Hybrid Regret Bounds for Combinatorial Semi-Bandits and Adversarial Linear Bandits

Shinji Ito
NEC Corporation
i-shinji@nec.com

Abstract

This study aims to develop bandit algorithms that automatically exploit tendencies of certain environments to improve performance, without any prior knowledge regarding the environments. We first propose an algorithm for combinatorial semi-bandits with a hybrid regret bound that includes two main features: a best-of-three-worlds guarantee and multiple data-dependent regret bounds. The former means that the algorithm will work nearly optimally in all environments in an adversarial setting, a stochastic setting, or a stochastic setting with adversarial corruptions. The latter implies that, even if the environment is far from exhibiting stochastic behavior, the algorithm will perform better as long as the environment is "easy" in terms of certain metrics. The metrics w.r.t. the easiness referred to in this paper include cumulative loss for optimal actions, total quadratic variation of losses, and path-length of a loss sequence. We also show hybrid data-dependent regret bounds for adversarial linear bandits, which include a first path-length regret bound that is tight up to logarithmic factors.

1 Introduction

In this work, we consider two fundamental problem settings w.r.t. online decision problems: *combinatorial semi-bandits* [42, 66, 4] and *linear bandits* [13, 17, 30]. In both problem settings, a player is given an action set $\mathcal{A} \in \mathbb{R}^d$, a compact subset of a d-dimensional vector space. In each round t, the player chooses an action $a_t \in \mathcal{A}$ and then incurs loss $\ell_t^\top a_t$, where $\ell_t \in \mathbb{R}^d$ is a *loss vector* chosen by the environment. The action set in the combinatorial semi-bandit is assumed to be a subset of $\{0,1\}^d$, each element of which corresponds to a subset of $[d] = \{1, 2, \ldots, d\}$. After choosing $a_t \in \mathcal{A}$, the player can observe ℓ_{ti} for all i such that $a_{ti} = 1$ in semi-bandits. Linear bandits are problems with even more limited feedback, ones in which the learner can only observe the incurred loss $\ell_t^\top a_t$. For combinatorial semi-bandits problems with $\ell_t \in [0,1]^d$, there is a known algorithm with an $O(\sqrt{mdT})$-regret bound [4], where $m = \max_{a \in \mathcal{A}} \|a\|_1$. For linear bandits such that $|\ell_t^\top a| \leq 1$ holds for any $a \in \mathcal{A}$, algorithms with $\tilde{O}(d\sqrt{T})$-regret (where factors in $\log T$ and $\log d$ are ignored) have been developed [13, 17, 30]. These algorithms are optimal in terms of worst-case analysis. In fact, matching lower bounds of $\Omega(\sqrt{mdT})$ for combinatorial semi-bandits [3, 11] and of $\Omega(d\sqrt{T})$ for linear bandits [22] are known.

The worst-case optimal algorithms, however, tend to be too conservative in actual practice. This is because true worst-case environments are quite rare in real-world applications. Rather, the environments may have structures that are convenient for the learner, and it is desirable that the algorithm takes advantage of such structures to improve performance. To exploit such structures, two main categories of approaches have been studied: adapting to (nearly) stochastic environments and developing data-dependent regret bounds.

An example of the first category is in reference to *best-of-both-worlds (BOBW)* algorithms [12, 62, 5, 61, 68, 70], which means that they work well for both adversarial and stochastic settings.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

These algorithms enjoy $\tilde{O}(\sqrt{T})$-regret bounds in an adversarial setting and, simultaneously, achieve $O(\frac{(\log T)^c}{\Delta})$-regret in a stochastic setting with i.i.d. losses, where $c \geq 1$ is a constant and Δ represents the *suboptimality gap* defined by $\min_{a \in \mathcal{A} \setminus a^*} \mathbf{E}[\ell_t^\top (a - a^*)]$ for an optimal action $a^* \in \mathcal{A}$. As shown in Table 1, for combinatorial semi-bandits, Zimmert et al. [71] provide a BOBW algorithm, both bounds of which are tight up to constant factors as matching lower bounds are known for adversarial settings [3, 11] as well as for stochastic settings [42].

Studies on the adversarial robustness of stochastic bandit algorithms [48, 27, 70] can be considered to provide another approach in the first category, in which an adversary can corrupt stochastically-generated losses subject to the constraint that the total amount of the corruption is at most a parameter $C \geq 0$, referred to as the *corruption level*. This model includes both adversarial and stochastic settings and is closely related to studies on BOBW algorithms. In fact, the special cases of $C = 0$ and $C = \Omega(T)$ correspond to stochastic and adversarial settings, respectively. For the fundamental multi-armed bandit problem, Zimmert and Seldin [70] have proposed the Tsallis-INF algorithm, which achieves BOBW with tight regret bounds up to small constant factors, and which simultaneously is very robust w.r.t. corruptions; the degradation in the regret is only $O(\min\{C, \sqrt{\frac{CK \log T}{\Delta}}\})$. Such algorithms are called *best-of-three-worlds (BOTW)* algorithms. As shown in Table 1, the semi-bandit algorithm by Zimmert et al. [71] is also a BOTW algorithm (Sto. +Adv. refers to the stochastic setting with C-adversarial corruptions).[1] For linear bandits, Lee et al. [45] have recently developed a BOTW algorithm as well.

In studies on data-dependent regret bounds [55, 68, 29, 57, 15, 14, 34] of the other category, we define measures of the difficulty of problem instances, which we refer to as *difficulty indicators*, and aim to develop algorithms so that the regret will be smaller for smaller instance-difficulty indicators. Examples of difficulty indicators dealt with in this paper include L^*: the cumulative losses for an optimal action, Q_q: the total quadratic variation of losses, and V_q: the path-length of losses, definitions of which are given in Table 3. Tables 1 and 2 show data-dependent regret bounds, including difficulty indicators for, respectively, combinatorial semi-bandits and adversarial linear bandits. For example, the semi-bandit algorithm with an $O(\sqrt{dL^* \log T})$-regret bound given by Wei and Luo [68] achieves much smaller regret than the worst-case optimal $O(\sqrt{dmT})$-bound when $L^* = o(mT)$, i.e., when there exists an action for which the cumulative losses are much smaller than mT.

Given the various algorithms above, a new challenge arises: how, in practice, can we choose an appropriate algorithm? If the environment is expected to behave in an almost i.i.d. stochastic manner, either BOBW or BOTW algorithms would work well. If the environment is far from exhibiting stochastic behavior but is "easy" in terms of some difficulty indicator, algorithms with a corresponding data-dependent regret bound may work better. In practice, however, it is hard to tell what kind of environment we are working in until we have actually tried out the algorithm.

1.1 Contributions of this work

This work addresses the above-mentioned issue of algorithm selection by developing *hybrid* algorithms. Its main contribution is to develop a semi-bandit algorithm (Algorithm 1) that enjoys multiple data-dependent regret bounds as well. This can be seen as an an extension of the work on the multi-armed bandit problem by Ito [34] to combinatorial semi-bandits. In addition to this, for linear bandits, we provide a hybrid data-dependent regret bound.

For combinatorial semi-bandits problems, we propose Algorithm 1 with the regret bounds shown in Table 1. More explicit statements are provided in Theorem 1, Corollary 1 (for the adversarial setting), and Corollary 2 (for the stochastic setting with/without adversarial corruption). As can be seen in the table, the regret bound for the adversarial setting encompasses three different data-dependent regret bounds. Note that $O(\sqrt{dL^* \log T})$-bounds imply the nearly worst-case optimal bound of $O(\sqrt{dmT \log T})$, as $L^* \leq mT$ always follows from the model definition. The new regret bound is of $O(R + \sqrt{RCm})$ for the corrupted stochastic setting, where R stands for the bound for the stochastic setting. Note that the regret bounds for the stochastic settings (with corruption) can be improved for some special cases, such as for size-invariant ($\|a\|_1 = m$ for all $a \in \mathcal{A}$) or for a matroid constraint ($\mathcal{A}$ forms bases of a matroid), as described in Corollary 2. We would also like to stress that the

[1] This bound is not explicitly stated in their paper, but can be shown via a straightforward modification of the proof. A proof of this is given in Appendix B of the supplementary material.

Table 1: Regret bounds for combinatorial semi-bandits.

Reference	Regime	Regret bound
Audibert et al. [4]	Adv.	$O(\sqrt{dmT})$
Kveton et al. [42]	Sto.	$O\left(\frac{dm \log T}{\Delta}\right)$
Neu [54]	Adv.	$O(m\sqrt{dL^* \log(d/m)})$
Wei and Luo [68] (Sec. 3.1)	Adv.	$O(\sqrt{dQ_2 \log T})$
Wei and Luo [68] (Sec. 4.1)	Adv.	$O(\sqrt{dV_1 \log T})$
Wei and Luo [68] (Sec. 4.2)	Adv.	$O(\sqrt{dL^* \log T})$
Zimmert et al. [71]	Adv. Sto. Sto. + Adv.	$O(\sqrt{dmT})$ $O\left(\frac{dm \log T}{\Delta}\right)$ $O\left(\frac{dm \log T}{\Delta} + \sqrt{\frac{Cdm^2 \log T}{\Delta}}\right)$
[This work] (Algorithm 1)	Adv. Sto. Sto. + Adv.	$O(\sqrt{d \min\{L^*, Q_2, V_1\} \log T})$ $O\left(\frac{dm \log T}{\Delta}\right)$ $O\left(\frac{dm \log T}{\Delta} + \sqrt{\frac{Cdm^2 \log T}{\Delta}}\right)$

Table 2: Regret bounds for adversarial linear bandits. We assume that $\|a\|_p \leq 1$ for all $\mathcal{A}$ and $\|\ell_t\|_q \leq 1$ for all t, for some $p, q \in [1, \infty]$ such that $1/p + 1/q = 1$. $\tilde{O}(\cdot)$ ignores factors of $(\log T)^{O(1)}$ and $(\log d)^{O(1)}$.

Reference	Regret bound
Bubeck et al. [13]	$\tilde{O}(d\sqrt{T})$
Hazan and Kale [29]	$\tilde{O}(d\sqrt{\vartheta Q_2})$
Bubeck et al. [15] (Cor. 4, 8)	$\tilde{O}(d\sqrt{\vartheta V_2})$
Bubeck et al. [15] (Cor. 6)	$\tilde{O}(d^{3/2}\sqrt{\vartheta V_q})$
Ito et al. [36]	$\tilde{O}(d\sqrt{\min\{L^*, Q_q\}})$
[This work]	$\tilde{O}(d\sqrt{\min\{L^*, Q_q, V_q\}})$

Table 3: Definitions of parameters.

T	number of rounds
d	dimension of $\mathcal{A}$
m	$\max_{a \in \mathcal{A}} \|a\|_1$
ϑ	self-concordant parameter
C	corruption level
L^*	$\min_{a^*} \sum_{t=1}^T \ell_t^\top a^*$
Q_q	$\sum_{t=1}^T \|\ell_t - \bar{\ell}\|_q^2$ $\left(\bar{\ell} = \frac{1}{T}\sum_{t=1}^T \ell_t\right)$
V_q	$\sum_{t=1}^{T-1} \|\ell_t - \ell_{t+1}\|_q$
Δ	$\min_{a \in \mathcal{A} \setminus \{a^*\}} \mathbf{E}[\ell_t^\top (a - a^*)]$

proposed algorithm is parameter-free, i.e., it does not require any prior information w.r.t. parameters Δ, L^*, Q_2, V_1, and C.

The proposed semi-bandit algorithm is based on a follow-the-regularized-leader (FTRL) framework [70, 71], combined with an optimistic prediction for the losses [58, 57, 68]. More precisely, it uses a mixture regularizer [14, 55, 71, 24] consisting of the log-barrier in variables x_i and the Shannon entropy in the *complement* $(1 - x_i)$ of x_i with entry-wisely adaptive learning rates, by which BOTW is achieved. The regret analysis for the stochastic setting (with corruption) is based on self-bounding inequalities for the regret, similarly to what is seen in the analyses by Zimmert et al. [71], Zimmert and Seldin [70]. In addition to this, by choosing optimistic predictors with simple gradient descent methods, we can achieve multiple data-dependent regret bounds as well. The most relevant algorithms are given by Wei and Luo [68] and Zimmert et al. [71]. The proposed algorithm differs from that of Wei and Luo [68] in that the former employs follow-the-regularized-leader rather than online mirror descent methods and uses a mixture regularizer. The major differences with the work by Zimmert et al. [71] are that Algorithm 1 uses a log-barrier rather than Tsallis entropy, and that it takes advantages of optimistic predictors. As far as we have managed to determine, it appears difficult to combine Tsallis-entropy-based algorithms with optimistic predictors, as we discuss in Subsection 4.2. Our work overcomes this difficulty by using a log-barrier regularizer, rather than Tsallis entropy, in exchange for additional $O(\sqrt{\log T})$-factors in worst-case regret bounds.

