Beating Stochastic and Adversarial Semi-bandits
Optimally and Simultaneously

Julian Zimmert ¹ Haipeng Luo ² Chen-Yu Wei ²

Abstract
We develop the first general semi-bandit algorithm that simultaneously achieves $O(\log T)$ regret for stochastic environments and $O(\sqrt{T})$ regret for adversarial environments without knowledge of the regime or the number of rounds $T$. The leading problem-dependent constants of our bounds are not only optimal in some worst-case sense studied previously, but also optimal for two concrete instances of semi-bandit problems. Our algorithm and analysis extend the recent work of (Zimmert & Seldin, 2019) for the special case of multi-armed bandit, but importantly requires a novel hybrid regularizer designed specifically for semi-bandit. Experimental results on synthetic data show that our algorithm indeed performs well uniformly over different environments. We finally provide a preliminary extension of our results to the full bandit feedback.

1. Introduction
Multi-armed bandit is one of the most fundamental online learning problems with partial information feedback. In this problem a learner repeatedly selects one of the $d$ arms and observes its loss generated by the environment, with the goal of minimizing her regret, the difference between her total loss and the loss of the best fixed arm in hindsight. It is well known that in the stochastic environment (Lai & Robbins, 1985) where each arm's loss is drawn independently from a fixed distribution, the minimax optimal regret is of order $O(\log T)$ where $T$ is the number of rounds (dependence on all other parameters is omitted), while in the adversarial environment (Auer et al., 2002) where each arm’s loss can be completely arbitrary, the minimax optimal regret is of order $O(\sqrt{T})$.

Several recent works (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Wei & Luo, 2018; Zimmert & Seldin, 2019) develop “best-of-both-worlds” results for multi-armed bandit and propose adaptive algorithms that achieve $O(\log T)$ regret for stochastic environments while simultaneously ensuring worst-case robustness, that is, $O(\sqrt{T})$ regret even for adversarial environments. Importantly, this is achieved without any prior knowledge of the environment.

In this work, we extend such best-of-both-worlds results to the combinatorial bandit problem, a generalization of multi-armed bandit where the learner has to pick a subset of arms (called a combinatorial action) at each time (see Section 2 for formal definitions). In particular we consider the semi-bandit feedback, meaning that the learner observes the loss of each arm in the selected subset. Our main contributions include the following:

1. We propose a simple and general semi-bandit algorithm based on the Follow-the-Regularized-Leader (FTRL) framework with a novel regularizer. (Section 2.1)

2. For any combinatorial action set, we prove that our algorithm achieves $O(C_{sto} \log T)$ regret for stochastic environments and $O(C_{adv} \sqrt{T})$ regret for adversarial environments, where $C_{sto}$ and $C_{adv}$ are problem-dependent constants (that do not depend on $T$) and are optimal in some worst-case sense. This is the first best-of-both-worlds result for combinatorial bandit to the best of our knowledge. (Section 3.1)

3. For two common special cases of combinatorial action set: the set of all subsets of arms and the set of all subsets with a fixed size $m$ (so called $m$-set), we further show finer bounds for the problem-dependent constants $C_{sto}$ and $C_{adv}$, which are again optimal for each of these special cases. As a side result, our bounds imply that for $m$-set with $m > d/2$, semi-bandit feedback is no harder than full-information feedback in the adversarial case. (Sections 3.2 and 3.3)

4. We conduct experiments with synthetic data to show that our algorithm indeed adapts well to the nature of

¹Department of Computer Science, University of Copenhagen, Copenhagen, Denmark ²Department of Computer Science, University of Southern California, United States. Correspondence to: Julian Zimmert <zimmert@di.ku.dk>, Haipeng Luo <haipengl@usc.edu>, Chen-Yu Wei <chenyu.wei@usc.edu>.
the environment. Additionally, we present a simple intermediate setting where our algorithm outperforms all baselines. (Section 4)

5. We also provide a preliminary extension of our results to a special case with the more challenging full bandit feedback. (Section 6)

Our techniques are close to that of (Zimmert & Seldin, 2019): we make use of the FTRL algorithm, a well-known framework for adversarial environments, and show that with a simple time-decaying learning rate schedule (that is, \(1/\sqrt{t}\) for time \(t\)), the regret admits some self-bounding property under the stochastic environment which eventually leads to the logarithmic regret in this case. Importantly, however, our results require the use of a novel hybrid regularizer, designed specifically for semi-bandit. Roughly speaking, the idea is that for arms outside the optimal subset, the problem of selecting this arm or not behaves like a two-armed bandit problem, and we apply the regularizer of (Zimmert & Seldin, 2019) to these arms; and on the other hand for arms in the optimal subset, the problem behaves like the full-information expert problem (Freund & Schapire, 1997), and we thus apply the classic Shannon entropy as the regularizer to these arms.

1.1. Related work

Semi-bandit. The combinatorial semi-bandit problem is a natural generalization of multi-armed bandit and captures many real-life applications. There are many algorithms based on the well-known optimistic principle for stochastic semi-bandit (Gai et al., 2012; Chen et al., 2013; Kveton et al., 2015; Combes et al., 2015). Optimistic algorithms are provably not instance-optimal (Lattimore & Szepesvari, 2017) and recent work (Combes et al., 2017) developed a general instance-optimal algorithm for any structured stochastic bandits (including semi-bandit as a special case). Specifically, they obtain the optimal regret \(O(C \log T)\) where \(C\) is an instance-dependent term expressed as the solution of some optimization problem. The constant \(C_{sto}\) in our stochastic bound \(O(C_{sto} \log T)\) is also expressed as an optimization problem (see Theorem 1), but it is not clear how it compares to the instance-optimal constant \(C\) in general, except for the two special cases we show in Section 3. Two advantages of our algorithm compared to all these prior works are: a) our stochastic assumption is weaker than others (see Section 2) and b) our algorithm ensures worst-case robustness even when the stochastic assumption does not hold.

On the other hand, algorithms with \(O(\sqrt{T})\) regret for the adversarial semi-bandit setting are also well-studied (Audibert et al., 2013; Neu & Bartók, 2013; Combes et al., 2015; Neu, 2015; Wei & Luo, 2018). These algorithms are either based on Follow-the-Regularized-Leader (equivalently Online Mirror Descent) or Follow-the-Perturbed-Leader, both of which are standard frameworks for designing adversarial online learning algorithms (see (Hazan et al., 2016) for an introduction). Unlike our algorithm, it is easy to show that even if the environment is stochastic, the regret of these algorithms is still \(\Theta(\sqrt{T})\), indicating the lack of adaptivity. Moreover, even for the adversarial case the leading constant in previous bounds is only worst-case optimal but not instance-optimal. In contrast, our adversarial regret bound \(O(C_{adv}\sqrt{T})\) is instance-dependent through the term \(C_{adv}\), again expressed as the solution of some optimization problem (see Theorem 1). To the best of our knowledge, there is no known general instance-dependent lower bound for this term, but again we show the optimality of our bound for two special cases in Section 3.

Best-of-both-worlds. Algorithms that are optimal for both stochastic and adversarial environments were studied for multi-armed bandit (Bubeck & Slivkins, 2012; Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Wei & Luo, 2018; Zimmert & Seldin, 2019), and also for the easier full-information version (the expert problem) (Gaillard et al., 2014; Luo & Schapire, 2015; Koolen et al., 2016). Notably, among these works the recent two (Wei & Luo, 2018; Zimmert & Seldin, 2019) discovered that sophisticated testings or gap estimations used in earlier works are in fact not needed for such adaptivity. Instead, their algorithms are simply based on the FTRL framework with special regularizers. As mentioned, our work also follows this route by designing a new regularizer for the more general semi-bandit setting.

Hybrid regularizers. The idea of using hybrid regularizers for FTRL was first proposed in (Bubeck et al., 2018) for sparse bandit and bandit with some specific form of adaptive regret bound, and also recently used in (Luo et al., 2018) for the online portfolio problem. The form of the hybrid regularizers and the way they are used in the analysis, however, are all different among these two prior works and ours.

