Further Adaptive Best-of-Both-Worlds Algorithm for Combinatorial Semi-Bandits

Taira Tsuchiya
Kyoto University / RIKEN AIP

Shinji Ito
NEC Corporation

Junya Honda
Kyoto University / RIKEN AIP

Abstract

We consider the combinatorial semi-bandit problem and present a new algorithm with a best-of-both-worlds regret guarantee, in which the regrets are near-optimally bounded in the stochastic and adversarial regimes. In the stochastic regime, we prove a variance-dependent regret bound depending on the tight suboptimality gap introduced by Kveton et al. (2015) with a good leading constant. In the adversarial regime, we show that the same algorithm simultaneously obtains various data-dependent regret bounds. Our algorithm is based on the follow-the-regularized-leader framework with a refined regularizer and adaptive learning rate. Finally, we numerically test the proposed algorithm and confirm its superior or competitive performance over existing algorithms, including Thompson sampling under most settings.

1 INTRODUCTION

The combinatorial semi-bandit problem is an online decision-making problem, and it includes many practical problems such as multi-task bandits (Cesa-Bianchi and Lugosi, 2012), crowdsourcing (ul Hassan and Curry, 2016), learning spectrum allocations (Gai et al., 2012), shortest path problem (Gai et al., 2012), and recommender systems (Qin et al., 2014). In combinatorial semi-bandits, the learner and environment play the game sequentially. The learner is given an action set $\mathcal{A} \subset \{0,1\}^d$, where $d \in \mathbb{N}$ is the dimension of the action set. For every round $t \in [T] := \{1, \dots, T\}$, the environment chooses a loss $\ell(t) \in [0,1]^d$, and the learner then chooses an action $a(t) \in \mathcal{A}$ (also called a *super-arm*), incurs a loss $\langle \ell(t), a(t) \rangle$, and observes $\ell_i(t)$ for all $i \in [d]$ such that $a_i(t) = 1$. We refer to each index $i \in [d]$ as *base-arm* i. The goal of the learner is to minimize their cumulative loss over all rounds. The performance of the learner is evaluated based on regret R_T defined as the difference between the cumulative losses of the learner and the single optimal action a^* fixed in terms of the expected cumulative loss, *i.e.,* $a^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}[\sum_{t=1}^T \langle \ell(t), a \rangle]$ and

$$R_T = \mathbb{E}\left[\sum_{t=1}^T \langle \ell(t), a(t) - a^* \rangle\right],$$

where the expectation is taken w.r.t. the randomness of $\ell(t)$ and the internal randomness of the algorithm.

The combinatorial semi-bandit problem, or more broadly, a variety of online-decision making problems, have been investigated within mainly two regimes: *stochastic* and *adversarial* regimes. In the stochastic regime, the sequence of losses $(\ell(t))_{t=1}^T$ is sampled from a fixed distribution $\mathcal{D}$ in an i.i.d. manner with mean $\mu = \mathbb{E}_{\ell \sim \mathcal{D}}[\ell]$. In the adversarial regime, the losses are arbitrarily decided from $[0,1]^d$ (Kveton et al., 2015; Neu, 2015; Wang and Chen, 2018) or more generally from S^d for some bounded $S \subset \mathbb{R}$ (Wei and Luo, 2018; Zimmert et al., 2019) possibly depending on the past history of learner's actions.

There have been a considerable number of studies on combinatorial semi-bandits for both adversarial and stochastic regimes. In the adversarial regime, the regret bound of $O(\sqrt{mdT})$ was proved for $m = \max_{a \in \mathcal{A}} \|a\|_1$ (Audibert et al., 2014), which matches the lower bound of $\Omega(\sqrt{mdT})$ (Audibert et al., 2014). In the stochastic regime, many algorithms have been shown to achieve logarithmic regrets depending on the minimum suboptimality gap, which is defined by $\Delta = \min\{\mu^\top (a - a^*) : a \in \mathcal{A} \setminus \{a^*\}\}$. Kveton et al. (2015) and Wang and Chen (2018) derived gap-dependent regret bounds given by $O(dm \log(T)/\Delta)$ for general action sets and $O((d-m)\log(T)/\Delta)$ for matroid semi-bandits. Furthermore, Kveton et al. (2015) derived a refined bound given by $O(\sum_{i:a_i^*=1}(m/\Delta_{i,\min}) \log T)$ depending on $\Delta_{i,\min} = \min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\}, a_i = 1\} \geq \Delta$ of each base-arm i rather than on Δ.

It is unclear which regime's algorithms are better suited

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

to practical applications. Algorithms specialized for the stochastic regime occasionally suffer a linear regret, whereas algorithms for the adversarial regime work poorly in the stochastic regime. Because it is difficult to know it in practice, it is desirable to obtain a near-optimal performance both for the stochastic and adversarial regimes *without* knowing the underlying environment.

To this end, particullary in the classical multi-armed bandits, the Best-of-Both-Worlds (BOBW) algorithm has been developed, which performs near-optimally both in the stochastic and adversarial regimes. In a seminal study, Bubeck and Slivkins (2012) developed the first BOBW algorithm, and the celebrated Tsallis-INF algorithm was recently proposed by Zimmert and Seldin (2021). For combinatorial semi-bandits, we are aware of the works by Wei and Luo (2018), Zimmert et al. (2019), and Ito (2021a). Some BOBW algorithms achieve favorable regret guarantees also in the *stochastic regime with adversarial corruptions* (Lykouris et al., 2018), which is an intermediate regime between the stochastic and adversarial regimes. This intermediate regime is advantageous in practice since the stochastic assumption on losses often fails to hold, whereas the adversarial assumption is excessively pessimistic.

Adaptive algorithms that exploit the characteristics of a sequence of losses have been actively developed for both the adversarial and stochastic regimes. In the adversarial regime, *data-dependent regret bounds* have been recently investigated to enhance the adaptivity of the algorithm to a given structure of the loss data. Well-known examples are the first-order bounds depending on the cumulative loss $L^* = \min_{a\in\mathcal{A}} \mathbb{E}[\sum_{t=1}^T \langle \ell(t), a\rangle]$, second-order bounds depending on the empirical variations of losses $Q_2 = \mathbb{E}[\sum_{t=1}^T \|\ell(t) - \bar{\ell}\|^2]$ defined with $\bar{\ell} = T^{-1}\mathbb{E}[\sum_{t=1}^T \ell(t)]$, and path-length bounds depending on the variation of losses $V_1 = \mathbb{E}[\sum_{t=1}^{T-1} \|\ell(t) - \ell(t+1)\|_1]$. For the semi-bandit problem, Wei and Luo (2018) presented the first-order regret bound of $O(\sqrt{dL^* \log T})$, second-order bound of $O(\sqrt{dQ_2 \log T})$, and the path-length bound of $O(\sqrt{dV_1 \log T})$. Note that the data-dependent bounds developed by Wei and Luo (2018) cannot be achieved using the same algorithm. Table 1 summarizes notation used in this paper.

In the stochastic regime, one of the most promising approaches to making an algorithm more adaptive is to estimate and use distributional information. In the multi-armed bandit problem, algorithms that exploit the *variance* of losses have been developed (Audibert et al., 2007; Ito et al., 2022a), and (co)variance-aware algorithms for semi-bandits have also been investigated (Komiyama et al., 2015; Degenne and Perchet, 2016; Merlis and Mannor, 2019; Perrault et al., 2020; Vial et al., 2022; Liu et al., 2022). The variance-aware algorithm is highly advantageous in real-world applications since the variances of losses for each base-arm i, $\sigma_i^2 = \mathbb{E}_{\ell\sim\mathcal{D}}[(\ell_i - \mu_i)^2] \in [0, 1/4]$, are extremely small in many real-world applications, whereas the variance can be $1/4$ in the worst case scenario. For example, for a search engine, the click-through rate is usually below 0.05 (Komiyama et al., 2017), implying that the variance of the base-arm is much smaller than $1/4$. Additionally, in the shortest path problem (György et al., 2007), the congestion level of the road does not change substantially in many cases, and hence the variance is expected to be much smaller than in the worst-case scenario also of this problem. Indeed, variance-aware algorithms are known to be highly effective in the problem of online eco-routing for electric vehicles (Chen et al., 2022). Accordingly, we aim to achieve a variance-dependent regret bounds in the stochastic regime with multiple data-dependent regret bounds simultaneously in the adversarial regime by the same algorithm.

Table 1: Notation

Symbol	Meaning
$\mathcal{A} \in \{0,1\}^d$	Action set
$d \in \mathbb{N}$	Dimensionality of action set
$m \le d$	$m = \max_{a\in\mathcal{A}} \|a\|_1$
$a^* \in \mathcal{A}$	Optimal action
$I^* \subset [d]$	$\{i \in [d] : a_i^* = 1\}$, set of optimal base-arms
$J^* \subset [d]$	$[d] \setminus I^*$, set of sub-optimal base-arms
$\mu_i \in [0,1]$	$\mathbb{E}[\ell_i]$, mean of base-arm i
$\sigma_i^2 \in [0, 1/4]$	$\mathbb{E}[(\ell_i - \mu_i)^2]$, variance of base-arm i
$\Delta \in (0, m]$	$\min\{\langle \mu, a - a^*\rangle : a \in \mathcal{A} \setminus \{a^*\}\}$
$\Delta_{i,\min} \ge \Delta$	$\min\{\langle \mu, a - a^*\rangle : a \in \mathcal{A} \setminus \{a^*\}, a_i = 1\}$
$\Delta'_{i,\min} \ge \Delta$	$\min\{\langle \mu, a - a^*\rangle : a \in \mathcal{A} \setminus \{a^*\}, a_i = 0\}$
$w(\mathcal{A}) \le m$	Action-set-dependent constant (Section 5)
L^*	$\min_{a\in\mathcal{A}} \mathbb{E}[\sum_{t=1}^T \langle \ell(t), a\rangle]$
Q_2	$\mathbb{E}[\sum_{t=1}^T \|\ell(t) - \bar{\ell}\|^2]$ ($\bar{\ell} = T^{-1}\mathbb{E}[\sum_{t=1}^T \ell(t)]$)
V_1	$\mathbb{E}[\sum_{t=1}^{T-1} \|\ell(t) - \ell(t+1)\|_1]$
$C \in [0, T]$	$\mathbb{E}[\sum_{t=1}^T \|\ell(t) - \ell'(t)\|_\infty]$, corruption level

Contribution of This Study In this study, we establish a new BOBW algorithm for the combinatorial semi-bandit problem. The proposed algorithm is based on the Optimistic Follow-the-Regularized-Leader (OFTRL) framework (McMahan, 2011; Rakhlin and Sridharan, 2013b,a) with a refined regularizer and adaptive learning rate inspired by Ito et al. (2022a). Let $I^* = \{i \in [d] : a_i^* = 1\}$ and $J^* = [d] \setminus I^*$. OFTRL has a component called an optimistic prediction and the proposed algorithm considers two methods for its estimation: the Least Square (LS) and Gradient Descent (GD) based on past observations. Let $w(\mathcal{A}) \le m$ be an action-set-dependent constant defined in Section 5. We drop $\mathcal{A}$ when it is clear from context. The regret of the proposed algorithm with the LS and GD is then bounded as follows.

Theorem 1 (Informal)**.** For the stochastic regime, the pro-

Table 2: Regret upper bounds for combinatorial semi-bandits. $w = w(\mathcal{A}) \leq m$ is an action-set-dependent constant.

Reference	Regime	Regret bound
Audibert et al. (2014)	Adv.	$O(\sqrt{dmT})$
Kveton et al. (2015)	Stoc.	$534 \sum_{i \in J^*} \frac{m}{\Delta_{i,\min}} \log T + O(dm)$
Zimmert et al. (2019)	Adv.	$O(\sqrt{dmT})$
	Stoc.	$O\left(\frac{dm}{\Delta} \log T\right) =: \mathcal{R}^{\mathrm{ZLS}}$
	Stoc. w/ adv.	$\mathcal{R}^{\mathrm{ZLS}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{ZLS}}})$
Ito (2021a)	Adv.	$O(\sqrt{d \min\{L^*, Q_2, V_1\} \log T})$
	Stoc.	$O\left(\frac{dm}{\Delta} \log T\right) =: \mathcal{R}^{\mathrm{I}}$
	Stoc. w/ adv.	$\mathcal{R}^{\mathrm{I}} + \sqrt{Cm\mathcal{R}^{\mathrm{I}}}$
Proposed (LS)	Adv.	$\sqrt{4d \min\{L^*, mT - L^*, Q_2\} \log T}$
	Stoc.	$\left(\sum_{i \in J^*} \max\left\{\frac{4w\sigma_i^2}{\Delta_{i,\min}} + c \log\left(1 + \frac{w\sigma_i^2}{\Delta_{i,\min}}\right), 2(1+\epsilon)\right\} + O(\lvert I^* \rvert)\right) \log T =: \mathcal{R}^{\mathrm{LS}}$
	Stoc. w/ adv.	$\mathcal{R}^{\mathrm{LS}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{LS}}})$
Proposed (GD)	Adv.	$\sqrt{\frac{4d}{1-2\eta}\left(\min\left\{L^*, mT - L^*, Q_2, \frac{2V_1}{\eta}\right\} + \frac{d}{\eta}\right) \log T}$
	Stoc.	$\frac{1}{1-2\eta}\left(\sum_{i \in J^*} \max\left\{\frac{4w\sigma_i^2}{\Delta_{i,\min}} + c \log\left(1 + \frac{w\sigma_i^2}{\Delta_{i,\min}}\right), 2(1+\epsilon)\right\} + O(\lvert I^* \rvert)\right) \log T =: \mathcal{R}^{\mathrm{GD}}$
	Stoc. w/ adv.	$\mathcal{R}^{\mathrm{GD}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{GD}}})$

posed algorithm with LS achieves

$$R_T \leq \left(\sum_{i \in J^*} \max\left\{4\frac{w\sigma_i^2}{\Delta_{i,\min}} + c \log\left(1 + \frac{w\sigma_i^2}{\Delta_{i,\min}}\right), 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right) \log T + o(\log T) =: \mathcal{R}^{\mathrm{LS}},$$

where $\epsilon \in (0, 1/2]$ is an input parameter for the algorithm and $c = O((\log \epsilon^{-1})^2)$. Further, for the adversarial regime, the algorithm achieves

$$R_T \leq \sqrt{4d \min\{L^*, mT - L^*, Q_2\} \log T} + O(d \log T) + d^2 + d(1 + 2\delta).$$

Additionally, for the stochastic regime with adversarial corruptions, we have $R_T \leq \mathcal{R}^{\mathrm{LS}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{LS}}})$.

Theorem 2 (Informal)**.** For the stochastic regime, the proposed algorithm with GD estimations with a step size $\eta \in (0, 1/2)$ achieves

$$R_T \leq \frac{1}{1-2\eta}\left(\sum_{i \in J^*} \max\left\{4\frac{w\sigma_i^2}{\Delta_{i,\min}} + c \log\left(1 + \frac{w\sigma_i^2}{\Delta_{i,\min}}\right), 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right) \log T + o(\log T) =: \mathcal{R}^{\mathrm{GD}}.$$

For the adversarial regime, the algorithm achieves

$$R_T \leq \sqrt{\frac{4d}{1-2\eta}\left(\min\left\{L^*, mT - L^*, Q_2, \frac{2V_1}{\eta}\right\} + \frac{d}{\eta}\right) \log T} + O(d \log T) + d^2 + d(1 + 2\delta).$$

Additionally, for the stochastic regime with adversarial corruptions, we have $R_T \leq \mathcal{R}^{\mathrm{GD}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{GD}}})$.

A comparison with existing bounds is given in Section 5.