Remark 1. In the previously mentioned study by Wei and Luo [68], the regret bounds for combinatorial semi-bandits are not given explicitly. However, as stated just after Corollary 3 in their paper ("but they can be straightforwardly generalized to the semi-bandit case"), we can obtain the regret bounds in Table 1 via a simple calculation. To be more precise, their regret bounds in Table 1 dependent on Q_2 and V_1 can be refined by replacing them with $Q_2(I^*) := \sum_{t=1}^{T} \sum_{i \in I^*} (\ell_{ti} - \bar{\ell}_i)^2$ and $V_2(I^*) := \sum_{t=1}^{T-1} \sum_{i \in I^*} |\ell_{ti} - \ell_{t+1,i}|$, respectively, where we define $I^* = \{i \in [d] \mid a_i^* = 1\}$ for an optimal action a^*. This means that Algorithm 1 is not necessarily superior to their algorithms. On the other hand, it should be noted that their algorithms require prior knowledge w.r.t. $Q_2(I^*)$ and $V_2(I^*)$ to achieve corresponding regret bounds.

For linear bandits problems, we provide the hybrid data-dependent regret bounds shown in Table 2 and in Theorem 3, which holds for $p \in [2, \infty]$ and $q \in [1, 2]$ such that $1/p + 1/q = 1$, under the assumption of $\|a\|_p \leq 1$ for all $a \in \mathcal{A}$ and $\|\ell_t\|_q \leq 1$ for all t. The parameter $\vartheta \geq 1$ is associated with a self-concordant barrier over the convex hull of $\mathcal{A}$. It is known that any convex set has a self-concordant barrier with $\vartheta = O(d)$ [53], which is tight up to a constant factor. Substituting $\vartheta = d$ and noting $L^* \leq T$, we can see that the new regret bound includes previous bounds. Further, for the special case of $(p, q) = (\infty, 1)$, the new path-length regret bound of $\tilde{O}(d\sqrt{V_1})$ is tight up to a logarithmic factor in T, as a matching lower bound of $O(d\sqrt{V_1})$ is known [22, 15]. To our knowledge, this is the first (nearly) tight path-length regret bound for linear bandits. The regret bounds of $\tilde{O}(d\sqrt{L^*})$ and $\tilde{O}(d\sqrt{Q_q})$ are also nearly tight, as has been noted in the literature [36]. The approach for the new regret bound is quite simple: we combine regret bounds dependent on optimistic predictors [57, 36] and the algorithm of *tracking the best linear predictor* [31, 16].

2 Problem settings

This section introduces the problem settings of *combinatorial semi-bandits* and *linear bandits*. In both settings, a player is given, before the game starts, an *action set* $\mathcal{A} \in \mathbb{R}^d$ and the total number T of rounds. In each round $t \in [T]$, the player chooses an action $a_t \in \mathcal{A}$, while the environment chooses a loss vector $\ell_t \in \mathbb{R}^d$. After choosing the action, the player gets feedback on the loss, which will depend on the problem settings. Player performance is measured in terms of regret R_T defined as follows:

$$R_T(a^*) = \mathbf{E}\left[\sum_{t=1}^{T} \ell_t^\top (a_t - a^*)\right], \quad R_T = \max_{a^* \in \mathcal{A}} R_T(a^*), \tag{1}$$

where the expectation is taken w.r.t. the randomness of ℓ_t and the algorithm's internal randomness.

2.1 Combinatorial semi-bandits

This subsection provides settings of action sets and feedback information in combinatorial semi-bandits. The action set $\mathcal{A}$ is a subset of binary vectors $\{0, 1\}^d$, each element of which can be interpreted as a subset of $[d]$. Denote $m = \max_{a \in \mathcal{A}} \|a\|_1$. For each chosen action $a_t = [a_{t1}, a_{t2}, \ldots, a_{td}]^\top \in \mathcal{A}$, we denote $I_t = \{i \in [d] \mid a_{ti} = 1\}$. We further assume that $\ell_t \in [0, 1]^d$, similarly to what is seen in existing work [71, 40, 42, 66, 54].

In combinatorial semi-bandits, the player can get entry-wise bandit feedback. More precisely, after choosing an action a_t, which corresponds to a subset I_t of $[d]$, the player can observe ℓ_{ti} for each $i \in I_t$, while ℓ_{ti} for $i \in J_t := [d] \setminus I_t$ will not be revealed.

In addition to general action set $\mathcal{A} \in \{0, 1\}^d$, this paper analyzes two special cases of settings. One is *size-invariant semi-bandits*, in which all actions $a \in \mathcal{A}$ have the same size $\|a\|_1 = m$. The other one is *matroid semi-bandits* [40, 66], in which an action set $\mathcal{A}$ corresponds to the bases of a matroid. As all bases of an arbitrary matroid have the same size, matroid semi-bandits are special cases of size-invariant semi-bandits, which implies

$$\{\text{general semi-bandits}\} \supseteq \{\text{size-invariant semi-bandits}\} \supseteq \{\text{matroid semi-bandits}\}.$$

An important example of matroid semi-bandits is the problem over m-set, which is defined as $\mathcal{A} = \{a \in \{0, 1\} \mid \|a\|_1 = m\}$. The problem with full-combinatorial set $\mathcal{A} = \{0, 1\}^n$ can also be reduced to a special case of matroid semi-bandits with $(d, m) = (2n, n)$. Zimmert et al. [71] have provided improved regret bounds for such special cases of m-sets and full-combinatorial sets.

2.2 Linear bandits

In linear bandits, the action set $\mathcal{A}$ is assumed to be an arbitrary closed and bounded subset of $\mathbb{R}^d$. The special cases in which $\mathcal{A}$ consists of binary vectors in $\{0,1\}^d$ are called *combinatorial bandits* [17]. Similarly to what has been done in existing work [15, 29], we assume that there exists $p, q \in [1, \infty]$ for which $1/p + 1/q = 1$, $\|a\|_p \leq 1$ and $\|\ell_t\|_q \leq 1$ hold for all $a \in \mathcal{A}$ and ℓ_t. By rescaling $\mathcal{A}$ and $\{\ell_t\}$ as needed, any problem with bounded $\mathcal{A}$ and $\{\ell_t\}$ can be transformed into a problem satisfying this assumption.

The available feedback in linear bandits is even more limited than in combinatorial semi-bandits. After choosing an action $a_t \in \mathcal{A}$, the player can only observe the incurred loss $\ell_t^\top a_t$. In the special case of combinatorial bandits, the player can only observe the sum of losses $\sum_{i \in I_t} \ell_{ti}$ for the chosen subset, unlike in combinatorial semi-bandits in which ℓ_{ti} is revealed for each $i \in I_t$.

2.3 Assumptions regarding environments

The scope of this work includes the following three different settings in terms of the environments' determining losses ℓ_t:

$$\{\text{adversarial regimes}\} \supseteq \{\text{stochastic regimes with adversarial corruptions}\} \supseteq \{\text{stochastic regiems}\}.$$

Stochastic regimes In a stochastic regime, the loss vectors ℓ_t are supposed to follow an unknown distribution $\mathcal{D}$, i.i.d. for $t = 1, 2, \ldots, T$. Denote $\mu = \mathbf{E}_{\ell \sim \mathcal{D}}[\ell]$ and set $a^* \in \arg\min_{a \in \mathcal{A}} \mu^\top a$. The regret can then be expressed as $R_T = \mathbf{E}[\sum_{t=1}^T \mu^\top (a_t - a^*)]$. It is known that the optimal regret in this regime can be characterized by the *suboptimality gap* parameter Δ defined as $\Delta = \min_{a \in \mathcal{A} \setminus a^*} \mu^\top a - \mu^\top a^*$. Note that no prior information on the distribution $\mathcal{D}$, including the parameter Δ, is given to the player. When we consider stochastic regime, we assume that $\Delta > 0$, which implies that the optimal action $a^* \in \arg\min_{a \in \mathcal{A}} \mu^\top a$ is assumed to be unique.

Adversarial regimes In an adversarial regime, no stochastic models on ℓ_t are assumed, but the loss ℓ_t may be chosen in an adversarial manner. More precisely, the environment can choose ℓ_t depending on the actions and losses $\{(\ell_j, a_j)\}_{j=1}^{t-1}$ chosen up until the $(t-1)$-th round.

Stochastic regimes with adversarial corruptions A stochastic regime with adversarial corruptions is a regime intermediate between stochastic regimes and adversarial regimes. In such a regime, a temporary loss ℓ_t' is drawn from an unknown distribution $\mathcal{D}$, and then the environment may corrupt it to determine ℓ_t in each round, subject to the constraint $\sum_{t=1}^T \|\ell_t - \ell_t'\|_\infty \leq C$, where $C \geq 0$ is a parameter called the *corruption level* and corresponds to the total amount of corruptions. In this paper, we suppose that the corruptions on ℓ_t depend on ℓ_t' and historical data $\{(\ell_j', \ell_j, a_j)\}_{j=1}^{t-1}$, and that they do not depend on a_t, similarly to what is seen in existing models [48, 27, 70, 8].

The special cases of the stochastic regime with adversarial corruptions in which $C \geq \Omega(T)$ and $C = 0$ coincide, respectively, with the adversarial regime and the stochastic regime. This paper supposes that the player is not given parameter C in advance, i.e., it aims to adapt to any environment with an arbitrary corruption level C.

3 Related work

Combinatorial semi-bandits have been extensively studied for a wide range of applications, including adaptive routing [25], network optimization [40], spectrum allocations [25] and recommender systems [41, 65, 56]. For stochastic combinatorial semi-bandits, Kveton et al. [42, 40], Wang and Chen [66] provide tight regret bounds dependent on the suboptimality gap. Interestingly, these tight regret bounds differ depending on the assumption of the action set: for the general action set, the tight bound is of $O(\frac{dm \log T}{\Delta})$, while, in the matroid semi-bandit cases, the tight bound is of $O(\frac{(d-m) \log T}{\Delta})$. Similar tight bounds are reproduced in the work by [71] and in this work as well, together with worst-case optimal regret bounds for the adversarial setting. Chen et al. [19, 20] have considered a more extended framework including nonlinear reward functions. Linear bandits also have many applications, including end-to-end adaptive routing [7, 6] and various examples of combinatorial bandits [17].

In a bandits context, BOBW algorithms have been developed for various problem settings, including the multi-armed bandit [12, 62, 5, 61, 68, 70, 59], combinatorial semi-bandits [71], episodic Markov decision processes [37, 38], online learning with feedback graphs [24], and linear bandits [45]. Similar algorithms have been developed for full-information online learning problems as well, such as the problem of prediction with expert advice [2, 52, 23, 26] and online linear optimization [32]. For achieving BOBW, two main approaches can be found in these papers. One is to select an appropriate mode in an online manner, by determining whether the environment is i.i.d. or not. The other is to exploit self-bounding constraints, i.e., an approach which is meant to lead to improved bounds by combining a regret *lower* bound expressed with a suboptimality gap and a regret upper bound dependent on the action probability vectors. This work adopts the latter approach, similarly to certain existing work [70, 71, 68].

Since Lykouris et al. [48] initiated a study on stochastic bandits robust to adversarial corruptions, research in this direction has been extended to a variety of models, such as those for (adversarial) multi-armed bandits [27, 70, 28, 50], episodic Markov decision processes [49, 21, 38], Gaussian process bandits [8], the problem of prediction with expert advice [2, 33], online learning with feedback graphs [24], and linear bandits [9, 45]. There can be found studies on effective attacks and on the vulnerability of well-known algorithms [39, 46]. We note that some existing studies (e.g., [9, 39, 28, 46]) have considered *targeted corruption* models, in which the adversary may choose corruption on ℓ_t after observing the player's action a_t, unlike this work and some previous studies [48, 27, 70, 8]. The differences in corruption models can be summarized; see, e.g., the paper by Hajiesmaili et al. [28]. In addition, there are alternative definitions of regret, e.g., as one is defined with the losses *after* corruptions, as in this work and certain previous studies [48, 70, 8] and another is defined with the losses *without* corruptions [27, 45, 9]. As the gap between these two notions of regret is at most $O(C)$, regret bounds for one side has consequences for the other sides up to an additional $O(C)$-term. For further discussion on alternative notions of regret, see, e.g., Section 5.2 of the paper by Gupta et al. [27].