2. Problem Setting and Algorithm

The semi-bandit problem is a sequential game between a learner and an environment with \(d\) fixed arms. We call a subset of arms a combinatorial action,\(^1\) and the learner is given a fixed set of combinatorial actions \(X \subset \{0, 1\}^d\). At any time \(t = 1, 2, \ldots\), the learner chooses an action \(X_t \in X\) and at the same time the environment chooses a loss vector \(\ell_t \in [-1, 1]^d\). The learner suffers the loss \(\langle X_t, \ell_t \rangle\) and receives the feedback \(o_t = X_t \circ \ell_t\), where \(\circ\) stands for the element-wise multiplication. In other words,

\(^1\)In some works a combinatorial action is also referred as “an arm”, but here we exclusively use the term “arm” for one of the \(d\) elements and “combinatorial action” for a subset of these elements.
the learner observes and only observes the loss of each arm in the selected subset (so-called semi-bandit feedback).

The environment can be either stochastic or adversarial. In the stochastic case, we adopt and extend the broader “stochastically constrained adversarial setting” (Wei & Luo, 2018; Zimmert & Seldin, 2019) and assume that there is a fixed action \( x^* \in \mathcal{X} \) such that for any \( x \in \mathcal{X} \setminus \{x^*\} \) there exists a constant \( \Delta_x > 0 \) such that \( \mathbb{E}[\langle x - x^*, \ell_t \rangle] \geq \Delta_x \) for all \( t \). Note that this clearly subsumes the traditional stochastic setting where \( \ell_1, \ldots, \ell_T \) are i.i.d. samples of some fixed unknown distribution, and is much more general since neither independence nor identical distributions are required. In the adversarial case, on the other hand, \( \ell_t \) is an arbitrary way based on the history \( \ell_1, X_1, \ldots, \ell_{t-1}, X_{t-1} \) and possibly some internal randomness of the environment.

The performance of a learner is measured by pseudo-regret:

\[
\overline{\text{Reg}}_T := \mathbb{E} \left[ \sum_{t=1}^T \langle X_t - x^*, \ell_t \rangle \right]
\]

where \( x^* = \arg\min_{x \in \mathcal{X}} \mathbb{E} \left[ \sum_{t=1}^T \langle x, \ell_t \rangle \right] \) is the best action in hindsight and the expectation is with respect to the randomness of both the learner and the environment. Note that in the stochastic case we are overloading the notation \( x^* \) since clearly they are the same action.

It is well known that in terms of the dependence on \( T \), the optimal regret is \( \Theta(\log T) \) in the stochastic case and \( \Theta(\sqrt{T}) \) in the adversarial case (see for example (Audibert et al., 2013; Combes et al., 2017)).

**Notations.** We denote by \( \mathbb{E}_x[\cdot] \) the conditional expectation \( \mathbb{E}[\cdot | \mathcal{F}_{t-1}] \) where \( \mathcal{F}_t \) is the filtration \( \sigma(X_1, o_1, \ldots, X_t, o_t) \).

We also use a shorthand \( \mathbb{I}[\cdot] \) for the indicator function \( \mathbb{I}[X_{t+1} = 1] \) and write the characteristic function of a set \( A \) as \( \mathbb{I}_A(x) \) which is 0 if \( x \in A \) and +∞ otherwise.

### 2.1. Our algorithm

Our algorithm is based on the general FTRL framework.\(^2\)

In this framework, each time the algorithm computes the regularized leader \( x_t = \arg\min_{x \in \text{Conv}(\mathcal{X})} \langle x, \hat{L}_{t-1} \rangle + \eta_{t-1} \Psi(x) \) where \( \text{Conv}(\mathcal{X}) \) is the convex hull of \( \mathcal{X} \), \( \hat{L}_{t-1} = \sum_{s=1}^{t-1} \hat{\ell}_s \) is the cumulative estimated loss, \( \eta_t > 0 \) is some learning rate, and \( \Psi(x) : \text{Conv}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\} \) is some regularizer. Then the algorithm samples \( X_t \sim P(x_t) \) for some sampling rule \( P \) that provides a distribution over \( \mathcal{X} \) with mean \( \mathbb{E}_{X \sim P(x)}[X] = x \). As long as \( \text{Conv}(\mathcal{X}) \) can be described by a polynomial number of constraints, one can always find an efficient sampling rule \( P \) (also see concrete examples in Section 3). Finally, the algorithm constructs a loss estimator \( \hat{\ell}_t \) based on the observed information and proceeds to the next round.

The novelty of our algorithm lies in the use of the hybrid regularizer

\[
\Psi(x) = \sum_{i=1}^d -\sqrt{x_i} + \gamma(1 - x_i) \log(1 - x_i)
\]

for some parameter \( \gamma \leq 1 \) chosen later based on the action set \( \mathcal{X} \) (for most cases we simply use \( \gamma = 1 \)). This is a combination of the Tsallis entropy (with power 1/2) and the Shannon entropy, shown to be optimal for both stochastic and adversarial environments for multi-armed bandit (Zimmert & Seldin, 2019), and the Shannon entropy \( \sum_i (1 - x_i) \log(1 - x_i) \) on the complement of \( x \).

In addition, similar to (Zimmert & Seldin, 2019) our algorithm uses a very simple time-decaying learning rate schedule \( \eta_t = 1/\sqrt{t} \). The loss estimators \( \hat{\ell}_t \) is defined as \( \hat{\ell}_t = \frac{(\eta_{t+1} + \eta_t)\ell_t(i)}{x_{ti}} - 1 \) for all \( i \). It is clear that this is unbiased: \( \mathbb{E}_o[\hat{\ell}_t] = \ell_t \) just as common importance weighted estimators, and the extra shift of 1 is to ensure that the range of estimated losses is bounded from one side: \( \hat{\ell}_{t,i} \geq -1 \).

See Algorithm 1 for the complete pseudocode.

**Intuition of the new regularizer.** It is known that the classic Shannon entropy regularizer (Freund & Schapire, 1997) is optimal for both adversarial and stochastic environments in the full-information setting. In fact, the Shannon entropy on the complement of \( x \) is also optimal for full-information. This can be verified by considering the complementary problem: the problem with action set \( 1_d - \mathcal{X} \) and reversed losses \( -\ell_t \). Both problems describe the exact same game with the same information, and using Shannon entropy in the complementary problem is the same as using it on the complement of \( x \) in the original problem.

---

\(^2\)For linear objectives and Legendre regularizers, FTRL is equivalent to Online Mirror Descent. The same framework is also known under the names OMD, OSMD or INF.

---

**Algorithm 1** FTRL with hybrid regularizer for semi-bandits

Parameter: \( \eta < 1 \)

Initialize: \( \hat{L}_0 = (0, \ldots, 0), \eta_t = 1/\sqrt{t} \)

for \( t = 1, 2, \ldots \) do

compute

\[
x_t = \arg\min_{x \in \text{Conv}(\mathcal{X})} \langle x, \hat{L}_{t-1} \rangle + \eta_{t-1} \Psi(x)
\]

where \( \Psi(\cdot) \) is defined in Eq. (1)

sample \( X_t \sim P(x_t) \)

observe \( o_t = X_t \circ \hat{\ell}_t \)

construct estimator \( \hat{\ell}_t : \hat{\ell}_{ti} = \frac{(\eta_{t+1} + \eta_t)\ell_t(i)}{x_{ti}} - 1, \forall i \)

update \( \hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t \)

end for
The intuition behind combining Tsallis and Shannon entropy is that when \( x_i \) is close to 0, the learner is starved of information and has to act similarly to a regular bandit problem. The magnitude of the gradient and its slope in that regime are dominated by the Tsallis entropy, which again is known to be optimal for bandits.