The proposed algorithm is inspired by the algorithm proposed by Ito (2021a); however, their bound depends on Δ and *not* either on σ_i^2 or on $\Delta_{i,\min}$. The proposed algorithm takes care of the characteristics of the instances, and specifically, we modify the regularizer and optimistic prediction in OFTRL and refine the analysis. As a result, the bounds of the proposed algorithm depend on σ_i^2 and $\Delta_{i,\min}$, and a leading constant of our bounds are at least 81 times better than their bound. The resulting regret upper bound in Theorem 1 is at most approximately twice as large as the achievable lower bounds (Section 5). Note that one can prove the same order of upper bounds as in Theorem 2 for the algorithm in Ito (2021a) by using the analysis given in Section 5. Table 2 lists the regret bounds provided in this study and summarizes comparisons with existing work.

Our regret bounds are favorable compared to those reported in existing studies in that enjoying following properties:

1. Our algorithm enjoys BOBW guarantees and works well even in the stochastic regime with adversarial corruptions.
2. The leading constant of the regret bound in Theorem 1 (resp. Theorem 2) for the stochastic regime is only twice (resp. $2/(1-\eta)$) as large as an achievable lower bound.
3. The regret bounds in the stochastic regime depend on the tighter suboptimality gap $\Delta_{i,\min}$ rather than the minimal suboptimality gap Δ.
4. The regret bounds in the stochastic regime depend on the variances of base-arms, which can be tremendously small value under certain practical scenarios.
5. The regret in the adversarial regime enjoys data-dependent regret bounds.

Note that the first and fifth properties are already realized in existing studies, (*e.g.,* Zimmert et al. 2019; Ito 2021a.) We consider using a self-bounding technique (Zimmert and Seldin, 2021) to obtain BOBW guarantees. In the self-bounding technique, we first derive upper and lower bounds of the regret using a variable depending on the (base-)arm selection probability, and we then derive a regret bound by combining the upper and lower bounds. For bounding the regret with the tight suboptimality gap $\Delta_{i,\min}$ in the stochastic regime, we derive a new regret lower bound.

To prove the variance-dependent regret upper bound, we consider an algorithm inspired by the learning rate and regularizer developed by Ito et al. (2022a), which focuses on the classical multi-armed bandit problem. However, their theoretical analysis is based on the fact that the sum of the arm selection probabilities equals 1, which does not hold in the semi-bandit problem. Our analysis uses a new approach to handle this problem by deriving a regret upper bound that collaborates well with the new regret lower bound.

Further, we empirically investigate the performance of the proposed algorithm, whereas experiments are often missing in studies on the BOBW algorithm such as Wei and Luo (2018), Lee et al. (2021), and Ito (2021a). The results of this study show that the proposed algorithm empirically works the best in the adversarial regime and as well as Thompson sampling in the practical stochastic regime.

2 RELATED WORK

György et al. (2007) and Uchiya et al. (2010) initiated research on the combinatorial semi-bandit problem for the adversarial regime, and since then, many algorithms with $O(\sqrt{T})$-regret bounds have been developed (*e.g.,* Neu and Bartók 2013; Audibert et al. 2014; Neu 2015; Wei and Luo 2018).

Combinatorial semi-bandits have been also investigated in the stochastic regime, and algorithms in the literature are significantly different from those in the adversarial regime. Most are based on *index-based approaches*, where the algorithm estimates the loss means for each base-arm and *pessimistically* predicts the true value of the losses. Kveton et al. (2015) and Wang and Chen (2018) prove gap-dependent regret bounds depending on $\Delta_{i,\min}$ rather than Δ, and they also consider special action sets such as the size-invariant and matroid semi-bandits.

Since the seminal study conducted by Bubeck and Slivkins (2012), BOBW algorithms have been developed for many online-decision making problems beyond the multi-armed bandits (Zimmert and Seldin, 2021; Seldin and Lugosi, 2017; Rouyer and Seldin, 2020; Huang et al., 2022): the problem of prediction with expert advice (Gaillard et al., 2014; Luo and Schapire, 2015), dueling bandits (Saha and Gaillard, 2022), online learning with feedback graphs (Erez and Koren, 2021; Ito et al., 2022b), linear bandits (Lee et al., 2021), and episodic Markov decision processes (Jin and Luo, 2020; Jin et al., 2021). For combinatorial semi-bandits, we are aware of the works by Wei and Luo (2018), Zimmert et al. (2019), and Ito (2021a).

3 PRELIMINARES

This section introduces the preliminaries for this study. Let $\|x\|$, $\|x\|_1$, and $\|x\|_\infty$ be the Euclidian, ℓ_1, and ℓ_∞-norms for vector x, respectively, and $\mathbf{1}$ be the all-one vector.

3.1 Combinatorial Semi-Bandits

We consider the combinatorial semi-bandit problem with action set $\mathcal{A} \subset \{0,1\}^d$, where each element $a \in \mathcal{A}$ is called an action. We assume that for all $i \in [d]$, there exists $a \in \mathcal{A}$ such that $a_i = 1$. Define $m = \max_{a\in\mathcal{A}} \|a\|_1$.

In the combinatorial semi-bandit problem, the learner observes entry-wise bandit feedback. At each step $t \in [T]$, when the learner takes action $a(t) \in \mathcal{A}$, they observe the elements in $I_t = \{i \in [d] : a_i(t) = 1\}$, whereas the elements in $J_t = [d] \setminus I_t$ are not observed. We assume that $T \geq \max\{d, 55\}$.

This study also considers the special cases of action sets: *size-invariant semi-bandits* and *matroid semi-bandits*. For size-invariant semi-bandits, the size of action $\|a\|_1$ is fixed to m, *i.e.,* $\mathcal{A} \subset \{a \in \{0,1\}^d : \|a\|_1 = m\}$. For the matroid semi-bandits, a special case of size-invariant semi-bandits, an action set $\mathcal{A}$ corresponds to the bases of a matroid. The well-known *m-set semi-bandits*, in which $\mathcal{A} = \{a \in \{0,1\}^d : \|a\|_1 = m\}$, is an example of the matroid semi-bandit problem.

In this study, we assume that there exists a unique optimal action $a^* \in \mathcal{A}$. This assumption has been employed by many studies aiming at the development of BOBW algo-

rithms (Gaillard et al., 2014; Luo and Schapire, 2015; Wei and Luo, 2018; Zimmert and Seldin, 2021).

3.2 Considered Regimes

We consider three regimes as the assumptions for the losses. In the *stochastic regime*, the loss vectors $(\ell(t))$ follow an unknown distribution $\mathcal{D}$ in an i.i.d. manner for all $t \in [T]$. We define the expectation of the losses by $\mu = \mathbb{E}_{\ell\sim\mathcal{D}}[\ell]$.

By contrast, the *adversarial regime* does not assume any stochastic structure for the losses and the losses can be chosen in an arbitrarily manner. In this regime, the environment can choose $\ell(t)$ depending on the past history until the $(t-1)$-th round, *i.e.,* $\{(\ell(s), a(s))\}_{s=1}^{t-1}$.

We also consider an intermediate regime between the stochastic and adversarial regimes. One of the most representative intermediate regimes is the *stochastic regime with adversarial corruptions*. In this regime, a temporary loss $\ell'(t) \in [0,1]^d$ is sampled from an unknown distribution $\mathcal{D}$, and then the adversary corrupts $\ell'(t)$ to $\ell(t)$. We define the corruption level by $C = \mathbb{E}\left[\sum_{t=1}^{T} \|\ell(t) - \ell'(t)\|_\infty\right] \geq 0$. If $C = 0$, this regime coincides with the stochastic regime, and if $C = T$, this regime corresponds to the adversarial regime. We will see that the proposed algorithm works without the knowledge of the corruption level C.

3.3 Optimistic Follow-the-Regularized-Leader

We establish the algorithm based on the *Optimistic follow-the-regularized-leader (OFTRL)* framework, which has occasionally been used in the development of BOBW algorithms (Wei and Luo, 2018; Ito, 2021b). Let $\mathcal{X} = \mathrm{conv}(\mathcal{A})$ be the convex hull of the action set $\mathcal{A}$. OFTRL maintains $x(t) \in \mathcal{X}$, and it then chooses $a(t) \in \mathcal{A}$ so that $\mathbb{E}[a(t)|x(t)] = x(t)$. The OFTRL update rule is expressed as

$$x(t) \in \operatorname*{arg\,min}_{x\in\mathcal{X}} \left\{ \left\langle m(t) + \sum_{s=1}^{t-1} \widehat{\ell}(s), x \right\rangle + \psi_t(x) \right\}, \quad (1)$$

where $m(t) \in [0,1]^d$ corresponds to an optimistic prediction (also known as a hint vector) of the true loss vector $\ell(t)$, the vector $\widehat{\ell}(t) \in \mathbb{R}^d$ is an unbiased estimator of $\ell(t)$, and ψ_t is a convex regularizer function over $\mathcal{X}$.

4 PROPOSED ALGORITHM

This section describes details of the proposed algorithm (Logarithmic Barrier Implicit Normalized Forecaster considering Variances for semi-bandits; `LBINFV`) by specifying the optimistic prediction $m(t)$, estimator $\widehat{\ell}(t)$, and convex regularization ψ_t in (1).

We consider two different methods for estimating optimistic predictions; these methods result in regret upper bounds that differ by a constant factor in the stochastic regime and have different data-dependent bounds. One method is a *least square* (LS) estimation based on the losses thus far, *i.e.,* we define $m(t) = (m_1(t), \ldots, m_d(t))^\top \in [0,1]^d$ by

$$m_i(t) = \frac{1}{1 + N_i(t-1)} \left(\frac{1}{2} + \sum_{s=1}^{t-1} a_i(s)\, \ell_i(s) \right), \quad (2)$$

where $N_i(t)$ is the number of times the base-arm i is chosen until the t-th round, *i.e.,* $N_i(t) = |\{s \in [t] : a_i(t) = 1\}|$. The other method is based on the *gradient descent* (GD), where we define $m(t)$ by $m_i(1) = 1/2$ and

$$m_i(t+1) = \begin{cases} (1-\eta) m_i(t) + \eta \ell_i(t) & \text{if } i \in I(t) \\ m_i(t) & \text{otherwise} \end{cases} \quad (3)$$

for $i \in [d]$ with a step size $\eta \in (0, 1/2)$.

Let $a(t) \in \mathcal{A}$ be an action selected at round t and $I(t) = \{i \in [d] : a_i(t) = 1\}$ be the set of base-arms selected at round t. Note that $\{a_i(t) = 1\}$ is equivalent to $\{i \in I(t)\}$ and $\Pr[i \in I(t)|x_i(t)] = \Pr[a_i(t) = 1|x_i(t)] = x_i(t)$.

The design of LS is to reduce the leading constant in the regret, and GD is to derive a path-length bound. LS was developed by Ito et al. (2022a). The original idea of GD comes from online learning literature (Herbster and Warmuth, 2001), and Ito (2021a) developed the idea in semi-bandits.

We use an unbiased estimator $\widehat{\ell}(t) = (\widehat{\ell}_1(t), \ldots, \widehat{\ell}_d(t))^\top \in \mathbb{R}^d$ of $\ell(t)$ given by

$$\widehat{\ell}_i(t) = m_i(t) + \frac{a_i(t)}{x_i(t)} (\ell_i(t) - m_i(t)) \quad (4)$$

for $i \in [d]$. This is indeed an unbiased estimator of $\ell(t)$ since $\mathbb{E}[\widehat{\ell}_i(t)|x(t)] = m_i(t) + \frac{x_i(t)}{x_i(t)}(\ell_i(t) - m_i(t)) = \ell_i(t)$. The optimistic prediction $m(t)$ in (4) plays a role in reducing the variance of $\widehat{\ell}(t)$; the better $m(t)$ predicts $\ell(t)$, the smaller the variance in $\widehat{\ell}(t)$ becomes.

The regularizer function $\psi_t : \mathbb{R}^d \to \mathbb{R}$ is given by

$$\psi_t(x) = \sum_{i=1}^{d} \beta_i(t) \varphi(x_i), \quad (5)$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is defined as

$$\varphi(z) = z - 1 - \log z + \gamma \left(z + (1-z)\log(1-z) \right) \quad (6)$$

with $\gamma = \log T$ and regularization parameters $\beta_i(t) \geq 0$. Our regularizer in (5) comprises the logarithmic barrier $-\log x_i$ and the (negative) Shannon entropy $(1 - x_i)\log(1 - x_i)$ for the complement of $x_i \in [0,1]$. Such a regularizer is called a *hybrid* regularizer, and this type of regularizer was employed in existing studies for bounding a component of the regret (Zimmert et al., 2019; Ito et al.,

Algorithm 1: LBINFV for semi-bandits

input: action set $\mathcal{A}$, time horizon T
for $t = 1, 2, \dots, T$ **do**
  Compute $x(t) \in \mathcal{X}$ by (1) with $\widehat{\ell}(t)$ in (4) and ψ_t in (5).
  Sample $a(t)$ such that $\mathbb{E}[a(t)|x(t)] = x(t)$.
  Take action $a(t)$ and observe feedback $\ell_i(t)$ for i such that $a_i(t) = 1$.
  Update the regularization parameters $\beta_i(t)$ in (7) and optimistic prediction $m_i(t)$ using (2) or (3).

2022a,b). The affine part of the regularizer in (6), $z-1+\gamma z$, is introduced to simplify the analysis and yields smaller constant factors, which is also used by Ito et al. (2022a).

Regularization parameters $\beta_i(t)$ are defined as

$$\beta_i(t) = \sqrt{(1+\epsilon)^2 + \frac{1}{\gamma}\sum_{s=1}^{t-1}\alpha_i(s)}\,, \tag{7}$$

where $\epsilon \in (0, 1/2]$ is an input parameter and

$$\alpha_i(t) = a_i(t)(\ell_i(t) - m_i(t))^2 \min\left\{1, \frac{2(1-x_i(t))}{x_i(t)^2\gamma}\right\}. \tag{8}$$

We design $\alpha_i(t)$ in (8) so that it corresponds to an upper bound of the component of regret, which appears when we use a standard analysis of (O)FTRL with regularizer (5). We can introduce a $2(1-x_i(t))/(x_i(t)^2\gamma)$ part in $\alpha_i(t)$ thanks to the Shannon entropy part in regularizer (5). This part allows us to bound the regret corresponding to optimal base-arms. The $(\ell_i(t) - m_i(t))^2$ part of $\alpha_i(t)$ comes from the use of optimistic predictions and can be related to the base-arm variances by using the LS and GD methods to estimate $m(t)$. Algorithm 1 summarizes the proposed algorithms.

From the intuitive viewpoint, $\alpha_i(t)$ determines the strength of the regularization, and as $\alpha_i(t)$ increases, the algorithm further explores base-arm i. Since $(\ell_i(t) - m_i(t))^2$ in (8) represents the squared error of the optimistic prediction, the algorithm becomes more explorative when the loss is unpredictable or has a high variance. Also note that $\mu_i \simeq 1$ corresponds to the base-arm with the almost worst expected loss with the least variance. The factor $(1 - x_i(t))$ in (8) contributes to a fast elimination of such a base-arm since the regularization does not become strong when $x_i(t) = 1$ is observed.

5 REGRET ANALYSIS

This section derives the regret upper bounds of the proposed algorithm. We define the minimum suboptimality gaps that contain and do not contain base-arm i by

$$\Delta_{i,\min} = \min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\},\, a_i = 1\}\,;$$
$$\Delta'_{i,\min} = \min\{\langle \mu, a - a^* \rangle : a \in \mathcal{A} \setminus \{a^*\},\, a_i = 0\}\,.$$

We define constants $v(\mathcal{A})$ and $w(\mathcal{A})$ depending on the action set $\mathcal{A}$ by

$$v(\mathcal{A}) = \begin{cases} 2 & \mathcal{A} \text{ is a matroid} \\ 2\min\{|I^*|, d-m\} & \text{otherwise} \end{cases} \tag{9}$$

and

$$w(\mathcal{A}) = \begin{cases} 2 & \mathcal{A} \text{ is a matroid} \\ 2\min\{m, d-m\} & \mathcal{A} \text{ is size-invariant} \\ 2\min\{m, |J^*|\} & \text{otherwise}\,. \end{cases} \tag{10}$$

5.1 Regret Upper Bounds

This section introduces regret upper bounds of the proposed algorithm for each optimistic prediction method.