Data-dependent bounds have been studied for a variety of difficulty indicators. For a bandits context, Allenberg et al. [1] have developed a multi-armed bandit algorithm with a first-order regret bound, i.e., the bound dependent on L^* rather than on T. Hazan and Kale [29] provided algorithms so-called second-order regret bounds, which depend on Q_2. Similarly to what is seen in such full-information online learning problems as the problem of prediction with expert advice, there are algorithms with first- and second-order regret bounds [18, 26, 63, 47]. It is worth mentioning that one kind of second-order regret bound implies BOBW guarantees, as shown by Gaillard et al. [26]. Note that some known difficulty indicators are not dealt with in this work, e.g., the sparsity of loss vectors [43, 14].

4 Combinatorial semi-bandits

4.1 Preliminary: existing techniques

Convex combination and decomposition Let $\mathcal{X}$ denote the convex hull of the action set $\mathcal{A}$, i.e., the set of all vectors that can be expressed by a convex combination of vectors in $\mathcal{A}$. Our proposed algorithm manages vectors $x_t \in \mathcal{X}$, and chooses $a_t \in \mathcal{A}$ so that $\mathbf{E}[a_t|x_t] = x_t$. Such a_t can be generated via a convex decomposition of x_t. In fact, from Carathéodory's theorem, for any $x_t \in \mathcal{X}$, there exist $\{\lambda_k\}_{k=0}^d \subseteq [0, 1]$ and $\{a_t^{(k)}\}_{k=0}^d \subseteq \mathcal{A}$ such that $\sum_{k=0}^d \lambda_k = 1$ and $\sum_{k=0}^d \lambda_k a_t^{(k)} = x_t$. Hence, by choosing $a_t = a_t^{(k)}$ with probability λ_k, we have $\mathbf{E}[a_t|x_t] = x_t$. Such $\{\lambda_k\}_{k=0}^d$ and $\{a_t^{(k)}\}_{k=0}^d$ can be computed efficiently if there is an algorithm for solving linear optimization over $\mathcal{A}$, as shown, e.g., in Corollary 11.4 in [60]. Similar techniques are used in [35, 36] as well. More efficient algorithms for computing $\{\lambda_k\}_{k=0}^d$ have been developed for the special cases of bases of uniform matroids [71, 67] and general matroids [64]. In our regret analyses, we use the fact that $I_t = \{i \in [d] \mid a_{ti} = 1\}$ satisfies $\mathrm{Prob}[i \in I_t|x_t] = x_{ti}$.

Optimistic follow the regularized leader Our proposed algorithm is based on the framework of *optimistic follow the regularized leader* [57, 58, 44], in which the vector x_t is defined as

$$x_t \in \arg\min_{x \in \mathcal{X}} \left\{ \sum_{j=1}^{t-1} \hat{\ell}_j^\top x + m_t^\top x + \psi_t(x) \right\}, \tag{2}$$

where each $\hat{\ell}_j$ is an unbiased estimator of ℓ_j, m_t is an arbitrary *hint vector* estimating ℓ_t, and ψ_t is a regularizer that is a smooth convex function over $\mathcal{X}$. The regret for this algorithm can be evaluated by means of Bregman divergences defined by $D_t(x, y) = \psi_t(x) - \psi_t(y) + \nabla \psi_t(y)^\top (x - y)$.

Lemma 1. *If x_t is given by* (2)*, we then have* $\sum_{t=1}^T \hat{\ell}_t^\top (x_t - x^*) \leq \psi_{T+1}(x^*) - \psi_1(x_1') + \sum_{t=1}^T \left((\hat{\ell}_t - m_t)^\top (x_t - x_{t+1}') - D_t(x_{t+1}', x_t) + \psi_t(x_{t+1}') - \psi_{t+1}(x_{t+1}') \right)$*, where x_t' is defined as* $x_t' \in \arg\min_{x \in \mathcal{X}} \left\{ \sum_{j=1}^{t-1} \hat{\ell}_j^\top x + \psi_t(x) \right\}$.

All omitted proofs are given in the Appendix. A similar framework is used in [36] for linear bandits. Further, a special case in which $m_t = 0$ has been employed in [71].

Remark 2. In some existing work, a slightly different approach called online mirror descent has been used, e.g., in [68]. In online mirror descent, the update rule is expressed as $x_t \in \arg\min_{x \in \mathcal{X}} \left\{ \hat{\ell}_t^\top x + D_t(x, x_{t-1}) \right\}$. The relationship between follow the regularized leader and online mirror descent has been widely discussed [51, 44]. Amir et al. [2] have pointed out an essential difference: an algorithm in the follow-the-regularized-leader framework has improved performance in stochastic regimes, but none in online mirror descent has done so.

4.2 Proposed algorithm

In our proposed algorithm, we define an unbiased estimator $\hat{\ell}_t$ and regularizer ψ_t as follows:

$$\hat{\ell}_{ti} = m_{ti} + \frac{a_{ti}}{x_{ti}} (\ell_{ti} - m_{ti}), \quad \psi_t(x) = \sum_{i=1}^d \beta_{ti} \left(-\log x_i + \gamma (1 - x_i) \log (1 - x_i) \right), \tag{3}$$

where β_{ti} and γ are defined as

$$\alpha_{ti} = a_{ti} (\ell_{ti} - m_{ti})^2 \min \left\{ 1, \frac{1 - x_{ti}}{\gamma x_{ti}^2} \right\}, \quad \beta_{ti} = \sqrt{2 + \frac{1}{\log T} \sum_{j=1}^{t-1} \alpha_{ji}}, \quad \gamma = \log T. \tag{4}$$

Our hybrid regularizer given in (3) is designed to lead to improved regret bounds in stochastic settings. In order to show BOTW regret bounds, it is necessary that round-wise regret bounds (e.g., the stability term in the paper by Zimmert et al. [71]) converge to 0 when x_t approaches extreme points in $\{0, 1\}^d$. As can be seen in Lemma 2 below, the round-wise regret bounds of our algorithm can be expressed as $O(\sum_{i=1}^d \frac{\alpha_{ti}}{\beta_{ti}}) = O(\sum_{i=1}^d \frac{x_{ti}}{\beta_{ti}} \min\{1, \frac{1 - x_{ti}}{\gamma x_{ti}^2}\})$ in expectation. Without hybrid regularization, i.e., if $\gamma = 0$, we cannot obtain the above-mentioned convergence property, particularly when x_{ti} approaches 1 for some i's. On the other hand, thanks to hybrid regularization (with $\gamma > 0$), we can show that the round-wise regret converge to 0 when approaching any points in $\{0, 1\}^d$. The learning rate parameters β_{ti} given in (4) are designed so that two main parts of the regret bound, $\sum_{t=1}^T \frac{\alpha_{ti}}{\beta_{ti}}$ and $\log T \cdot \beta_{T+1,i}$, will be well-balanced.

The optimistic predictor m_t is updated as follows:

$$m_{1i} = 1/4 \quad (i \in [d]), \quad m_{t+1,i} = m_{ti} + a_{ti} (\ell_{ti} - m_{ti})/4 \quad (t \in [T], i \in [d]). \tag{5}$$

The proposed algorithm can be summarized in Algorithm 1, which is similar to the one proposed in [71] in that both are based on the follow-the-regularized-leader framework with a round-dependent regularizer. The main differences are as follows:

- Algorithm 1 employs an optimistic-prediction framework while the algorithm in [71] does not, i.e., m_t is fixed to the zero vector for each t.

Algorithm 1 Hybrid algorithm for combinatorial semi-bandits

Require: Action set $\mathcal{A}$, time horizon $T \in \mathbb{N}$
1: Initialize $m_t \in [0,1]^d$ by $m_{1i} = 1/2$ for all $i \in [d]$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Compute x_t as (2), where $\hat{\ell}_j$, ψ_t and β_{ti} are defined in (3) and (4), respectively.
4: Pick a_t so that $\mathbf{E}[a_t|x_t] = x_t$, output a_t, and get feedback of ℓ_{ti} for each i such that $a_{ti} = 1$.
5: Compute $\hat{\ell}_t$, $\beta_{t+1,i}$ and m_{t+1} on the basis of (3), (4) and (5), respectively.
6: **end for**

- Algorithm 1 uses a hybrid regularizer combining the log-barrier and Shannon entropy given in (3), while Zimmert et al. [71] adopt the combination of Tsallis entropy with power $1/2$ and Shannon entropy defined as $\psi_t(x) = \frac{1}{\sqrt{t}} \sum_{i=1}^d (-\sqrt{x_i} + \gamma(1 - x_i)\log(1 - x_i))$.
- Algorithm 1 maintains the strength γ_{ti} for regularization that is different for each entry, and it updates each on the basis of historical data $\{x_j, a_j, (\ell_{ji})_{i \in I_t}\}_{j=1}^{t-1}$.

The reason for using a log-barrier regularizer is that it allows us to exploit the optimistic prediction framework, i.e., it provides regret bounds dependent on $(\ell_t - m_t)$. When m_t are non-zero vectors, $\hat{\ell}_t$ can be $O(1/x_{ti})$ negative values, which would make it even more difficult to bound the regret if Tsallis entropy were used. For details, see, e.g., [69] and the paragraph just after (RV) in [70]. In contrast to this, a log-barrier regularizer works well even for $O(1/x_{ti})$ negative losses, which is convenient for combining with an optimistic prediction framework.

4.3 Regret analysis

This subsection provides regret bounds achieved by Algorithm 1. First, as we have $\mathbf{E}[\hat{\ell}|x_t] = \ell_t$ from (3) and $\mathbf{E}[a_t|x_t] = x_t$, the regret can be bounded as

$$R_T(a^*) \leq \mathbf{E}\left[\sum_{t=1}^T \ell_t^\top (x_t - x^*)\right] + T\|a^* - x^*\|_1 = \mathbf{E}\left[\sum_{t=1}^T \hat{\ell}_t^\top (x_t - x^*)\right] + T\|a^* - x^*\|_1 \quad (6)$$

for any $x^* \in \mathcal{X}$. We set $x^* = (1 - \frac{d}{T})a^* + \frac{d}{T}x_0$, where x_0 is a point in $\mathcal{X}$ such that $x_{0i} \geq 1/d$ for all $i \in [d]$. The existence of such a point follows from the assumption that for any $i \in [d]$ there exists $a \in \mathcal{A}$ satisfying $a_i = 1$. The term $\|a^* - x^*\|_1$ in (6) can then be bounded as $\|a^* - x^*\|_1 = \frac{d}{T}\|a^* - x_0\|_1 \leq \frac{d^2}{T}$. Further, the term $\sum_{t=1}^T \hat{\ell}_t^\top (x_t - x^*)$ can be bounded via Lemma 1 and the following lemma:

Lemma 2. *Suppose $\hat{\ell}_t$ and ψ_t are given by* (3)*, respectively. The following part of the bound in Lemma 1 can then be bounded as $(\hat{\ell}_t - m_t)^\top (x_t - x'_{t+1}) - D_t(x'_{t+1}, x_t) = O\left(\sum_{i=1}^d \frac{\alpha_{ti}}{\beta_{ti}}\right)$, where α_{ti} is defined in* (4).

