum_{t=1}^{T} f_t(a^*) \right]$
$Q \geq 0$	total quadratic variation in loss sequence: $Q = \min_{\bar{\ell} \in \mathbb{R}^d} \mathbf{E}\left[\sum_{t=1}^{T} \left\| \ell_t - \bar{\ell} \right\|_2^2 \right]$
$P \geq 0$	path-length of loss sequence: $P = \mathbf{E}\left[\sum_{t=1}^{T-1} \left\| \ell_t - \ell_{t+1} \right\|_2 \right]$

ther, the algorithm achieves first-order, second-order, and path-length bounds in adversarial environments. Simultaneously, it has variance-adaptive regret bounds for (corrupted) stochastic environments.

The proposed algorithm (Algorithm 1) follows the approach of SCRiBLe (Abernethy et al., 2008a, 2012) which stands for *Self-Concordant Regularization in Bandit Learning*. This approach uses a class of functions known as *self-concordant barriers* (Nesterov and Nemirovskii, 1994) as regularizers. Self-concordant barriers are characterized with a parameter $\vartheta \geq 1$ that can be assumed to satisfy $\vartheta = O(d)$, details of which are given in Section 2.3. The regret bounds of our algorithm can be expressed with parameters explained in Table 1, including the parameter ϑ, as follows:

Theorem 1 (informal) *In adversarial environments with $\varepsilon_t(a) = 0$, the regret for Algorithm 1 is bounded as $R_T = O\left(d \sqrt{\vartheta \min\{T, Q, P\} \log T} \right)$. Further, if $f_t(a) \geq 0$ for all $a \in \mathcal{A}$ and $t \in [T]$, we have $R_T = O\left(d \sqrt{\vartheta L^* \log T} \right)$. In stochastic environments (i.e., if $\ell_t = \ell^*$ for all t), we have $R_T = O\left(\left(\frac{d\sigma^2}{\Delta_{\min}} + 1 \right) d \vartheta \log T \right)$. In corrupted stochastic environments with the corruption level $C = \sum_{t=1}^{T} \left\| \ell_t - \ell^* \right\|_2$, we have $R_T = O\left(\left(\frac{d\sigma^2}{\Delta_{\min}} + 1 \right) d \vartheta \log T + \sqrt{C \cdot \left(\frac{\sigma^2}{\Delta_{\min}} + 1 \right) d^2 \vartheta \log T} \right)$.*

Table 2 provides a comparison of our regret bounds with those in existing studies. For stochastic settings, the tight asymptotic regret given $\mathcal{A}$ and ℓ^* can be characterized with $c^* = c(\mathcal{A}, \ell^*)$, a definition of which can be found in, e.g., the paper by Lattimore and Szepesvari (2017). They have provided an algorithm that achieves an asymptotically optimal regret bound of $R_T = c^* \log T + o(\log T)$. However, such asymptotically optimal algorithms are not necessarily optimal in environments with small variance σ^2. In the case of $c^* = \Omega\left(\left(\frac{d\sigma^2}{\Delta_{\min}} + 1 \right) d\vartheta \right)$, the proposed algorithm provides a better regret bound. We would also like to emphasize the fact that our stochastic regret bound includes only a single $\log T$ factor, while the bound by the BOTW algorithm of Lee et al. (2021) includes a $(\log T)^2$ factor. On the other hand, the results by Lee et al. (2021) are better than ours in that their regret bounds are c^*-dependent and hold with high probability while ours

Table 2: Regret bounds for stochastic and adversarial linear bandits. Bounds depending on L^* are applicable when $f_t(a) \geq 0$. Bounds depending on Q or P are applicable when $\varepsilon_t(a) = 0$.

Algorithm	Stochastic	Adversarial
Bubeck et al. (2012)		$O\left(\sqrt{dT\log(|\mathcal{A}|)}\right)$
Abernethy et al. (2008a)		$O\left(d\sqrt{\vartheta T \log T}\right)$
Ito (2021a)		$O\left(d\sqrt{\min\{T, L^*, Q, P\}}(\log T)^2\right)$
Lattimore and Szepesvari (2017)	$c^* \log T + o(\log T)$	
Lee et al. (2021)	$O\left(c^* \log(T|\mathcal{A}|)\log T\right)$	$O\left(\sqrt{dT}\log(T|\mathcal{A}|\log T)\right)$
Dann et al. (2023a, Theorem 3)	$O\left(\frac{d^2\vartheta}{\Delta_{\min}}\log T\right)$	$O\left(d\sqrt{\vartheta L^* \log T}\right)$
Dann et al. (2023a, Corollary 7)	$O\left(\frac{d^2}{\Delta_{\min}}\log T\right)$	$O\left(d\sqrt{T\log T}\right)$
Dann et al. (2023a, Corollary 12)	$O\left(\frac{d\log(|\mathcal{A}|)}{\Delta_{\min}}\log T\right)$	$O\left(\sqrt{dT\log(|\mathcal{A}|)}\right)$
[This work]	$O\left((\frac{d\sigma^2}{\Delta_{\min}}+1)d\vartheta\log T\right)$	$O\left(d\sqrt{\vartheta\min\{T, L^*, Q, P\}\log T}\right)$

are gap-dependent and hold in expectation. Note that the extra $\log T$ factor is unavoidable for a high-probability bound as mentioned in their Appendix D.

For adversarial environments, Ito (2021a) has provided an algorithm with data-dependent regret bounds that depend on L^*, Q, and P simultaneously. In this regard, our regret bounds here have an additional factor of $\sqrt{\vartheta}$ but are better in terms of the dependency w.r.t. $\log T$. Our regret bounds can be better than those with BOTW algorithm by Lee et al. (2021) if the loss sequence satisfies $\min\{L^*, Q, P\} = O\left(T\log(T|\mathcal{A}|)/\sqrt{d\vartheta\log T}\right)$.

For corrupted stochastic environments, the BOTW algorithm by Lee et al. (2021) achieves a regret bound of $O\left(\frac{d\log(T|\mathcal{A}|)\log T}{\Delta_{\min}} + C\right)$, while our bound is $O\left(\mathcal{R}^{\text{sto}} + \sqrt{\mathcal{R}^{\text{sto}}C}\right)$, where $\mathcal{R}^{\text{sto}}$ satisfies $\mathcal{R}^{\text{sto}} \leq O\left((\frac{\sigma^2}{\Delta_{\min}}+1)d^2\vartheta\log T\right)$.[2] As $\sqrt{\mathcal{R}^{\text{sto}}C} \leq \frac{1}{2}(\mathcal{R}^{\text{sto}} + C)$ follows from the AM-GM inequality, our algorithm also implies $R_T = O\left(\mathcal{R}^{\text{sto}} + C\right)$, which is superior to the bound by Lee et al. (2021) when $\sigma^2 + \Delta_{\min} = O\left(\frac{\log(T|\mathcal{A}|)}{d\vartheta}\right)$. We would like to stress here that the impact of corruption on the performance of our algorithm is only of a square-root factor in C, while algorithms in existing studies (Lee et al., 2021; Li et al., 2019; Bogunovic et al., 2021) include at least a linear factor in C. Comparison of such results w.r.t. corrupted settings, however, requires particular care, as there are differences in the details of problem settings.

Remark 2 In this paper, regret is defined in terms of loss *including* corruption, while existing studies define regret in terms of loss *without* corruption. As the difference between these two notions of regret is at most $O(C)$, our algorithm enjoys $O\left(\mathcal{R}^{\text{sto}} + C\right)$-bound even in terms of the latter definition of regret. Such a difference in models has been discussed, e.g., by Gupta et al. (2019); Ito (2021b). We also note that the definition of C of Lee et al. (2021) slightly differs from ours and is smaller, as it measures the amount of corruption of the loss of each action, not the loss vector itself.

---

2. This comparison requires special care because of subtle differences in the models as discussed in Remark 2.

Remark 3 Concurrently with this paper, two papers working on BOTW algorithms for linear bandits were published and will be presented at COLT 2023. One is by Dann et al. (2023b), which provides three different results on linear bandits as shown in Table 2 and also includes a general reduction approach for BOTW algorithms for sequential decision-making problems. The other is by Kong et al. (2023), which achieves $O\left(\frac{d\log(|\mathcal{A}|T)\log T}{\Delta_{\min}}\right)$-regret for stochastic environments and $O\left(\sqrt{dT\log(|\mathcal{A}|T)\log T}\right)$-regret for adversarial environments. Among these, the result of Dann et al. (2023b, Theorem 3) is very similar to ours in that they take the approach of extending SCRiBLe.

The main innovations for achieving high-level adaptability (i.e., the BOTW property) are with regard to the sampling method for actions. In a previous study by Abernethy et al. (2008b), they compute a point x_t in the convex hull $\text{conv}(\mathcal{A})$ of the action set $\mathcal{A}$ using a follow-the-regularized-leader (FTRL) approach, and they then pick a_t from the *Dikin ellipsoid* $W_1(x_t) \subseteq \text{conv}(\mathcal{A})$ that is defined from the self-concordant barrier for $\text{conv}(\mathcal{A})$. Here, the action a_t must be sampled so that its expectation matches x_t. In addition, the larger the variance of a_t, the better estimator $\hat{\ell}_t$ for ℓ_t that we can construct, i.e., the smaller variance of $\hat{\ell}_t$. In this paper, in order to improve the variance of the loss estimator, we introduce a new technique that we refer to as *scaled-up sampling* (see Figure 1). In this approach, we construct a scaled-up set $W' \subseteq \text{conv}(\mathcal{A})$ of the Dikin ellipsoid $W_1(x_t)$ with a reference point $z_t \in \mathcal{A}$, for which we let $\alpha_t \geq 1$ denote the scaling factor. Rather than sampling from $W_1(x_t)$ as is done in the previous study, we pick a_t from W' with probability α_t^{-1}, and otherwise set $a_t = z_t$ (the expectation of a_t then matches x_t as well). Consequently, the variance of a_t becomes α_t times larger and the variance of the loss-vector estimator becomes α_t^{-1} times smaller, which contributes to the improvement of the regret upper bound. In stochastic environments in particular, intuitively, x_t approaches an extreme point (a truly optimal solution), allowing for a smaller $W_1(x_t)$ and a larger value of α_t, which leads to a significant improvement in regret. We note that a result of the concurrent work by Dann et al. (2023b, Theorem 3), referred to as Variance-Reduced SCRiBLe (VR-SCRiBLe), is also based on the framework of SCRiBLe (Abernethy et al., 2008b), in which the variance of the loss estimator is reduced by a different sampling scheme.