On the other hand, when \( x_i \) is close to 1, the game resembles a full-information game, and Shannon entropy on the complement becomes the dominating part of the regularizer in that regime. This allows us to effectively regularize arms in the optimal combinatorial set differently than arms outside the optimal set, without knowing which arms are in the optimal set.

3. Main Results

In this section we present the general regret guarantees of our algorithm, followed by more concrete results for two special cases.

3.1. Arbitrary action set

To state the general regret bound of our algorithm for any arbitrary action set \( \mathcal{X} \), we define the following two functions:

\[
 f(x) = \sum_{i:x_i^* = 0} \sqrt{x_i} \\
g(x) = \sum_{i:x_i^* = 1} (\gamma^{-1} - \gamma \log(1-x_i))(1-x_i)
\]

and the immediate regret function \( r: [0, \infty)^{|\mathcal{X}|} \to \mathbb{R} \) as

\[
r(\alpha) = \sum_{x \in \mathcal{X} \setminus \{x^*\}} \alpha_x \Delta_x
\]

(recall the definition of \( x^* \) and \( \Delta_x \) from Section 2). We also define \( \overline{\alpha} = \sum_{x \in \mathcal{X}} \alpha_x x \) for any \( \alpha \in [0, \infty)^{|\mathcal{X}|} \), and let \( \Delta(\mathcal{X}) \) denote the simplex of distributions over \( \mathcal{X} \).

**Theorem 1.** For any \( \gamma \leq 1 \) the pseudo regret of Algorithm 1 is upper bounded by

\[
\operatorname{Reg}_T \leq \mathcal{O}(C_{sto} \log T) + \mathcal{O}(C_{adv})
\]

in the stochastic case and

\[
\operatorname{Reg}_T \leq \mathcal{O}(C_{adv} \sqrt{T})
\]

in the adversarial case, where \( C_{sto}, C_{add} \) and \( C_{adv} \) are defined as

\[
C_{sto} := \max_{\alpha \in [0, \infty)^{|\mathcal{X}|}} f(\overline{\alpha}) - r(\alpha), \\
C_{adv} := \sum_{t=1}^{\infty} \max_{\alpha \in \Delta(\mathcal{X}) \setminus \{x^*\}} \left( \frac{100}{\sqrt{t}} g(\overline{\alpha}) - r(\alpha) \right), \\
C_{adv} := \max_{x \in \text{Conv}(\mathcal{X})} f(x) + g(x).
\]

Moreover, it always holds that \( C_{sto} = \mathcal{O}\left(\frac{m^2}{\Delta_{\min}}\right) \), \( C_{add} = \mathcal{O}\left(\frac{m^2}{\gamma^2 \Delta_{\min}}\right) \), and \( C_{adv} = \mathcal{O}\left(\frac{1}{\gamma} \sqrt{md}\right) \), where \( m = \max_{x \in \mathcal{X}} \|x\|_1 \) and \( \Delta_{\min} = \min_{x \in \mathcal{X} \setminus \{x^*\}} \Delta_x \).

We defer the proof to Section 5. The dependence of our bounds on \( T \) is optimal in both cases. The leading problem-dependent constants \( C_{sto} \) and \( C_{adv} \) are expressed as solutions to some optimization problems. Recent works (Combes et al., 2015; Lattimore & Szepesvari, 2017; Combes et al., 2017) also expressed the instance-optimal leading constant in the stochastic case in a similar way, but it is not clear how to compare them in general.

The concrete upper bounds on these constants stated at the end of the theorem immediately imply that for \( \gamma = 1 \), our bounds are worst-case optimal according to (Kveton et al., 2015) and (Audibert et al., 2013). Here, worst-case optimality refers to minimax regret among all environments with the same value \( m \) for \( \max_{x \in \mathcal{X}} \|x\|_1 \) and also the same value \( \Delta_{\min} \) for \( \min_{x \in \mathcal{X} \setminus \{x^*\}} \Delta_x \) in the stochastic case.

However, for concrete instances, one can hope to achieve even better bounds. In the next two subsections we show that our algorithm is also optimal in two special cases, by exploiting the structure of the problem and providing better bounds on the constants \( C_{sto}, C_{add} \) and \( C_{adv} \). For better interpretability, in the stochastic case we consider the more traditional setting where \( \ell_1, \ldots, \ell_T \) are i.i.d. samples of some unknown distribution \( D \). It is clear that we can define \( \Delta_x = \mathbb{E}_{\ell \sim D}[\langle x - x^*, \ell \rangle] \) in this case.

3.2. Special case: full combinatorial set

The simplest semi-bandit problem is when \( \mathcal{X} = \{0,1\}^d \), that is, the learner can pick any subset of arms. In this case \( \text{Conv}(\mathcal{X}) = \{0,1\}^d \) and a trivial sampling rule is \( P(x) = \prod_{i=1}^{d} \text{Ber}(x_i) \) where \( \text{Ber}(\cdot) \) stands for Bernoulli distribution.

It is clear that in this case each dimension/arm can be treated completely independently. Note, however, that the problem of each dimension is not exactly a two-armed bandit problem since the loss of “not choosing the arm” is known to be 0, and the problem is asymmetric between positive and negative losses. Specifically, we prove the following regret guarantee of our algorithm, where in the stochastic case with a slight abuse of notation we define \( \Delta_i = \mathbb{E}_{\ell \sim D}[\ell_i] \).

**Theorem 2.** If \( \mathcal{X} = \{0,1\}^d \), the pseudo-regret of Algorithm 1 with \( \gamma = 1 \) is

\[
\operatorname{Reg}_T \leq \mathcal{O}\left(\sum_{\Delta_i > 0} \frac{\log(T)}{\Delta_i} \right) + \mathcal{O}\left(\sum_{\Delta_i < 0} \frac{1}{|\Delta_i|} \right)
\]
in the stochastic case and
\[ \text{Reg}_T \leq O \left( d\sqrt{T} \right) \]
in the adversarial case. Moreover, both bounds are optimal.

Proof. Note that in this case the algorithm is equivalent to the following: for each coordinate, run a copy of Algorithm 1 for a one-dimensional problem with \( \mathcal{X} = \{0, 1\} \) as the action set. We can thus apply Theorem 1 to such one-dimensional problems and finally sum up the regrets of each coordinate. Below we focus on a fixed coordinate \( i \).

In particular, in the stochastic case, if \( \Delta_i > 0 \), it implies \( x^*_i = 0 \) and thus \( g(\cdot) \equiv 0 \) and \( C_{add} = \sum_i \max_{\alpha \in [0,1]} -\alpha \Delta_i = 0 \). For \( C_{sto} \) we simply apply the general bound from Theorem 1 and obtain \( C_{sto} = O \left( 1/\Delta_i \right) \) (since \( m = d = 1 \) and \( \Delta_{\min} = \Delta_i \)). This gives the bound \( O \left( \frac{\log(T)}{\Delta_i} \right) \) for \( \Delta_i > 0 \).

On the other hand if \( \Delta_i < 0 \) then \( x^*_i \neq 1 \) and \( f(\cdot) \equiv 0 \), so \( C_{sto} = \max_{\alpha \geq 0} \alpha \Delta_i = 0 \). For \( C_{add} \) we simply apply the general bound from Theorem 1 and obtain \( C_{add} = O \left( 1/\Delta_i \right) \) (since \( m = \gamma = 1 \) and \( \Delta_{\min} = \Delta_i \)). This gives the bound \( O \left( \frac{\log(T)}{\Delta_i} \right) \) for \( \Delta_i < 0 \).