Theorem 3 (Formal version of Theorem 1)**.** Consider Algorithm 1 using the least square method in (2) for optimistic predictions. Then, for the stochastic regime,

$$R_T \le \left(\sum_{i \in J^*} \max\left\{4\frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1 + \frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}}\right),\, 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T$$
$$+ O\left(\sum_{i \in I^*} \frac{v(\mathcal{A})}{\Delta'_{i,\min}}\sqrt{\log T}\right) + o(\sqrt{\log T})\,, \tag{11}$$

where $\epsilon \in (0, 1/2]$ is an input parameter for the algorithm and $c = O((\log \epsilon^{-1})^2)$. Further, for the adversarial regime,

$$R_T \le \sqrt{4d \min\{L^*, mT - L^*, Q_\infty\}\log T} + O(d\log T) + d^2 + d(1+2\delta)\,. \tag{12}$$

Additionally, in the stochastic regime with adversarial corruptions, we have $R_T \le \mathcal{R}^{\mathrm{LS}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{LS}}})$, where $\mathcal{R}^{\mathrm{LS}}$ is the RHS of (11).

Theorem 4 (Formal version of Theorem 2)**.** Consider Algorithm 1 using the gradient descent method with a step size $\eta \in (0, 1/2)$ in (3) for optimistic predictions. Then, for the stochastic regime,

$$R_T \le \frac{1}{1-2\eta}\left(\sum_{i \in J^*} \max\left\{4\frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1 + \frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}}\right),\, 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T + O\left(\left(\sum_{i \in I^*}\frac{v(\mathcal{A})}{\Delta'_{i,\min}} + \frac{d}{\sqrt{\eta(1-2\eta)}}\right)\sqrt{\log T}\right) + o(\sqrt{\log T})\,. \tag{13}$$

Further, for the adversarial regime,

$$R_T \leq \sqrt{\frac{4d}{1-2\eta}\left(\min\left\{L^*, mT-L^*, Q_2, \frac{2V_1}{\eta}\right\}+\frac{d}{\eta}\right)\log T} + O(d\log T) + d^2 + d(1+2\delta)\,. \tag{14}$$

Additionally, in the stochastic regime with adversarial corruptions, we have $R_T \leq \mathcal{R}^{\mathrm{GD}} + O(\sqrt{Cm\mathcal{R}^{\mathrm{GD}}})$, where $\mathcal{R}^{\mathrm{GD}}$ is the RHS of (13).

Note that the proposed algorithm does not require any prior knowledge on $\sigma_i^2, \Delta_i, L^*, Q_\infty$, and C. Theorem 4 indicates that the leading constant worsens by a factor of $1/(1-2\eta)$ in the stochastic regime compared to the bound in Theorem 3. This is at the expense of the path-length bound depending on V_1 in the adversarial regime.

5.2 Comparison with Existing Regret Bounds

The regret upper bounds for the stochastic regime in Theorems 3 and 4 improve on the existing regret upper bounds in three aspects: (i) dependence on the tight suboptimality gap $\Delta_{i,\min}$, (ii) the dependence on the variance of base-arms σ_i^2, and (iii) the leading constants particularly in the stochastic regime. For the suboptimality gap, our upper bounds are of the same order as the regret upper bound by Kveton et al. (2015), which is an algorithm specialized for the stochastic regime, and our bounds are up to d times better than the regret upper bounds by Zimmert et al. (2019) and Ito (2021a). For the variance dependency, in the stochastic regime, bounds in Theorems 3 and 4 improve the results in Ito (2021a) by replacing a constant in their bound with variance σ_i^2, which can be considerably small under certain practical scenarios such as ad allocations. Finally, it is worth noting that the leading constants are also significantly improved. The leading constant of our bounds are at least 81 times better than the bound by Ito (2021a). Moreover, the resulting regret upper bound in Theorem 3 and (resp. 4) are approximately at most twice (resp. $2/(1-2\eta)$) as large as the achievable lower bounds, which can be confirmed by comparing the bounds with Proposition 1 of Ito et al. (2022a).

5.3 Key Technique and Analysis

To obtain the regret bound depending on $\Delta_{i,\min}$ in the stochastic regime and the stochastic regime with adversarial corruptions, we prove the following regret *lower* bound.

Lemma 1. *In the stochastic regime with adversarial corruptions, for any algorithm and any action set $\mathcal{A}$, the regret is bounded from below as*

$$R_T \geq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{1}{v(\mathcal{A})}\sum_{i\in I^*}\Delta'_{i,\min}(1-a_i(t)) + \frac{1}{w(\mathcal{A})}\sum_{i\in J^*}\Delta_{i,\min}a_i(t)\right)\right] - 2Cm\,.$$

Note that if $a^* \in \mathcal{A}$ is unique, $i \in I^*$ implies that $\Delta'_{i,\min} > 0$, and $i \in J^*$ implies that $\Delta_{i,\min} > 0$. This regret lower bound improves Lemma 4 of Ito (2021a) for general action sets.

To prove the variance-dependent regret bounds, we make use of the learning rate inspired by Ito et al. (2022a), in which the classical multi-armed bandit problem is considered. However, their theoretical analysis is based on the fact that the sum of the arm selection probabilities equals 1, which does not hold in the semi-bandits. To handle this problem, we introduce a technique developed in Ito (2021a) and sophisticate the analysis to derive a regret upper bound that collaborates well with the regret lower bound.

In the following, we provide a sketch of analysis commonly used to prove Theorems 3 and 4, and see that that the regret lower bound in Lemma 1 indeed helps us obtain the desired regret bound. In the subsequent analysis, we will mainly focus on terms that are dominant for sufficiently large T, and will not include the other terms. Let $\gamma = \log T$. Using the similar analysis given by Ito et al. (2022a), we first show in Lemma 3 that the regret of the proposed algorithm is roughly bounded as

$$R_T = O\left(\gamma\sum_{i=1}^{d}\mathbb{E}\left[\beta_i(T+1)\right]\right) = O\left(\sum_{i=1}^{d}\sqrt{\mathbb{E}\left[\gamma\sum_{t=1}^{T}\alpha_i(t)\right]}\right).$$

Define (P_i) and (Q_i) by

$$P_i = \mathbb{E}\left[\sum_{t=1}^{T}x_i(t)\right], \quad Q_i = \mathbb{E}\left[\sum_{t=1}^{T}(1-x_i(t))\right],$$

which will be used in the self-bounding argument in the following. Using this and combining the analysis given by Ito et al. (2022a) and Ito (2021a), we can show that the regret is further bounded as

$$\frac{R_T}{\gamma} = O\left(\sum_{i\in J^*}\sqrt{\beta_0^2+\frac{\sigma_i^2P_i}{\gamma}} + \sum_{i^*\in I^*}\sqrt{\frac{Q_i}{\gamma^{3/2}}}\right). \tag{15}$$

For the stochastic regime, using the upper bound (15) and lower bound (Lemma 1 with $C=0$), the regret can be further roughly bounded as

$$\frac{R_T}{\gamma} = 2\frac{R_T}{\gamma} - \frac{R_T}{\gamma}$$

$$
\begin{aligned}
&\leq O\left(\sum_{i\in J^*}\sqrt{\beta_0^2+\frac{\sigma_i^2 P_i}{\gamma}}+\sum_{i\in I^*}\sqrt{\frac{Q_i}{\gamma^{3/2}}}\right)\\
&\quad-\frac{1}{\gamma}\left(\frac{1}{v(\mathcal{A})}\sum_{i\in I^*}\Delta'_{i,\min}Q_i+\frac{1}{w(\mathcal{A})}\sum_{i\in J^*}\Delta_{i,\min}P_i\right)\\
&=O\Bigg(\sum_{i\in J^*}\left(\sqrt{\beta_0^2+\frac{\sigma_i^2 P_i}{\gamma}}-\frac{\Delta_{i,\min}}{w(\mathcal{A})}\frac{P_i}{\gamma}\right)\\
&\qquad+\sum_{i\in I^*}\left(\sqrt{\frac{Q_i}{\gamma^{3/2}}}-\frac{\Delta'_{i,\min}}{v(\mathcal{A})}\frac{Q_i}{\gamma}\right)\Bigg)\\
&\leq O\left(\sum_{i\in J^*}\frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}}+|I^*|\frac{1}{\sqrt{\gamma}}\frac{v(\mathcal{A})}{\Delta'_{i,\min}}\right),
\end{aligned}
$$

where the first inequality follows by (15) and Lemma 1 with $C=0$, and in the last inequality we considered the worst case in terms of $(P_i)_{i\in J^*}$ and $(Q_i)_{i\in I^*}$. This result corresponds to the desired bounds in Theorems 3 and 4. A more complete and detailed analysis is given in the appendix.

6 EXPERIMENTS

This section presents the results of the numerical investigation of the empirical performance of the proposed LBINFV algorithm with $\epsilon=0.2$. The proposed algorithm with LS and GD (with $\eta=1/4$) estimations for the optimistic predictions are denoted by LBINFV-LS and LBINFV-GD, respectively. We use the following baselines. The algorithms for the stochastic regime are CombUCB1 (Kveton et al., 2015) and Thompson sampling (TS) (Komiyama et al., 2015; Wang and Chen, 2018). The algorithms with BOBW guarantees are HYBRID (Zimmert et al., 2019) and LBINF (Ito, 2021a).

To compare the performance, we consider the m-set semi-bandits with $T=10^4$. In the m-set semi-bandit setting, it is known that we can sample $a(t)$ satisfying $\mathbb{E}[a(t)|x(t)]=x(t)$ at an $O(d\log d)$ computational cost (Zimmert et al., 2019, Appendix B.2), and we employ this sampling technique. We repeat the simulations 20 times. The source code to reproduce all figures in this paper is available at https://github.com/tsuchhiii/bobw-variance.

6.1 Setup

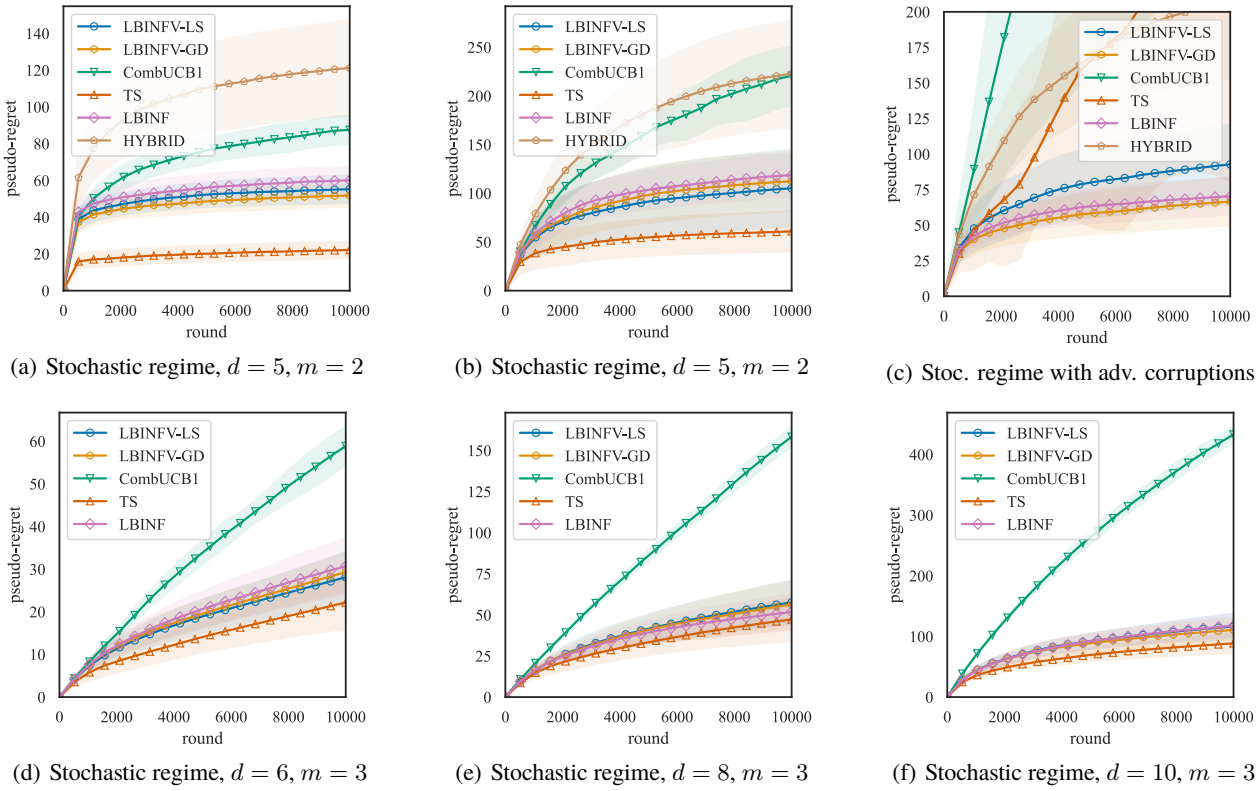
Synthetic data In the synthetic data experiments, we set $d=5$ and $m=2$ and consider the stochastic regime and stochastic regime with adversarial corruptions. In the stochastic regime, we consider two instances, where each base-arm is associated with a Bernoulli distribution. We set expectations μ for each instance to $(0.5,0.5,0.9,0.9,0.9)$ and $(0.5,0.5,0.6,0.6,0.6)$, respectively. In the stochastic regime with adversarial corruptions, we consider an instance considered by Zimmert et al. (2019). The environment alternates between two stochastic settings, (i) and (ii), where the losses are sampled from Bernoulli distributions with the following time-varying loss means. In setting (i), the expected losses are 0 for the optimal base-arms $i\in I^*$, and Δ' for the suboptimal base-arms $i\in J^*$. In setting (ii), the expected losses are $1-\Delta'$ for the optimal base-arms, and 1 for the suboptimal arms. We set $\Delta'=0.1$. The number of rounds between alternations increases exponentially with a factor of 1.6 after each alternation. Note that this instance also belongs to the stochastically constrained adversarial regime (Wei and Luo, 2018; Zimmert and Seldin, 2021).

Semi-synthetic data In semi-synthetic data experiments, we consider the stochastic regime. We used the KDD Cup 2012 track 2 dataset (Tencent Inc., 2012), which was used in the studies on multiple-play bandit problem (Komiyama et al., 2015; Lagrée et al., 2016; Komiyama et al., 2017), which is equivalent to the m-set semi-bandit problem. The dataset includes session logs of the Tencent search engine, soso.com. We use the estimated *reward* means of Komiyama et al. (2017) although the rewards therein are estimated under a different context, where the reward mean for base-arm i is defined by $1-\mu_i$ corresponding to the click-through rate for example. The details of the parameters are summarized in Table 3 in Appendix D. One characteristic of this type of dataset is that the reward mean for each base-arm is extremely small (smaller than 0.05 in most cases). Hence, each σ_i^2 is supposed to be extremely small, and algorithms with adaptivity to variances are desirable.

6.2 Numerical Results

Figure 1 shows an empirical comparison of the proposed algorithm against the baselines. The experimental results from the synthetic data in (a) and (b) indicate that the proposed LBINFV-LS and LBINFV-GD algorithms achieve the best performance in the stochastic regime, except for Thompson sampling. Further, under the setting in (a), where the variances of the base-arms are small, the proposed algorithm shows a significant improvement compared to HYBRID. Additionally, these figures also confirm that LBINFV-LS performs better in the stochastic regime than LBINF. This indicates that the modification of the regularizer and the optimistic prediction contribute not only to the better leading constant of the regret upper bound but also to the empirical performance.