In proving the high-level adaptability, we use the self-bounding technique (Zimmert and Seldin, 2021; Wei and Luo, 2018). We first show that the proposed algorithm, an FTRL method with scaled-up sampling and an adaptive learning rate, achieves a regret bound of $R_T = O\left(d\sqrt{\vartheta\sum_{t=1}^T \alpha_t^{-1}\log T}\right)$. We further show that $\alpha_t^{-1} = O(\Delta(x_t)/\Delta_{\min})$ holds in any stochastic environment, where $\Delta(x_t)$ denotes the round-wise regret caused by choosing x_t. Combining these two facts, we can obtain $R_T = O\left(d\sqrt{\vartheta\sum_{t=1}^T \Delta(x_t)\Delta_{\min}^{-1}\log T}\right) = O\left(d\sqrt{\vartheta R_T\Delta_{\min}^{-1}\log T}\right)$, which immediately leads to $R_T = O(d^2\vartheta\Delta_{\min}^{-1}\log T)$ in stochastic environments. As has been done in previous analyses using the self-bounding technique, we can prove improved regret bounds for *the stochastically constrained adversarial regime* (Zimmert and Seldin, 2021) as well, which includes corrupted stochastic environments.

To achieve low-level adaptability (i.e., data-dependent bounds in adversarial environments and variance-adaptive bounds in stochastic environments), we employ the framework of *optimistic online learning* (Rakhlin and Sridharan, 2013). This framework incorporates *optimistic prediction* m_t for ℓ_t into online learning algorithms, thereby providing regret bounds depending on $(\langle \ell_t - m_t, a_t\rangle)^2$
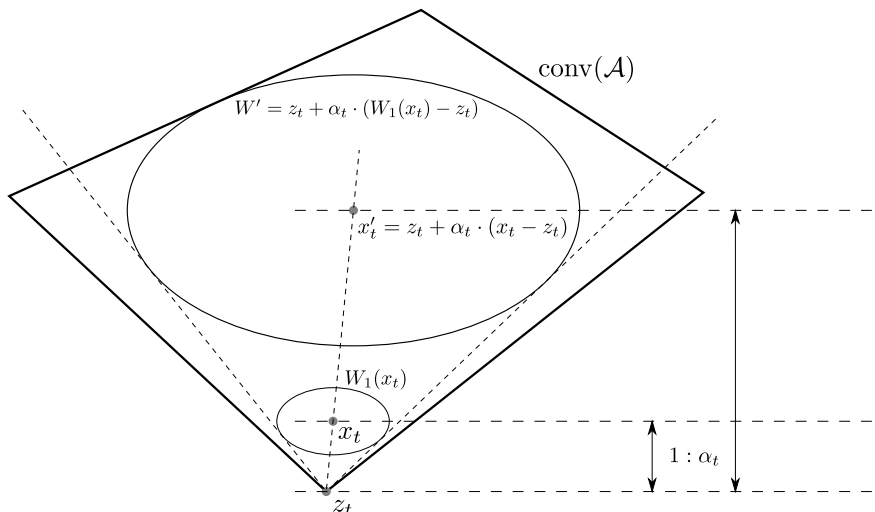
5

Figure 1: Illustration of scaled-up sampling.

rather than $(\langle \ell_t, a_t \rangle)^2$. The proposed algorithm determines m_t by means of the technique of *tracking the best linear predictor*, which leads the hybrid data-dependent bounds and variance-adaptive bounds. Similar approaches can be found in (Bubeck et al., 2019; Ito, 2021a; Ito et al., 2022). This approach of optimistic online learning can also be combined with the VR-SCRiBLe (Dann et al., 2023b, Theorem 3) to achieve the same regret guarantees as ours.

1.2. Limitation of this work and future work

We should note the issue of computational complexity w.r.t. the proposed algorithm. In the proof of $O(\log T)$-regret bounds for (corrupted) stochastic environments, we need the assumption that the reference point z_t, illustrated in Figure 1, is chosen so that the scaling factor α_t is (approximately) maximized. We have not, however, found an efficient method for computing such a point z_t. A naive method for computing such a z_t requires a computational time of at least $\Omega(|\mathcal{A}|)$, which is highly expensive, e.g., as in most examples of combinatorial bandits (Cesa-Bianchi and Lugosi, 2012). This issue can be resolved by the sampling scheme used in RV-SCRiBLe (Dann et al., 2023b, Theorem 3).

There is still some room for improvement in terms of regret bounds as well. As can be seen from Example 4 by Lattimore and Szepesvari (2017), the gap between c^* and $d/\Delta_{\min}$ can be arbitrarily large, which implies that our stochastic regret bound is much larger than the lower bound in the worst case. We also note that our regret bounds only hold in expectation while regret guarantees by Lee et al. (2021) hold with high probability. If we pursue high probability bounds as well, we cannot avoid an extra $O(\log T)$ factor, as discussed in their Appendix D.

Another issue is that our analysis relies on a uniqueness assumption, i.e., we need to assume that the optimal action is unique in the proof of regret bounds for stochastic environments. Such an assumption is common in regret analyses based on the self-bounding technique (Zimmert and Seldin, 2021; Zimmert et al., 2019; Wei and Luo, 2018). In addressing this issue, the analysis methods developed by Jin et al. (2023); Ito (2021c) may be useful.

2. Preliminary

2.1. Problem setup

This section introduces the setup of the linear bandit problems dealt with in this paper. Before a game starts, the player is given the *time horizon T* and an *action set $\mathcal{A} \subseteq \mathbb{R}^d$*, a closed and bounded set of d-dimensional vectors. Without loss of generality, we assume that $\mathcal{A}$ is not included in any proper affine subspace of $\mathbb{R}^d$. We also assume that all points in $\mathcal{A}$ have L_2 norm of at most 1, i.e., $\mathcal{A} \subseteq B_2^d(1)$, where $B_2^d(r)$ denotes an L_2 ball of the radius r: $B_2^d(r) = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq r\}$. In each round $t = 1, 2, \ldots, T$, the environment determines a *loss function $f_t : \mathcal{A} \to [-1, 1]$*, and the player then chooses an *action $a_t \in \mathcal{A}$* without knowing f_t. After that, the player observes the incurred loss $f_t(a_t)$. The loss function f_t can be chosen depending on the actions $(a_s)_{s=1}^{t-1}$ selected so far. We assume that the conditional expectation of f_t given $(a_s)_{s=1}^{t-1}$ is an affine function, i.e., there exists $\ell_t \in \mathbb{R}^d$, which is referred to as a *loss vector*, and $\xi_t \in \mathbb{R}$ such that f_t is expressed as

$$f_t(a) = \langle \ell_t, a \rangle + \varepsilon_t(a), \quad \text{where } \mathbf{E}\left[\varepsilon_t(a) | (a_s)_{s=1}^{t-1}\right] = \xi_t \text{ for all } a \in \mathcal{A}. \tag{1}$$

This paper also assumes that $\ell_t \in B_2^d(1)$. By imposing further conditions on f_t, we can express a variety of regimes, as are discussed below:

Stochastic regime In a *stochastic regime*, it is assumed that f_t follows an unknown distribution $\mathcal{D}$ for all $t \in [T]$ independently. This assumption implies that ℓ_t and ξ_t do not change over all rounds, i.e., there exists a *true loss vector $\ell^* \in \mathbb{R}^d$* and ξ^* such that $\ell_t = \ell^*$ and $\xi_t = \xi^*$ hold for all $t \in [T]$. Note here that standard stochastic settings also assume that functions of ε_t represent *zero-mean* noise, i.e., $\xi^* = 0$. This assumption is, however, not necessary in the algorithm proposed in this paper. Moreover, the proposed algorithm does not even require the assumption that ε_t follows an identical distribution, details of which will be discussed in Section 4.1.

Adversarial regime In the *adversarial regime*, by way of contrast to the stochastic regime, $(\ell_t)_{t=1}^T$ is an arbitrary sequence. More precisely, ℓ_t can be chosen in an adversarial way depending on $(a_s)_{s=1}^{t-1}$. Though adversarial environments considered in previous studies are often free from noise, i.e., $\varepsilon_t(a) = 0$ is assumed, most algorithms work well as long as the noise follows bounded zero-mean distributions. The proposed algorithm in this paper does not require this assumption as well.