In the adversarial case, we simply apply the general bound of Theorem 1 and obtain \( C_{adv} = O(1) \). This finishes the proof for the regret upper bounds. The optimality of the adversarial bound is trivial since it matches the full-information lower bound. Obtaining a matching lower bound in the stochastic regime is a simple adaptation of the regular two-armed bandit lower bound. We believe this result is well known, but provide a proof in the appendix in absence of a reference.

\[ \square \]

3.3. Special case: \( m \)-set

Another common instance of semi-bandit is when the learner can only select subsets with a fixed size. Specifically, let \( m \in \{1, \ldots, d - 1\} \) be some fixed parameter and define the \( m \)-set as
\[ \mathcal{X} = \left\{ x \in \{0, 1\}^d \mid \sum_{i=1}^d x_i = m \right\}. \tag{2} \]

Note that we are overloading the notation \( m = \max_{x \in \mathcal{X}} |x|_1 \) since clearly they are the same in this case. It is well-known that the convex hull of \( m \)-set is \( \text{Conv}(\mathcal{X}) = \left\{ x \in \{0, 1\}^d \mid \sum_{i=1}^d x_i = m \right\} \), and in the appendix we provide a simple sampling rule \( P \) with \( O(d \log(d)) \) time complexity.

In the stochastic case, we assume without loss of generality that the expected losses of arms are increasing in \( i \). Overloading the notation again we define the stochastic gaps as
\[ \Delta_i = \mathbb{E}_{t \sim \mathcal{D}} [\ell_i - \ell_m] \text{ for all } i. \]
Note that the uniqueness of \( x^*_i \) also implies \( \Delta_i \neq 0 \) for all \( i > m \). The next theorem shows that our algorithm is again optimal for both environments. As a side result, we also show that when \( m > d/2 \), semi-bandit feedback is no harder than full-information feedback in the adversarial case, which is unknown previously to the best of our knowledge.

**Theorem 3.** If \( \mathcal{X} \) is the \( m \)-set defined by Eq. (2) for some \( m \), then the pseudo-regret of Algorithm 1 with
\[ \gamma = \begin{cases} 1 & \text{if } m \leq d/2 \\ \min\{1, 1/\sqrt{\log(d/(d-m))}\} & \text{else} \end{cases} \]
is bounded as
\[ \text{Reg}_T \leq O \left( \sum_{i=m+1}^d \frac{\log(T)}{\Delta_i} \right) + O \left( \sum_{i=m+1}^d \frac{(\log d)^2}{\Delta_i} \right) \]
in the stochastic case and
\[ \text{Reg}_T \leq \begin{cases} O \left( \sqrt{mdT} \right) & \text{if } m \leq d/2 \\ O \left( (d - m) \sqrt{\log(d/(d-m)/T)} \right) & \text{else} \end{cases} \]
in the adversarial case. Moreover both bounds are optimal.

**Proof sketch.** We provide a proof sketch here and defer some details to Appendix B.

\( C_{adv} \): The optimization problem is concave in \( x \) and symmetric for all \( i \) with the same value of \( x^*_i \). Therefore the optimal solution takes the form
\[ \arg\max_{x \in \text{Conv}(\mathcal{X})} f(x) + g(x) = \begin{cases} \lambda & \text{if } x^*_i = 0 \\ 1 - d - m \lambda / m & \text{if } x^*_i = 1 \end{cases} \]
for some \( \lambda \in \left[0, \min\{1, \frac{m}{d-m}\}\right] \). In Appendix B we show that the function is increasing in \( \lambda \), and that inserting \( \lambda = \min\{1, \frac{m}{d-m}\} \) leads to the stated adversarial bound.

\( C_{sto} \): With the definitions of the gaps, we can express \( \Delta_x = \sum_{i: x_i \neq x^*_i} |\Delta_i| \), which is lower bounded by \( \sum_{i: x_i = 0, x^*_i = 1} \Delta_i = \sum_{i: x_i^* = 1} \Delta_i x_i \). So the immediate regret function \( r(\alpha) \) can be bounded as
\[ r(\alpha) = \sum_{x \neq x^*} \Delta_x \alpha_x \geq \sum_{x \neq x^*} \sum_{i: x_i = 0} \Delta_i \alpha_x x_i = \sum_{i: x_i = 0} \Delta_i \sum_{x \neq x^*} \alpha_x x_i \]
\[ = \sum_{i: x_i = 0} \Delta_i \left( \sum_{x \neq x^*} \alpha_x x_i \right) = \sum_{i: x_i = 0} \Delta_i \bar{\alpha}_i. \]

The optimization problem can now be bounded as
\[ C_{sto} = \max_{\alpha \in [0,\infty)^d} \sum_{i: x_i = 0} \sqrt{\alpha_i} - \sum_{x \neq x^*} \alpha_x \Delta_x \]
\[ \leq \max_{\pi \in [0,\infty)^d} \sum_{i: x_i = 0} \left( \sqrt{\pi_i} - \sqrt{\bar{\alpha}_i} \right) = \sum_{i: x_i = 0} \frac{1}{4} \Delta_i. \]
We compare our novel algorithm with four baselines we choose \(E\) which is the same as previously. We can thus bound
\[
C_{add} = \sum_{i:x_i = 1} (\gamma^{-1} - \gamma \log (1 - \pi_i))(1 - \pi_i)
\]
where the first inequality is by the concavity of \(g\); the second equality is by the fact \(\sum_{i:x_i = 1} 1 - \pi_i = \sum_{i:x_i = 0} \pi_i\) since \(\pi\) is in the convex hull of \(m\)-set.

Recall the lower bound \(\sum_{i:x_i = 0} \Delta \pi_i\) of \(r(\alpha)\) as derived previously. We can thus bound \(C_{add}\) as
\[
\sum_{i:x_i = 0} \sum_{\alpha = 1} ^{\infty} \frac{100}{\sqrt{t}} \left(\gamma^{-1} - \gamma \log \left(\frac{A}{m}\right)\right) A - \Delta_i A
\]

Solving the one-dimensional optimization problems above independently for each \(i\) (see Appendix B) proves \(C_{add} \leq \mathcal{O}\left(\sum_{i:x_i = 0} \left(\frac{\log d^2}{\Delta_i}\right)\right)\).

**Optimality:** The optimality for the stochastic case is implied by (Anantharam et al., 1987; Combes et al., 2017). For the adversarial case, only a matching lower bound \(\Omega(\sqrt{mdT})\) for \(m \leq d/2\) is known (Theorem 2 of (Lattimore et al., 2018)). We close this gap by making a simple observation that when \(m > d/2\), our bound in fact matches the lower bound of the same problem with full-information feedback. This clearly implies the optimality of our bound since semi-bandit feedback is harder.

Indeed, Koolen et al. (2010) prove the lower bound \(\Omega(m^{1/2}T \log (d/m))\) for full-information \(m\)-set when \(m \leq d/2\). When \(m > d/2\), one can simply work on the complementary problem with action set \(1 \cdots X\) and reversed losses. This is exactly a \((d - m)\)-set problem and thus a lower bound \(\Omega((d - m)^{1/2}T \log (d/(d - m)))\) applies. This exactly matches our upper bound.

4. **Empirical Comparisons**

We compare our novel algorithm with four baselines from the literature. For stochastic algorithms, we choose COMBUCB (Kveton et al., 2015) and THOMPSON SAMPLING (Gopalan et al., 2014); for adversarial algorithms, we choose EXP2 (Audibert et al., 2013) and LOGBARRIER (Wei & Luo, 2018), which are respectively FTRL with generalized Shannon entropy and log-barrier regularizer. For each adversarial algorithm, we tune the time-independent part of the learning rate by choosing from the grid of \(\{2^i | i \in \{-5, -4, \ldots, 5\}\}\), and the optimal value happens to be identical for both adversarial and stochastic environment in our experiments. Specifically the final learning rates \(\eta\) for our algorithm, EXP2 and LOGBARRIER are respectively \(1/\sqrt{T}, 1/(4\sqrt{T})\) and \(4\sqrt{\log(t)/T}\).