The proposed algorithm achieves the best performance in the adversarial regime, whereas CombUCB1 and Thompson sampling highly degrade their performance. We can also see from (a) and (b) that the performance of LBINFV-GD becomes slightly worse than that of LBINF-LS in most cases, as suggested by the theoretical results, whereas in (c) the performance of LBINFV-GD is better than that of LBINFV-LS, which seemingly occurs because the adversarial instance in this experiment is



(a) Stochastic regime, $d = 5$, $m = 2$ (b) Stochastic regime, $d = 5$, $m = 2$ (c) Stoc. regime with adv. corruptions

(d) Stochastic regime, $d = 6$, $m = 3$ (e) Stochastic regime, $d = 8$, $m = 3$ (f) Stochastic regime, $d = 10$, $m = 3$

Figure 1: Regret-round plots of algorithms used for synthetic and semi-synthetic data. The solid lines indicate the average over 20 independent trials. The thin fillings represent the standard error.

a regime with a small path-length and the former algorithm has the path-length bound.

The experimental results using the semi-synthetic data in (d)–(f) indicate that `LBINFV-LS` and `LBINFV-GD` perform comparably well to `Thompson sampling`. These results can be attributed to the fact that the variance is small for semi-synthetic data. Furthermore, (d)–(f), where the variances of the base-arms are extremely small, indicates that `CombUCB1` performs significantly worse than the other variance-aware algorithms. This observation indicates the importance of variance-aware algorithms in practical applications.

7 CONCLUSION AND FUTURE WORK

This study considered the combinatorial semi-bandit problem and presented the new BOBW algorithm with various adaptive guarantees. The new algorithm enjoys a variance-dependent regret bound depending on the tight suboptimality gap with a good leading constant in the stochastic regime and multiple data-dependent regret bounds. We numerically investigated the performance of the proposed algorithm and confirmed that the proposed algorithm performs competitively to Thompson sampling and achieve the best results in the adversarial regime.

One limitation of the proposed algorithm lies in its computational complexity: (i) sampling action $a(t)$ based on $x(t)$ and (ii) efficiently computing $x(t)$ in (1). Limitation (i) has long been a problem in semi-bandits using the (O)FTRL framework. Although polynomial-time algorithms exist (*e.g.,* Schrijver 1998, Corollary 14.1g), they are not very practical. For limitation (ii), it is not easy to efficiently compute $x(t)$ in existing studies, where Shannon entropy regularization for $1 - x_i$ is combined with the typical regularizers. If we can safely remove the Shannon entropy regularization for $1 - x_i(t)$, $x(t)$ then has a closed form, and an analysis for such a variant is important future work.

Acknowledgements

TT was supported by JST, ACT-X Grant Number JPMJAX210E, Japan and JSPS, KAKENHI Grant Number JP21J21272, Japan. JH was supported by JSPS, KAKENHI Grant Number JP21K11747, Japan.

References

Audibert, J.-Y., Bubeck, S., and Lugosi, G. (2014). Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2007). Tun-

ing bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pages 150–165.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 42.1–42.23.

Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422. JCSS Special Issue: Cloud Computing 2011.

Chen, X., Xue, J., Lei, Z., Qian, X., and Ukkusuri, S. V. (2022). Online eco-routing for electric vehicles using combinatorial multi-armed bandit with estimated covariance. *Transportation Research Part D: Transport and Environment*, 111:103447.

Degenne, R. and Perchet, V. (2016). Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems*, volume 29, pages 2972–2980.

Erez, L. and Koren, T. (2021). Towards best-of-all-worlds online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 34, pages 28511–28521.

Gai, Y., Krishnamachari, B., and Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478.

Gaillard, P., Stoltz, G., and van Erven, T. (2014). A second-order bound with excess losses. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 176–196.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. (2007). The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(79):2369–2403.

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

Huang, J., Dai, Y., and Huang, L. (2022). Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9173–9200.

Ito, S. (2021a). Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 2654–2667.

Ito, S. (2021b). Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 2552–2583.

Ito, S., Tsuchiya, T., and Honda, J. (2022a). Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 1421–1422.

Ito, S., Tsuchiya, T., and Honda, J. (2022b). Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35.

Jin, T., Huang, L., and Luo, H. (2021). The best of both worlds: stochastic and adversarial episodic MDPs with unknown transition. In *Advances in Neural Information Processing Systems*, volume 34, pages 20491–20502.

Jin, T. and Luo, H. (2020). Simultaneously learning stochastic and adversarial episodic MDPs with known transition. In *Advances in Neural Information Processing Systems*, volume 33, pages 16557–16566.

Komiyama, J., Honda, J., and Nakagawa, H. (2015). Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1152–1161.

Komiyama, J., Honda, J., and Takeda, A. (2017). Position-based multiple-play bandit problem with unknown position bias. In *Advances in Neural Information Processing Systems*, volume 30, pages 4998–5008.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015). Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 535–543.

Lagrée, P., Vernade, C., and Cappe, O. (2016). Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems*, volume 29, pages 1597–1605.

Lee, C.-W., Luo, H., Wei, C.-Y., Zhang, M., and Zhang, X. (2021). Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 6142–6151.

Liu, X., Zuo, J., Wang, S., Joe-Wong, C., Lui, J., and Chen, W. (2022). Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms. In *Advances in Neural Information Processing Systems*.

Luo, H. and Schapire, R. E. (2015). Achieving all with no parameters: AdaNormalHedge. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1286–1304.

Lykouris, T., Mirrokni, V., and Paes Leme, R. (2018). Stochastic bandits robust to adversarial corruptions. In

Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pages 114–122.

McMahan, B. (2011). Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 525–533.

Merlis, N. and Mannor, S. (2019). Batch-size independent regret bounds for the combinatorial multi-armed bandit problem. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2465–2489. PMLR.

Neu, G. (2015). First-order regret bounds for combinatorial semi-bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1360–1375.

Neu, G. and Bartók, G. (2013). An efficient algorithm for learning with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 234–248.

Perrault, P., Valko, M., and Perchet, V. (2020). Covariance-adapting algorithm for semi-bandits with application to sparse outcomes. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3152–3184.

Qin, L., Chen, S., and Zhu, X. (2014). Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469.

Rakhlin, A. and Sridharan, K. (2013a). Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 993–1019.

Rakhlin, S. and Sridharan, K. (2013b). Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, volume 26, pages 3066–3074.

Rouyer, C. and Seldin, Y. (2020). Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 3227–3249.

Saha, A. and Gaillard, P. (2022). Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 19011–19026.

Schrijver, A. (1998). *Theory of linear and integer programming*. John Wiley & Sons.

Seldin, Y. and Lugosi, G. (2017). An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1743–1759.

Tencent Inc. (2012). KDD Cup - 2012 track 2, Kaggle.

Uchiya, T., Nakamura, A., and Kudo, M. (2010). Algorithms for adversarial bandit problems with multiple plays. In *Algorithmic Learning Theory*, pages 375–389.

ul Hassan, U. and Curry, E. (2016). Efficient task assignment for spatial crowdsourcing: A combinatorial fractional optimization approach with semi-bandit learning. *Expert Systems with Applications*, 58:36–56.

Vial, D., Sanghavi, S., Shakkottai, S., and Srikant, R. (2022). Minimax regret for cascading bandits. In *Advances in Neural Information Processing Systems*.

Wang, S. and Chen, W. (2018). Thompson sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5114–5122.

Wei, C.-Y. and Luo, H. (2018). More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 1263–1291.

Zimmert, J., Luo, H., and Wei, C.-Y. (2019). Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7683–7692.

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

Further Adaptive Best-of-Both-Worlds Algorithm for Combinatorial Semi-Bandits : Supplementary Materials

A COMMON ANALYSIS

A.1 General Regret Upper Bound

Define $\beta_0 = 1 + \epsilon$. Let D_t be the Bregman divergence induced by ψ_t, *i.e.,*

$$D_t(y, x) = \psi_t(y) - \psi_t(x) - \langle \nabla \psi_t(x), y - x \rangle \ .$$

Then, the regret for OFTRL is bounded as follows.

Lemma 2 (Lemma 2 of Ito et al. 2022a)**.** *If $x(t)$ is given by the OFTRL update (1), for any $x^* \in \mathcal{X} \cap \mathbb{R}_+^d$ we have*

$$\sum_{t=1}^{T} \left\langle \widehat{\ell}(t), x(t) - x^* \right\rangle \leq \underbrace{\psi_{T+1}(x^*) - \psi_1(y(1)) + \sum_{t=1}^{T} (\psi_t(x(t+1)) - \psi_{t+1}(x(t+1)))}_{\textit{penalty term}} + \underbrace{\sum_{t=1}^{T} \left(\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1) \right\rangle - D_t(y(t+1), x(t)) \right)}_{\textit{stability term}}, \tag{16}$$

where we define $y(t) \in \arg\min_{x \in \mathcal{X}} \left\{ \left\langle \sum_{s=1}^{t-1} \widehat{\ell}(s), x \right\rangle + \psi_t(x) \right\}$.

In the RHS of the above inequality (16), we refer to the sum of the first three terms as the *penalty term* and the remaining term as the *stability term*.

First, we prove the following lemma.

Lemma 3. *The regret of the proposed algorithm is bounded as*

$$R_T \leq \gamma \sum_{i=1}^{d} \mathbb{E} \left[2\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)} \right] + d^2 + d(1 + 2\delta) \,, \tag{17}$$

where $\delta > 0$ is defined by

$$\delta = (1+\epsilon)^3 \log \frac{1+\epsilon}{\epsilon} - (1+\epsilon)^2 - \frac{1+\epsilon}{2} \leq \frac{27}{8} \log \frac{3}{2\epsilon} - \frac{3}{2} = O\left(\log \frac{1}{\epsilon} \right) .$$

Proof. Using $x_0 \in \mathcal{X}$ such that $(x_0)_i \geq 1/d$ for all $i \in [d]$, let

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

Using this and the equality $\mathbb{E}[\widehat{\ell} | x_t] = \ell$, we have

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell(t), x(t) - a^* \rangle \right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle \ell(t), x(t) - x^* \rangle + \sum_{t=1}^{T} \langle \ell(t), x^* - a^* \rangle \right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}(t), x(t) - x^* \right\rangle + \frac{d}{T} \sum_{t=1}^{T} \langle \ell(t), x_0 - a^* \rangle \right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{\ell}(t), x(t) - x^* \right\rangle \right] + d^2 \,, \end{aligned} \tag{18}$$

where in the last inequality we used $\sum_{t=1}^{T} \langle \ell(t), x_0 - a^* \rangle \leq T\|x_0 - a^*\|_1 \leq Td$.

The first term in (18) is bounded by (16) in Lemma 2, the components of which we will bound in the following. We first consider the penalty term. The remaining part of the proof follows a similar argument as that in Ito et al. (2022a), and we include the argument for completeness.

Bounding the penalty term in (16) Using the definition of the regularizer $\psi_t(x) = \sum_{i=1}^{d} \beta_i(t)\varphi(p_i)$ defined in (5), we have

$$\psi_t(x^*) = \sum_{i=1}^{d} \beta_i(t)\varphi(x_i^*) \leq \sum_{i=1}^{d} \beta_i(t) \max_{x\in[1/T,1]} \varphi(x) \leq \sum_{i=1}^{d} \beta_i(t) \max\{\varphi(1/T), \varphi(1)\}\,, \tag{19}$$

where the first inequality follows since the definition of x^* implies $x_i^* \geq \frac{d}{T}(x_0)_i \geq 1/T$ for $i \in [d]$ and the second inequality holds since φ is a convex function. Further, from the definition of φ in (6), we have

$$\begin{aligned}\max\{\varphi(1/T), \varphi(1)\} &= \max\left\{\frac{1}{T} - 1 + \log T + \gamma\left(\frac{1}{T} + \left(1 - \frac{1}{T}\right)\log\left(1 - \frac{1}{T}\right)\right), \gamma\right\} \\ &\leq \max\left\{\frac{1+\gamma}{T} - 1 + \log T, \gamma\right\} = \gamma\,,\end{aligned}$$

where the last inequality follows from $\gamma = \log T$. From this and (19), we have

$$\psi_{T+1}(x^*) \leq \gamma \sum_{i=1}^{d} \beta_i(T+1)\,. \tag{20}$$

Further, as we have $\beta_i(t) \leq \beta_i(t+1)$ from (7) and $\varphi(x) \geq 0$ for any $x \in (0, 1]$, we have

$$\begin{aligned}&-\psi_1(y(1)) + \sum_{t=1}^{T}(\psi_t(y(t+1)) - \psi_{t+1}(y(t+1))) \\ &= -\sum_{i=1}^{d}\left(\beta_i(1)\varphi(y_i(1)) + \sum_{t=1}^{T}(\beta_i(t+1) - \beta_i(t))\varphi(y_i(t+1))\right) \leq 0\,.\end{aligned} \tag{21}$$

Combining (20) and (21), we can bound the penalty term in (16) as

$$\psi_{T+1}(x^*) - \psi_1(y(1)) + \sum_{t=1}^{T}(\psi_t(y(t+1)) - \psi_{t+1}(y(t+1))) \leq \gamma \sum_{i=1}^{d} \beta_i(T+1)\,. \tag{22}$$

Bounding the stability term in (16) The Bregman divergence $D_t(x, y)$ is expressed as

$$\begin{aligned}D_t(x, y) &= \sum_{i=1}^{d}\left(\beta_i(t)D^{(1)}(x_i, y_i) + \beta_i(t)\gamma D^{(2)}(x_i, y_i)\right) \\ &\geq \sum_{i=1}^{d} \max\left\{\beta_i(t)D^{(1)}(x_i, y_i), \beta_i(t)\gamma D^{(2)}(x_i, y_i)\right\}\,,\end{aligned}$$

where $D^{(1)}$ and $D^{(2)}$ are Bregman divergences induced by $\varphi^{(1)}(x) = -\log x$ and $\varphi^{(2)}(x) = (1-x)\log(1-x)$, respectively. We hence have

$$\begin{aligned}&\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1)\right\rangle - D_t(y(t+1), x(t)) \\ &\leq \sum_{i=1}^{d}\left((\widehat{\ell}_i(t) - m_i(t))(x_i(t) - y_i(t+1)) - \beta_i(t)\max\left\{D^{(1)}(y_i(t+1), x_i(t)), \gamma D^{(2)}(y_i(t+1), x_i(t))\right\}\right) \\ &\leq \sum_{i=1}^{d}\left(\min\left\{\beta_i(t)g\left(\frac{p_i(t)(\widehat{\ell}_i(t) - m_i(t))}{\beta_i(t)}\right), \beta_i(t)\gamma(1 - x_i(t))h\left(\frac{\widehat{\ell}_i(t) - m_i(t)}{\gamma\beta_i(t)}\right)\right\}\right)\,,\end{aligned} \tag{23}$$

where the last inequality follows from Lemma 5 of Ito et al. (2022a), and g and h are defined as

$$g(x) = x - \log(x+1) \leq \frac{1}{2}x^2 + \delta|x|^3 \quad \left(x \geq -\frac{1}{\beta_0}\right), \tag{24}$$

$$h(x) = \exp(x) - x - 1 \leq x^2 \quad (x \leq 1). \tag{25}$$

Note that $g(0) = h(0) = 0$ and it holds from (4) that

$$\widehat{\ell}_j(t) - m_j(t) = \begin{cases} (\ell_j(t) - m_j(t))/x_j(t) & \text{if } j \in I(t) \\ 0 & \text{otherwise}. \end{cases} \tag{26}$$