Stochastic regime with adversarial corruption The *stochastic regime with adversarial corruption* is an intermediate regime between stochastic and adversarial regimes. It is parametrized by a true loss vector $\ell^* \in B_2^d(1)$ and by a *corruption level $C \geq 0$*. In this regime, the sequence of $(\ell_t)_{t=1}^T$ is subject to the constraint that $\sum_{t=1}^T \|\ell_t - \ell^*\|_2 \leq C$. This can be interpreted as a situation in which an adversary adds a corruption of $c_t = \ell_t - \ell^*$ to the loss function defined by ℓ^* and the magnitude of c_t sums up to C at most, i.e., $\sum_{t=1}^T \|c_t\|_2 \leq C$. If we set the condition level C to zero, this regime coincides with the stochastic regime. On the other hand, if $C = \Omega(T)$, then the regime is adversarial as there are no constraints on ℓ_t except for $\|\ell_t\|_2 \leq 1$.

2.2. Follow the regularized leader

In the proposed algorithm, we use the framework of (optimistic) *follow-the-regularized-leader (FTRL)* methods. In this framework, we choose a point x_t in a closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$ by

solving the following optimization problem:

$$x_t \in \underset{x \in \mathcal{X}}{\arg\min} \left\{ \left\langle m_t + \sum_{s=1}^{t-1} \hat{\ell}_s, x \right\rangle + \psi_t(x) \right\}, \tag{2}$$

where $\hat{\ell}_s$ is the (estimated) loss vector, $m_t \in \mathbb{R}^d$ is an *optimistic prediction*, and $\psi_t(x)$ is a *regularization term*, which is a differentiable convex function over $\mathcal{X}$. Note that the original FTRL framework here does not employ optimistic prediction, i.e., the value of m_t is fixed to 0. The technique of optimistic prediction m_t has been introduced to further improve the performance of FTRL, e.g., by Rakhlin and Sridharan (2013).

In the analysis of FTRL, we use the *Bregman divergence D_ψ* associated with some differentiable convex function ψ defined as follows:

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle, \tag{3}$$

where $\nabla\psi(y)$ denotes the gradient of ψ at y. We can easily see that $D_\psi(x, y) \geq 0$ for any x and y, which follows from the convexity of ψ. The following lemma provides an upper bound of the regret for FTRL:

Lemma 4 *We assume that $\psi_1(x) \geq 0$ and $\psi_{t+1}(x) \geq \psi_t(x)$ hold for all x and t. If x_t is given by* (2), *it holds for any $x^* \in \mathrm{int}(\mathcal{X})$ that*

$$\sum_{t=1}^{T} \left\langle \hat{\ell}_t, x_t - x^* \right\rangle \leq \sum_{t=1}^{T} \left(\left\langle \hat{\ell}_t - m_t, x_t - \tilde{x}_{t+1} \right\rangle - D_{\psi_t}(\tilde{x}_{t+1}, x_t) \right) + \psi_{T+1}(x^*), \tag{4}$$

where $\tilde{x}_t$ is defined by $\tilde{x}_t \in \arg\min_{x \in \mathcal{X}} \left\{ \left\langle \sum_{s=1}^{t-1} \hat{\ell}_s, x \right\rangle + \psi_t(x) \right\}$.

This lemma can be shown via a standard analysis for FTRL, e.g., as in Chapter 28 of Lattimore and Szepesvári (2019). We can also refer to, e.g., the proof of Lemma 1 by Ito et al. (2022).

2.3. Self-concordant barriers

In our proposed algorithm, we use *self-concordant barriers* to define regularization terms, just as Abernethy et al. (2008b) did. Self-concordant barriers are defined as follows:

Definition 5 A convex function $\psi : \mathrm{int}(\mathcal{X}) \to \mathbb{R}$ of class C^3 is called a *self-concordant function* if (i) $|D^3\psi(x)[h, h, h]| \leq 2(D^2\psi(x)[h, h])^{3/2}$ holds for any $x \in \mathrm{int}(\mathcal{A})$ and $h \in \mathbb{R}^d$, and (ii) $\psi(x_i)$ tends to infinity along every sequence $x_1, x_2, \ldots \in \mathrm{int}(\mathcal{X})$ converging to a boundary point of $\mathrm{int}(\mathcal{X})$, where $D^k\psi(x)[h_1, \ldots, h_k]$ denotes the value of the k-th differential of ψ at x along the directions $h_1, \ldots, h_k$. Let $\vartheta \geq 0$ be a non-negative real number. A self-concordant function $\psi : \mathrm{int}(\mathcal{X}) \to \mathbb{R}$ is called a *ϑ-self-concordant barrier* for $\mathcal{X}$ if $|D\psi(x)[h]| \leq \vartheta^{1/2}(D^2\psi(x)[h, h])^{1/2}$ holds for any $x \in \mathrm{int}(\mathcal{X})$ and $h \in \mathbb{R}^d$.

Remark 6 For any convex set $\mathcal{X} \subseteq \mathbb{R}^d$, there exists a d-self-concordant barrier for $\mathcal{X}$ (Lee and Yue, 2021). This barrier is, however, not always efficiently computable. On the other hand, for any d-dimensional polytope, we can compute an ϑ-self-concordant barrier with $\vartheta = O(d)$ in polynomial time (Lee and Sidford, 2014, 2019).

Given a self-concordant barrier $\psi : \text{int}(\mathcal{X}) \to \mathbb{R}$, for any $x \in \text{int}(\mathcal{X})$ and $h \in \mathbb{R}^d$, we assume that $\nabla^2 \psi(x)$ has full rank. Denote

$$\|h\|_{x,\psi} = \sqrt{h^\top \nabla^2 \psi(x) h}, \quad \|h\|_{x,\psi}^* = \sqrt{h^\top (\nabla^2 \psi(x))^{-1} h} \tag{5}$$

and define the *Dikin's ellipsoid* $W_r(x) \subseteq \mathbb{R}^d$ of ψ centered at x of the radius $r > 0$ as follows:

$$W_r(x) = \left\{ y \in \mathbb{R}^d \mid \|y - x\|_{x,\psi} \leq r \right\}. \tag{6}$$

The three lemmas below are used in the design and analysis of our proposed algorithm.

Lemma 7 (Theorem 2.1.2 by Nesterov and Nemirovskii (1994)) *If ψ is a self-concordant barrier for a closed convex set $\mathcal{X}$, every Dikin's ellipsoid of ψ of radius 1 is contained in $\mathcal{X}$, i.e., $W_1(x) \subseteq \mathcal{X}$ holds for any $x \in \text{int}(\mathcal{X})$.*

Let $\pi_{z,\mathcal{X}}(x)$ denote the Minkowsky function of $\mathcal{X}$ whose pole is at z:

$$\pi_{z,\mathcal{X}}(x) = \inf \left\{ r > 0 \mid z + r^{-1}(x - z) \in \mathcal{X} \right\}. \tag{7}$$

We have an upper bound on ψ expressed with this Minkowsky function, as follows:

Lemma 8 (Propositoin 2.3.2 by Nesterov and Nemirovskii (1994)) *If ψ is a ϑ-self-concordant barrier for $\mathcal{X}$, it holds for any x and y in $\text{int}(\mathcal{X})$ that $\psi(x) \leq \psi(y) + \vartheta \log \frac{1}{1 - \pi_{y,\mathcal{X}}(x)}$.*

If we use a self-concordant barrier ψ, we can use the following lemma to bound the *stability term* $\left(\left\langle \hat{\ell}_t - m_t, x_t - x'_{t+1} \right\rangle - D_{\psi_t}(x'_{t+1}, x_t) \right)$ in Lemma 4.

Lemma 9 *Let ψ be a self-concordant function on $\mathcal{X}$ and $x, y \in \text{int}(\mathcal{X})$. Let $\beta > 0$ and $\ell \in \mathbb{R}^d$. Suppose that $\|\ell\|_{x,\psi}^* \leq \beta/3$. We then have $\langle \ell, x - y \rangle - \beta D_\psi(y, x) \leq \frac{2}{\beta} \|\ell\|_{x,\psi}^{*2}$.*

3. Algorithm

Let $\mathcal{X}$ be the convex hull of $\mathcal{A}$ and ψ be a ϑ-self-concordant barrier for $\mathcal{X}$. In the proposed algorithm, we compute x_t by solving the optimization problem (2) with $\psi_t(x) = \beta_t \psi(x)$, where β_t is a *learning rate parameter* satisfying $6d \leq \beta_1 \leq \beta_2 \leq \cdots$. The manner of computing a_t, $\hat{\ell}_t$, m_t, and β_t will be presented below.