We test the algorithms on concrete instances of the \(m\)-set problem with parameters: \(d = 10, m = 5, T = 10^7\). We specify the mean of each arm’s loss at each time below and with mean \(\mu_i\) the actual loss of arm \(i\) at time \(t\) will be \(-1\) with probability \((1 - \mu_i)/2\) and \(+1\) with probability \((1 + \mu_i)/2\), independent of everything else. We create the following two environments:

**Stochastic environment.** In this case the losses are drawn from a fixed distribution with \(\mu_i = -\Delta\) if \(i \leq 5\) and \(\mu_i = \Delta\) otherwise, where \(\Delta = 1/8\).

**“Adversarial” environment.** Since it is difficult to create truly adversarial data, here we in fact use a stochastically constrained adversarial setting defined in Section 2. The construction is similar to that of (Zimmert & Seldin, 2019). Specifically, the environment is decomposed into phases
\[
1, \ldots, t_1, t_1 + 1, \ldots, t_2, \ldots, t_{n-1}, \ldots, T
\]

The length of phase \(s\) is about \(T_s = 1.6^s\), and the means of losses are set to
\[
\mu_i = \begin{cases} 
-\Delta/2 \pm (1 - \Delta/2) & \text{if } i \leq 5 \\
+\Delta/2 \pm (1 - \Delta/2) & \text{else}
\end{cases}
\]

where \(\pm\) represents \(+\) if \(i\) belongs to an odd phase and \(-\) otherwise. This model is not only a nice toy example, but could also be justified by real world applications. For example, in a network routing problem, an adversary might periodically attack the network, making the delay of every edge increase by roughly the same amount.

We measure the performance of the algorithms by the average pseudo-regret over at least 20 runs. For COMBUCB and THOMPSON SAMPLING in the adversarial environment, we increase the number of runs to 500 and 1000 respectively due to the high variance of the pseudo-regret. Figure 1 shows the average pseudo-regret of all algorithms at each time, where plot a) uses the stochastic data and plot b) uses the adversarial data. We use log-log scale after \(10^4\) rounds. Shaded areas in the plot show the confidence intervals.

The plots clearly confirm our theoretical results. Our algorithm outperforms EXP2 and LOGBARRIER (in the later stage) in both environments. In the stochastic case our algorithm is competitive with COMBUCB, while THOMPSON
We then further bound these two terms respectively in the following two lemmas using mostly standard FTRL analysis (see Appendix A for the proofs).

**Lemma 1.** The regularization penalty is bounded as

\[
\text{Reg}_{\text{pen}} \leq \sum_{t=1}^{T} \frac{3}{2\sqrt{t}} \left( \sum_{i:x_i^* = 0} \sqrt{\mathbb{E}[x_{ti}]} - \sum_{i:x_i^* = 1} \gamma (1 - \mathbb{E}[x_{ti}]) \log (1 - \mathbb{E}[x_{ti}]) \right).
\]

**Lemma 2.** The stability term is bounded as

\[
\text{Reg}_{\text{stab}} \leq \sum_{t=1}^{T} \frac{16\sqrt{2}}{\sqrt{t}} \left( \sum_{i:x^*_i = 0} \sqrt{\mathbb{E}[x_{ti}]} + \sum_{i:x^*_i = 1} \gamma^{-1} (1 - \mathbb{E}[x_{ti}]) \right) + c.
\]

where \( c = 58m/\gamma^2 \) (recall \( m = \max_{x \in X} ||x||_1 \)).

We now proceed to the proof of Theorem 1.

**Proof of Theorem 1.** Using Lemma 1 and Lemma 2 in Eq. (3) and the definition of functions \( f \) and \( g \), we can bound the regret by

\[
\overline{\text{Reg}}_T \leq \sum_{t=1}^{T} \frac{25}{\sqrt{t}} (f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])) + c \quad (4)
\]

\[
\leq 50\sqrt{T} \max_{x \in \text{Conv}(X)} (f(x) + g(x)) + c
\]

\[
= \mathcal{O} \left( C_{adv} \sqrt{T} \right),
\]

which concludes the adversarial case.

For the stochastic case we use a self-bounding technique similar to (Wei & Luo, 2018; Zimmert & Seldin, 2019). First by the definition of the function \( r \) and the stochastic assumption we have

\[
\overline{\text{Reg}}_T = \mathbb{E} \left[ \sum_{t=1}^{T} \langle \mathbb{E}[x_t] - x^*, \ell_t \rangle \right] \geq \sum_{t=1}^{T} r(P(\mathbb{E}[x_t])).
\]

Together with Eq. (4) we have

\[
\sum_{t=1}^{T} \frac{25}{\sqrt{t}} (f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])) + c - \sum_{t=1}^{T} r(P(\mathbb{E}[x_t])) \geq 0.
\]

Combining the above with Eq. (4) again we bound \( \overline{\text{Reg}}_T \) by

\[
\sum_{t=1}^{T} \left( 50\sqrt{2} (f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t]) - r(P(\mathbb{E}[x_t]))) + 2c. \right)
\]
We next decompose the summation above into two terms and upper bound them as $C_{sto} \log T$ and $C_{add}$ respectively:

$$
\sum_{t=1}^{T} \frac{50}{\sqrt{t}} f(\mathbb{E}[x_t]) - \frac{1}{2} r(P(\mathbb{E}[x_t]))
\leq \sum_{t=1}^{T} \max_{\alpha \in \Delta(X)} \frac{50}{\sqrt{t}} f(\pi) - \frac{1}{2} r(\alpha)
\leq \sum_{t=1}^{T} \max_{\alpha \in [0, \infty)^{|\mathcal{X}|}} \frac{100}{\sqrt{t}} g(\pi) - \frac{1}{2} r(\alpha) = O(C_{add}),
$$

where $(\ast)$ follows since $r$ is linear and $f$ satisfies for any scalar $a \geq 0$: $f(ax) = \sqrt{af(x)}$. On the other hand,

$$
\sum_{t=1}^{T} \frac{50}{\sqrt{t}} g(\mathbb{E}[x_t]) - \frac{1}{2} r(P(\mathbb{E}[x_t]))
\leq \sum_{t=1}^{\infty} \max_{\alpha \in \Delta(X)} \left( \frac{100}{\sqrt{t}} g(\pi) - r(\alpha) \right) = O(C_{add}),
$$

where the last inequality uses the fact: for all $t > 0$, $\max_{\alpha \in \Delta(X)} \left( \frac{100}{\sqrt{t}} g(\pi) - r(\alpha) \right) \geq 0$. This is because a particular $\alpha$ that puts all the weights to $x^*$ attains the value of 0.

The above finishes the proof for the general regret bounds. Due to space limit we defer the concrete upper bounds on the constants $C_{sto}, C_{add}$ and $C_{adv}$ to Appendix A. □

6. Extensions to Full Bandit Feedback

The most natural extension of our work is to consider the full bandit feedback setting, where each time after playing action $X_t$ the learner only observes $(X_t, \ell_t)$. Again both stochastic and adversarial versions of the problem are well-studied in the literature, but there is no best-of-both-worlds result. Here, we provide a preliminary result for the simplest case $\mathcal{X} = \{0, 1\}^d$. Similar to Section 3.2, in the stochastic case we assume $\ell_t \sim D$ and define $\Delta_i = \mathbb{E}_{\ell \sim D}[\ell_i]$.

**Theorem 4.** For the full bandit feedback setting with $\mathcal{X} = \{0, 1\}^d$, FTRL with regularizer $\Psi(x) = \sum_{i=1}^{d} \sqrt{x_i} + \sqrt{1-x_i}$, learning rate $\eta_t = 1/\sqrt{t}$ and loss estimators $\hat{\ell}_t = \frac{(X_t, \ell_t)X_t}{x_t} - \frac{(X_t, \ell_t)(1-x_t)}{1-x_t}$ ensures:

$$
\overline{\text{Reg}_T} \leq O \left( \sum_{i: \Delta_i \neq 0} \frac{\log(T)}{|\Delta_i|} \right)
$$

in the stochastic case and

$$
\overline{\text{Reg}_T} \leq O \left( d\sqrt{T} \right)
$$

in the adversarial case. Moreover, both bounds are optimal.