Therefore, the LHS of (23) is further bounded as

$$\begin{aligned}
&\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1) \right\rangle - D_t(y(t+1), x(t)) \\
&\leq \sum_{j \in I(t)} \min\left\{ \beta_j(t)\, g\left(\frac{\ell_j(t) - m_j(t)}{\beta_j(t)}\right), \beta_j(t)\gamma(1-x_j(t))\, h\left(\frac{\ell_j(t) - m_j(t)}{\gamma\beta_j(t)x_j(t)}\right) \right\} \\
&\leq \begin{cases} \sum_{j \in I(t)} \left(\frac{(\ell_j(t) - m_j(t))^2}{2\beta_j(t)} + \frac{\delta|\ell_j(t) - m_j(t)|^3}{\beta_j(t)^2} \right) & \text{if } \gamma x_j(t) \leq 1 \\ \sum_{j \in I(t)} \min\left\{ \frac{(\ell_j(t) - m_j(t))^2}{2\beta_j(t)} + \frac{\delta|\ell_j(t) - m_j(t)|^3}{\beta_j(t)^2}, \frac{(1-x_j(t))(\ell_j(t) - m_j(t))^2}{\gamma x_j(t)^2 \beta_j(t)} \right\} & \text{otherwise} \end{cases} \\
&\leq \sum_{j \in I(t)} \min\left\{ \frac{(\ell_j(t) - m_j(t))^2}{2\beta_j(t)} + \frac{\delta|\ell_j(t) - m_j(t)|^3}{\beta_j(t)^2}, \frac{(1-x_j(t))(\ell_j(t) - m_j(t))^2}{\gamma x_j(t)^2 \beta_j(t)} \right\} \\
&\leq \sum_{j \in I(t)} \left(\frac{1}{2\beta_j(t)} + \frac{\delta}{\beta_j(t)^2} \right) (\ell_j(t) - m_j(t))^2 \min\left\{ 1, \frac{2(1-x_j(t))}{\gamma x_j(t)^2} \right\} = \sum_{i=1}^{d} \left(\frac{1}{2\beta_i(t)} + \frac{\delta}{\beta_i(t)^2} \right) \alpha_i(t),
\end{aligned} \tag{27}$$

where the first inequality follows from (23) and (26), the second inequality follows from (24), (25), and the fact that $|\frac{\ell_j(t) - m_j(t)}{\beta_j(t)}| \leq \frac{1}{\beta_0} \leq 1$, and the third inequality holds since $\gamma x_j(t) \leq 1$ means $\frac{1-x_j(t)}{\gamma x_j(t)^2} \geq \frac{1-1/\gamma}{\gamma(1/\gamma)^2} = \gamma - 1 \geq \frac{1}{2} + \delta$, which implies

$$\frac{(\ell_j(t) - m_j(t))^2}{2\beta_j(t)} + \frac{\delta|\ell_j(t) - m_j(t)|^3}{\beta_j(t)^2} \leq \frac{(1-x_j(t))(\ell_j(t) - m_j(t))^2}{\gamma x_j(t)^2 \beta_j(t)}.$$

We hence have

$$\sum_{t=1}^{T} \left(\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1) \right\rangle - D_t(y(t+1), x(t)) \right) \leq \sum_{i=1}^{d} \sum_{t=1}^{T} \left(\frac{1}{2\beta_i(t)} + \frac{\delta}{\beta_i(t)^2} \right) \alpha_i(t). \tag{28}$$

We can show that a part of (28) is bounded as

$$\sum_{t=1}^{T} \frac{\alpha_i(t)}{2\beta_i(t)} \leq \gamma \left(\sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{t=1}^{T} \alpha_i(t)} - \sqrt{\beta_0^2 - \frac{1}{\gamma}} \right) \leq \gamma \left(\beta_i(T+1) - \beta_0 \right). \tag{29}$$

The first inequality in (29) holds since

$$\begin{aligned}
&\sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{s=1}^{t} \alpha_i(s)} - \sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_i(s)} \\
&= \frac{\alpha_i(t)}{\gamma \left(\sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{s=1}^{t} \alpha_i(s)} + \sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_i(s)} \right)} \geq \frac{\alpha_i(t)}{2\gamma \sqrt{\beta_0^2 + \frac{1}{\gamma} \sum_{s=1}^{t-1} \alpha_i(s)}} = \frac{\alpha_i(t)}{2\gamma\beta_i(t)},
\end{aligned}$$

where the inequality follows by $\alpha_i(t) \leq 1$. The second inequality in (29) follows from

$$\sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{t=1}^{T} \alpha_i(t)} - \sqrt{\beta_0^2 - \frac{1}{\gamma}} \leq \sqrt{\beta_0^2 - \frac{1}{\gamma} + \frac{1}{\gamma} \sum_{t=1}^{T} \alpha_i(t)} - \beta_0 + \frac{1}{\gamma} \leq \beta_i(T+1) - \beta_0 + \frac{1}{\gamma},$$

where the first inequality follows from $\sqrt{x} - \sqrt{x-y} \leq y/\sqrt{x}$ for $x \geq y \geq 0$ and $\beta_0 \geq 1$. Similarly, we can show

$$\sum_{t=1}^{T} \frac{\alpha_i(t)}{\beta_i(t)^2} = \sum_{t=1}^{T} \frac{\alpha_i(t)}{\beta_0^2 + \frac{1}{\gamma}\sum_{s=1}^{t-1}\alpha_i(s)} = \gamma \sum_{t=1}^{T} \frac{\alpha_i(t)}{\gamma\beta_0^2 + \sum_{s=1}^{t-1}\alpha_i(s)}$$
$$\leq \gamma \log\left(1 + \frac{1}{\gamma\beta_0^2 - 1}\sum_{t=1}^{T}\alpha_i(t)\right) \leq 2\gamma \log \frac{\beta_i(T+1)}{\beta_i(1)} + 2\,. \tag{30}$$

The first inequality in (30) follows since

$$\log\left(1 + \frac{1}{\gamma\beta_0^2 - 1}\sum_{s=1}^{t}\alpha_i(s)\right) - \log\left(1 + \frac{1}{\gamma\beta_0^2 - 1}\sum_{s=1}^{t-1}\alpha_i(s)\right)$$
$$= -\log\left(1 - \frac{\alpha_i(t)}{\gamma\beta_0^2 - 1 + \sum_{s=1}^{t}\alpha_i(s)}\right) \geq -\log\left(1 - \frac{\alpha_i(t)}{\gamma\beta_0^2 + \sum_{s=1}^{t-1}\alpha_i(s)}\right) \geq \frac{\alpha_i(t)}{\gamma\beta_0^2 + \sum_{s=1}^{t-1}\alpha_i(s)}\,,$$

where the first inequality follows from $\alpha_i(t) \leq 1$ and the last inequality follows from $-\log(1-x) \geq x$ for $x < 1$. The second inequality in (30) follows from

$$\log\left(1 + \frac{1}{\gamma{\beta_0}^2 - 1}\sum_{t=1}^{T}\alpha_i(t)\right) < \log\left(1 + \frac{1}{\gamma{\beta_0}^2}\sum_{t=1}^{T}\alpha_i(t)\right) + \log\frac{\gamma{\beta_0}^2}{\gamma\beta_0^2 - 1}$$
$$= \log\left(\frac{\beta_i(T+1)^2}{\beta_0^2}\right) + \log\left(1 + \frac{1}{\gamma\beta_0^2 - 1}\right) \leq 2\log\frac{\beta_i(T+1)}{\beta_0} + \frac{2}{\gamma}\,,$$

where the last inequality follows from $\log(1+1/(x-1)) \geq 2/x$ for $x \geq 3/2$. Bounding the RHS of (27) with (29) and (30) yields

$$\sum_{t=1}^{T}\left(\left\langle \widehat{\ell}(t) - m(t), x(t) - y(t+1)\right\rangle - D_t(y(t+1), x(t))\right)$$
$$\leq \gamma \sum_{i=1}^{d}\left(\beta_i(T+1) - \beta_i(1) + 2\delta\log\frac{\beta_i(T+1)}{\beta_i(1)}\right) + d(1+2\delta)\,. \tag{31}$$

Finally, by bounding the RHS of (16) by sequentially using (18), (22) and (31), we have

$$R_T \leq \gamma\sum_{i=1}^{d}\mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta\log\frac{\beta_i(T+1)}{\beta_i(1)}\right] + d^2 + d(1+2\delta)\,,$$

which completes the proof. □

A.2 Proof of Lemma 1

Proof. We can bound the regret from below as

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell(t), a(t) - a^*\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell'_t, a(t) - a^*\rangle + \sum_{t=1}^{T}\langle \ell(t) - \ell'_t, a(t) - a^*\rangle\right]$$
$$\geq \mathbb{E}\left[\sum_{t=1}^{T}\langle \mu, a(t) - a^*\rangle - \sum_{t=1}^{T}\|\ell(t) - \ell'_t\|_\infty\|a(t) - a^*\|_1\right]$$
$$\geq \mathbb{E}\left[\sum_{t=1}^{T}\langle \mu, a(t) - a^*\rangle - 2m\sum_{t=1}^{T}\|\ell(t) - \ell'_t\|_\infty\right]$$
$$\geq \mathbb{E}\left[\sum_{t=1}^{T}\langle \mu, a(t) - a^*\rangle\right] - 2mC\,, \tag{32}$$

where the first inequality follows from the Hölder's inequality and $\mathbb{E}[\ell'_t] = \mu$, the second inequality follows since $\|a(t) - a^*\|_1 \leq 2m$, and the last inequality follows from the definition of $C = \sum_{t=1}^T \|\ell(t) - \ell'_t\|_\infty$. We then bound $\mathbb{E}\left[\sum_{t=1}^T \langle \mu, a(t) - a^* \rangle\right]$.

We consider the case of general action sets and recall that $I^* = \{i \in [d] : a_i^* = 1\}$ and $J^* = [d] \setminus I^*$. Since $\sum_{i \in I^*}(1 - a_i(t)) \leq \min\{|I^*|, d - m\}$ and $\sum_{i \in J^*} a_i(t) \leq \min\{|J^*|, m\}$, we have

$$\begin{aligned}
\langle \mu, a(t) - a^* \rangle &= \frac{1}{2} \langle \mu, a(t) - a^* \rangle + \frac{1}{2} \langle \mu, a(t) - a^* \rangle \\
&\geq \frac{1}{2 \min\{|I^*|, d - m\}} \sum_{i \in I^*} (1 - a_i(t)) \langle \mu, a(t) - a^* \rangle + \frac{1}{2 \min\{|J^*|, m\}} \sum_{i \in J^*} a_i(t) \langle \mu, a(t) - a^* \rangle \\
&\geq \frac{1}{2 \min\{|I^*|, d - m\}} \sum_{i \in I^*} \Delta'_{i,\min}(1 - a_i(t)) + \frac{1}{2 \min\{m, |J^*|\}} \sum_{i \in J^*} \Delta_{i,\min} a_i(t) \,,
\end{aligned}$$

where the last inequality follows since for any $i \in I^*$ we have $\langle \mu, a(t) - a^* \rangle \geq \Delta'_{i,\min}$, and for any $i \in J^*$ we have $\langle \mu, a(t) - a^* \rangle \geq \Delta_{i,\min}$. Combining this inequality with (32) completes the proof. □

Note that in the stochastic regime with adversarial corruptions, from Lemma 1 it holds that

$$\begin{aligned}
R_T &\geq \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{v(\mathcal{A})} \sum_{i \in I^*} \Delta'_{i,\min}(1 - a_i(t)) + \frac{1}{w(\mathcal{A})} \sum_{i \in J^*} \Delta_{i,\min} a_i(t) \right)\right] - 2Cm \\
&= \frac{1}{v(\mathcal{A})} \sum_{i \in I^*} \Delta'_{i,\min} Q_i + \frac{1}{w(\mathcal{A})} \sum_{i \in J^*} \Delta_{i,\min} P_i - 2Cm \,, \qquad (33)
\end{aligned}$$

where the equality follows from the law of iterated expectations.

B PROOF OF THEOREM 3

B.1 Preliminaries

Before proving the regret upper bounds in Theorem 3, we prepare some lemmas. We bound the sum over $i \in [d]$ in (17) by considering different upper bounds for the optimal and sub-optimal base-arms. Recall that $\alpha_i(t)$ and $m_i(t)$ are given by (7) and (2), respectively. We use a following lemma to bound $\sum_{t=1}^T \alpha_i(t)$ for sub-optimal base-arms $i \in J^*$.

Lemma 4. *It holds for any $i \in [d]$ and $m_i^* \in [0, 1]$ that*

$$\sum_{t=1}^T \alpha_i(t) \leq \sum_{t=1}^T a_i(t)(\ell_i(t) - m_i(t))^2 \leq \sum_{t=1}^T a_i(t)(\ell_i(t) - m_i^*)^2 + \log(1 + N_i(T)) + \frac{5}{4} \,.$$

To prove this lemma, we use the following lemma.

Lemma 5 (Lemma 8 of Ito et al. 2022a)**.** *Suppose $\ell(s) \in [0, 1]$ for any $s \in [t]$ and define $m(t) \in [0, 1]$ by*

$$m(t) = \frac{1}{t}\left(\frac{1}{2} + \sum_{s=1}^{t-1} \ell(s)\right) .$$

Then, for any $m^* \in [0, 1]$ we have

$$\sum_{t=1}^T ((\ell(t) - m(t))^2 - (\ell(t) - m^*)^2) \leq \frac{5}{4} + \log T \,.$$

Proof of Lemma 4. From the definition of $\alpha_i(t)$, we have

$$\sum_{t=1}^T \alpha_i(t) \leq \sum_{t=1}^T a_i(t)(\ell_i(t) - m_i(t))^2$$

$$\leq \sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i^*)^2 + \frac{5}{4} + \log\left(1 + \sum_{t=1}^{T} a_i(t)\right)$$
$$= \sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i^*)^2 + \frac{5}{4} + \log\left(1 + N_i(T)\right) ,$$

where the second inequality follows from Lemma 5 and the definition of $m_i(t)$ given in (2). □

From Lemma 4, in the stochastic regime it holds that

$$\mathbb{E}\left[\sum_{t=1}^{T} \alpha_i(t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} x_i(t)\sigma_i^2 + \log(1 + N_i(T))\right] + \frac{5}{4} \leq \sigma_i^2 P_i + \log(1 + P_i) + \frac{5}{4} , \tag{34}$$

where the first inequality follows from Lemma 4 with $m_i^* = \mu_i$ and in the last inequality we define the expected number of times that the base-arm i is chosen by

$$P_i = \mathbb{E}\left[N_i(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[i \in I(t)\right]\right] = \mathbb{E}\left[\sum_{t=1}^{T} a_i(t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} x_i(t)\right] . \tag{35}$$

On the other hand, for the analysis of the optimal base-arms $i^* \in I^*$, we give a bound on $\sum_{t=1}^{T} \alpha_i(t)$ using the following lemma.