Action a_t and unbiased estimator $\hat{\ell}_t$ for loss vector   After computing x_t, we choose the action $a_t \in \mathcal{A}$ so that $\mathbf{E}[a_t | x_t] = x_t$. Let $\{e_1, \ldots, e_d\}$ and $\{\lambda_1, \ldots, \lambda_d\}$ be the set of eigenvectors and eigenvalues of $\nabla^2 \psi(x_t)$. Define $\mathcal{E}_t := \{x_t + \lambda_i^{-1/2} e_i \mid i \in [d]\} \cup \{x_t - \lambda_i^{-1/2} e_i \mid i \in [d]\}$. Note that here $\mathcal{E}_t \subseteq \mathcal{X}$ holds since $\mathcal{E}_t \subseteq W_1(x_t)$ follows from the definition of $\mathcal{E}_t$ and since $W_1(x) \subseteq \mathcal{X}$ follows from Lemma 7. In the algorithm by Abernethy et al. (2008b), the action a_t is chosen from $\mathcal{E}_t$ uniformly at random. Unlike this existing method, our proposed algorithm chooses an action from a set $\mathcal{E}'_t$ scaled up from $\mathcal{E}_t$ with a reference point $z_t \in \mathcal{A}$, or chooses $a_t = z_t$ with some probability. More precisely, after computing $\mathcal{E}_t$ and choosing a point $z_t \in \mathcal{A}$, we set $\mathcal{E}'_t$ by

$$\mathcal{E}'_t = \{ z_t + \alpha_t (x - z_t) \mid x \in \mathcal{E}_t \}, \tag{8}$$

where $\alpha_t \geq 1$ is defined as the largest real number such that $\mathcal{E}'_t$ is included in $\mathcal{X}$. How to choose z_t is discussed in the next pragraph. If we denote $r_t = \alpha_t^{-1} \in (0, 1]$, we can express r_t as follows:

$$r_t = \inf \left\{ r > 0 \mid z_t + r^{-1}(x - z_t) \in \mathcal{X} \quad (x \in \mathcal{E}_t) \right\} = \max_{x \in \mathcal{E}_t} \pi_{z_t, \mathcal{X}}(x), \tag{9}$$

where π is the Minkowsky function defined by (7). We choose $z_t \in \mathcal{A}$ so that the value of r_t is as small as possible. Let x'_t denote the center of $\mathcal{E}'_t$, i.e., define $x'_t = z_t + r_t^{-1}(x_t - z_t)$. We then set $b_t = 1$ with probability r_t and $b_t = 0$ with probability $1 - r_t$. If $b_t = 0$, we choose $a'_t = z_t$. If $b_t = 1$, we choose a'_t from $\mathcal{E}'_t$ uniformly at random. In other words, we pick i_t uniformly at random from $[d]$ and $v_t = \pm 1$ with probability $1/2$, and set $a'_t = z_t + r_t^{-1}(x_t + v_t \lambda_{i_t}^{-1/2} e_{i_t} - z_t)$. We then output $a_t \in \mathcal{A}$ so that its expectation coincides with $a'_t \in \mathcal{X} = \text{conv}(\mathcal{A})$. After obtaining feedback of $f_t(a_t)$, we define $\hat{\ell}_t$ by

$$\hat{\ell}_t = m_t + d b_t v_t \lambda_{i_t}^{1/2}(f_t(a_t) - \langle m_t, a_t \rangle) e_{i_t}. \tag{10}$$

We can show that the conditional expectation of a_t is equal to x_t and that $\hat{\ell}_t$ is an unbiased estimator of ℓ_t, i.e., we have $\mathbf{E}[a_t | x_t] = x_t$ and $\mathbf{E}[\hat{\ell}_t | x_t] = \ell_t$, proofs of which are given in Section D in the appendix. We note that, thanks to the scaled-up sampling, the mean square of $\hat{\ell}_t - m_t$ is improved by a factor of $1/\alpha_t$, which plays a central role in our proof of BOTW regret bounds.

Reference point z_t We will see that the smaller value of r_t is, the smaller variance of $\hat{\ell}_t - m_t$ is, resulting in an improvement in regret. To take maximum advantage of this effect, we choose z_t so that r_t is as small as possible. More precisely, for a constant $\kappa \geq 1$, we assume that z_t satisfies

$$r_t = \max_{x \in \mathcal{E}_t} \pi_{z_t, \mathcal{X}}(x) \leq \kappa \cdot \min_{z \in \mathcal{A}} \max_{x \in \mathcal{E}_t} \pi_{z, \mathcal{X}}(x) \tag{11}$$

for all $t \in [T]$. This assumption is used in our proof of $O(\log T)$-regret in stochastic environments.

Learning rate parameter β_t In the regret analysis in Section 4, we will show that the regret for the proposed algorithm is bounded as $R_T = O\left(\mathbf{E}\left[d^2 \sum_{t=1}^{T} \frac{g_t(m_t)}{\beta_t} + \beta_{T+1} \vartheta \log T\right]\right)$, where $g_t(m)$ is defined as

$$g_t(m) = b_t \cdot (\langle a_t, m \rangle - f_t(a_t))^2. \tag{12}$$

Intuitively, $g_t(m_t)/\beta_t$ comes from the part of $\left\langle \hat{\ell}_t - m_t, x_t - \tilde{x}_{t+1} \right\rangle - D_{\psi_t}(\tilde{x}_{t+1}, x_t)$ in (4), which is called *stability terms*, and $\beta_{T+1} \log T$ comes from the part of $\psi_{T+1}(x^*)$, called *penalty terms*. To balance stability and penalty terms, we set β_t by

$$\beta_t = 6d + 2d\sqrt{\frac{\sum_{s=1}^{t-1} g_s(m_s)}{\vartheta \log T}}, \tag{13}$$

which leads to $R_T = O\left(d\,\mathbf{E}\left[\sqrt{\vartheta \log T \cdot \sum_{t=1}^{T} g_t(m_t)}\right] + d \log T\right)$.

Optimistic prediction m_t To minimize the part of $\sum_{t=1}^{T} g_t(m_t)$, we choose m_t by using online projected gradient descent for g_t. We set $m_1 = 0$ and update m_t as follows:

$$m'_{t+1} = m_t - \eta b_t \cdot (\langle a_t, m_t \rangle - f_t(a_t)) a_t, \quad m_{t+1} = \min\left\{1, \frac{1}{\|m'_{t+1}\|_2}\right\} m'_{t+1}, \tag{14}$$

where $\eta \in (0, 1/4)$ is the learning rate parameter for updating m_t.

The proposed algorithm can be summarized as Algorithm 1 in Section B in the appendix.

Computational complexity   The procedure in each round can be performed in polynomial time in d, except for the computation of z_t. Indeed, given a self-concordant barrier for $\mathcal{X}$, we can solve an arbitrary linear optimization problem over $\mathcal{X}$ (and thus also over $\mathcal{A}$), with the aid of, e.g., interior point methods (Nesterov and Nemirovskii, 1994). This implies that convex optimization problems (2) can be solved in polynomial time as well. Further, for any $a'_t \in \mathcal{X}$, we can find an expression of convex combination of points in $\mathcal{A}$ in polynomial time if $\mathcal{A}$ is a finite set (Mirrokni et al., 2017), (Schrijver, 1998, Corollary 14.1g), which means that we can randomly choose $a_t \in \mathcal{A}$ so that $\mathbf{E}[a_t|a'_t] = a'_t$. As for the calculation of z_t satisfying (11), it is not clear if there is a computationally efficient way at this point. Because we can compute the value of $\max_{x \in \mathcal{E}_t} \pi_{z,\mathcal{X}}(x)$ for any $z \in \mathcal{A}$ in polynomial time in d, we can find z_t minimizing this value in $O(\mathrm{poly}(d)|\mathcal{A}|)$ time, which can be exponential in d. We also note that we can compute z_t when the action set has some kind of symmetry. For example, if the action set is given as $\mathcal{A} = \{-1,1\}^d$ and ϕ is defined as $\phi(x) = -\sum_{i=1}^d \ln(1 - x_i^2)$, a point in $\mathcal{A}$ closest to x_t is an optimal solution to the optimization problem corresponding to the right-hand side of (11).

4. Analysis

4.1. Regret bounds for the proposed algorithm

Theorem 10 (Regret bounds in the adversarial regime)   *Let L^*, Q and P be parameters defined as in Table 1. The regret for Algorithm 1 is bounded as*

$$R_T = O\left(d\sqrt{\vartheta \log T \left(\min\{Q, P\} + \mathbf{E}\left[\sum_{t=1}^T (\varepsilon_t(a_t))^2\right]\right)} + d\vartheta \log T\right). \tag{15}$$

Further, if $f_t(a) \geq 0$ for any $a \in \mathcal{A}$ and $t \in [T]$, we have

$$R_T = O\left(d\sqrt{\vartheta L^* \log T} + d\vartheta \log T\right). \tag{16}$$

Note that the regret bounds in Theorem 10 are valid regardless of the choice of z_t. In fact, we can demonstrate these regret bounds even if we sample a_t from $\mathcal{E}_t$, as is similarly done with the algorithm by Abernethy et al. (2008b), which corresponds to $r_t = 1$. By way of contrast, to show $O(\log T)$-regret bounds for stochastic environments, we need the assumption of (11). Under this assumption, we have the following regret bounds:

Theorem 11 (Regret bounds in the corrupted stochastic regime)   *Let $\ell^* \in \mathbb{R}^d$ and denote $C = \sum_{t=1}^T \|\ell_t - \ell^*\|_2$. Define $a^* \in \arg\min_{a \in \mathcal{A}} \langle \ell^*, a \rangle$ and $\Delta_{\min} = \min_{a \in \mathcal{A} \setminus \{a^*\}} \langle \ell^*, a - a^* \rangle$. We have $R_T = O\left(d\sqrt{\left(C + \sum_{t=1}^T \sigma_t^2\right) \vartheta \log T} + d\vartheta \log T\right)$, where we define $\sigma_t^2 = \max_{a \in \mathcal{A}} \mathbf{E}[(\varepsilon_t(a))^2]$. Further, if $a^* \in \arg\min_{a \in \mathcal{A}} \langle \ell^*, a \rangle$ exists uniquely, under the assumption of (11), we have*

$$R_T(a^*) = O\left(\left(\frac{\kappa d\sigma^2}{\Delta_{\min}} + 1\right) d\vartheta \log T + \sqrt{\left(\frac{\kappa\sigma^2}{\Delta_{\min}} + 1\right) Cd^2\vartheta \log T}\right), \tag{17}$$

where $\sigma^2 = \max_{t \in [T]} \sigma_t^2$.