**Proof sketch.** The optimization of FTRL decomposes over the coordinates in this case and it is clear that the stated algorithm is equivalent to the following: for each coordinate $i$, apply the algorithm of (Zimmert & Seldin, 2019) to a two-armed bandit problem where the loss of arm 1 at time $t$ is $\ell_t \Delta_i + \sum_{j \neq i} X_j \ell_j$ and the loss of arm 2 is $\sum_{j \neq i} X_j \ell_j$.³ In the stochastic case this exactly fits into the stochastically constrained adversarial setting of (Zimmert & Seldin, 2019) with gap $|\Delta|$ and therefore applying their Theorem 2 and summing up the regrets of each coordinate finish the proof for the stated regret bounds. The optimality of the stochastic bound follows from (Combes et al., 2017) and the optimality of the adversarial bound is trivial since even with full information $\Omega(d\sqrt{T})$ is unavoidable. □

For general action sets, however, the problem becomes significantly harder, because all known adversarial algorithms (e.g. (Cesa-Bianchi & Lugosi, 2012)) require implicit or explicit exploration of order $1/\sqrt{T}$, which prohibits $\log(T)$ regret for the stochastic case. We leave this as a future direction.

7. Conclusions

In this work we provide the first best-of-both-worlds results for combinatorial bandits, via an FTRL-based algorithm with a novel hybrid regularizer. Our bounds are worst-case optimal and also optimal for two concrete instances as well. Empirical comparisons also confirm our theory.

Other than the clear open problem on full bandit feedback mentioned in Section 6, another open question is whether our stochastic bound is instance-optimal as in (Combes et al., 2017), and if not, whether there is a best-of-both-worlds algorithm that is instance-optimal in the stochastic case. One can also ask the same question for the adversarial case, however, little to nothing is known for the instance-optimality of the adversarial case, let alone best-of-both-worlds results.

References

Anantharam, V., Varaiya, P., and Walrand, J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: iid rewards. IEEE Transactions on Automatic Control, 32(11), 1987.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. Regret in online combinatorial optimization. Mathematics of Operations Research, 39(1), 2013.

³The losses are well defined since they do not depend on $X_t$. Although the range of the losses do not fall into $[0, 1]$ as assumed in (Zimmert & Seldin, 2019), it is straightforward to verify that their results hold as long as the difference of losses falls into $[0, 1]$. 

A
Beating Stochastic and Adversarial Semi-bandits Optimally and Simultaneously

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 32(1), 2002.

Bertsekas, D. P., Nedi, A., Ozdaglar, A. E., et al. Convex analysis and optimization. Athena Scientific, 2003.

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

Bubeck, S., Perchet, V., and Rigollet, P. Bounded regret in stochastic multi-armed bandits. In Conference on Learning Theory, 2013.

Bubeck, S., Cohen, M. B., and Li, Y. Sparsity, variance and curvature in multi-armed bandits. In International Conference on Algorithmic Learning Theory, 2018.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. Journal of Computer and System Sciences, 78(5), 2012.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In International Conference on Machine Learning, 2013.

Combes, R., Shahi, M. S. T. M., Proutiere, A., et al. Combinatorial bandits revisited. In Advances in Neural Information Processing Systems, 2015.

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

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 1997.

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

Gaillard, P., Stoltz, G., and Van Erven, T. A second-order bound with excess losses. In Conference on Learning Theory, 2014.

Gopalan, A., Mannor, S., and Mansour, Y. Thompson sampling for complex online problems. In International Conference on Machine Learning, 2014.

Hazan, E. et al. Introduction to online convex optimization. Foundations and Trends® in Optimization, 2(3-4), 2016.

Koolen, W. M., Warmuth, M. K., Kivinen, J., et al. Hedging structured concepts. In Conference on Learning Theory, 2010.

Koolen, W. M., Grünwald, P., and van Erven, T. Combining adversarial guarantees and stochastic fast rates in online learning. In Advances in Neural Information Processing Systems, 2016.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight regret bounds for stochastic combinatorial semi-bandits. In Artificial Intelligence and Statistics, 2015.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1), 1985.

Lattimore, T. and Szepesvari, C. The end of optimism? an asymptotic analysis of finite-armed linear bandits. 2017.

Lattimore, T., Kveton, B., Li, S., and Szepesvari, C. Toprank: A practical algorithm for online stochastic ranking. In Advances in Neural Information Processing Systems, 2018.

Luo, H. and Schapire, R. E. Achieving all with no parameters: Adanormalhedge. In Conference on Learning Theory, 2015.

Luo, H., Wei, C.-Y., and Zheng, K. Efficient online portfolio with logarithmic regret. In Advances in Neural Information Processing Systems, 2018.

Neu, G. First-order regret bounds for combinatorial semi-bandits. In Conference on Learning Theory, 2015.

Neu, G. and Bartók, G. An efficient algorithm for learning with semi-bandit feedback. In International Conference on Algorithmic Learning Theory, 2013.

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

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In International Conference on Machine Learning, 2014.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In Computational Learning Theory, 2018.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In Artificial Intelligence and Statistics, 2019.
A. Omitted details for the Proof of Theorem 1

In this section we provide omitted details for the proof of Theorem 1. We first prove Lemmas 1 and 2, then continue on Section 5 and prove the upper bounds for $C_{sto}$, $C_{add}$ and $C_{adv}$.

A.1. Regularization penalty

In order to bound the regularization penalty, we make use of the following standard result for FTRL.

Lemma 3. The penalty term defined in Eq. (3) is upper bounded by

$$\text{Reg}_{pen} \leq \mathbb{E} \left[ \frac{-\Psi(x_i) + \Psi(x^*)}{\eta_t} + \sum_{t=2}^{T} (\eta_t^{-1} - \eta_{t-1}^{-1}) (-\Psi(x_t) + \Psi(x^*)) \right].$$

Proof. We proceed as follows:

$$\sum_{t=1}^{T} \left( -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \langle x^*, \hat{e}_t \rangle \right)$$

$$= \sum_{t=1}^{T} \left( \min_{x \in \text{Conv}(x)} \left\{ \langle x, \hat{L}_t \rangle + \eta_t^{-1}\Psi(x) \right\} - \left( \langle x_t, \hat{L}_{t-1} \rangle + \eta_{t-1}^{-1}\Psi(x_t) \right) \right) - \sum_{t=1}^{T} \langle x^*, \hat{e}_t \rangle$$

(by the definitions of $\Phi_t$ and $x_t$)

$$\leq \langle x^*, \hat{L}_T \rangle + \eta_T^{-1}\Psi(x^*) + \sum_{t=1}^{T} \left( \langle x_{t+1}, \hat{L}_t \rangle + \eta_t^{-1}\Psi(x_{t+1}) \right) - \sum_{t=1}^{T} \left( \langle x_t, \hat{L}_{t-1} \rangle + \eta_{t-1}^{-1}\Psi(x_t) \right) - \langle x^*, \hat{L}_T \rangle$$

$$= \eta_T^{-1}\Psi(x^*) + \sum_{t=2}^{T} \eta_{t-1}^{-1}\Psi(x_t) - \sum_{t=1}^{T} \eta_{t-1}^{-1}\Psi(x_t)$$

(by telescoping and $\hat{L}_0 = 0$)

$$= \frac{-\Psi(x_1) + \Psi(x^*)}{\eta_t} + \sum_{t=2}^{T} (\eta_t^{-1} - \eta_{t-1}^{-1}) (-\Psi(x_t) + \Psi(x^*)).$$

Finally using $\mathbb{E} |\ell_t| = \mathbb{E}[\hat{e}_t]$ and plugging in the definition of $\text{Reg}_{pen}$ finish the proof. \hfill \Box