Lemma 6. *It holds for any $i^* \in [d]$ that*

$$\mathbb{E}[\alpha_{i^*}(t)] \leq 2\mathbb{E}\left[\min\left\{x_{i^*}(t), \frac{1 - x_{i^*}(t)}{\sqrt{\gamma}}\right\}\right] \leq 2\mathbb{E}\left[\frac{1 - x_{i^*}(t)}{\sqrt{\gamma}}\right] .$$

Proof. From the definition of $\alpha_i(t)$ in (7), we have

$$\mathbb{E}[\alpha_i(t)|x_i(t)] = \mathbb{E}\left[a_i(t)(\ell_i(t) - m_i(t))^2 \min\left\{1, \frac{2(1 - x_i(t))}{\gamma x_i(t)^2}\right\} \,\middle|\, x_i(t)\right]$$
$$\leq \mathbb{E}\left[a_i(t) \min\left\{1, \frac{2(1 - x_i(t))}{\gamma x_i(t)^2}\right\} \,\middle|\, x_i(t)\right]$$
$$= \min\left\{x_i(t), \frac{2(1 - x_i(t))}{\gamma x_i(t)}\right\}$$
$$\leq \begin{cases} x_i(t) & (x_i(t) < \frac{1}{\sqrt{\gamma}}) \\ \frac{2(1-x_i(t))}{\sqrt{\gamma}} & (x_i(t) \geq \frac{1}{\sqrt{\gamma}}) \end{cases} \leq \frac{2}{\sqrt{\gamma}}(1 - x_i(t)) ,$$

where the first inequality follows from the condition of $\ell_i(t), m_i(t) \in [0, 1]$ and the last inequality is due to $\sqrt{\gamma} \geq 2$ that follows from the assumption of $T \geq 55$. □

B.2 Proof for the Stochastic Regime

Proof of (11) in Theorem 3. We bound the RHS of (17) separately considering sub-optimal and optimal base-arms.

Sub-optimal base-arms side First, we let $i \in J^*$ be a sub-optimal base-arm. From (34), the component of the RHS of (17) is bounded as

$$\mathbb{E}\left[2\beta_i(T + 1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T + 1)}{\beta_i(1)}\right]$$
$$= \mathbb{E}\left[2\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T} \alpha_i(t)} - \beta_0 + \delta \log\left(1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T} \alpha_i(t)\right)\right]$$
$$\leq 2\sqrt{\beta_0^2 + \frac{1}{\gamma}\left(\sigma_i^2 P_i + \log(1 + P_i) + \frac{5}{4}\right)} - \beta_0 + \delta \log\left(1 + \frac{1}{\gamma\beta_0^2}\left(\sigma_i^2 P_i + \log(1 + P_i) + \frac{5}{4}\right)\right)$$

$$\leq 2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} + \frac{1}{\gamma\beta_0}\left(\log(1+P_i) + \frac{5}{4}\right) - \beta_0 + \delta\log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\delta}{\gamma\beta_0^2}\left(\log(1+P_i) + \frac{5}{4}\right)$$
$$= 2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \beta_0 + \delta\log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\xi}{\gamma}\left(\log(1+P_i) + \frac{5}{4}\right), \tag{36}$$

where the first inequality follows from (34), the second inequality follows from $\sqrt{x+y} \leq \sqrt{x} + \frac{y}{2\sqrt{x}}$ that holds for any $x > 0$ and $y \geq 0$, $\log(1+x+y) \leq \log(1+x) + y$ that holds for any $x, y \geq 0$, and in the last equality we define $\xi = \frac{1}{\beta_0} + \frac{\delta}{\beta_0^2} = \frac{1}{1+\epsilon} + \frac{\delta}{(1+\epsilon)^2}$.

Optimal base-arm side Next, we let $i \in I^*$ be an optimal base-arm. We define the complement version of P_i by

$$Q_i = \mathbb{E}\left[\sum_{t=1}^{T}(1 - x_i(t))\right] \tag{37}$$

for $i \in [d]$. Then from Lemma 6 we have

$$\begin{aligned}
&\mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta\log\frac{\beta_i(T+1)}{\beta_i(1)}\right] \\
&= \mathbb{E}\left[2\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T}\alpha_i(t)} - \beta_0 + \delta\log\left(1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)\right)\right] \\
&\leq \mathbb{E}\left[2\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T}\alpha_i(t)} - \beta_0 + 2\delta\left(\sqrt{1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)} - 1\right)\right] \\
&= 2\left(\beta_0 + \delta\right)\mathbb{E}\left[\sqrt{1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)} - 1\right] + \beta_0 \\
&\leq 2(\beta_0 + \delta)\left(\sqrt{1 + \frac{2}{\gamma^{3/2}\beta_0^2}\mathbb{E}\left[\sum_{t=1}^{T}(1 - x_i(t))\right]} - 1\right) + \beta_0\,. \\
&\leq 2(\beta_0 + \delta)\sqrt{\frac{2}{\gamma^{3/2}\beta_0^2}\mathbb{E}\left[\sum_{t=1}^{T}(1 - x_i(t))\right]} + \beta_0\,. \\
&\leq 2(1+\delta)\sqrt{\frac{2}{\gamma^{3/2}}Q_i} + \beta_0\,,
\end{aligned} \tag{38}$$

where the first inequality follows from the inequality of $\log(1+x) \leq 2(\sqrt{1+x} - 1)$ for $x > 0$, the second inequality follows from Lemma 6, the third inequality follows from $\sqrt{1+x} - 1 \leq \sqrt{x}$ for $x \geq 0$, and the last inequality follows from $\beta_0 \geq 1$.

Putting together the upper and lower bounds and applying a self-bounding technique Bounding the RHS of (17) using (36) and (38) yields the regret upper bound depending on $(P_i)_{i\in J^*}$ and $(Q_i)_{i\in I^*}$ as

$$\begin{aligned}
\frac{R_T}{\gamma} &\leq \sum_{i\in J^*}\left(2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \beta_0 + \delta\log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\xi}{\gamma}\left(\log(1+P_i) + \frac{5}{4}\right)\right) \\
&\quad + 2\left(1+\delta\right)\sum_{i\in I^*}\sqrt{\frac{2}{\gamma^{3/2}}Q_i} + \beta_0|I^*| + \frac{d^2 + d(1+2\delta)}{\gamma} \\
&= \sum_{i\in J^*}\bar{f}_i\left(\frac{P_i}{\gamma}\right) + 2\left(1+\delta\right)\sum_{i\in I^*}\sqrt{\frac{2}{\gamma^{3/2}}Q_i} + \beta_0|I^*| + \frac{1}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right),
\end{aligned} \tag{39}$$

where we define convex function $\bar{f}_i : \mathbb{R}_+ \to \mathbb{R}$ by

$$\bar{f}_i(x) = 2\sqrt{\beta_0^2 + \sigma_i^2 x} + \delta \log\left(1 + \frac{\sigma_i^2 x}{\beta_0^2}\right) + \frac{\xi}{\gamma}\log(1+\gamma x) - \beta_0 \,. \tag{40}$$

In the stochastic regime, setting $C = 0$ in (33) yields the regret lower bound depending on $(P_i)_{i\in J^*}$ and $(Q_i)_{i\in I^*}$ as

$$R_T \geq \frac{1}{v(\mathcal{A})}\sum_{i\in I^*}\Delta'_{i,\min}Q_i + \frac{1}{w(\mathcal{A})}\sum_{i\in J^*}\Delta_{i,\min}P_i \,. \tag{41}$$

Combining (39) and (41), we have

$$\begin{aligned}
\frac{R_T}{\log T} &= \frac{R_T}{\gamma} = 2\frac{R_T}{\gamma} - \frac{R_T}{\gamma} \\
&\leq 2\frac{R_T}{\gamma} - \frac{1}{\gamma}\left(\frac{1}{v(\mathcal{A})}\sum_{i\in I^*}\Delta'_{i,\min}Q_i + \frac{1}{w(\mathcal{A})}\sum_{i\in J^*}\Delta_{i,\min}P_i\right) \\
&\leq \sum_{i\in J^*}\left(2\bar{f}_i\left(\frac{P_i}{\gamma}\right) - \frac{\Delta_{i,\min}}{w(\mathcal{A})}\frac{P_i}{\gamma}\right) + \sum_{i\in I^*}\left(4(1+\delta)\sqrt{\frac{2}{\gamma^{1/2}}\frac{Q_i}{\gamma}} - \frac{\Delta'_{i,\min}}{v(\mathcal{A})}\frac{Q_i}{\gamma}\right) \\
&\quad + 2\beta_0|I^*| + \frac{2}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) \\
&\leq \sum_{i\in J^*}\max_{x\geq 0}\left\{2\bar{f}_i(x) - \frac{\Delta_{i,\min}}{w(\mathcal{A})}x\right\} + \sum_{i\in I^*}\max_{x\geq 0}\left\{4(1+\delta)\sqrt{\frac{2}{\gamma^{1/2}}x} - \frac{\Delta'_{i,\min}}{v(\mathcal{A})}x\right\} \\
&\quad + 2\beta_0|I^*| + \frac{2}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) \\
&\leq \sum_{i\in J^*}\max_{x\geq 0}\left\{2\bar{f}_i(x) - \frac{\Delta_{i,\min}}{w(\mathcal{A})}x\right\} + \sum_{i\in I^*}\frac{16(1+\delta)^2 v(\mathcal{A})}{\sqrt{\gamma}\Delta'_{i,\min}} \\
&\quad + 2\beta_0|I^*| + \frac{2}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) \,,
\end{aligned} \tag{42}$$

where the second inequality follows from (39) and the last inequality follows from $a\sqrt{x} - bx \leq a^2/(2b)$ for $a, b, x \geq 0$.

In the following, we evaluate the first term of (42).

Bounding the first term of (42) We will prove the following statement:

$$\max_{x\geq 0}\left\{2\bar{f}_i(x) - \frac{\Delta_{i,\min}}{w(\mathcal{A})}x\right\} \leq h\left(w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}}\right) + O\left(\frac{\log(1+\gamma)}{\gamma}\right) \,, \tag{43}$$

where $h : \mathbb{R}_+ \to \mathbb{R}$ is defined as

$$h(z) = \begin{cases} 2\beta_0 & \text{if } 0 \leq z \leq \frac{\beta_0}{2(1+\delta/\beta_0)} \,, \\ 2z\left(1 + \sqrt{1 + 2\frac{\delta}{z}}\right) - 2\delta + 4\delta\left(\log\frac{z}{\beta_0} + \log\left(1 + \sqrt{1+2\frac{\delta}{z}}\right)\right) + \frac{\beta_0^2}{z} - 2\beta_0 & \text{if } z > \frac{\beta_0}{2(1+\delta/\beta_0)} \,. \end{cases} \tag{44}$$

Let $\bar{\Delta}_i = \Delta_{i,\min}/w(\mathcal{A})$ for the notational simplicity. As f_i is concave, the maximum of $2f_i(x) - \bar{\Delta}_i x$ is attained by $x_i^* \in \mathbb{R}$ satisfying $2f'(x_i^*) = \bar{\Delta}_i$. Define $\tilde{x}_i \geq 0$ by

$$\tilde{x}_i \coloneqq \max\left\{\left(\frac{4\sigma_i}{\bar{\Delta}_i}\right)^2, \frac{8\delta}{\bar{\Delta}_i}, \frac{16\xi}{\gamma\bar{\Delta}_i}\right\} \,.$$

We then have

$$2f_i'(\tilde{x}_i) \leq \frac{2\sigma_i}{\sqrt{\left(\frac{4\sigma_i}{\bar{\Delta}_i}\right)^2}} + \frac{2\delta\sigma_i^2}{{\beta_0}^2 + \sigma_i^2\frac{8\delta}{\bar{\Delta}_i}} + \frac{2\xi}{1+\gamma\frac{16\xi}{\gamma\bar{\Delta}_i}} \leq \frac{\bar{\Delta}_i}{2} + \frac{\bar{\Delta}_i}{4} + \frac{\bar{\Delta}_i}{8} < \bar{\Delta}_i \,,$$

which implies $\tilde{x}_i \geq x_i^*$. Hence, we have

$$\begin{aligned}
&\max_{x\geq 0}\left\{2f_i(x) - \bar{\Delta}_i x\right\} = 2f_i(x_i^*) - \bar{\Delta}_i x_i^* \\
&= 4\sqrt{\beta_0^2 + \sigma_i^2 x_i^*} + 2\delta\log\left(1 + \frac{\sigma_i^2 x_i^*}{\beta_0^2}\right) + 2\frac{\xi}{\gamma}\log(1+\gamma x_i^*) - \bar{\Delta}_i x_i^* - 2\beta_0 \\
&\leq 4\sqrt{\beta_0^2 + \sigma_i^2 x_i^*} + 2\delta\log\left(1 + \frac{\sigma_i^2 x_i^*}{\beta_0^2}\right) + 2\frac{\xi}{\gamma}\log(1+\gamma \tilde{x}_i) - \bar{\Delta}_i x_i^* - 2\beta_0 \\
&\leq \max_{x\geq 0}\left\{4\sqrt{\beta_0^2 + \sigma_i^2 x} + 2\delta\log\left(1 + \frac{\sigma_i^2 x}{\beta_0^2}\right) - \bar{\Delta}_i x\right\} + 2\frac{\xi}{\gamma}\log(1+\gamma \tilde{x}_i) - 2\beta_0 \\
&= \max_{x\geq 0}\left\{g_i(x) - \bar{\Delta}_i x\right\} - 2\beta_0 + O\left(\frac{\log(1+\gamma)}{\gamma}\right), \qquad (45)
\end{aligned}$$

where we define

$$g_i(x) = 4\sqrt{\beta_0^2 + \sigma_i^2 x} + 2\delta\log\left(1 + \frac{\sigma_i^2 x}{\beta_0^2}\right).$$

From (45) and (42), we have

$$\limsup_{T\to\infty}\frac{R_T}{\log T} \leq \sum_{i\in J^*}\left(\max_{x\geq 0}\left\{g_i(x) - \bar{\Delta}_i x\right\} - 2\beta_0\right) + 2\beta_0|I^*|. \qquad (46)$$

In the following, we write $z_i = \frac{\sigma_i^2}{\bar{\Delta}_i}$. As we have

$$g_i'(x) = \frac{2\sigma_i^2}{\sqrt{\beta_0^2 + \sigma_i^2 x}} + \frac{2\delta\sigma_i^2}{\beta_0^2 + \sigma_i^2 x} \leq 2\sigma_i^2\left(\frac{1}{\beta_0} + \frac{\delta}{\beta_0^2}\right),$$

If $z_i = \frac{\sigma_i^2}{\bar{\Delta}_i} \leq \frac{1}{2(1/\beta_0 + \delta/\beta_0^2)} = \frac{\beta_0}{2(1+\delta/\beta_0)}$, the maximum of $g_i(x) - \bar{\Delta}_i x$ is attained by $x = 0$, implying

$$\max_{x\geq 0}\left\{g_i(x) - \bar{\Delta}_i x\right\} = g_i(0) = 4\beta_0 \quad \text{if} \quad z_i \coloneqq \frac{\sigma_i^2}{\bar{\Delta}_i} \leq \frac{\beta_0}{2(1+\delta/\beta_0)}. \qquad (47)$$

Otherwise, we have

$$\begin{aligned}
g_i(x) - \bar{\Delta}_i x &= 4\beta_0\sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}} + 2\delta\log\left(1 + \frac{\sigma_i^2 x}{\beta_0^2}\right) - \frac{\beta_0^2\bar{\Delta}_i}{\sigma_i^2}\left(1 + \frac{\sigma_i^2 x}{\beta_0^2}\right) + \frac{\beta_0^2}{z_i} \\
&= 4\beta_0\sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}} + 4\delta\log\left(\sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}}\right) - \frac{\beta_0^2}{z_i}\left(\sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}}\right)^2 + \frac{\beta_0^2}{z_i}.
\end{aligned}$$

From this, by setting $y = \sqrt{1 + \frac{\sigma_i^2 x}{\beta_0^2}}$, we obtain

$$\max_{x\geq 0}\left\{g_i(x) - \bar{\Delta}_i x\right\} \leq \max_{y\geq 0}\left\{4\beta_0 y + 4\delta\log y - \frac{\beta_0^2}{z_i}y^2\right\} + \frac{\beta_0^2}{z_i}. \qquad (48)$$

We here use the following:

$$\max_{y\geq 0}\left\{ay + b\log y - cy^2\right\} = \frac{1}{2}\left(\frac{a}{4c}\left(a + \sqrt{a^2 + 8bc}\right) - b\right) + b\log\frac{a + \sqrt{a^2 + 8bc}}{4c},$$

which holds for any $a, b, c > 0$. We hence have

$$\max_{y\geq 0}\left\{4\beta_0 y + 4\delta\log y - \frac{\beta_0^2}{z_i}y^2\right\}$$

$$= \frac{1}{2}\left(\frac{4\beta_0 z_i}{4\beta_0^2}\left(4\beta_0 + \sqrt{(4\beta_0)^2 + 32\frac{\delta\beta_0^2}{z_i}}\right) - 4\delta\right) + 4\delta\log\frac{4\beta_0 + \sqrt{(4\beta_0)^2 + 32\delta\beta_0^2/z_i}}{4\beta_0^2/z_i}$$
$$= 2\left(z_i\left(1 + \sqrt{1 + 2\frac{\delta}{z_i}}\right) - \delta\right) + 4\delta\left(\log\frac{z_i}{\beta_0} + \log\left(1 + \sqrt{1 + 2\frac{\delta}{z_i}}\right)\right). \tag{49}$$

Combining (45) with (47), (48), and (49), we obtain

$$\max_{x\geq 0}\left\{2f_i(x) - \bar{\Delta}_i x\right\} \leq h\left(\frac{\sigma_i^2}{\bar{\Delta}_i}\right) + O\left(\frac{\log(1+\gamma)}{\gamma}\right) = h\left(w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}}\right) + O\left(\frac{\log(1+\gamma)}{\gamma}\right), \tag{50}$$

where $h : \mathbb{R}_+ \to \mathbb{R}$ is defined by (44). From (42) and (50), we complete the proof of (43).