Remark 12 In standard settings of the stochastic regime, it is assumed that f_t follows an identical distribution for different rounds and $\mathbf{E}[\varepsilon_t(a)|(a_s)_{s=1}^{t-1}] = 0$ for all a. Such assumptions are not, however, needed in Theorem 11. In other words, even when $\xi_t = \mathbf{E}[\varepsilon_t(a)|(a_s)_{s=1}^{t-1}]$ is non-zero and changes depending on t, we still have the $O(\log T)$-regret bounds given in Theorem 11.

4.2. Proof sketch

Regret bounds in Theorems 10 and 11 are derived from the following lemma:

Lemma 13 *The regret for Algorithm 1 is bounded as follows:*

$$R_T = O\left(d \cdot \mathbf{E}\left[\sqrt{\vartheta \log T \sum_{t=1}^{T} g_t(m_t)}\right] + d\vartheta \log T\right). \qquad (18)$$

In proving this lemma, we use Lemmas 4, 8 and 9. From Lemma 9, the stability term $\left\langle \hat{\ell}_t - m_t, x_t - \tilde{x}_{t+1} \right\rangle - D_{\psi_t}(\tilde{x}_{t+1}, x_t)$ in Lemma 4 is bounded by $\frac{2}{\beta_t}\|\hat{\ell}_t - m_t\|_{x,\psi}^{*2} = \frac{2}{\beta_t}d^2 g_t(m_t)$. From Lemma 8, we can bound the penalty term $\psi_{T+1}(x^*)$ in Lemma 4 as $\psi_{T+1}(x^*) \leq \beta_{T+1}\vartheta \log T$. Combining these bounds, we obtain $R_T = O\left(\mathbf{E}\left[d^2 \sum_{t=1}^{T} \frac{g_t(m_t)}{\beta_t} + \beta_{T+1}\vartheta \log T\right]\right)$. From this and the definition of β_t given in (13), we have the regret bound in Lemma 13. A complete proof of this lemma is given in Section D in the appendix.

From the result of *tracking linear experts* (Herbster and Warmuth, 2001), we obtain the following upper bound on $\sum_{t=1}^{T} g_t(m_t)$.

Lemma 14 *If m_t is given by (14), it holds for any sequence $(u_t)_{t=1}^{T+1} \in (B_2^d(1))^{T+1}$ that*

$$\sum_{t=1}^{T} g_t(m_t) \leq \frac{1}{1-2\eta}\sum_{t=1}^{T} g_t(u_t) + \frac{1}{\eta(1-2\eta)}\left(\sqrt{2}\sum_{t=1}^{T}\|u_{t+1} - u_t\|_2 + \frac{1}{2}\|u_{T+1}\|_2^2\right). \qquad (19)$$

This lemma is a special case of Theorem 11.4 by Cesa-Bianchi and Lugosi (2006).

Proof sketch of Theorem 10 By substituting $u_t = \bar{\ell} \in \arg\min_\ell \sum_{t=1}^{T}\|\ell_t - \ell\|_2^2$ for all t in (19), we obtain $\mathbf{E}\left[\sum_{t=1}^{T} g_t(m_t)\right] = O\left(Q + \mathbf{E}\left[\sum_{t=1}^{T}(\varepsilon_t(a_t))^2\right] + 1\right)$. Similarly, by substituting $u_t = \ell_t$ for all t in (19), we obtain $\mathbf{E}\left[\sum_{t=1}^{T} g_t(m_t)\right] = O\left(P + \mathbf{E}\left[\sum_{t=1}^{T}(\varepsilon_t(a_t))^2\right] + 1\right)$. Combining these with Lemma 13, we obtain (15) in Theorem 10. Further, if $f_t(a) \geq 0$, by substituting $u_t = 0$, we obtain $\mathbf{E}\left[\sum_{t=1}^{T} g_t(m_t)\right] = O\left(L^* + R_T + 1\right)$, which leads to a regret bound of $R_T = O\left(d\sqrt{\vartheta \log T\left(L^* + R_T\right)} + d\vartheta \log T\right)$. This implies that (16) in Theorem 10 holds.

Proof sketch of Theorem 11 By setting $u_t = \ell^*$ for all t in (19), we obtain $\mathbf{E}\left[\sum_{t=1}^{T} g_t(m_t)\right] = O\left(\mathbf{E}\left[C + \sum_{t=1}^{T} r_t\sigma_t^2 + 1\right]\right)$. As we have $r_t \leq 1$, from this bound and Lemma 13, we have $R_T = O\left(d\sqrt{(C + \sum_{t=1}^{T}\sigma_t^2)\vartheta \log T} + d\vartheta \log T\right)$. We also have the following regret bound:

$$R_T = O\left(d\sqrt{\left(C + \sigma^2 \mathbf{E}\left[\sum_{t=1}^{T} r_t\right]\right)\vartheta \log T} + d\vartheta \log T\right). \qquad (20)$$

From the assumption of (11), r_t is bounded as

$$r_t \leq \kappa \cdot \min_{z \in \mathcal{A}} \left\{ \max_{x \in \mathcal{E}_t} \pi_{z, \mathcal{X}}(x) \right\} \leq \kappa \cdot \min_{z \in \mathcal{A}} \left\{ \max_{x \in W_1(x_t)} \pi_{z, \mathcal{X}}(x) \right\}, \tag{21}$$

where the second inequality follows from $\mathcal{E}_t \subseteq W_1(x_t)$. The following lemma provides an upper bound on the right-hand side of this:

Lemma 15 *Suppose $a^* \in \arg\min_{a \in \mathcal{A}} \langle \ell^*, a \rangle$ uniquely exists. It holds for any $y \in \mathrm{int}(\mathcal{X})$ that*

$$\max_{x \in W_1(y)} \pi_{a^*, \mathcal{X}}(x) \leq 2 \frac{\Delta(y)}{\Delta_{\min}}, \quad \text{where} \quad \Delta(y) = \langle \ell^*, y - a^* \rangle, \quad \Delta_{\min} = \min_{a \in \mathcal{A} \setminus \{a^*\}} \Delta(a). \tag{22}$$

By combining this lemma with (20) and (21), we obtain a bound depending on $\sum_{t=1}^{T} \Delta(x_t)$ as follows: $R_T = O\left(d\sqrt{\left(C + \frac{\kappa \sigma^2}{\Delta_{\min}} \mathbf{E}\left[\sum_{t=1}^{T} \Delta(x_t) \right] \right) \vartheta \log T} + d\vartheta \log T \right)$. On the other hand, regret is bounded from below as $R_T(a^*) \geq \mathbf{E}\left[\sum_{t=1}^{T} \Delta(x_t) \right] - 2C$. By combining these two bounds on R_T, we obtain

$$R_T(a^*) = O\left(d\sqrt{\vartheta \log T \cdot \left(C + \frac{\kappa \sigma^2}{\Delta_{\min}} (R_T + C) \right)} + d\vartheta \log T \right)$$

$$= O\left(d\sqrt{\frac{\vartheta \kappa \sigma^2 \log T}{\Delta_{\min}} R_T(a^*)} + d\sqrt{\left(\frac{\kappa \sigma^2}{\Delta_{\min}} + 1 \right) C\vartheta \log T} + d\vartheta \log T \right).$$

As $X = O(\sqrt{AX} + B)$ implies $X = O(A + B)$, we have

$$R_T(a^*) = O\left(\frac{d^2 \vartheta \kappa \sigma^2 \log T}{\Delta_{\min}} + d\sqrt{\left(\frac{\kappa \sigma^2}{\Delta_{\min}} + 1 \right) C\vartheta \log T} + d\vartheta \log T \right)$$

$$= O\left(\left(\frac{d \kappa \sigma^2}{\Delta_{\min}} + 1 \right) d\vartheta \log T + d\sqrt{\left(\frac{\kappa \sigma^2}{\Delta_{\min}} + 1 \right) C\vartheta \log T} \right),$$

which means that (17) holds. A complete proof is given in Section G of the appendix.

References

Jacob Abernethy, Elad E Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, pages 263–273, 2008a.

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

Jacob Duncan Abernethy, Elad Hazan, and Alexander Rakhlin. An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory*, 2008b.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory: 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007. Proceedings 18*, pages 150–165. Springer, 2007.

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

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, volume 23, pages 41.1–41.14, 2012.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, pages 508–528, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. Best of both worlds policy optimization. In *International Conference on Machine Learning*, 2023a.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *Conference on Learning Theory*, 2023b.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 34, 2021.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. In *Advances in Neural Information Processing Systems*, pages 11478–11489, 2020.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578, 2019.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

Mark Herbster and Manfred K Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1(281-309):10–1162, 2001.

Junya Honda, Shinji Ito, and Taira Tsuchiya. Follow-the-perturbed-leader achieves best-of-both-worlds for bandit problems. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201, pages 726–754, 2023.

Shinji Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. *Advances in Neural Information Processing Systems*, 34:2654–2667, 2021a.

Shinji Ito. On optimal robustness to adversarial corruption in online decision problems. *Advances in Neural Information Processing Systems*, 34:7409–7420, 2021b.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pages 2552–2583, 2021c.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Conference on Learning Theory*, pages 1421–1422, 2022.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in Neural Information Processing Systems*, 33:16557–16566, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34:20491–20502, 2021.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: Ftrl with general regularizers and multiple optimal arms. *arXiv preprint arXiv:2302.13534*, 2023.

Fang Kong, Canzhe Zhao, and Shuai Li. Best-of-three-worlds analysis for linear bandits with follow-the-regularized-leader algorithm. In *Conference on Learning Theory*, 2023.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *International Conference on Machine Learning*, pages 5805–5814, 2021.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737, 2017.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint, Revision: 1699*, 2019.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pages 6142–6151, 2021.

Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 424–433. IEEE, 2014.

Yin Tat Lee and Aaron Sidford. Solving linear programs with sqrt (rank) linear system solves. *arXiv preprint arXiv:1910.08033*, 2019.

15

Yin Tat Lee and Man-Chung Yue. Universal barrier is n-self-concordant. *Mathematics of Operations Research*, 46(3):1129–1148, 2021.

Yingkai Li, Edmund Y Lou, and Liren Shan. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

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

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245, 2021.

Vahab Mirrokni, Renato Paes Leme, Adrian Vladu, and Sam Chiu-wai Wong. Tight bounds for approximate carathéodory and beyond. In *International Conference on Machine Learning*, pages 2440–2448, 2017.

Arkadi Nemirovski. Interior point polynomial time methods in convex programming. *Lecture notes*, 42(16):3215–3224, 2004.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. In *Advances in Neural Information Processing Systems*, pages 10328–10337, 2020.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013.

Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295, 2014.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3367–3375, 2021.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201, pages 1484–1515, 2023a.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Further adaptive best-of-both-worlds algorithm for combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*, 2023b.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291, 2018.

Lin Yang, Mohammad Hajiesmaili, Mohammad Sadegh Talebi, John C. S. Lui, and Wing Shing Wong. Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm. In *Advances in Neural Information Processing Systems*, volume 33, pages 19943–19952, 2020.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692, 2019.

Appendix A. Related Work

Best-of-Both-Worlds Bandit Algorithms   Best-of-both-worlds algorithms have been developed for various settings of multi-armed bandit (MAB) problems, including the standard MAB problem (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Zimmert and Seldin, 2021; Ito et al., 2022; Honda et al., 2023), combinatorial semi-bandits (Zimmert et al., 2019; Ito, 2021a; Tsuchiya et al., 2023b), partial monitoring problems (Tsuchiya et al., 2023a), episodic Markov decision processes (Jin and Luo, 2020; Jin et al., 2021; Dann et al., 2023a), and linear bandits (Lee et al., 2021). While most of these studies focuses only on high-level adaptability, the algorithms by Ito et al. (2022); Tsuchiya et al. (2023b) for the MAB problem and combinatorial semi-bandit problems have low-level adaptability as well, similarly to our proposed algorithm. In fact, their algorithms are best-of-three-worlds algorithms with multiple data-dependent regret bounds as well as variance-adaptive regret bounds. Their algorithms are also similar to ours in that it is based on the optimistic follow-the-regularized-leader approach with an adaptive learning rate. As the class of linear bandits problem includes the multi-armed bandit problem, the results in this paper can be interpreted as an extension of their results. Regret bounds by Ito et al. (2022) are, however, better than ours in terms of the dependency on the dimensionality of the action set (or the number of arms) and in that they depend on arm-wise sub-optimality gaps.

Adversarial Corruption   There are several studies on the stochastic environment with adversarial corruption in the linear bandit problem (Li et al., 2019; Bogunovic et al., 2021; Lee et al., 2021) and related problems such as the multi-armed bandits (Lykouris et al., 2018; Gupta et al., 2019; Zimmert and Seldin, 2021; Yang et al., 2020) and the linear Markov decision processes (Lykouris et al., 2021). These studies and this paper have different assumptions and regret. This paper and Lee et al. (2021) assume that corruption depends only on information in the past rounds and is an affine function of the chosen action. On the other hand, Li et al. (2019); Bogunovic et al. (2021) allow corruption to be any (possibly non-linear) function. Furthermore, Bogunovic et al. (2021) consider the corruption that depends on the action chosen in that round. We also note that the definitions of the corruption level in these studies are slightly different. While this paper includes the corruption in regret, Li et al. (2019); Bogunovic et al. (2021); Lee et al. (2021) do not. It is known that we can convert one to the other by an additional $O(C)$-regret. Moreover, the regret bounds in these existing studies have linear terms with respect to C. Thus, our regret bound for the corrupted stochastic regime have the same dependence of the corruption as in these studies, but not vice versa.

Misspecified Linear Contextual Bandits   The corrupted stochastic regime is a special case of the misspecified linear contextual bandits without knowledge of the misspecification (Lattimore et al., 2020; Foster et al., 2020; Pacchiano et al., 2020; Takemura et al., 2021; Krishnamurthy et al., 2021).[3] This problem assumes that the expected loss functions can be approximated by a linear function. While the approximation error can be any function of the information in the past and the current rounds in general, the corrupted stochastic regime assumes that the approximation error is an affine function of the action chosen in the current round. It is an open question whether the proposed algorithm can obtain a regret upper bound similar to the known regret bounds for this problem when the approximation error can be non-linear.

---

3. Note that some studies assume oblivious adversary (Lattimore et al., 2020; Foster et al., 2020; Krishnamurthy et al., 2021), i.e., the approximation errors do not depend on the actions chosen in the past.

Appendix B. Pseudocode of the proposed algorithm

Algorithm 1

Require: T: time horizon, d: dimensionality of action set, $\mathcal{A} \subseteq \mathbb{R}^d$: action set, ψ: self-concordant barrier over $\mathcal{X} = \mathrm{conv}(\mathcal{A})$, $\vartheta \geq 1$: self-concordance parameter of ψ, $\eta \in (0, 1/4)$: learning rate for optimistic prediction

1: Set $m_1 = 0$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3:      Set β_t by (13) and compute x_t defined by (2).
4:      Let $\{e_1, \ldots, e_d\}$ and $\{\lambda_1, \ldots, \lambda_d\}$ be the set of eigenvectors and eigenvalues of $\nabla^2 \psi(x_t)$ and set $\mathcal{E}_t := \{x_t + \lambda_i^{-1/2} e_i \mid i \in [d]\} \cup \{x_t - \lambda_i^{-1/2} e_i \mid i \in [d]\}$.
5:      Choose $z_t \in \mathcal{A}$ and set $r_t \in (0, 1]$ by (9).
6:      Set $b_t = 1$ with probability r_t and set $b_t = 0$ with probability $(1 - r_t)$.
7:      **if** $b_t = 1$ **then**
8:          Choose i_t from $[d]$ uniformly at random and set $v_t \pm 1$ with probability $1/2$.
9:          Set $a'_t = z_t + r_t^{-1}(x_t + v_t \lambda_{i_t}^{-1/2} e_{i_t} - z_t)$.
10:      **else**
11:          Set $a'_t = z_t$.
12:      **end if**
13:      Output $a_t \in \mathcal{A}$ so that $\mathbf{E}[a_t] = a'_t$ and get feedback of $f_t(a_t)$.
14:      Compute $\hat{\ell}_t$ defined by (10) and update m_t by (14).
15: **end for**

Appendix C. Proof of Lemma 9

For the convex function ψ and $x \in \mathrm{dom}(\psi)$, denote *the Newton decrement* at point x by $\lambda(x, \psi)$, i.e., $\lambda(x, \psi) = \|\nabla \psi(x)\|_{x,\psi}^*$.

Lemma 16 (Theorem 2.2.1 by Nesterov and Nemirovskii (1994)) *Let $\mathcal{S}$ be an open non-empty convex subset of a finite-dimensional real vector space. Let ψ be a self-concordant function on $\mathcal{S}$ and $x \in \mathcal{S}$. Then, for each $y \in \mathcal{S}$ such that $\|x - y\|_{x,\psi} < 1$, we have*

$$(1 - \|x - y\|_{x,\psi})^2 \nabla^2 \psi(y) \preceq \nabla^2 \psi(x) \preceq (1 - \|x - y\|_{x,\psi})^{-2} \nabla^2 \psi(y)$$

Lemma 17 ((2.21) by Nemirovski (2004)) *Let ψ be a self-concordant function on $\mathcal{X}$. If $\lambda(x, \psi) < 1$, we have*

$$\|x - x^*\|_{x,\psi} \leq \frac{\lambda(x, \psi)}{1 - \lambda(x, \psi)},$$

where $x^* \in \arg\min_y \psi(y)$.

Lemma 18 *Let ψ be a self-concordant function on $\mathcal{X}$ and $x, y \in \mathrm{int}(\mathcal{X})$. Suppose that $\|x - y\|_{x,\psi} \leq 1/2$. Then, we have*

$$\langle \ell, x - y \rangle - \beta D_\psi(y, x) \leq \frac{2}{\beta} \|\ell\|_{x,\psi}^{*2}$$

for all $\ell \in \mathbb{R}^d$ and $\beta > 0$.

Proof Using the Cauchy-Schwarz inequality and the AM-GM inequality, we have

$$\langle \ell, x - y \rangle \leq \|\ell\|_{x,\psi}^* \|x - y\|_{x,\psi} \leq \frac{2}{\beta} \|\ell\|_{x,\psi}^{*2} + \frac{\beta}{8} \|x - y\|_{x,\psi}^2.$$

Thus, it is sufficient to show $D_\psi(y, x) \geq \frac{1}{8} \|x - y\|_{x,\psi}^2$.