This lemma can be shown via standard techniques used, e.g., in [71, 68]. Combining Lemmas 1, 2 and (6), we obtain the following regret bound:

Theorem 1. *For Algorithm 1, the regret is bounded as $R_T = O\left(\log T \cdot \mathbf{E}\left[\sum_{i=1}^d \beta_{T+1,i}\right] + d^2\right)$ where β_{ti} is defined as* (4)*. Consequently, we have*

$$R_T = O\left(\sum_{i=1}^d \sqrt{\log T\, \mathbf{E}\left[\sum_{t=1}^T a_{ti}(\ell_{ti} - m_{ti})^2\right]} + d^2 + d\log T\right) \quad (7)$$

as well as

$$R_T = O\left(\sum_{i=1}^d \sqrt{\log T\, \mathbf{E}\left[\sum_{t=1}^T \min\left\{x_{ti}, \frac{1 - x_{ti}}{\sqrt{\log T}}\right\}\right]} + d^2 + d\log T\right). \quad (8)$$

Note that this theorem holds for arbitrary $m_t \in [0,1]^d$. The specific choice of m_t in (5) and (7) leads to the following bound in the adversarial regime:

Corollary 1 (Data-dependent bounds for adversarial regimes). *For Algorithm 1, the regret is bounded as* $R_T = O\left(\sqrt{d \log T \cdot \min\left\{\sum_{t=1}^T \ell_t^\top a^*, \sum_{t=1}^T \left\|\ell_t - \bar{\ell}\right\|_2^2, \sum_{t=1}^{T-1} \left\|\ell_t - \ell_{t+1}\right\|_1\right\}} + d \log T + d^2\right)$ *for any* $a^* \in \mathcal{A}$*, where* $\bar{\ell} = \frac{1}{T}\sum_{t=1}^T \ell_t$.

Further, from (8), we see that Algorithm 1 may offer improved performance in stochastic regimes (with adversarial corruptions), as follows:

Corollary 2 (Improved regret bounds for stochastic regimes with adversarial crruptions). *If the environment is in a stochastic regime with adversarial corruptions (defined in Subsection 2.3), Algorithm 1 has the following regret bound:* $R_T(a^*) = O\left(\frac{B(\mathcal{A}) \log T}{\Delta} + \sqrt{B(\mathcal{A}) \frac{Cm \log T}{\Delta}} + d^2\right)$, *where* $B(\mathcal{A}) \geq 0$ *is a constant dependent on the action set, bounded as*

$$B(\mathcal{A}) \leq \left\{ \begin{array}{ll} dm & (\textit{general cases}) \\ (d - m + m/\sqrt{\log T}) \min\{m, d-m\} & (\textit{size-invariant semi-bandits}) \\ d - m + m/\sqrt{\log T} & (\textit{matroid semi-bandits}) \end{array} \right. . \tag{9}$$

Remark 3. From Corollary 2, we can obtain regret bounds for the stochastic regime as well, by substituting $C = 0$. In a stochastic regime, the BOBW algorithm proposed by Zimmert et al. [71] has been shown to enjoy similar but slightly different regret bounds, e.g., $B(\mathcal{A}) \leq (d + m/\log T)m$ for general cases (which is slightly worse than in (9)), and $B(\mathcal{A}) \leq (d-m)(1 + (\log d)^2/\log T)$ for the cases of uniform matroids (which in general is not comparable to (9)). For a stochastic regime with adversarial corruptions, their algorithm achieves $O(\frac{B(\mathcal{A}) \log T}{\Delta} + Cm)$-regret for such a modified $B(\mathcal{A})$, though it is not known if the bound can be improved to an $O(\sqrt{C})$-type as in Corollary 2.

Remark 4. For matroid semi-bandits, we can state a more refined regret bound. For the optimal action $a^* \in \arg\min_{a \in \mathcal{A}} \mu^\top a$ set $J^* = \{i \in [d] \mid a_i^* = 0\}$ and denote $\Delta = \min_{a \in \mathcal{A}: a_i = 1} \mu^\top a - \mu^\top a^*$ for each $i \in J^*$. We then have $R_T(a^*) = O\left(\sum_{i \in J^*} \frac{\log T}{\Delta_i} + \frac{m\sqrt{\log T}}{\Delta} + \sqrt{Cm\left(\sum_{i \in J^*} \frac{\log T}{\Delta_i} + \frac{m\sqrt{\log T}}{\Delta}\right)}\right)$.

5 Linear bandits

5.1 Predictor-dependent regret bounds

Regret bounds dependent on m_t have been developed for linear bandit problems, similarly to what is seen for semi-bandits. This paper focuses on regret bounds in the following form:

$$R_T \leq D \cdot \mathbf{E}\left[\sqrt{\sum_{t=1}^T ((\ell_t - m_t)^\top a_t)^2}\right], \tag{10}$$

where D is a parameter dependent on $\mathcal{A}, d$ and T. Rakhlin and Sridharan [57] proposed the SCRiBLe algorithm, which achieves a regret bound as in (10) with $D = O(\vartheta d \log T)$, given a ϑ-self-concordant barrier over the convex hull of $\mathcal{A}$, if an appropriate learning rate is chosen. Even without self-concordant barriers, for general action sets, an algorithm proposed by Ito et al. [36] achieves a regret bound with $D = O(d \log T \cdot \log(dT))$, as is shown in Theorem 2 in their paper.

From (10), we can achieve small regret by choosing m_t so that $\sum_{t=1}^T g_t(m_t)$, where we define $g_t(m) = \frac{1}{2}((\ell_t - m)^\top a_t)^2$. When choosing m_t, we can use the information of $\{(\ell_j, \ell_j^\top a_j)\}_{j=1}^{t-1}$.

5.2 Tracking linear experts

In this subsection, we revisit the work by Herbster and Warmuth [31]. Let $p' = \min\{p, 2 \log d\}$ and let $q' \geq q$ be such that $1/p' + 1/q' = 1$. Define $\Phi_p(m) = \|m\|_p^2$. Let $\mathcal{L} = \{m \in \mathbb{R}^d \mid \|m\|_q \leq 1\}$. Consider $m_t \in \mathcal{L}$ defined as follows:

$$m_1 = 0, \quad m_{t+1} \in \arg\min_{m \in \mathcal{L}} \left\{\Phi_{p'}(m) + \left(\nabla \Phi_{q'}(m_t) - \frac{1}{2(p'-1)}\left(m_t^\top a_t - \ell_t^\top a_t\right) a_t\right)^\top m\right\}. \tag{11}$$

For m_t defined by (11), we have the following:

Theorem 2 ([31], Theorem 11.4 in [16])**.** *Suppose* $p \geq 2$. *If* m_t *is chosen by* (11)*, for any sequence* $\{u_t\}_{t=1}^T \subseteq \mathcal{L}$*, we have* $\sum_{t=1}^T g_t(m_t) \leq 2 \sum_{t=1}^T g_t(u_t) + 4(p'-1) \left(\sum_{t=1}^T \|u_t - u_{t+1}\|_q + 1 \right)$.

5.3 Hybrid data-dependent regret bound

By combining (10) and Theorem 2, we obtain the following regret bound:

Theorem 3. *Suppose* $p \geq 2$. *Suppose an algorithm enjoys a regret bound as* (10) *and* m_t *is chosen by* (11)*. We then have* $R_T = O\left(D \cdot \mathbf{E}\left[\sqrt{\min\{Q_q, p' \cdot V_q\} + p'} \right] \right)$ *for any* $u \in \mathcal{L}$*. Further, if* $\ell_t^\top a \geq 0$ *holds for all* $a \in \mathcal{A}$ *and all* $t \in [T]$*, it holds for all* $a^* \in \mathcal{A}$ *that* $R_T(a^*) = O\left(D \cdot \sqrt{\mathbf{E}\left[L^*\right] + p'} \right)$.[2]

By combining this theorem and (10) with $D = \tilde{O}(d)$ [36], we obtain the regret bound in Table 2.

On Potential Societal Impact This study is primarily theoretical in nature, and we do not see any negative social consequences. Researchers working on bandit theory may benefit from this paper. In the long run, we expect that the proposed algorithms, which are robust to adversarial attacks, have the potential to contribute to the realization of a safer and more secure society.

Acknowledgments and Disclosure of Funding

The author was supported by JST, ACT-I, Grant Number JPMJPR18U5, Japan.

References

[1] C. Allenberg, P. Auer, L. Györfi, and G. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *International Conference on Algorithmic Learning Theory*, pages 229–243, 2006.

[2] I. Amir, I. Attias, T. Koren, Y. Mansour, and R. Livni. Prediction with corrupted expert advice. *Advances in Neural Information Processing Systems*, 33, 2020.

[3] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

[4] J.-Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

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

[6] B. Awerbuch and R. Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

[7] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, pages 45–53, 2004.

[8] I. Bogunovic, A. Krause, and J. Scarlett. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1071–1081. PMLR, 2020.

[9] I. Bogunovic, A. Losalka, A. Krause, and J. Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021.

[10] R. A. Brualdi. Comments on bases in dependence structures. *Bulletin of the Australian Mathematical Society*, 1(2):161–167, 1969.

[2] Definitions of Q_q, V_q and L^* are given in Table 3. $p' = \min\{p, 2\log d\}$ as given in Section 5.2.

[11] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[12] S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1, 2012.

[13] S. Bubeck, N. Cesa-Bianchi, and S. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pages 41.1–41.14, 2012.

[14] S. Bubeck, M. Cohen, and Y. Li. Sparsity, variance and curvature in multi-armed bandits. In *Algorithmic Learning Theory*, pages 111–127. PMLR, 2018.

[15] S. Bubeck, Y. Li, H. Luo, and C.-Y. Wei. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, pages 508–528. PMLR, 2019.

[16] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[17] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[18] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

[19] W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159, 2013.

[20] W. Chen, Y. Wang, Y. Yuan, and Q. Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

[21] Y. Chen, S. Du, and K. Jamieson. Improved corruption robust algorithms for episodic reinforcement learning. In *International Conference on Machine Learning*, pages 1561–1570, 2021.

[22] V. Dani, S. M. Kakade, and T. P. Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2008.

[23] S. De Rooij, T. Van Erven, P. D. Grünwald, and W. M. Koolen. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1):1281–1316, 2014.

[24] L. Erez and T. Koren. Best-of-all-worlds bounds for online learning with feedback graphs. In *Advances in Neural Information Processing Systems, to appear*, 2021.

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

[26] P. Gaillard, G. Stoltz, and T. Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR, 2014.

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

[28] M. Hajiesmaili, M. S. Talebi, J. Lui, W. S. Wong, et al. Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm. *Advances in Neural Information Processing Systems*, 33, 2020.

[29] E. Hazan and S. Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

[30] E. Hazan and Z. Karnin. Volumetric spanners: An efficient exploration basis for learning. *Journal of Machine Learning Research*, 17(1):4062–4095, 2016.

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

[32] R. Huang, T. Lattimore, A. György, and C. Szepesvári. Following the leader and fast rates in linear prediction: Curved constraint sets and other regularities. In *Advances in Neural Information Processing Systems*, pages 4970–4978, 2016.

[33] S. Ito. On optimal robustness to adversarial corruption in online decision problems. In *Advances in Neural Information Processing Systems, to appear*, 2021.

[34] S. Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pages 2552–2583. PMLR, 2021.

[35] S. Ito, D. Hatano, H. Sumita, K. Takemura, T. Fukunaga, N. Kakimura, and K.-I. Kawarabayashi. Oracle-efficient algorithms for online linear optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, pages 10589–10598, 2019.

[36] S. Ito, S. Hirahara, T. Soma, and Y. Yoshida. Tight first- and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33:2028–2038, 2020.

[37] T. Jin and H. Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in Neural Information Processing Systems*, 33, 2020.

[38] T. Jin, L. Huang, and H. Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *arXiv preprint arXiv:2106.04117*, 2021.

[39] K.-S. Jun, L. Li, Y. Ma, and X. Zhu. Adversarial attacks on stochastic bandits. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3644–3653, 2018.

[40] B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 420–429, 2014.

[41] B. Kveton, Z. Wen, A. Ashkan, and M. Valko. Learning to act greedily: Polymatroid semi-bandits. *arXiv preprint arXiv:1405.7752*, 2014.

[42] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 535–543, 2015.

[43] J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *The Journal of Machine Learning Research*, 17(1):8106–8137, 2016.

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

[45] C.-W. Lee, H. Luo, C.-Y. Wei, M. Zhang, and X. Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pages 6142–6151, 2021.

[46] F. Liu and N. Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pages 4042–4050. PMLR, 2019.

[47] H. Luo and R. E. Schapire. Achieving all with no parameters: Adanormalhedge. In *Conference on Learning Theory*, pages 1286–1304. PMLR, 2015.

[48] T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

[49] T. Lykouris, M. Simchowitz, A. Slivkins, and W. Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

[50] S. Masoudian and Y. Seldin. Improved analysis of the tsallis-inf algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 3330–3350. PMLR, 2021.

[51] B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 525–533. JMLR Workshop and Conference Proceedings, 2011.