Proof of Lemma 1. We directly plug into Lemma 3 the learning rate $\eta_t = 1/\sqrt{t}$ and the regularizer $\Psi(x) = \sum_{i=1}^{d} \sqrt{x_i} - \gamma(1 - x_i) \log(1 - x_i)$. Since $\gamma \leq 1$ and $-(1 - x) \log(1 - x) \leq \frac{x}{2}$ for $x \in [0, 1]$, we get

$$-\Psi(x_t) + \Psi(x^*) = \sum_{i=1}^{d} \sqrt{x_{ti}} - \gamma(1 - x_{ti}) \log(1 - x_{ti}) - \sum_{i:x_i^* = 1} \sqrt{1}$$

$$\leq \sum_{i:x_i^* = 0} \frac{3}{2} \sqrt{x_{ti}} - \sum_{i:x_i^* = 1} \gamma(1 - x_{ti}) \log(1 - x_{ti})$$

$$\leq \frac{3}{2} \left( \sum_{i:x_i^* = 0} \sqrt{x_{ti}} - \sum_{i:x_i^* = 1} \gamma(1 - x_{ti}) \log(1 - x_{ti}) \right).$$

It further holds that $\eta_t = \eta_t^{-1}$ and

$$\eta_t^{-1} - \eta_{t-1}^{-1} = \sqrt{t} - \sqrt{t-1} \leq \frac{1}{2\sqrt{t-1}} \leq \frac{1}{\sqrt{t}} = \eta_t.$$
Inserting everything into Lemma 3:

\[ \text{Reg}_{\text{pen}} \leq \mathbb{E} \left[ \frac{-\Psi(x_1) + \Psi(x^*)}{\eta_1} + \frac{T}{\sum_{t=2}^T} \eta_t^{-1} \eta_t^{-1} \right] \]

\[ \leq \mathbb{E} \left[ \sum_{t=1}^T \eta_t \left( -\Psi(x_t) + \Psi(x^*) \right) \right] \]

\[ \leq \mathbb{E} \left[ \sum_{t=1}^T \frac{3}{2\sqrt{t}} \left( \sum_{i:x_i^*=0} \sqrt{x_{ti}} - \sum_{i:x_i^*=1} \gamma(1 - x_{ti}) \log(1 - x_{ti}) \right) \right] \]

\[ \leq \sum_{t=1}^T \frac{3}{2\sqrt{t}} \left( \sum_{i:x_i^*=0} \sqrt{\mathbb{E}[x_{ti}]} - \sum_{i:x_i^*=1} \gamma(1 - \mathbb{E}[x_{ti}]) \log(1 - \mathbb{E}[x_{ti}]) \right) \cdot \]

where the last step follows from Jensen’s inequality and the concavity of functions \( \sqrt{x} \) and \(-(1 - x) \log(1 - x)\).

A.2. Stability term

Bounding the stability term defined in Eq. (3) requires tools from convex analysis. First we extend the domain of \( \Psi \) to \( \mathbb{R}^d \) by setting \( \Psi(x) = \infty \), \( \forall x \in \mathbb{R}^d \setminus [0, 1]^d \). Recall the convex conjugate of a convex function \( f \) is defined as

\[ f^*(\cdot) = \max_{x \in \mathbb{R}^d} \langle x, \cdot \rangle - f(x), \]

and the Bregman divergence associated with \( f \) is defined as

\[ D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle. \]

By the above definition, \( \Phi_t \) can be written as \( (\Psi_t + \mathbb{I}_{\text{Conv}(\mathcal{X})})^* \). Note that \( \Psi_t^* \) differs from \( \Phi_t \) because it does not constrain its maximizer to be within \( \text{Conv}(\mathcal{X}) \). The following properties hold (see, e.g., Chapter 7 of (Bertsekas et al., 2003)):

\[ \nabla \Phi_t(x) = \arg\max_{x \in \text{Conv}(\mathcal{X})} \langle x, \cdot \rangle - \Psi_t(x), \quad (5) \]

\[ \nabla \Psi_t^*(x) = \arg\max_{x \in [0, 1]^d} \langle x, \cdot \rangle - \Psi_t(x). \quad (6) \]

For \( \Psi_t \) and \( \Psi_t^* \), we have

\[ \nabla \Psi_t = (\nabla \Psi_t^*)^{-1}, \]

\[ \nabla^2 \Psi_t(x) = (\nabla^2 \Psi_t^*(\nabla \Psi_t(x)))^{-1}. \]

Furthermore, by Taylor’s theorem, for any \( x, y \in \mathbb{R}^d \) there exists a \( z \in \text{Conv}(\{x, y\}) \) such that

\[ D_{\Psi_t^*}(x, y) = \frac{1}{2}||x - y||^2_{\nabla^2 \Psi_t^*(z)}. \]

The explicit expressions for \( \nabla \Psi_t, \nabla^2 \Psi_t \) and a convenient upper bound for \( (\nabla^2 \Psi_t)^{-1} \) in the domain \((0, 1)^d\) are

\[ \Psi_t(x) = \eta_t^{-1} \left( \sum_{i=1}^d -\sqrt{x_i} + \gamma(1 - x_i) \log(1 - x_i) \right), \]

\[ \nabla \Psi_t(x) = \eta_t^{-1} \left( -\frac{1}{2\sqrt{x_i}} - \gamma \log(1 - x_i) - \gamma \right)_{i=1}^{d} \]

\[ \nabla^2 \Psi_t(x) = \eta_t^{-1} \text{diag} \left[ \left( \frac{1}{4\sqrt{x_i}^3} + \gamma \frac{1}{1 - x_i} \right)_{i=1}^{d} \right], \]

\[ (\nabla^2 \Psi_t(x))^{-1} \preceq \eta_t \text{diag} \left[ \left( \min \left( \frac{4}{\sqrt{x_i}}, \gamma^{-1}(1 - x_i) \right) \right)_{i=1}^{d} \right], \]
We then proceed as follows:

where \( (v_i)_{i=1}^{d} \) denotes \( (v_1, \ldots, v_d) \), \( \text{diag}[(v_i)_{i=1}^{d}] \) denotes a diagonal matrix with \( (v_i)_{i=1}^{d} \) on the diagonal, and \( A \preceq B \) for two matrices \( A \) and \( B \) means \( B - A \) is positive semidefinite. Note \( \nabla \Psi_t \) is a bijection from \( (0, 1)^d \) to \( \mathbb{R}^d \). Therefore \( \nabla \Psi_t^*(L) \in (0, 1)^d \) for any \( L \in \mathbb{R}^d \), and all \( x_i \)'s we consider here are in the domain \( (0, 1)^d \).

The following Lemma will be useful to show that the stability term can be bounded independently of the action set \( X \).

**Lemma 4.** For any \( L \), let \( \tilde{L} = \nabla \Psi_t(\nabla \Phi_t(L)) \). Then it holds for any \( \ell \in \mathbb{R}^d \):

\[
D_{\Phi_t}(L + \ell, L) \leq D_{\Psi_t^*}(L + \ell, \tilde{L}).
\]

**Proof.** First we state two equalities that follow from the previously stated properties.

\[
\nabla \Psi_t^*(\tilde{L}) = \nabla \Psi_t^*(\nabla \Psi_t(\nabla \Phi_t(L))) \quad \text{Eq. (7)}
\]
\[
\Psi_t^*(\tilde{L}) = \left\langle \nabla \Psi_t^*(\tilde{L}), \tilde{L} \right\rangle - \nabla \Psi_t(\nabla \Phi_t(L)), \quad \text{Eq. (6)}
\]
\[
\nabla \Phi_t(L) = \left\langle \nabla \Phi_t(L), \tilde{L} \right\rangle - \Psi_t(\nabla \Phi_t(L)) = \left\langle \nabla \Phi_t(L), \tilde{L} \right\rangle - \Psi_t(\nabla \Phi_t(L)) \quad \text{Eq. (5)}
\]
\[
 \Phi_t(L) = \Phi_t(L) + \left\langle \nabla \Phi_t(L), \tilde{L} - L \right\rangle.
\]