By Taylor's theorem, we have $D_\psi(y, x) = \frac{1}{2} \|x - y\|_{\xi,\psi}^2$ for some $\xi = x + \alpha(y - x)$ where $\alpha \in (0, 1)$. It follows from Lemma 16 that

$$\|x - y\|_{\xi,\psi}^2 \geq (1 - \|\xi - x\|_{x,\psi})^2 \|x - y\|_{x,\psi}^2 = (1 - \alpha\|x - y\|_{x,\psi})^2 \|x - y\|_{x,\psi}^2 \geq \frac{1}{4} \|x - y\|_{x,\psi}^2.$$

∎

Proof of Lemma 9 Let $f(y) = D_\psi(y, x) - \langle \ell, x - y \rangle / \beta$. Since ψ is self-concordant, there exists $y^* \in \text{int}(\mathcal{X})$ such that $y^* \in \arg\min_{y \in \mathcal{X}} f(y)$. If we have $\lambda(x, f) \leq 1/3$, by Lemma 17, we obtain

$$\|x - y^*\|_{x,\psi} = \|x - y^*\|_{x,f} \leq \frac{\lambda(x, f)}{1 - \lambda(x, f)} \leq 1/2.$$

Thus, we obtain

$$\langle \ell, x - y \rangle - \beta D_\psi(y, x) \leq \langle \ell, x - y^* \rangle - \beta D_\psi(y^*, x) \leq \frac{2}{\beta} \|\ell\|_{x,\psi}^{*2},$$

where the first inequality holds due to $y^* \in \arg\min_{y \in \mathcal{X}} f(y)$ and the second inequality is derived from Lemma 18. Hence, it suffices to show $\lambda(x, f) \leq 1/3$. By the definition of f, we have $\nabla f(x) = \ell/\beta$. Thus, we obtain

$$\lambda(x, f) = \|\nabla f(x)\|_{x,f}^* = \|\nabla f(x)\|_{x,\psi}^* = \|\ell\|_{x,\psi}^*/\beta \leq 1/3,$$

where the inequality is obtained by the assumption. ∎

Appendix D. Proof of Lemma 13

We first show that $\mathbf{E}[a_t|x_t] = x_t$. The expectation of a_t is

$$\begin{aligned}
\mathbf{E}[a_t|x_t] = \mathbf{E}[a_t'|x_t] &= (1 - r_t)\,\mathbf{E}[a_t'|b_t = 0] + r_t\,\mathbf{E}[a_t'|b_t = 1] \\
&= (1 - r_t)\,z_t + r_t\,\mathbf{E}\left[z_t + r_t^{-1}\left(x_t + v_t\lambda_{i_t}^{-1/2}e_{i_t} - z_t\right)\right] \\
&= (1 - r_t)\,z_t + r_t(z_t + r_t^{-1}(x_t - z_t)) = z_t + (x_t - z_t) = x_t, \quad (23)
\end{aligned}$$

where the forth equality follows from $\mathbf{E}[v_t] = 0$.

Let us next show that $\hat{\ell}_t$ defined by (10) is an unbiased estimator of ℓ_t. We have

$$
\begin{aligned}
\mathbf{E}\left[\hat{\ell}_t - m_t | x_t\right] &= r_t\,\mathbf{E}\left[dv_t\lambda_{i_t}^{1/2}(f_t(a_t) - \langle m_t, a_t\rangle)e_{i_t}\right]\\
&= r_t\,\mathbf{E}\left[dv_t\lambda_{i_t}^{1/2}(\langle \ell_t - m_t, a_t\rangle + \xi_t)e_{i_t}\right]\\
&= r_t\,\mathbf{E}\left[dv_t\lambda_{i_t}^{1/2}e_{i_t}\left\langle z_t + r_t^{-1}(x_t + v_t\lambda_{i_t}^{-1/2}e_{i_t} - z_t), \ell_t - m_t\right\rangle\right]\\
&= r_t\,\mathbf{E}\left[dv_t\lambda_{i_t}^{1/2}e_{i_t}\left\langle r_t^{-1}v_t\lambda_{i_t}^{-1/2}e_{i_t}, \ell_t - m_t\right\rangle\right] + r_t\,\mathbf{E}\left[dv_t\lambda_{i_t}^{1/2}e_{i_t}\left\langle z_t + r_t^{-1}(x_t - z_t), \ell_t - m_t\right\rangle\right]\\
&= d\,\mathbf{E}\left[e_{i_t}e_{i_t}^\top(\ell_t - m_t)\right] = \ell_t - m_t, \quad\quad (24)
\end{aligned}
$$

where we used $v_t^2 = 1$, $\mathbf{E}[v_t] = 0$, and the fact that v_t and m_t are independent in the fifth equality.

Suppose that $\min_{x\in\mathcal{X}}\psi(x) = 0$ holds without loss of generality. Let $x_0 \in \arg\min_{x\in\mathcal{X}}\psi(x)$. Given $a^* \in \mathcal{A}$, define x^* by

$$
x^* = \left(1 - \frac{1}{T}\right)a^* + \frac{1}{T}x_0 = a^* + \frac{1}{T}(x_0 - a^*).
$$

From this, (23) and (24), we have

$$
\begin{aligned}
R_T(a^*) &= \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, a_t - a^*\rangle\right] = \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, a_t - x^*\rangle\right] + \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, x^* - a^*\rangle\right]\\
&= \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, a_t - x^*\rangle\right] + \frac{1}{T}\mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, x_0 - a^*\rangle\right] \le \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, a_t - x^*\rangle\right] + 1\\
&= \mathbf{E}\left[\sum_{t=1}^T \langle \ell_t, x_t - x^*\rangle\right] + 1 = \mathbf{E}\left[\sum_{t=1}^T \left\langle \hat{\ell}_t, x_t - x^*\right\rangle\right] + 1 \quad\quad (25)
\end{aligned}
$$

Then, as we have $x_0 + (1 - 1/T)^{-1}(x^* - x_0) = x_0 + (a^* - x_0) = a^* \in \mathcal{A}$, we have $\pi_{x_0}(x^*) \le 1 - 1/T$. Hence, from Lemma 8, we have

$$
\psi(x^*) = \psi(x^*) - \psi(x_0) \le \vartheta\log\left(\frac{1}{1 - \pi_{x_0,\mathcal{X}}(x^*)}\right) \le \vartheta\log\left(\frac{1}{1 - (1 - 1/T)}\right) = \vartheta\log T.
$$

From this, (25) and Lemma 4, we have

$$
R_T(a^*) \le \mathbf{E}\left[\sum_{t=1}^T \left(\left\langle \hat{\ell}_t - m_t, x_t - x'_{t+1}\right\rangle - \beta_t D(x'_{t+1}, x_t)\right) + \beta_{T+1}\vartheta\log T\right] + 1. \quad\quad (26)
$$

The part of $\left\langle \hat{\ell}_t - m_t, x_t - x'_{t+1}\right\rangle - \beta_t D(x'_{t+1}, x_t)$ can bounded by using Lemma 9. From the definition (10), we have

$$
\begin{aligned}
\|\hat{\ell}_t - m_t\|_{x_t,\psi}^{*2} &= (\hat{\ell}_t - m_t)^\top(\nabla^2\psi(x_t))^{-1}(\hat{\ell}_t - m_t)\\
&= b_t d^2(f_t(a_t) - \langle m_t, a_t\rangle)^2\lambda_{i_t}e_{i_t}^\top(\nabla^2\psi(x_t))^{-1}e_{i_t}\\
&= b_t d^2(f_t(a_t) - \langle m_t, a_t\rangle)^2 = d^2 g_t(m_t) \le 4b_t d^2 \le 4d^2. \quad\quad (27)
\end{aligned}
$$

Hence, if $\beta_t \geq 6d$, we have $\|\hat{\ell}_t - m_t\|_{x_t,\psi}^* \leq \beta_t/3$, and, consequently, we can apply Lemma 9 to bound the stability term as follows:

$$\left\langle \hat{\ell}_t - m_t, x_t - x'_{t+1} \right\rangle - \beta_t D(x'_{t+1}, x_t) \leq \frac{2}{\beta_t} \|\hat{\ell}_t - m_t\|_{\psi,x_t}^{*2} = \frac{2d^2 g_t(m_t)}{\beta_t},$$

where $g_t(m)$ is defined in (12). Then, from this and (26), we have

$$R_T(a^*) \leq \mathbf{E}\left[2\sum_{t=1}^{T} \frac{d^2 g_t(m_t)}{\beta_t} + \beta_{T+1} \vartheta \log T \right] + 1. \tag{28}$$

If β_t is given by (13), we then have

$$\frac{d^2 g_t(m_t)}{\beta_t} = d\frac{\sqrt{\vartheta \log T} g_t(m_t)}{6\sqrt{\vartheta \log T} + 2\sqrt{\sum_{s=1}^{t-1} g_s(m_s)}} \leq d\frac{\sqrt{\vartheta \log T} g_t(m_t)}{\sqrt{\sum_{s=1}^{t-1} g_s(m_s) + 36\vartheta \log T} + \sqrt{\sum_{s=1}^{t-1} g_s(m_s)}}$$

$$\leq d\frac{\sqrt{\vartheta \log T} g_t(m_t)}{\sqrt{\sum_{s=1}^{t} g_s(m_s)} + \sqrt{\sum_{s=1}^{t-1} g_s(m_s)}} = \sqrt{\vartheta \log T}\left(\sqrt{\sum_{s=1}^{t} g_s(m_s)} - \sqrt{\sum_{s=1}^{t-1} g_s(m_s)} \right).$$

which yields

$$\sum_{t=1}^{T} \frac{d^2 g_t(m_t)}{\beta_t} \leq \sqrt{\vartheta \log T} \sum_{t=1}^{T}\left(\sqrt{\sum_{s=1}^{t} g_s(m_s)} - \sqrt{\sum_{s=1}^{t-1} g_s(m_s)} \right) = \sqrt{\vartheta \log T \cdot \sum_{t=1}^{T} g_t(m_t)}. \tag{29}$$

We also have

$$\beta_{T+1} \vartheta \log T = 2\sqrt{\vartheta \log T \cdot \sum_{t=1}^{T} g_t(m_t)} + 6d\vartheta \log T$$

from the definition (13) of β_t. Combining this with (28) and (29), we obtain

$$R_T(a^*) \leq 4d\,\mathbf{E}\left[\sqrt{\vartheta \log T \cdot \sum_{t=1}^{T} g_t(m_t)} \right] + 6d\vartheta \log T + 1,$$

which completes the proof.