[52] J. Mourtada and S. Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20:1–28, 2019.

[53] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

[54] G. Neu. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory*, pages 1360–1375. PMLR, 2015.

[55] R. Pogodin and T. Lattimore. On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In *Uncertainty in Artificial Intelligence*, pages 894–904, 2020.

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

[57] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

[58] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. In *Conference on Learning Theory*, pages 559–594, 2011.

[59] C. Rouyer and Y. Seldin. Tsallis-inf for decoupled exploration and exploitation in multi-armed bandits. In *Conference on Learning Theory*, pages 3227–3249. PMLR, 2020.

[60] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, 1998.

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

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

[63] J. Steinhardt and P. Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, pages 1593–1601. PMLR, 2014.

[64] D. Suehiro, K. Hatano, S. Kijima, E. Takimoto, and K. Nagano. Online prediction under submodular constraints. In *International Conference on Algorithmic Learning Theory*, pages 260–274. Springer, 2012.

[65] K. Takemura and S. Ito. An arm-wise randomization approach to combinatorial linear semi-bandits. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1318–1323. IEEE, 2019.

[66] S. Wang and W. Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122. PMLR, 2018.

[67] M. K. Warmuth and D. Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9(Oct):2287–2320, 2008.

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

[69] J. Zimmert and T. Lattimore. Connections between mirror descent, thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

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

A Appendix

A.1 Proof of Lemma 1

From the definition of x_t and x'_t shown in (2) and the first-order optimality condition, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \hat{\ell}_t^\top x^* + \psi_{T+1}(x^*) \geq \sum_{t=1}^{T} \hat{\ell}_t^\top x'_{T+1} + \psi_{T+1}(x'_{T+1}) \\
&= \left(\sum_{t=1}^{T-1} \hat{\ell}_t + m_T \right)^\top x'_{T+1} + \psi_T(x'_{T+1}) + \psi_{T+1}(x'_{T+1}) - \psi_T(x'_{T+1}) + (\hat{\ell}_T - m_T)^\top x'_{T+1} \\
&\geq \left(\sum_{t=1}^{T-1} \hat{\ell}_t + m_T \right)^\top x_T + \psi_T(x_T) + \psi_{T+1}(x'_{T+1}) - \psi_T(x'_{T+1}) + (\hat{\ell}_T - m_T)^\top x'_{T+1} + D_T(x'_{T+1}, x_T) \\
&= \sum_{t=1}^{T-1} \hat{\ell}_t^\top x_T + \psi_T(x_T) + m_T^\top x_T + \psi_{T+1}(x'_{T+1}) - \psi_T(x'_{T+1}) + (\hat{\ell}_T - m_T)^\top x'_{T+1} + D_T(x'_{T+1}, x_T) \\
&\geq \sum_{t=1}^{T-1} \hat{\ell}_t^\top x'_T + \psi_T(x'_T) + m_T^\top x_T + \psi_{T+1}(x'_{T+1}) - \psi_T(x'_{T+1}) + (\hat{\ell}_T - m_T)^\top x'_{T+1} + D_T(x'_{T+1}, x_T) \\
&\geq \cdots \geq \psi_1(x'_1) + \sum_{t=1}^{T} \Big(\psi_{t+1}(x'_{t+1}) - \psi_t(x'_{t+1}) + m_t^\top x_t + (\hat{\ell}_t - m_t)^\top x'_{t+1} + D_t(x'_{t+1}, x_t) \Big),
\end{aligned}
$$

where the first and the third inequalities follow from the definition of x'_t, the second inequality follows from the definition (2) of x_t and the first-order optimality condition, and the last inequality is obtained by applying similar arguments recursively. This inequality immediately implies the bound in Lemma 1. $\square$

For the special case in which $m_t = 0$, a proof can be found in the literature, e.g., in Exercise 28.12 of the book by Lattimore and Szepesvári [44].

A.2 Proof of Lemma 2

For the sake of simplicity, we here assume $T \geq 3$ and, consequently, have $\gamma = \log T \geq 1$. For each $i \in [d]$, we denote

$$
\psi_{ti}^{(1)}(x_{ti}) = -\beta_{ti} \log(x_i), \quad \psi_{ti}^{(2)}(x) = \gamma \beta_{ti} (1 - x_i) \log(1 - x_i) \tag{12}
$$

and let $D_{ti}^{(1)}$ and $D_{ti}^{(2)}$ be the Bregman divergences over $\mathbb{R}_{>0}$, corresponding to $\psi_{ti}^{(1)}$ and $\psi_{ti}^{(2)}$, respectively. As ψ_t can be expressed as $\psi_t(x) = \sum_{i=1}^d (\psi_{ti}^{(1)}(x_i) + \psi_{ti}^{(2)}(x_i))$, from the linearity of Bregman divergences, we have $D_t(x, y) = \sum_{i=1}^d (D_{ti}^{(1)}(x_i, y_i) + D_{ti}^{(2)}(x_i, y_i))$. We hence have

$$
\begin{aligned}
&(\hat{\ell}_t - m_t)^\top (x_t - x'_{t+1}) - D_t(x'_{t+1}, x_t) \\
&= \sum_{i=1}^{d} \Big((\hat{\ell}_{ti} - m_{ti})(x_{ti} - x'_{t+1,i}) - D_{ti}^{(1)}(x'_{t+1,i}, x_{ti}) - D_{ti}^{(1)}(x'_{t+1,i}, x_{ti}) \Big) \\
&\leq \sum_{i=1}^{d} \min \Big\{ (\hat{\ell}_{ti} - m_{ti})(x_{ti} - x'_{t+1,i}) - D_{ti}^{(1)}(x'_{t+1,i}, x_{ti}), (\hat{\ell}_{ti} - m_{ti})(x_{ti} - x'_{t+1,i}) - D_{ti}^{(2)}(x'_{t+1,i}, x_{ti}) \Big\}.
\end{aligned} \tag{13}
$$

We show below that

$$
(\hat{\ell}_{ti} - m_{ti})^\top (x_{ti} - x) - D_{ti}^{(1)}(x, x_{ti}) \leq \beta_{ti} h^{(1)} \left(\frac{a_{ti}(\ell_{ti} - m_{ti})}{\beta_{ti}} \right) \tag{14}
$$

$$
(\hat{\ell}_{ti} - m_{ti})^\top (x_{ti} - x) - D_{ti}^{(2)}(x, x_{ti}) \leq \gamma \beta_{ti} (1 - x_{ti}) h^{(2)} \left(\frac{a_{ti}(\ell_{ti} - m_{ti})}{\gamma \beta_{ti} x_{ti}} \right) \tag{15}
$$

hold for any $x \in \mathbb{R}_{>0}$, where we define $h^{(1)}(z) = z - \log(z+1)$ and $h^{(2)}(z) = \exp(z) - z - 1$.

Let us first show (14). From the first-order optimality condition, the left-hand side of (14) is maximized when x satisfies

$$-\hat{\ell}_{ti} + m_{ti} - \nabla\psi_{ti}^{(1)}(x) + \nabla\psi_{ti}^{(1)}(x_{ti}) = 0, \tag{16}$$

which can be rewritten as

$$\frac{1}{x} = \frac{1}{x_{ti}} + \frac{\hat{\ell}_{ti} - m_{ti}}{\beta_{ti}}, \quad \text{and equivalently,} \quad \frac{x_{ti}}{x} = 1 + \frac{x_{ti}(\hat{\ell}_{ti} - m_{ti})}{\beta_{ti}}. \tag{17}$$

For such x, the left-hand side of (14) can be expressed as

$$\begin{aligned}
&(\hat{\ell}_{ti} - m_{ti})(x_{ti} - x) - D_{ti}^{(1)}(x, x_{ti}) \\
&= (\hat{\ell}_{ti} - m_{ti})(x_{ti} - x) + \beta_{ti}\left(\log x - \log x_{ti} - \frac{1}{x_{ti}}(x - x_{ti})\right) \\
&= \left(\hat{\ell}_{ti} - m_{ti} + \frac{\beta_{ti}}{x_{ti}}\right)(x_{ti} - x) + \beta_{ti}\left(-\log\left(\frac{x_{ti}(\hat{\ell}_{ti} - m_t)}{\beta_{ti}} + 1\right)\right) \\
&= \frac{\beta_{ti}}{x}(x_{ti} - x) + \beta_{ti}\left(-\log\left(\frac{x_{ti}(\hat{\ell}_{ti} - m_{ti})}{\beta_{ti}} + 1\right)\right) \\
&= \beta_{ti}\left(-\log\left(\frac{x_{ti}(\hat{\ell}_{ti} - m_{ti})}{\beta_{ti}} + 1\right) + \frac{x_{ti}(\hat{\ell}_{ti} - m_{ti})}{\beta_{ti}}\right) \\
&= \beta_{ti} h^{(1)}\left(\frac{x_{ti}(\hat{\ell}_{ti} - m_{ti})}{\beta_{ti}}\right) = \beta_{ti} h^{(1)}\left(\frac{a_{ti}(\ell_{ti} - m_{ti})}{\beta_{ti}}\right),
\end{aligned}$$

where the second, third, and fourth equalities follow from (17), and the last equality follows from the definition of $\hat{\ell}_t$ in (3).

Let us next show (15). From the first-order optimality condition, the left-hand side of (15) is maximized when x satisfies

$$\log(1 - x) = \log(1 - x_{ti}) + \frac{\hat{\ell}_{ti} - m_{ti}}{\gamma\beta_{ti}}. \tag{18}$$

The left-hand side of (15) can then be expressed as

$$\begin{aligned}
&(\hat{\ell}_{ti} - m_{ti})(x_{ti} - x) - D_{ti}^{(2)}(x, x_{ti}) \\
&= (\hat{\ell}_{ti} - m_{ti} + \gamma\beta_{ti}(\log(1 - x_{ti}) + 1))(x_{ti} - x_i) - \gamma\beta_{ti}\left((1 - x_i)\log(1 - x_i) - (1 - x_{ti})\log(1 - x_{ti})\right) \\
&= \gamma\beta_{ti}\left((\log(1 - x_i) + 1)\cdot(x_{ti} - x_i) - ((1 - x_i)\log(1 - x_i) - (1 - x_{ti})\log(1 - x_{ti}))\right) \\
&= \gamma\beta_{ti}\left((x_{ti} - 1)\log(1 - x_i) + (1 - x_{ti})\log(1 - x_{ti}) + x_{ti} - x_i\right) \\
&= \gamma\beta_{ti}(1 - x_{ti})\left(\log\frac{1 - x_{ti}}{1 - x_i} - 1 + \frac{1 - x_i}{1 - x_{ti}}\right) \\
&= \gamma\beta_{ti}(1 - x_{ti})\left(\exp\left(\frac{\hat{\ell}_{ti} - m_{ti}}{\gamma\beta_{ti}}\right) - \frac{\hat{\ell}_{ti} - m_{ti}}{\gamma\beta_{ti}} - 1\right) \\
&= \gamma\beta_{ti}(1 - x_{ti})h^{(2)}\left(\frac{\hat{\ell}_{ti} - m_{ti}}{\gamma\beta_{ti}}\right) = \gamma\beta_{ti}(1 - x_{ti})h^{(2)}\left(\frac{a_{ti}(\ell_{ti} - m_{ti})}{\gamma\beta_{ti}x_{ti}}\right).
\end{aligned}$$

Combining (13), (14) and (15), we obtain

$$\begin{aligned}
&(\hat{\ell}_t - m_t)^\top(x_t - x'_{t+1}) - D_t(x'_{t+1}, x_t) \\
&\leq \sum_{i=1}^{d} \beta_{ti} \min\left\{h^{(1)}\left(\frac{a_{ti}(\ell_{ti} - m_{ti})}{\beta_{ti}}\right), \gamma(1 - x_{ti})h^{(2)}\left(\frac{a_{ti}(\ell_{ti} - m_{ti})}{\gamma\beta_{ti}x_{ti}}\right)\right\}.
\end{aligned} \tag{19}$$