Bounding h For $z > \frac{\beta_0}{2(1+\delta/\beta_0)}$, $h(z)$ in (44) is bounded as

$$h(z) \leq 2z\left(1 + 1 + \frac{\delta}{z}\right) - 2\delta + 4\delta\left(\log z + \log\left(1 + \sqrt{1 + 2\frac{\delta}{z}}\right)\right) + \frac{\beta_0^2}{\beta_0}\cdot 2\left(1 + \frac{\delta}{\beta_0}\right) - 2\beta_0$$
$$= 4z + 4\delta\left(\log z + \log\left(1 + \sqrt{1 + 2\frac{\delta}{z}}\right) + \frac{1}{2}\right)$$
$$\leq 4z + c\log(1+z) \quad \left(c = O\left(\delta^2\right) = O\left(\left(\log\epsilon^{-1}\right)^2\right)\right),$$

where the last inequality follows from $\log(1+z) = \Omega(1/\delta)$ that holds for $z > \frac{\beta_0}{2(1+\delta/\beta_0)}$. Hence, for any $z \geq 0$, $h(z)$ is bounded as

$$h(z) \leq \max\left\{4z + c\log(1+z), 2\beta_0\right\}. \tag{51}$$

From this and (50), recalling that $\beta_0 = 1 + \epsilon$, we obtain

$$R_T \leq \left(\sum_{i\in J^*}\max\left\{4\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}} + c\log\left(1 + \frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}\right), 2(1+\epsilon)\right\} + 2(1+\epsilon)|I^*|\right)\log T$$
$$+ \sum_{i\in I^*}\frac{16(1+\delta)^2 v(\mathcal{A})}{\Delta'_{i,\min}}\sqrt{\log T} + o(\sqrt{\log T}),$$

which completes the proof of (11) in Theorem 3. □

B.3 Proof for the Stochastic Regime with Adversarial Corruptions

We here show a regret bound for the stochastic regime with adversarial corruptions given in Theorem 3, which is the following regret bound:

$$R_T \leq \mathcal{R}^{\mathrm{LS}} + O\left(\sqrt{Cm\mathcal{R}^{\mathrm{LS}}}\right),$$

where $\mathcal{R}^{\mathrm{LS}}$ is the RHS of (11) and C is the corruption level defined in Section 3.

Proof. In stochastic regimes with adversarial corruptions, using Lemma 4 with $m_i^* = \mu_i$ we have

$$\mathbb{E}\left[\sum_{t=1}^T \alpha_i(t)\right] \leq \mathbb{E}\left[\sum_{t=1}^T a_i(t)(\ell_i(t) - \mu_i)^2 + \log(1 + N_i(T))\right] + \frac{5}{4}$$
$$= \mathbb{E}\left[\sum_{t=1}^T x_i(t)(\ell_i(t) - \ell'_i(t) + \ell'_i(t) - \mu_i)^2 + \log(1 + N_i(T))\right] + \frac{5}{4}$$
$$= \mathbb{E}\left[\sum_{t=1}^T x_i(t)\left((\ell_i(t) - \ell'_i(t))^2 + \sigma_i^2\right) + \log(1 + N_i(T))\right] + \frac{5}{4}$$

$$\leq \sigma_i^2 P_i + \log(1+P_i) + \frac{5}{4} + P_i' \,, \tag{52}$$

where we define

$$P_i' = \mathbb{E}\left[\sum_{t=1}^{T} x_i(t)(\ell_i(t) - \ell_i'(t))^2\right] . \tag{53}$$

Hence, in a similar argument to that of showing (36), by using (52) instead of (34), we obtain

$$\begin{aligned}
&\mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)}\right] \\
&= \mathbb{E}\left[2\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T}\alpha_i(t)} - \beta_0 + \delta \log\left(1 + \frac{1}{\gamma\beta_0^2}\sum_{t=1}^{T}\alpha_i(t)\right)\right] \\
&\leq 2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \beta_0 + \delta\log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\xi}{\gamma}\left(\log(1+P_i) + \frac{5}{4}\right) + 2\sqrt{\frac{P_i'}{\gamma}} + \delta\log\left(1 + \frac{P_i'}{\gamma\beta_0^2}\right) \\
&\leq 2\sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\gamma}} - \beta_0 + \delta\log\left(1 + \frac{\sigma_i^2 P_i}{\gamma\beta_0^2}\right) + \frac{\xi}{\gamma}\left(\log(1+P_i) + \frac{5}{4}\right) + \left(2 + \frac{\delta}{\beta_0}\right)\sqrt{\frac{P_i'}{\gamma}} \,,
\end{aligned} \tag{54}$$

where the last inequality follows from $\log(1+x) \leq \sqrt{x}$ for $x \geq 0$. Combining this with (17) and (38), via a similar argument to that of showing (42), we have

$$\frac{R_T}{\gamma} \leq \sum_{i\in J^*} \bar{f}_i\left(\frac{P_i}{\gamma}\right) + \beta_0|I^*| + \frac{1}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) + \left(2 + \frac{\delta}{\beta_0}\right)\sum_{i\in J^*}\sqrt{\frac{P_i'}{\gamma}} \,, \tag{55}$$

where we recall that $\bar{f}_i$ is defined in (40) by

$$\bar{f}_i(x) = 2\sqrt{\beta_0^2 + \sigma_i^2 x} + \delta\log\left(1 + \frac{\sigma_i^2 x}{\beta_0^2}\right) + \frac{\xi}{\gamma}\log(1+\gamma x) - \beta_0 \,.$$

We further have

$$\begin{aligned}
\sum_{i\in J^*}\sqrt{\frac{P_i'}{\gamma}} &\leq \sqrt{\frac{|J^*|}{\gamma}\sum_{i\in J^*} P_i'} = \sqrt{\frac{|J^*|}{\gamma}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in J^*} x_i(t)(\ell_i(t) - \ell_i'(t))^2\right]} \\
&\leq \sqrt{\frac{m|J^*|}{\gamma}\mathbb{E}\left[\sum_{t=1}^{T}\|\ell(t) - \ell'(t)\|_\infty^2\right]} \leq \sqrt{\frac{m|J^*|}{\gamma}\mathbb{E}\left[\sum_{t=1}^{T}\|\ell(t) - \ell'(t)\|_\infty\right]} = \sqrt{\frac{m|J^*|}{\gamma}C} \,,
\end{aligned} \tag{56}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the first equality follows from the definition of P_i' in (53), and the second inequality follows from the fact that $\sum_{i\in J^*} x_i(t) \leq m$. Combining (55) and (56), we obtain

$$\frac{R_T}{\gamma} \leq \sum_{i\in J^*} \bar{f}_i\left(\frac{P_i}{\gamma}\right) + \beta_0|I^*| + \frac{1}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right) + \left(2 + \frac{\delta}{\beta_0}\right)\sqrt{\frac{m|J^*|}{\gamma}C} \,. \tag{57}$$

From (57) and Lemma 1, for any $\lambda \in (0,1]$, letting $\bar{\Delta}_i = \Delta_{i,\min}/w(\mathcal{A})$ we have

$$\begin{aligned}
\frac{R_T}{\log T} &= (1+\lambda)\frac{R_T}{\gamma} - \lambda\frac{R_T}{\gamma} \\
&\leq \sum_{i\in J^*}\max_{x\geq 0}\left\{(1+\lambda)\bar{f}_i(x) - \lambda\bar{\Delta}_i x\right\} + \sum_{i\in I^*}\frac{(1+\lambda)^2}{\lambda}\frac{4(1+\delta)^2 v(\mathcal{A})}{\sqrt{\gamma}\Delta_{i,\min}'} + 2\left(2 + \frac{\delta}{\beta_0}\right)\sqrt{\frac{m|J^*|}{\gamma}C} + \frac{2\lambda C m}{\gamma} \\
&\quad + (1+\lambda)\left(\beta_0|I^*| + \frac{1}{\gamma}\left(d^2 + d(1+2\delta) + \frac{5}{4}\xi|J^*|\right)\right) \,,
\end{aligned} \tag{58}$$

which can be shown in a way similar to the argument of (42). Further, we have

$$\begin{aligned}
&\max_{x\geq 0}\left\{(1+\lambda)\bar{f}_i(x)-\lambda\bar{\Delta}_i x\right\} = \frac{1+\lambda}{2}\max_{x\geq 0}\left\{2\bar{f}_i(x)-\frac{2\lambda\bar{\Delta}_i}{1+\lambda}x\right\}\\
&\leq \frac{1+\lambda}{2}h\left(\frac{(1+\lambda)\sigma_i^2}{2\lambda\bar{\Delta}_i}\right)+O\left(\frac{\log(1+\gamma)}{\gamma}\right)\\
&\leq \max\left\{\frac{(1+\lambda)^2}{\lambda}\frac{\sigma_i^2}{\bar{\Delta}_i}+c\log\left(1+\frac{\sigma_i^2}{\lambda\bar{\Delta}_i}\right),(1+\lambda)\beta_0\right\}+O\left(\frac{\log(1+\gamma)}{\gamma}\right)\\
&\leq \max\left\{4\frac{\sigma_i^2}{\bar{\Delta}_i}+c\log\left(1+\frac{\sigma_i^2}{\bar{\Delta}_i}\right),2\beta_0\right\}+(1+c)\left(\frac{1}{\lambda}-1\right)\frac{\sigma_i^2}{\bar{\Delta}_i}+O\left(\frac{\log(1+\gamma)}{\gamma}\right), \qquad (59)
\end{aligned}$$

where $h(z)$ is defined as (44), the first inequality follows from (50), the second inequality comes from (51) and $\lambda\in(0,1]$, and the last inequality follows from

$$\begin{aligned}
\frac{(1+\lambda)^2}{\lambda} &= \lambda+2+\frac{1}{\lambda}\leq 3+\frac{1}{\lambda}=4+\left(\frac{1}{\lambda}-1\right),\\
\log\left(1+\frac{\sigma_i^2}{\lambda\bar{\Delta}_i}\right) &\leq \frac{1}{\lambda}\log\left(1+\frac{\sigma_i^2}{\bar{\Delta}_i}\right)\leq \log\left(1+\frac{\sigma_i^2}{\bar{\Delta}_i}\right)+\left(\frac{1}{\lambda}-1\right)\frac{\sigma_i^2}{\bar{\Delta}_i}.
\end{aligned}$$

Using (58), (59), and $\lambda\leq 1$, we obtain

$$\begin{aligned}
\frac{R_T}{\log T} &\leq \sum_{i\in J^*}\max\left\{4\frac{\sigma_i^2}{\bar{\Delta}_i}+c\log\left(1+\frac{\sigma_i^2}{\bar{\Delta}_i}\right),2\beta_0\right\}+2\beta_0|I^*|\\
&\quad+2\left(2+\frac{\delta}{\beta_0}\right)\sqrt{\frac{m|J^*|}{\gamma}C}+2\lambda\frac{Cm}{\gamma}+(1+c)\left(\frac{1}{\lambda}-1\right)\sum_{i\in J^*}\frac{\sigma_i^2}{\bar{\Delta}_i}\\
&\quad+\sum_{i\in I^*}\frac{(1+\lambda)^2}{\lambda}\frac{4(1+\delta)^2v(\mathcal{A})}{\sqrt{\gamma}\Delta'_{i,\min}}+O\left(\frac{\log(1+\gamma)}{\gamma}\right). \qquad (60)
\end{aligned}$$

By choosing $\lambda=\sqrt{\frac{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)+2Cm}}$, we have

$$\lambda\leq\sqrt{\frac{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}{2Cm}}\quad\text{and}\quad\frac{1}{\lambda}-1=\sqrt{1+\frac{2Cm}{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}}-1\leq\sqrt{\frac{2Cm}{\gamma\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}},$$

which imply that

$$2\left(2+\frac{\delta}{\beta_0}\right)\sqrt{\frac{m|J^*|}{\gamma}C}+\frac{2\lambda Cm}{\gamma}+(1+c)\left(\frac{1}{\lambda}-1\right)\sum_{i\in J^*}\frac{\sigma_i^2}{\bar{\Delta}_i}=O\left(\sqrt{\frac{Cm}{\gamma}\sum_{i\in J^*}\left(\frac{\sigma_i^2}{\bar{\Delta}_i}+1\right)}\right).$$

From this and (60), recalling that $\gamma=\log T$, $\beta_0=1+\epsilon$ and $\bar{\Delta}_i=\Delta_{i,\min}/w(\mathcal{A})$, we obtain

$$\begin{aligned}
R_T &\leq \left(\sum_{i\in J^*}\max\left\{4w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}}+c\log\left(1+w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}}\right),2(1+\epsilon)\right\}+2(1+\epsilon)|I^*|\right)\log T\\
&\quad+O\left(\sqrt{Cm\sum_{i\in J^*}\left(w(\mathcal{A})\frac{\sigma_i^2}{\Delta_{i,\min}}+1\right)\log T}\right)+\sum_{i\in I^*}\frac{(1+\lambda)^2}{\lambda}\frac{16(1+\delta)^2v(\mathcal{A})}{\Delta'_{i,\min}}\sqrt{\log T}+o(\sqrt{\log T}),
\end{aligned}$$

which completes the proof for the stochastic regime with adversarial corruptions. □

B.4 Proof for the Adversarial Regime

Proof of (12) in Theorem 3. First, we prove $R_T \leq \sqrt{4dQ_2 \log T} + O(d \log T) + d^2 + d(1+2\delta)$. For any $m^* \in [0,1]^d$, bounding the RHS of Lemma 3 we have

$$\begin{aligned}
R_T &\leq \gamma \sum_{i=1}^{d} \mathbb{E}\left[2\beta_i(T+1) - \beta_i(1) + 2\delta \log \frac{\beta_i(T+1)}{\beta_i(1)}\right] + d^2 + d(1+2\delta) \\
&\leq 2\gamma \sum_{i=1}^{d} \mathbb{E}\left[\beta_i(T+1)\right] + O(d\gamma + d^2) \\
&= 2\gamma \sum_{i=1}^{d} \mathbb{E}\left[\sqrt{\beta_0^2 + \frac{1}{\gamma}\sum_{t=1}^{T} \alpha_i(t)}\right] + O(d\gamma + d^2) \\
&\leq 2\gamma \sum_{i=1}^{d} \mathbb{E}\left[\sqrt{\beta_0^2 + \frac{1}{\gamma}\left(\sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i^*)^2 + \log(1+N_i(T)) + \frac{5}{4}\right)}\right] + O(d\gamma + d^2) \\
&\leq 2\gamma \sum_{i=1}^{d} \mathbb{E}\left[\sqrt{\frac{1}{\gamma}\sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i^*)^2}\right] + O(d\gamma + d^2) \\
&\leq 2\mathbb{E}\left[\sqrt{d\gamma \sum_{i=1}^{d}\sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i^*)^2}\right] + O(d\gamma + d^2) \qquad (61) \\
&\leq 2\mathbb{E}\left[\sqrt{d\gamma \sum_{t=1}^{T} \|\ell(t) - m^*\|_2^2}\right] + O(d\gamma + d^2)\,,
\end{aligned}$$

where the second inequality follows from $\beta_i(T+1) = O(T)$, the third inequality follows from Lemma 4, and the fifth inequality follows from the Cauchy-Schwarz inequality. Since m^* is arbitrary, we obtain the desired results by $m^* = \bar{\ell}$.