Appendix E. Proof of Theorem 10

Fix $\eta \in (0, 1/4)$ arbitrarily. By substituting $u_t = \bar{\ell} \in \arg\min_\ell \sum_{t=1}^{T} \|\ell_t - \ell\|_2^2$ for all t in (19), we obtain

$$\sum_{t=1}^{T} g_t(m_t) = O\left(\sum_{t=1}^{T} g_t(\bar{\ell}_T) + 1 \right) = O\left(\sum_{t=1}^{T} b_t \left(\langle \ell_t - \bar{\ell}_T, a_t \rangle + \varepsilon_t(a_t) \right)^2 + 1 \right)$$

$$= O\left(\sum_{t=1}^{T} \left(\|\ell_t - \bar{\ell}_T\|_2^2 + (\varepsilon_t(a_t))^2 \right) + 1 \right).$$

Similarly, by substituting $u_t = \ell_t$, we obtain

$$\sum_{t=1}^{T} g_t(m_t) = O\left(\sum_{t=1}^{T}(\varepsilon_t(a_t))^2 + \sum_{t=1}^{T-1}(\|\ell_t - \ell_{t+1}\|_2) + 1\right).$$

By combining these with Lemma 13 and applying Jensen's inequality, we obtain (15). Further, if $f_t(a) \geq 0$, by substituting $u_t = 0$, we obtain

$$\sum_{t=1}^{T} g_t(m_t) = O\left(\sum_{t=1}^{T} g_t(0) + 1\right) = O\left(\sum_{t=1}^{T} b_t(f_t(a_t))^2 + 1\right) = O\left(\sum_{t=1}^{T} f_t(a_t) + 1\right).$$

By combining this with Lemma 13, we obtain

$$R_T(a^*) = O\left(d\sqrt{\vartheta \log T \left(\mathbf{E}\left[\sum_{t=1}^{T} f_t(a_t)\right] + 1\right)}\right)$$

$$= O\left(d\sqrt{\vartheta \log T \left(R_T(a^*) + \mathbf{E}\left[\sum_{t=1}^{T} f_t(a^*)\right] + 1\right)}\right)$$

which implies that (16) holds. ∎

Appendix F. Proof of Lemma 15

As $\mathcal{X}$ is the convex hull of $\mathcal{A}' = \{a^*\} \cup \text{conv}(\mathcal{A} \setminus \{a^*\})$, any point $y \in \mathcal{X}$ can be expressed as a convex combination of a^* and a point in $\text{conv}(\mathcal{A} \setminus \{a^*\})$, which means that there exists $\lambda \in [0, 1]$ and $x' \in \text{conv}(\mathcal{A} \setminus \{a^*\})$ such that $x = \lambda x' + (1 - \lambda)a^*$. For such x, we have

$$\pi_{a^*, \mathcal{X}}(x) \leq \lambda. \tag{30}$$

In fact, we have

$$a^* + \lambda^{-1}(x - a^*) = a^* + \lambda^{-1}\lambda(x' - a^*) = x' \in \mathcal{X},$$

which means that (30) holds. We further have

$$\Delta(x) = \lambda\Delta(x') + (1 - \lambda)\Delta(a^*) = \lambda\Delta(x') \geq \lambda\Delta_{\min},$$

where the last inequality follows from the fact that $x' \in \text{conv}(\mathcal{A} \setminus \{a^*\})$ and the definition of $\Delta_{\min}$. Combining this with (30), we obtain

$$\pi_{a^*, \mathcal{X}}(x) \leq \frac{\Delta(x)}{\Delta_{\min}}. \tag{31}$$

We next show

$$\max_{x \in W_1(y)} \Delta(x) \leq 2\Delta(y). \tag{32}$$

As $W_1(y)$ is an ellipsoid centered at y, it holds that

$$\langle \ell^*, y \rangle - \min_{x \in W_1(y)} \langle \ell^*, x \rangle = \max_{x \in W_1(y)} \langle \ell^*, x \rangle - \langle \ell^*, y \rangle .$$

We hence have

$$
\begin{aligned}
\max_{x \in W_1(y)} \Delta(x) &= \max_{x \in W_1(y)} \langle \ell^*, x \rangle - \langle \ell^*, y \rangle + \Delta(y) \\
&= \langle \ell^*, y \rangle - \min_{x \in W_1(y)} \langle \ell^*, x \rangle + \Delta(y) \\
&\leq \langle \ell^*, y \rangle - \min_{x \in \mathcal{X}} \langle \ell^*, x \rangle + \Delta(y) = 2\Delta(y),
\end{aligned}
$$

where the inequality follows from the fact that $W_1(y) \subseteq \mathcal{X}$. Combining (31) and (32), we obtain

$$\max_{x \in W_1(y)} \pi_{a^*, \mathcal{X}}(x) \leq \max_{x \in W_1(y)} \frac{\Delta(x)}{\Delta_{\min}} \leq 2\frac{\Delta(y)}{\Delta_{\min}}.$$

Appendix G. Proof of Theorem 11

From Lemma 14 with $u_t = \ell^*$, we have

$$
\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{T} g_t(m_t)\right] &= O\left(\mathbf{E}\left[\sum_{t=1}^{T} g_t(\ell^*)\right] + 1\right) \leq O\left(\mathbf{E}\left[\sum_{t=1}^{T} b_t(\langle \ell_t - \ell^*, a_t \rangle + \varepsilon_t(a_t))^2\right] + 1\right) \\
&\leq O\left(\mathbf{E}\left[\sum_{t=1}^{T} \left(\|\ell_t - \ell^*\|_2^2 + r_t \sigma_t^2\right)\right] + 1\right) \leq O\left(\mathbf{E}\left[C + \sum_{t=1}^{T} r_t \sigma_t^2\right] + 1\right).
\end{aligned}
$$

From this and Lemma 13, we have

$$R_T = O\left(d\sqrt{\vartheta \log T \cdot \mathbf{E}\left[C + \sum_{t=1}^{T} \sigma_t^2 r_t\right]} + d\vartheta \log T\right). \tag{33}$$

Under the assumption of (11), we have

$$r_t \leq \kappa \cdot \min_{z \in \mathcal{A}}\left\{\max_{x \in \mathcal{E}_t} \pi_{z, \mathcal{X}}(x)\right\} \leq \kappa \cdot \min_{z \in \mathcal{A}}\left\{\max_{x \in W_1(x_t)} \pi_{z, \mathcal{X}}(x)\right\} \leq 2\kappa \frac{\Delta(x_t)}{\Delta_{\min}},$$

where second inequality follows from $\mathcal{E}_t \subseteq W_1(x_t)$ and the last inequality follows from Lemma 15. From this, (33) and $\sigma^2 = \max_{t \in [T]} \sigma_t^2$, we have

$$R_T = O\left(d\sqrt{\vartheta \log T \cdot \mathbf{E}\left[C + \frac{\kappa \sigma^2}{\Delta_{\min}} \sum_{t=1}^{T} \Delta(x_t)\right]} + d\vartheta \log T\right). \tag{34}$$

24

On the other hand, R_T is bounded from below as follows:

$$
\begin{aligned}
R_T(a^*) = \sum_{t=1}^T \mathbf{E}\left[\langle \ell_t, a_t - a^* \rangle\right] &= \sum_{t=1}^T \mathbf{E}\left[\langle \ell^*, a_t - a^* \rangle + \langle \ell_t - \ell^*, a_t - a^* \rangle\right] \\
&\geq \sum_{t=1}^T \mathbf{E}\left[\langle \ell^*, a_t - a^* \rangle - 2\|\ell_t - \ell^*\|_2\right] = \mathbf{E}\left[\sum_{t=1}^T \langle \ell^*, x_t - a^* \rangle - 2C\right] \\
&= \mathbf{E}\left[\sum_{t=1}^T \Delta(x_t) - 2C\right].
\end{aligned}
\tag{35}
$$

Combining this with (34), we obtain

$$
\begin{aligned}
R_T(a^*) &= O\left(d\sqrt{\vartheta \log T \cdot \left(C + \frac{\kappa\sigma^2}{\Delta_{\min}}(R_T + C)\right)} + d\vartheta \log T\right) \\
&= O\left(d\sqrt{\frac{\vartheta\kappa\sigma^2 \log T}{\Delta_{\min}} R_T(a^*)} + d\sqrt{\left(\frac{\kappa\sigma^2}{\Delta_{\min}} + 1\right) C\vartheta \log T} + d\vartheta \log T\right).
\end{aligned}
$$

As $X = O(\sqrt{AX} + B)$ implies $X = O(A + B)$, we have

$$
\begin{aligned}
R_T(a^*) &= O\left(\frac{d^2\vartheta\kappa\sigma^2 \log T}{\Delta_{\min}} + d\sqrt{\left(\frac{\kappa\sigma^2}{\Delta_{\min}} + 1\right) C\vartheta \log T} + d\vartheta \log T\right) \\
&= O\left(\left(\frac{d\kappa\sigma^2}{\Delta_{\min}} + 1\right) d\vartheta \log T + d\sqrt{\left(\frac{\kappa\sigma^2}{\Delta_{\min}} + 1\right) C\vartheta \log T}\right).
\end{aligned}
$$