As we have $h^{(1)}(z) \leq 2z^2$ for $|z| \leq \frac{1}{\sqrt{2}}$ and $h^{(2)}(z) \leq z^2$ for $|z| \leq \sqrt{2}$, from $\beta_{ti} \geq \sqrt{2}$, we have

$$
\begin{aligned}
&\min\left\{ h^{(1)}\left(\frac{a_{ti}(\ell_{ti}-m_{ti})}{\beta_{ti}}\right), \gamma(1-x_{ti})h^{(2)}\left(\frac{a_{ti}(\ell_{ti}-m_{ti})}{\gamma\beta_{ti}x_{ti}}\right)\right\} \\
&\leq \begin{cases} \frac{2a_{ti}(\ell_{ti}-m_{ti})^2}{\beta_{ti}^2} & (\gamma\beta_{ti}x_{ti} < \frac{1}{\sqrt{2}}) \\ \min\left\{\frac{2a_{ti}(\ell_{ti}-m_{ti})^2}{\beta_{ti}^2}, \frac{a_{ti}(1-x_{ti})(\ell_{ti}-m_{ti})^2}{\gamma\beta_{ti}^2 x_{ti}^2}\right\} & (\gamma\beta_{ti}x_{ti} \geq \frac{1}{\sqrt{2}}) \end{cases} \\
&\leq \frac{2a_{ti}(\ell_{ti}-m_{ti})^2}{\beta_{ti}^2}\min\left\{1, \frac{1-x_{ti}}{\gamma x_{ti}^2}\right\} = \frac{2\alpha_{ti}}{\beta_{ti}^2}.
\end{aligned}
$$

The last inequality can be confirmed as follows: If $\frac{1-x_{ti}}{\gamma x_{ti}^2} \leq 1$, then $1 - x_{ti} \leq \gamma x_{ti}^2 \leq \gamma x_{ti}$, which implies $x_{ti} \geq \frac{1}{\gamma+1} \geq \frac{1}{2\gamma} \geq \frac{1}{\sqrt{2}\gamma\beta_{ti}}$. Combining this with (19), we obtain

$$
(\hat{\ell}_t - m_t)^\top (x_t - x'_{t+1}) - D_t(x'_{t+1}, x_t) \leq 2\sum_{i=1}^{d} \frac{\alpha_{ti}}{\beta_{ti}}.
$$

□

A.3 Proof of Theorem 1

From (6), the definition of x^*, and Lemmas 1 and 2, we have

$$
\begin{aligned}
R_T(a^*) &\leq \mathbf{E}\left[\psi_{T+1}(x^*) - \psi_1(x'_1) + 2\sum_{t=1}^{T}\left(\psi_t(x'_{t+1}) - \psi_{t+1}(x'_{t+1}) + \sum_{i=1}^{d}\frac{\alpha_{ti}}{\beta_{ti}}\right)\right] + d^2 \\
&\leq \mathbf{E}\left[\sum_{i=1}^{d}\beta_{T+1,i}\log\frac{1}{x_i^*} + \frac{\gamma}{\mathrm{e}}\sum_{i=1}^{d}\beta_{1i} + 2\sum_{t=1}^{T}\sum_{i=1}^{d}\left(\frac{\alpha_{ti}}{\beta_{ti}} + \frac{\gamma}{\mathrm{e}}(\beta_{t+1,i} - \beta_{ti})\right)\right] + d^2 \\
&\leq \mathbf{E}\left[\left(\log T + \frac{\gamma}{\mathrm{e}}\right)\sum_{i=1}^{d}\beta_{T+1,i} + 2\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\alpha_{ti}}{\beta_{ti}}\right] + d^2, \qquad (20)
\end{aligned}
$$

where the second inequality follows from $\log x'_{ti} - \gamma(1-x'_{ti})\log(1-x'_{ti}) \leq \gamma/\mathrm{e}$ as $x'_{ti} \in (0,1)$ and the last inequality follows from the definition of x^*. We further have

$$
\sum_{t=1}^{T}\frac{\alpha_{ti}}{\beta_{ti}} \leq 2\log T \cdot \beta_{T+1,i}. \qquad (21)
$$

In fact, for β'_{ti} defined by $\beta'_{ti} = \sqrt{\frac{1}{\log T}\sum_{j=1}^{t-1}\alpha_{ji}} \quad (\leq \beta_{t-1,i})$, we have

$$
\beta'_{t+1,i} - \beta'_{ti} = \frac{1}{\beta'_{t+1,i} + \beta'_{ti}}\frac{\alpha_{ti}}{\log T} \geq \frac{\alpha_{ti}}{2\beta_{ti}\log T},
$$

which implies

$$
\sum_{t=1}^{T}\frac{\alpha_{ti}}{\beta_{ti}} \leq 2\log T \cdot \sum_{t=1}^{T}(\beta'_{t+1,i} - \beta'_{ti}) = 2\log T \cdot \beta'_{T+1,i} \leq 2\log T \cdot \beta_{T+1,i}.
$$

Combining (20) and (21), and substituting $\gamma = \log T$, we obtain

$$
R_T \leq 6\log T \cdot \mathbf{E}\left[\sum_{i=1}^{d}\beta_{T+1,i}\right] + d^2. \qquad (22)
$$

From this and the definitions of α_{ti} and β_{ti} in (4), we have (7) and (8). In fact, as we have

$$
\beta_{T+1,i} = \sqrt{2 + \frac{1}{\log T}\sum_{t=1}^{T}\alpha_{ti}} \leq \sqrt{2} + \sqrt{\frac{1}{\log T}\sum_{t=1}^{T}\alpha_{ti}} \qquad (23)
$$

from (4), by combining this with (22), we obtain

$$
\begin{aligned}
R_T &\le 6 \log T \cdot \mathbf{E}\left[\sum_{i=1}^{d}\left(\sqrt{2} + \sqrt{\frac{1}{\log T}\sum_{t=1}^{T}\alpha_{ti}}\right)\right] + d^2 \\
&\le 6 \cdot \mathbf{E}\left[\sum_{i=1}^{d}\left(\sqrt{\log T \sum_{t=1}^{T}\alpha_{ti}}\right)\right] + d^2 + 9d\log T. \qquad (24)
\end{aligned}
$$

From this and $\alpha_{ti} \le a_{ti}(\ell_{ti} - m_{ti})^2$, which immediately follows from the definition of α_{ti} in (4), we have

$$
R_T \le 6 \cdot \mathbf{E}\left[\sum_{i=1}^{d}\sqrt{\log T \sum_{t=1}^{T} a_{ti}(\ell_{ti} - m_{ti})^2}\right] + d^2 + 9d\log T. \qquad (25)
$$

Further, as we have $(\ell_{ti} - m_{ti})^2 \le 1$ and $\mathbf{E}[a_{ti}|x_{ti}] = x_{ti}$, we have

$$
\mathbf{E}[\alpha_{ti}] \le \mathbf{E}\left[a_{ti}\min\left\{1, \frac{1 - x_{ti}}{\gamma x_{ti}^2}\right\}\right] = \mathbf{E}\left[\min\left\{x_{ti}, \frac{1 - x_{ti}}{\gamma x_{ti}}\right\}\right] \le 2\,\mathbf{E}\left[\min\left\{x_{ti}, \frac{1 - x_{ti}}{\sqrt{\gamma}}\right\}\right]. \qquad (26)
$$

In fact, if $\min\left\{x_{ti}, \frac{1-x_{ti}}{\sqrt{\gamma}}\right\} = \frac{1-x_{ti}}{\sqrt{\gamma}}$, we have $x_{ti} \ge \frac{1}{1+\sqrt{\gamma}} \ge \frac{1}{2\sqrt{\gamma}}$, which implies $\frac{1-x_{ti}}{\gamma x_{ti}} \le 2\frac{1-x_{ti}}{\sqrt{\gamma}}$. Combining (24) and (26), substituting $\gamma = \log T$, and applying Jensen's inequality, we obtain

$$
R_T \le 9 \cdot \sum_{i=1}^{d}\sqrt{\log T\,\mathbf{E}\left[\sum_{t=1}^{T}\min\left\{x_{ti}, \frac{1 - x_{ti}}{\sqrt{\log T}}\right\}\right]} + d^2 + 9d\log T. \qquad (27)
$$

A.4 Proof of Corollary 1

We start with showing the following lemma:

Lemma 3. *If m_t is given by* (5)*, it holds for any $\{u_t\}_{t=1}^{T} \subseteq [0,1]^d$ that*

$$
\sum_{t=1}^{T} a_{ti}(\ell_{ti} - m_{ti})^2 \le 2\sum_{t=1}^{T} a_{ti}(\ell_{ti} - u_{ti})^2 + \sum_{t=1}^{T} 16|u_{ti} - u_{t+1,i}| + 1 \qquad (28)
$$

for any $i \in [d]$.

Proof. From the definition (5) of m_t, we have

$$
\begin{aligned}
&a_{ti}(\ell_{ti} - m_{ti})^2 - a_{ti}(\ell_{ti} - u_{ti})^2 = a_{ti}(2\ell_{ti} - m_{ti} - u_{ti})(u_{ti} - m_{ti}) \\
&\le 2a_{ti}(\ell_{ti} - m_{ti})(u_{ti} - m_{ti}) \\
&= 2a_{ti}(\ell_{ti} - m_{ti})(m_{t+1,i} - m_{ti}) + 2a_{ti}(\ell_{ti} - m_{ti})(u_{ti} - m_{t+1,i}) \\
&= \frac{1}{2}a_{ti}(\ell_{ti} - m_{ti})^2 + 8(m_{t+1,i} - m_{ti})(u_{ti} - m_{t+1,i}) \\
&\le \frac{1}{2}a_{ti}(\ell_{ti} - m_{ti})^2 + 4((u_{ti} - m_{ti})^2 - (u_{ti} - m_{t+1,i})^2),
\end{aligned}
$$

where the third inequality follows from $m_{t+1,i} - m_{ti} = \frac{1}{4}a_{ti}(\ell_{ti} - m_{ti})$. We hence have

$$
a_{ti}(\ell_{ti} - m_{ti})^2 \le 2a_{ti}(\ell_{ti} - u_{ti})^2 + 8((u_{ti} - m_{ti})^2 - (u_{ti} - m_{t+1,i})^2).
$$

By taking the summation of this for $t \in [T]$, we obtain

$$
\begin{aligned}
&\sum_{t=1}^{T} a_{ti}(\ell_{ti} - m_{ti})^2 \leq 2\sum_{t=1}^{T} a_{ti}(\ell_{ti} - u_{ti})^2 + 8\sum_{t=1}^{T}((u_{ti} - m_{ti})^2 - (u_{ti} - m_{t+1,i})^2) \\
&\leq 2\sum_{t=1}^{T} a_{ti}(\ell_{ti} - u_{ti})^2 + 8\sum_{t=1}^{T-1}((u_{t+1,i} - m_{t+1,i})^2 - (u_{ti} - m_{t+1,i})^2) + (u_{1i} - m_{1i})^2 \\
&\leq 2\sum_{t=1}^{T} a_{ti}(\ell_{ti} - u_{ti})^2 + 8\sum_{t=1}^{T-1}(u_{t+1,i} + u_{ti} - 2m_{t+1,i})(u_{t+1,i} - u_{ti}) + 1 \\
&\leq 2\sum_{t=1}^{T} a_{ti}(\ell_{ti} - u_{ti})^2 + \sum_{t=1}^{T+1} 16|u_{t+1,i} - u_{ti}| + 1.
\end{aligned}
$$

□

Note that $\{u_t\}_{t=1}^T$ in this lemma does not appear in the algorithm and is used only in the analysis. Lemma 3 can be seen as a special case of Theorem 11.4 in [16].