Next, we prove $R_T \leq \sqrt{4dL^* \log T} + O(d \log T) + d^2 + d(1+2\delta)$. By setting $m^* = 0$ in (61), we have

$$\begin{aligned}
R_T &\leq 2\mathbb{E}\left[\sqrt{d\gamma \sum_{t=1}^{T}\sum_{i \in I(t)} \ell_i(t)^2}\right] + O(d\gamma + d^2) \\
&\leq 2\mathbb{E}\left[\sqrt{d\gamma \sum_{t=1}^{T}\sum_{i \in I(t)} \ell_i(t)}\right] + O(d\gamma + d^2) \\
&= 2\mathbb{E}\left[\sqrt{d\gamma \sum_{t=1}^{T} \ell(t)^\top a(t)}\right] + O(d\gamma + d^2) \\
&= 2\mathbb{E}\left[\sqrt{d\gamma \left(\sum_{t=1}^{T} (\ell(t)^\top a(t) - \ell(t)^\top a^*) + \sum_{t=1}^{T} \ell(t)^\top a^*\right)}\right] + O(d\gamma + d^2) \\
&\leq 2\sqrt{d\gamma \left(\mathbb{E}\left[\sum_{t=1}^{T} (\ell(t)^\top a(t) - \ell(t)^\top a^*)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \ell(t)^\top a^*\right]\right)} + O(d\gamma + d^2) \\
&= 2\sqrt{d\gamma\,(R_T + L^*)} + O(d\gamma + d^2)\,,
\end{aligned}$$

where the third inequality follows from Jensen's inequality. By solving this inequation in R_T, we obtain

$$R_T \leq 2\sqrt{d\gamma L^*} + O(d\gamma + d^2)\,,$$

which is the desired bound.

Finally, we prove $R_T \leq \sqrt{4d(mT - L^*)\log T} + O(d\log T) + d^2 + d(1+2\delta)$. By setting $m^* = \mathbf{1}$ in (61) and repeating a similar argument as for proving $R_T \leq \sqrt{4dL^*\log T} + O(d\log T) + d^2 + d(1+2\delta)$ we have

$$\begin{aligned}
R_T &\leq 2\mathbb{E}\left[\sqrt{d\gamma \sum_{t=1}^{T}\sum_{i\in I(t)} (\ell_i(t)-1)^2}\right] + O(d\gamma + d^2) \\
&\leq 2\mathbb{E}\left[\sqrt{d\gamma \sum_{t=1}^{T}\sum_{i\in I(t)} (1-\ell_i(t))}\right] + O(d\gamma + d^2) \\
&\leq 2\mathbb{E}\left[\sqrt{d\gamma \left(mT - \sum_{t=1}^{T}\ell(t)^\top a^* - \sum_{t=1}^{T}\langle \ell(t), a(t)-a^*\rangle\right)}\right] + O(d\gamma + d^2) \\
&\leq 2\sqrt{d\gamma\left(mT - L^* - R_T\right)} + O(d\gamma + d^2)\,,
\end{aligned}$$

where the third inequality follows since $\|a_i(t)\|_1 \leq m$ and the forth inequality follows from Jensen's inequality. By solving this inequation in R_T, we obtain

$$R_T \leq 2\sqrt{d\gamma(mT - L^*)} + O(d\gamma + d^2)\,,$$

which completes the proof. □

C PROOF OF THEOREM 4

We can prove Theorem 4 by using a similar argument as for Theorem 3. We first discuss the key lemma for this argument, the very similar argument of which is given in Ito (2021b).

C.1 Preliminary

Here, we present the key lemma for proving Theorem 4.

Lemma 7. *Assume that $m_i(t)$ is given by (3). Then for any $i \in [d]$ and $u_i(1), \ldots, u_i(T) \in [0,1]$ we have*

$$\begin{aligned}
\sum_{t=1}^{T}\alpha_i(t) &\leq \sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i(t))^2 \\
&\leq \frac{1}{1-2\eta}\sum_{t=1}^{T} a_i(t)(\ell_i(t) - u_i(t))^2 + \frac{1}{\eta(1-2\eta)}\left(\frac{1}{4} + 2\sum_{t=1}^{T-1}|u_i(t+1) - u_i(t)|\right)\,.
\end{aligned}$$

Proof. Take $i \in [d]$ satisfying $a_i(t) = 1$. Then it holds that

$$\begin{aligned}
&(\ell_i(t) - m_i(t))^2 - (\ell_i(t) - u_i(t))^2 \\
&\leq 2(\ell_i(t) - m_i(t))(u_i(t) - m_i(t)) \\
&= 2(\ell_i(t) - m_i(t))(m_i(t+1) - m_i(t)) + 2(\ell_i(t) - m_i(t))(u_i(t) - m_i(t+1)) \\
&= 2\eta(\ell_i(t) - m_i(t))^2 + \frac{2}{\eta}(m_i(t+1) - m_i(t))(u_i(t) - m_i(t+1)) \\
&\leq 2\eta(\ell_i(t) - m_i(t))^2 + \frac{1}{\eta}\left((u_i(t) - m_i(t))^2 - (u_i(t) - m_i(t+1))^2\right)\,,
\end{aligned}$$

where the inequalities follow from $y^2 - x^2 = 2y(y-x) - (x-y)^2 \leq 2y(y-x)$ for $x, y \in \mathbb{R}$ and the last equality follows from the definition of $m(t)$ in (3). Hence, we have

$$(\ell_i(t) - m_i(t))^2 \leq \frac{1}{1-2\eta}\left((\ell_i(t) - u_i(t))^2 + \frac{1}{\eta}\left((u_i(t) - m_i(t))^2 - (u_i(t) - m_i(t+1))^2\right)\right)\,. \tag{62}$$

From the definition of $\alpha_i(t)$ in (8) and (62), we have

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_i(t) &\leq \sum_{t=1}^{T} a_i(t)(\ell_i(t) - m_i(t))^2 \\
&\leq \frac{1}{1-2\eta} \sum_{t=1}^{T} (\ell_i(t) - u_i(t))^2 + \frac{1}{\eta(1-2\eta)} \sum_{t=1}^{T} \left\{ (u_i(t) - m_i(t))^2 - (u_i(t) - m_i(t+1))^2 \right\} \\
&= \frac{1}{1-2\eta} \sum_{t=1}^{T} (\ell_i(t) - u_i(t))^2 \\
&\quad + \frac{1}{\eta(1-2\eta)} \left(\sum_{t=1}^{T} \left\{ (u_i(t+1) - m_i(t+1))^2 - (u_i(t) - m_i(t+1))^2 \right\} + (u_i(1) - m_i(1))^2 \right) \\
&\leq \frac{1}{1-2\eta} \sum_{t=1}^{T} (\ell_i(t) - u_i(t))^2 \\
&\quad + \frac{1}{\eta(1-2\eta)} \left(\sum_{t=1}^{T} (u_i(t+1) + u_i(t) - 2m_i(t+1))(u_i(t+1) - u_i(t)) + \frac{1}{4} \right) \\
&\leq \frac{1}{1-2\eta} \sum_{t=1}^{T} a_i(t)(\ell_i(t) - u_i(t))^2 + \frac{1}{\eta(1-2\eta)} \left(\frac{1}{4} + 2 \sum_{t=1}^{T-1} |u_i(t+1) - u_i(t)| \right),
\end{aligned}$$

which completes the proof. □

C.2 Proof for the Stochastic Regime

Proof of (13) in Theorem 4. From Lemma 7, setting $u_i(t) = \mu_i$ for all $i \in [d]$ and $t \in [T]$ in Lemma 7 and taking the expectation yield that

$$\mathbb{E}\left[\sum_{t=1}^{T} \alpha_i(t) \right] \leq \frac{1}{1-2\eta} \mathbb{E}\left[\sum_{t=1}^{T} a_i(t)(\ell_i(t) - \mu_i)^2 \right] + \frac{1}{4\eta(1-2\eta)} = \frac{1}{1-2\eta} \sigma_i^2 P_i + \frac{1}{4\eta(1-2\eta)},$$

where P_i is defined in (35). By using this inequality instead of (34) and repeating the same argument as that in Appendix B.2, we obtain

$$\begin{aligned}
R_T \leq \frac{1}{1-2\eta} &\left(\sum_{i \in J^*} \max \left\{ 4 \frac{w(\mathcal{A})\, \sigma_i^2}{\Delta_{i,\min}} + c \log \left(1 + \frac{w(\mathcal{A})\, \sigma_i^2}{\Delta_{i,\min}} \right), 2(1+\epsilon) \right\} + 2(1+\epsilon)|I^*| \right) \log T \\
&+ O\left(d \sqrt{\frac{\log T}{\eta(1-2\eta)}} \right) + \sum_{i \in I^*} \frac{16(1+\delta)^2 v(\mathcal{A})}{\Delta'_{i,\min}} \sqrt{\log T} + o(\sqrt{\log T}),
\end{aligned}$$

which is the desired bound. □

C.3 Proof for the Stochastic Regime with Adversarial Corruptions

Here we show a regret bound for the stochastic regime with adversarial corruptions given in Theorem 4:

$$R_T \leq \mathcal{R}^{\mathrm{GD}} + O\left(\sqrt{C m \mathcal{R}^{\mathrm{GD}}} \right).$$

Proof. Letting $u_i(t) = \mu_i$ for all $i \in [d]$ and $t \in [T]$ in Lemma 7 and taking the expectation yield that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \alpha_i(t) \right] &\leq \frac{1}{1-2\eta} \mathbb{E}\left[\sum_{t=1}^{T} a_i(t)(\ell_i(t) - \mu_i)^2 \right] + \frac{1}{4\eta(1-2\eta)} \\
&\leq \frac{1}{1-2\eta} \sigma_i^2 P_i + P'_i + \frac{1}{4\eta(1-2\eta)},
\end{aligned}$$

where P_i is defined in (35) and the last inequality is obtained by a similar argument as for (52). By using this inequality instead of (34) and repeating a similar argument as that in Appendix B.3, we obtain

$$R_T \leq \frac{1}{1-2\eta}\left(\sum_{i\in J^*}\max\left\{4\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}+c\log\left(1+\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}\right),\,2(1+\epsilon)\right\}+2(1+\epsilon)|I^*|\right)\log T$$
$$+O\left(d\sqrt{\frac{\log T}{\eta(1-2\eta)}}\right)+O\left(\sqrt{Cm\sum_{i\in J^*}\left(\frac{w(\mathcal{A})\sigma_i^2}{\Delta_{i,\min}}+1\right)\log T}\right)+\sum_{i\in I^*}\frac{16(1+\delta)^2 v(\mathcal{A})}{\Delta'_{i,\min}}\sqrt{\log T}+o(\sqrt{\log T})\,,$$

which completes the proof. □

C.4 Proof for the Adversarial Regime

Proof of (14) in Theorem 4. From Lemma 7, we immediately obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{d}\alpha_i(t)\leq\frac{1}{1-2\eta}\sum_{t=1}^{T}\sum_{i=1}^{d}a_i(t)(\ell_i(t)-u_i(t))^2+\frac{1}{\eta(1-2\eta)}\left(\frac{d}{4}+2\sum_{t=1}^{T-1}\|u(t+1)-u(t)\|_1\right)\tag{63}$$

for any $u(t)=(u_1(t),\dots,u_d(t))^\top\in[0,1]^d$.

First, we prove $R_T\leq\sqrt{\frac{\gamma}{\eta(1-2\eta)}(d+8V_1)}+O(d\gamma+d^2)$. Letting $u(t)=\ell(t)$ in (63) we can bound the regret as

$$\begin{aligned}R_T&\leq 2\gamma\sum_{i=1}^{d}\mathbb{E}\left[\sqrt{\beta_0^2+\frac{1}{\gamma}\sum_{t=1}^{T}\alpha_i(t)}\right]+O(d\gamma+d^2)\\&\leq 2\mathbb{E}\left[\sqrt{\gamma\sum_{t=1}^{T}\sum_{i=1}^{d}\alpha_i(t)}\right]+O(d\gamma+d^2)\\&\leq\frac{2}{\sqrt{\eta(1-2\eta)}}\mathbb{E}\left[\sqrt{\gamma\left(\frac{d}{4}+2\sum_{t=1}^{T-1}\|\ell(t+1)-\ell(t)\|_1\right)}\right]+O(d\gamma+d^2)\\&\leq\sqrt{\frac{\gamma}{\eta(1-2\eta)}(d+8V_1)}+O(d\gamma+d^2)\,,\end{aligned}\tag{64}$$

where the second inequality follows from the Cauchy-Schwarz inequality, the third inequality follows by setting $u_i(t)=\ell_i(t)$ for all $i\in[d]$ and $t\in[T]$ in (63), and the last inequality follows from Jensen's inequality. This becomes the desired path-length bound.

Next, we prove we prove $R_T\leq\sqrt{\frac{\gamma}{1-2\eta}\min\{L^*,mT-L^*,Q_2\}}+O(d\gamma+d^2)$. For any $m^*\in[0,1]^d$, letting $u(t)=m^*$ for all $t\in[T]$ in (63), we have

$$\sum_{t=1}^{T}\sum_{i=1}^{d}\alpha_i(t)\leq\frac{1}{1-2\eta}\sum_{t=1}^{T}\sum_{i=1}^{d}a_i(t)(\ell_i(t)-m_i^*)^2+\frac{d}{4\eta(1-2\eta)}\,.$$

Using this inequality, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}\alpha_i(t)\right]&\leq\frac{1}{1-2\eta}\min_{m^*\in[0,1]^d}\left\{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d}a_i(t)(\ell_i(t)-m_i^*)^2\right]\right\}+\frac{d}{4\eta(1-2\eta)}\\&\leq\frac{1}{1-2\eta}\min\{R_T+L^*,mT-L^*-R_T,Q_2\}+\frac{d}{4\eta(1-2\eta)}\,,\end{aligned}$$

where in the last inequality we set $m^*=0$ and (resp. $m^*=1$) and use the same argument as that in Appendix B.4 for deriving the term with R_T+L^* (resp. $mT-L^*-R_T$), and $m^*=\bar{\ell}$ for deriving the term with Q_2, and this complete the proof. □

Table 3: Reward means for the semi-synthetic data.

Instance	d	m	Reward means $\mathbf{1} - \mu$
(d)	6	3	$(0.0315, 0.0208, 0.0193, 0.0182, 0.0179, 0.0177)$
(e)	8	3	$(0.0370, 0.0275, 0.0266, 0.0266, 0.0231, 0.0192, 0.0143, 0.0107)$
(f)	10	3	$(0.0774, 0.0709, 0.0669, 0.0631, 0.0430, 0.0393, 0.0296, 0.0217, 0.00797, 0.00219)$

D EXPERIMENTAL DETAILS

Table 3 lists the reward means used in the experiments.