Proof of Corollary 1. We first show $R_T = O\left(\sqrt{d \log T\, \mathbf{E}\left[\sum_{t=1}^T \ell_t^\top a^*\right]} + d^2 + d\log T\right)$. By substituting $u_{ti} = 0$ for all $t \in [T]$ and $i \in [d]$, from (28), we obtain

$$
\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} a_{ti}(\ell_{ti} - m_{ti})^2\right] &\leq 2\,\mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} a_{ti}\ell_{ti}^2\right] + 1 \leq 2\,\mathbf{E}\left[\sum_{t=1}^{T} \ell_t^\top a_t\right] + 1 \\
&= 2\left(R_T(a^*) + \mathbf{E}\left[\sum_{t=1}^{T} \ell_t^\top a^*\right]\right) + 1,
\end{aligned}
$$

where the first inequality follows from (28) with $u_{ti} = 0$, the second inequality follows from $\ell_t \in [0,1]^d$ and $a_t \in \{0,1\}^d$, and the last equality follows from the definition of $R_T(a^*)$ in (1). Combining this with (25) and applying Jensen's inequality, we obtain

$$
\begin{aligned}
R_T(a^*) &\leq 6 \cdot \mathbf{E}\left[\sum_{i=1}^{d}\sqrt{\log T \sum_{t=1}^{T} a_{ti}(\ell_{ti} - m_{ti})^2}\right] + d^2 + 9d\log T \\
&\leq 6 \cdot \mathbf{E}\left[\sqrt{d\log T \sum_{i=1}^{d}\sum_{t=1}^{T} a_{ti}(\ell_{ti} - m_{ti})^2}\right] + d^2 + 9d\log T \\
&\leq 6 \cdot \sqrt{d\log T\, \mathbf{E}\left[\sum_{i=1}^{d}\sum_{t=1}^{T} a_{ti}(\ell_{ti} - m_{ti})^2\right]} + d^2 + 9d\log T \\
&\leq 6 \cdot \sqrt{2d\log T\left(R_T(a^*) + \mathbf{E}\left[\sum_{t=1}^{T} \ell_t^\top a^*\right] + 1\right)} + d^2 + 9d\log T,
\end{aligned}
$$

where the second inequality follows from the Cauchy–Schwarz inequality. By solving the quadratic inequation $(R_T(a^*) - d^2 - 9d\log T)^2 \leq 72 d\log T\left(R_T(a^*) + \mathbf{E}\left[\sum_{t=1}^T \ell_t^\top a^*\right] + 1\right)$ in $R_T(a^*)$, we obtain $R_T(a^*) = O\left(\sqrt{d\log T\, \mathbf{E}\left[\sum_{t=1}^T \ell_t^\top a^*\right]} + d^2 + d\log T\right)$.

Similarly, we can show that $R_T = O\left(\sqrt{d\log T\, \mathbf{E}\left[\sum_{t=1}^T \sum_{i=1}^d a_{ti}(\ell_{ti} - \bar{\ell}_i)^2\right]} + d^2 + d\log T\right) = O\left(\sqrt{d\log T\, \mathbf{E}\left[\sum_{t=1}^T \|\ell_t - \bar{\ell}\|_2^2\right]} + d^2 + d\log T\right)$ by substituting $u_t = \bar{\ell}$ for all $t \in [T]$, to (28).

We can show that $R_T = O\left(\sqrt{d \log T \, \mathbf{E}\left[\sum_{t=1}^{T-1} \|\ell_t - \ell_{t+1}\|_1\right]} + d^2 + d \log T\right)$ as well by substituting $u_t = \ell_t$ for all $t \in [T]$, into (28). $\square$

A.5 Proof of Corollary 2

We provide improved regret upper bounds via the following regret *lower* bounds:

Lemma 4. *For $a^* \in \mathcal{A}$, denote $I^* = \{i \in [d] \mid a^*_i = 1\}$ and $J^* = [d] \setminus I^*$. In a stochastic regime with adversarial corruptions, for any algorithm, the regret is bounded from below as*

$$R_T(a^*) \geq \frac{\Delta}{B'(\mathcal{A})} \mathbf{E}\left[\sum_{t=1}^{T}\left(\sum_{i \in I^*}(1 - a_{ti}) + \sum_{i \in J*} a_{ti}\right)\right] - 2Cm, \tag{29}$$

where $B'(\mathcal{A}) > 0$ is defined as

$$B'(\mathcal{A}) = \begin{cases} 2m & (\textit{general cases}) \\ 2\min\{m, d-m\} & (\textit{size-invariant semi-bandits}) \\ 2 & (\textit{matroid semi-bandits}) \end{cases}. \tag{30}$$

Further, for matroid semi-bandits, we have

$$R_T(a^*) \geq \frac{1}{2} \mathbf{E}\left[\sum_{t=1}^{T}\left(\Delta \sum_{i \in I^*}(1 - a_{ti}) + \sum_{i \in J*} \Delta_i a_{ti}\right)\right] - 2Cm, \tag{31}$$

where we define $\Delta_i = \min_{a \in \mathcal{A}: a_i = 1} \mu^\top a - \mu^\top a^*$.

Proof. Let us recall the conditions in a stochastic regime with adversarial corruptions:

$$\mathbf{E}[\ell'_t] = \mu \quad (t \in [T]), \tag{32}$$

$$\sum_{t=1}^{T} \|\ell_t - \ell'_t\|_\infty \leq C, \tag{33}$$

$$a^* \in \arg\min_{a \in \mathcal{A}} \mu^\top a, \tag{34}$$

$$\Delta = \min_{a \in \mathcal{A} \setminus \{a^*\}} \mu^\top a - \mu^\top a^* > 0. \tag{35}$$

From these conditions, we have

$$\begin{aligned} R_T(a^*) &= \mathbf{E}\left[\sum_{t=1}^{T} \ell_t^\top (a_t - a^*)\right] = \mathbf{E}\left[\sum_{t=1}^{T} \ell_t'^\top (a_t - a^*) + \sum_{t=1}^{T} (\ell_t - \ell'_t)^\top (a_t - a^*)\right] \\ &\geq \mathbf{E}\left[\sum_{t=1}^{T} \mu^\top (a_t - a^*) - \sum_{t=1}^{T} \|\ell_t - \ell'_t\|_\infty \|a_t - a^*\|_1\right] \\ &\geq \mathbf{E}\left[\sum_{t=1}^{T} \mu^\top (a_t - a^*) - 2m \sum_{t=1}^{T} \|\ell_t - \ell'_t\|_\infty\right] \\ &\geq \mathbf{E}\left[\sum_{t=1}^{T} \mu^\top (a_t - a^*)\right] - 2Cm \geq \mathbf{E}\left[\sum_{t=1}^{T} \Delta \cdot \mathbf{1}[a_t \neq a^*]\right] - 2Cm, \end{aligned} \tag{36}$$

where the first, the third and the last inequality follows from (32), (33) and (35), respectively, and the second inequality follows from $\|a\|_1 \leq m$ for all $a \in \mathcal{A}$. As we have $\|a_t - a^*\|_1 \leq 2m$ for any $a_t \in \mathcal{A}$, we have

$$\mathbf{1}[a_t \neq a^*] \geq \frac{1}{2m} \|a_t - a^*\|_1 = \frac{1}{2m}\left(\sum_{i \in I^*}(1 - a_{ti}) + \sum_{i \in J*} a_{ti}\right). \tag{37}$$

Combining this with (36), we obtain (29) in which $\mathcal{B}'(\mathcal{A}) = 2m$. Similarly for the case of size-invariant semi-bandits, as we have $\|a_t - a^*\|_1 \leq 2\min\{m, d-m\}$ for any a_t, we obtain (29) in which $\mathcal{B}'(\mathcal{A}) = 2\min\{m, d-m\}$.

For the matroid case, we have

$$\mu^\top (a - a^*) \geq \sum_{i \in J^*} \Delta_i a_i \tag{38}$$

for any $a \in \mathcal{A}$. This can be shown via the symmetric basis-exchange property of matroid bases [10]. Denote $I = \{i \in [d] \mid a_i = 1\}$ and $k = |I \setminus I^*|$. We consider the following sequence of bases $\{I_j\}_{j=0}^k$:

- Set $I_0 = I^*$.
- For $j = 0, 1, \ldots, k-1$: choose $i_j \in I_j \setminus I^*$ arbitrarily. From the symmetric basis-exchange property, there exists $i'_j \in I^* \setminus I_j$ such that both $(I_j \cup i'_j) \setminus \{i_j\}$ and $(I^* \cup i_j) \setminus \{i'_j\}$ are bases. Let $I_{j+1} = (I_j \cup i'_j) \setminus \{i_j\}$.

As we have $|I_{j+1} \setminus I^*| = |I_j \setminus I^*| - 1$ for $j \in [k-1]$, we have $|I_k \setminus I^*| = 0$, and consequently, $I_k = I^*$ holds. Similarly, we can see that $\{i_j\}_{j=1}^k = I \setminus I^*$. The value of $\mu^\top (a - a^*)$ can be expressed as

$$\mu^\top (a - a^*) = \sum_{j=1}^k (\mu_{i_j} - \mu_{i'_j}) \geq \sum_{j=1}^k \Delta_{i_j} = \sum_{i \in I \setminus I^*} \Delta_i = \sum_{i \in J^*} \Delta_i a_i, \tag{39}$$

where the inequality follows from the definition of Δ_i and the fact that $(I^* \cup i_j) \setminus \{i'_j\}$ is a base. From this, $\Delta = \min_{i \in J^*} \Delta_i$, and the fact that $\sum_{i \in J^*} a_i = \sum_{i \in I^*} (1 - a_i)$, we have

$$\mu^\top (a - a^*) \geq \frac{\Delta}{2} \sum_{i \in J^*} a_i + \frac{1}{2} \sum_{i \in J^*} \Delta_i a_i = \frac{\Delta}{2} \sum_{i \in I^*} (1 - a_i) + \frac{1}{2} \sum_{i \in J^*} \Delta_i a_i. \tag{40}$$

Combining this with (36), we obtain (31). Further, it follows from $\Delta \leq \Delta_i$ for any $i \in J^*$ that (29) with $B'(\mathcal{A}) = 2$ holds. ☐

Proof of Corollary 2. From (27) and (29), for any $\lambda > 0$, we have

$$\begin{aligned}
R_T(a^*) &= (1+\lambda) R_T(a^*) - \lambda R_T(a^*) \\
&\leq (1+\lambda) \left(9 \cdot \sum_{i=1}^d \sqrt{\log T \, \mathbf{E}\left[\sum_{t=1}^T \min\left\{ x_{ti}, \frac{1 - x_{ti}}{\sqrt{\log T}} \right\} \right]} + d^2 + 9d \log T \right) \\
&\quad - \lambda \left(\frac{\Delta}{B'(\mathcal{A})} \mathbf{E}\left[\sum_{t=1}^T \left(\sum_{i \in I^*} (1 - a_{ti}) + \sum_{i \in J*} a_{ti} \right) \right] - 2Cm \right) \\
&\leq \sum_{i \in J^*} \left(9(1+\lambda)\sqrt{\log T} \sqrt{\mathbf{E}\left[\sum_{t=1}^T x_{ti} \right]} - \frac{\lambda \Delta}{B'(\mathcal{A})} \mathbf{E}\left[\sum_{t=1}^T x_{ti} \right] \right) \\
&\quad + \sum_{i \in I^*} \left(9(1+\lambda)(\log T)^{1/4} \sqrt{\mathbf{E}\left[\sum_{t=1}^T (1 - x_{ti}) \right]} - \frac{\lambda \Delta}{B'(\mathcal{A})} \mathbf{E}\left[\sum_{t=1}^T (1 - x_{ti}) \right] \right) \\
&\quad + (1+\lambda)(d^2 + 9d \log T) + 2\lambda C m \\
&\leq \sum_{i \in J^*} \frac{81(1+\lambda)^2 B'(\mathcal{A}) \log T}{4\lambda\Delta} + \sum_{i \in I^*} \frac{81(1+\lambda)^2 B'(\mathcal{A}) \sqrt{\log T}}{4\lambda\Delta} \\
&\quad + (1+\lambda)(d^2 + 9d \log T) + 2\lambda C m, \\
&= \left(|J^*| + |I^*| \frac{1}{\sqrt{\log T}} \right) \frac{81(1+\lambda)^2 B'(\mathcal{A}) \log T}{4\lambda\Delta} + (1+\lambda)(d^2 + 9d \log T) + 2\lambda C m,
\end{aligned}$$

where the first inequality follows from (27) and (29), the second inequality follows from $\mathbf{E}[a_t|x_t] = x_t$, and the third inequality follows from the inequality $a\sqrt{x} - bx = -b(\sqrt{x} - \frac{a}{2b})^2 + \frac{a^2}{4b} \leq \frac{a^2}{4b}$ that holds for any $x \geq 0$, $a \geq 0$, and $b > 0$. By setting $B(\mathcal{A}) = \frac{1}{2}\left(|J^*| + |I^*|\frac{1}{\sqrt{\log T}}\right) B'(\mathcal{A})$, we have

$$
\begin{aligned}
R_T(a^*) &\leq \frac{81(1+\lambda)^2 B(\mathcal{A})\log T}{2\lambda\Delta} + (1+\lambda)(d^2 + 9d\log T) + 2\lambda C m \\
&\leq \frac{81B(\mathcal{A})\log T}{\Delta} + d^2 + 9d\log T + \lambda\left(\frac{81B(\mathcal{A})\log T}{2\Delta} + d^2 + 9d\log T + Cm\right) + \frac{1}{\lambda}\frac{81B(\mathcal{A})\log T}{2\Delta}.
\end{aligned}
$$

By choosing $\lambda = \Theta\left(\sqrt{\left(\frac{B(\mathcal{A})\log T}{\Delta}\right) / \left(\frac{B(\mathcal{A})\log T}{\Delta} + d^2 + d\log T + Cm\right)}\right)$, we obtain $R_T(a^*) = O\left(\frac{B(\mathcal{A})\log T}{\Delta} + \sqrt{B(\mathcal{A})\frac{Cm\log T}{\Delta}} + d^2\right)$. From conditions in (30) and the definition of $B(\mathcal{A})$, we can confirm that (9) holds. $\square$

The regret bound in Remark 4 can be similarly shown, by combining (27) and (31).

A.6 Proof of Theorem 3

Combining (10) and Theorem 2, we obtain

$$
R_T \leq D \cdot \mathbf{E}\left[\sqrt{2\sum_{t=1}^{T}((\ell_t - u_t)^\top a_t)^2 + 8p' \sum_{t=1}^{T-1}\|u_t - u_{t+1}\|_q + 8p'}\right] \tag{41}
$$

for any $\{u_t\}_{t=1}^T \subseteq \mathcal{L}$. Considering a special case of $u_t = \bar{\ell}$ for all $t \in [T]$, we have

$$
\begin{aligned}
R_T &\leq D \cdot \mathbf{E}\left[\sqrt{2\sum_{t=1}^{T}((\ell_t - \bar{\ell})^\top a_t)^2 + 8p'}\right] \leq D \cdot \mathbf{E}\left[\sqrt{2\sum_{t=1}^{T}\|\ell_t - \bar{\ell}\|_q^2 \|a_t\|_p^2 + 8p'}\right] \\
&\leq D \cdot \mathbf{E}\left[\sqrt{2\sum_{t=1}^{T}\|\ell_t - \bar{\ell}\|_q^2 + 8p'}\right] = D \cdot \mathbf{E}\left[\sqrt{2Q_q + 8p'}\right].
\end{aligned}
$$

Similarly, by considering a special case of $u_t = \ell_t$ for all $t \in [T]$, we obtain

$$
R_T \leq D \cdot \mathbf{E}\left[\sqrt{8p' \sum_{t=1}^{T-1}\|\ell_t - \ell_{t+1}\|_q + 8p'}\right] = D \cdot \mathbf{E}\left[\sqrt{8p' V_q + 8p'}\right].
$$

Further, if $\ell_t^\top a_t \in [0,1]$ for all $t \in [T]$, by substituting $u_t = 0$ for all $t \in [T]$, we obtain

$$
\begin{aligned}
R_T(a^*) &\leq D \cdot \mathbf{E}\left[\sqrt{2\sum_{t=1}^{T}(\ell_t^\top a_t)^2 + 8p'}\right] \leq D \cdot \mathbf{E}\left[\sqrt{2\sum_{t=1}^{T}\ell_t^\top a_t + 8p'}\right] \\
&\leq D \cdot \sqrt{2\left(R_T(a^*) + \mathbf{E}\left[\sum_{t=1}^{T}\ell_t^\top a*\right]\right) + 8p'} = D \cdot \sqrt{2\left(R_T(a^*) + \mathbf{E}\left[L^*\right]\right) + 8p'},
\end{aligned}
$$

where the second inequality follows from $\ell_t^\top a_t \in [0,1]$ and the second inequality follows the definition of $R_T(a^*)$ and Jensen's inequality. This implies $(R_T(a^*))^2 \leq D^2(2\left(R_T(a^*) + \mathbf{E}\left[L^*\right]\right) + 8p')$. By solving this quadratic inequation in $R_T(a^*)$, we obtain $R_T(a^*) = O(D \cdot (\sqrt{\mathbf{E}[L^*]} + p'))$. $\square$

B Regret Bounds for an Existing Method in Corrupted Stochastic Settings

In this section, we see that the algorithm by Zimmert et al. [71] achieves a regret bound of

$$
R_T(a^*) = O\left(\frac{m}{\Delta}(d\log T + m) + \sqrt{\frac{Cm^2}{\Delta}(d\log T + m)}\right) \tag{42}
$$

in stochastic regimes with adversarial corruptions.

B.1 Natation and known results

From (4) in the paper [71], the regret for their algorithm is bounded as

$$R_T(a^*) \leq \sum_{t=1}^{T} \frac{25}{\sqrt{t}} \left(f(\mathbf{E}[a_t]) + g(\mathbf{E}[a_t]) \right) + c, \tag{43}$$

where f, g and c are defined by

$$f(x) = \sum_{i \in J^*} \sqrt{x_i}, \quad g(x) = \sum_{i \in I^*} (\gamma^{-1} - \gamma \log(1 - x_i))(1 - x_i), \quad c = \frac{58m}{\gamma^2} \tag{44}$$

and $\gamma \in (0, 1]$ is an input parameter. Note that I^* and J^* are defined in the same way as in Lemma 4. Define Δ_x for $x \in \mathcal{A}$, $r(\cdot)$, and $P(\cdot)$ in the same way as in the paper by Zimmert et al. [71]. Similarly, we further define C_{sto} and $C_{add}(u)$ by

$$C_{sto} := \max_{\alpha \in [0,\infty]^{\mathcal{A}}} \left(f(\bar{\alpha}) - r(\alpha) \right), \quad C_{add}(u) := \sum_{t=1}^{\infty} \max_{\alpha \in \Delta(\mathcal{A})} \left(\frac{u}{\sqrt{t}} f(\bar{\alpha}) - r(\alpha) \right) \tag{45}$$

for any $u > 0$, where $\bar{\alpha} \in \mathbb{R}^d$ is defined by $\bar{\alpha} = \sum_{x \in \mathcal{A}} \alpha_x x$. As can be seen from Section A.3 of [71], C_{sto} and C_{add} are bounded as

$$C_{sto} = O\left(\frac{md}{\Delta} \right), \quad C_{add}(u) = O\left(\frac{m^2 u^2}{\gamma^2 \Delta} \right) \tag{46}$$

in general. We note that Δ in this paper corresponds to $\Delta_{\min}$ in [71] and that Zimmert et al. [71] have provided even better bounds for C_{sto} and $C_{add}(u)$ in some special cases including problems with full combinatorial set or m-set. We only consider the case in which C_{sto} and $C_{add}(u)$ are bounded as in (46) for the sake of simplicity.

B.2 Regret analysis in stochastic regime with adversarial corruptions

In stochastic regimes with adversarial corruptions, the regret is bounded from below as follows:

$$R_T(a^*) \geq \sum_{t=1}^{T} r(P(\mathbf{E}[a_t])) - 2Cm, \tag{47}$$

which can be shown in a similar way to Lemma 4. Combining (43) with (47), for any $\lambda > 0$, we obtain

$$\begin{aligned}
&R_T(a^*) = (1+\lambda) R_T(a^*) - \lambda R_T(a^*) \\
&\leq (1+\lambda) \left(\sum_{t=1}^{T} \frac{25}{\sqrt{t}} \left(f(\mathbf{E}[a_t]) + g(\mathbf{E}[a_t]) \right) + c \right) - \lambda \left(\sum_{t=1}^{T} r(P(\mathbf{E}[a_t])) - 2Cm \right) \\
&= \sum_{t=1}^{T} \left(\frac{25(1+\lambda)}{\sqrt{t}} f(\mathbf{E}[a_t]) - \frac{\lambda}{2} r(P(\mathbf{E}[a_t])) \right) + \sum_{t=1}^{T} \left(\frac{50(1+\lambda)}{\sqrt{t}} g(\mathbf{E}[a_t]) - \frac{\lambda}{2} r(P(\mathbf{E}[a_t])) \right) \\
&\quad + (1+\lambda)c + 2\lambda Cm.
\end{aligned} \tag{48}$$

Each term in the right-hand side can be bounded via similar arguments to in the proof of Theorem 1 by Zimmert et al. [71]. In fact, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\frac{25(1+\lambda)}{\sqrt{t}} f(\mathbf{E}[a_t]) - \frac{\lambda}{2} r(P(\mathbf{E}[a_t])) \right) \leq \sum_{t=1}^{T} \max_{\alpha \in \Delta(\mathcal{A})} \left(\frac{25(1+\lambda)}{\sqrt{t}} f(\bar{\alpha}) - \frac{\lambda}{2} r(\alpha) \right) \\
&\leq \sum_{t=1}^{T} \max_{\alpha \in [0,\infty]^{\mathcal{A}}} \left(\frac{25(1+\lambda)}{\sqrt{t}} f\left(\frac{50^2(1+\lambda)^2}{\lambda^2 t} \bar{\alpha} \right) - \frac{\lambda}{2} r\left(\frac{50^2(1+\lambda)^2}{\lambda^2 t} \alpha \right) \right) \\
&= \sum_{t=1}^{T} \frac{50^2(1+\lambda^2)}{2\lambda t} \max_{\alpha \in [0,\infty]^{\mathcal{A}}} \left(f(\bar{\alpha}) - r(\alpha) \right) = O\left(\frac{(1+\lambda)^2}{\lambda} C_{sto} \log T \right) \\
&= O\left(\left(\lambda + \frac{1}{\lambda} \right) C_{sto} \log T \right),
\end{aligned} \tag{49}$$

where the first equality follows from the fact that r is linear and that $f(ux) = \sqrt{u} f(x)$ holds for any scalar $u \geq 0$, and the second equality follows from (46). Similarly, we can see that

$$\sum_{t=1}^{T} \left(\frac{50(1+\lambda)}{\sqrt{t}} g(\mathbf{E}[a_t]) - \frac{\lambda}{2} r(P(\mathbf{E}[a_t])) \right) = \frac{\lambda}{2} \sum_{t=1}^{T} \left(\frac{100(1+\lambda)}{\lambda\sqrt{t}} g(\mathbf{E}[a_t]) - r(P(\mathbf{E}[a_t])) \right)$$
$$\leq \frac{\lambda}{2} C_{add} \left(\frac{100(1+\lambda)}{\lambda\sqrt{t}} \right) = O\left(\frac{\lambda}{2} \left(\frac{1+\lambda}{\lambda^2} \right)^2 \frac{m^2}{\gamma^2 \Delta} \right) = O\left(\left(\lambda + \frac{1}{\lambda} \right) \frac{m^2}{\gamma^2 \Delta} \right), \tag{50}$$

where the inequality follows for the definition of $C_{add}(u)$ in (45) and the second inequality follows from (46). Combining (48), (49) and (50), we obtain

$$R_T(a^*) = O\left(\lambda \left(C_{sto} \log T + \frac{m^2}{\gamma^2 \Delta} + Cm + c \right) + \frac{1}{\lambda} \left(C_{sto} \log T + \frac{m^2}{\gamma^2 \Delta} \right) + c \right).$$

By choosing $\lambda = \sqrt{\frac{C_{sto} \log T + m^2/\gamma^2\Delta}{C_{sto} \log T + m^2/\gamma^2\Delta + Cm + c}}$, we obtain

$$\begin{aligned}
R_T(a^*) &= O\left(\sqrt{\left(C_{sto} \log T + \frac{m^2}{\gamma^2 \Delta} + Cm + c \right) \left(C_{sto} \log T + \frac{m^2}{\gamma^2 \Delta} \right)} + c \right) \\
&= O\left(C_{sto} \log T + \frac{m^2}{\gamma^2 \Delta} + c + \sqrt{(Cm + c) \left(C_{sto} \log T + \frac{m^2}{\gamma^2 \Delta} \right)} \right) \\
&= O\left(\frac{dm}{\Delta} \log T + \frac{m^2}{\gamma^2 \Delta} + c + \sqrt{(Cm + c) \left(\frac{dm}{\Delta} \log T + \frac{m^2}{\gamma^2 \Delta} \right)} \right),
\end{aligned}$$

where the last equality follows from (46). If we set $\gamma = 1$, we have $c = O(m)$ and hence the regret is bounded as

$$R_T = O\left(\frac{m}{\Delta} (d \log T + m) + \sqrt{\frac{Cm^2}{\Delta} (d \log T + m)} \right),$$

which means that (42) holds.