We then proceed as follows:

\[
\begin{align*}
D_{\Psi_t^*}(L + \ell, \tilde{L}) & = \Psi_t^*(L + \ell) - \Psi_t^*(\tilde{L}) - \left\langle \nabla \Psi_t^*(L), \ell \right\rangle & \text{(definition of Bregman divergence)} \\
 & = \Psi_t^*(L + \ell) - \Phi_t(L) - \left\langle \nabla \Phi_t(L), \tilde{L} - L \right\rangle - \left\langle \nabla \Phi_t(L), \ell \right\rangle & \text{(by Eq. (12) and (13))} \\
 & = \Psi_t^*(L + \ell) - \Phi_t(L) - \left\langle \nabla \Phi_t(L), \tilde{L} - L + \ell \right\rangle \\
 & \geq \left\langle \nabla \Phi_t(L + \ell), \tilde{L} + \ell \right\rangle - \Psi_t(\nabla \Phi_t(L + \ell)) - \Phi_t(L) - \left\langle \nabla \Phi_t(L), \tilde{L} - L + \ell \right\rangle & \left( \Psi_t^* \right. \text{is defined as the maximum)} \\
 & = (\nabla \Phi_t(L + \ell), L + \ell) - \Psi_t(\nabla \Phi_t(L + \ell)) + \left\langle \nabla \Phi_t(L + \ell), \tilde{L} - L \right\rangle - \Phi_t(L) - \left\langle \nabla \Phi_t(L), \tilde{L} - L + \ell \right\rangle \\
 & = \Phi_t(L + \ell) + \left\langle \nabla \Phi_t(L + \ell), \tilde{L} - L \right\rangle - \Phi_t(L) - \left\langle \nabla \Phi_t(L), \tilde{L} - L + \ell \right\rangle & \text{(by the definition of } \Phi_t \text{ and Eq. (5)}) \\
 & = D_{\Phi_t}(L + \ell, L) + \left\langle \nabla \Phi_t(L + \ell) - \nabla \Phi_t(L), \tilde{L} - L \right\rangle \\
 & = D_{\Phi_t}(L + \ell, L) + \left\langle \nabla \Phi_t(L + \ell) - \nabla \Phi_t(L), \nabla \Psi_t(\nabla \Phi_t(L)) - L \right\rangle \\
 & \geq D_{\Phi_t}(L + \ell, L).
\end{align*}
\]

The last step is by the first-order optimality condition: for the maximizer \( \nabla \Phi_t(L) := \arg \max_{x \in \text{Conv}(X)} \langle x, L \rangle - \Psi_t(x) \) it must hold that \( \langle y - \nabla \Phi_t(L), L - \nabla \Psi_t(\nabla \Phi_t(L)) \rangle \leq 0 \) for any \( y \in \text{Conv}(X) \). \[ \square \]

The next Lemma will be useful to bound the eigenvalues of the Hessian of \( \Psi_t^* \).

**Lemma 5.** If \( \eta_t \leq \min \{ \frac{\sqrt{2} - 1}{2}, \frac{\log(2)}{4} \} \), then for any \( x \in (0, 1)^d \) and \( \ell \) such that \( -1 \leq \ell_i \leq 2x_i \) for all \( i \), we have

\[
2x_i - 1 \leq \nabla \Psi_t^* (x) - \hat{\ell} \right_1 \leq 2x_i.
\]

**Proof.** The functions \( \nabla \Psi_t \) and \( \nabla \Psi_t^* \) are symmetric and independent in each dimension. Therefore it is sufficient to consider \( d = 1 \) and drop the index \( i \).

For the upper bound we can assume \( x < \frac{1}{2} \); otherwise the statement is trivial since the range of \( \nabla \Psi_t^* \) is \( (0, 1)^d \). Now assume
Beating Stochastic and Adversarial Semi-bandits Optimally and Simultaneously

the opposite holds: \( \nabla \Psi_t^*(\nabla \Psi_t(x) - \hat{\ell}) > 2x \), then we have

\[
\hat{\ell} = \nabla \Psi_t(x) - \nabla \Psi_t(x) + \hat{\ell} = \nabla \Psi_t(x) - \nabla \Psi_t(\nabla \Psi_t^*(\nabla \Psi_t(x) - \hat{\ell}))
\]
\[
< \nabla \Psi_t(x) - \nabla \Psi_t(2x)
\]
\[
= \eta_t^{-1} \left( -\frac{1}{2\sqrt{x}} - \gamma \log(1-x) + \frac{1}{2\sqrt{2x}} + \gamma \log(1-2x) \right)
\]
\[
< -\eta_t^{-1} \left( \frac{\sqrt{2} - 1}{2\sqrt{2}} \right) \frac{1}{\sqrt{x}}
\]
\[
< -\eta_t^{-1} \left( \frac{\sqrt{2} - 1}{2} \right).
\]

\( (x \leq \frac{1}{2}) \)

The last line is a contradiction to the conditions \( \eta_t \leq \frac{\sqrt{2} - 1}{2} \) and \( \hat{\ell} \geq -1 \).

For the lower bound we can assume \( x > \frac{1}{2} \), otherwise the statement is again trivial. Assume the opposite holds: \( \nabla \Psi_t^*(\nabla \Psi_t(x) + \hat{\ell}) < 2x - 1 \), then we have

\[
\hat{\ell} = \nabla \Psi_t(x) - \nabla \Psi_t(\nabla \Psi_t^*(\nabla \Psi_t(x) - \hat{\ell}))
\]
\[
> \nabla \Psi_t(x) - \nabla \Psi_t(2x - 1)
\]
\[
= \eta_t^{-1} \left( -\frac{1}{2\sqrt{x}} - \gamma \log(1-x) + \frac{1}{2\sqrt{2x} - 1} + \gamma \log(2-2x) \right)
\]
\[
> \eta_t^{-1} \log(2) > \eta_t^{-1} \frac{\gamma \log(2)}{4} \frac{2}{x}.
\]

\( (\gamma \leq 1 \text{ and } x > 1/2) \)

which again leads to a contradiction to the conditions \( \eta_t \leq \frac{\gamma \log(2)}{4} \) and \( \hat{\ell} \leq \frac{2}{x} \). This finishes the proof. \( \square \)

Finally we are ready to prove Lemma 2.

**Proof of Lemma 2.** Let \( \tilde{x}_t = \nabla \Psi^*(\nabla \Psi(x_t) - \hat{\ell}_t) \). Define \( A_t = \bigotimes_{i=1}^d [x_{t_i}, \tilde{x}_{t_i}] \). For any \( t_0 \geq 0 \), we bound the stability
We choose \( t_0 = 58 \gamma^{-2} \) such that \( \eta t \leq \min\left\{ \frac{\sqrt{2-1}}{2}, \frac{\gamma \log(2)}{4} \right\} \) for any \( t \geq t_0 \). By the construction of \( \hat{\ell} \) we clearly have \(-1 \leq \hat{\ell}_t \leq \frac{2}{x_t}\). We can then apply Lemma 5 to conclude that \( \tilde{x}_t \in [2x_t - 1, 2x_t] \). Therefore, with the form of Hessian (11) we have:

\[
\forall x \in A_t : \quad \nabla^2 \Psi(x)^{-1} \leq \text{diag} \left[ \left( \min\left\{ 4\sqrt{(2x_t)^2}, 2\gamma^{-1}(1 - x_t) \right\} \right)_{i=1,...,d} \right],
\]

(1) The difference of potentials is always negative and the loss \( \langle X_t, \ell_t \rangle \) is bounded by \( m = \max_{x \in X} |x|_1 \).

(2) By the tower rule of conditional expectation, the unbiasedness of \( \hat{\ell} \) and the sampling assumption, it holds that

\[
\mathbb{E} [\langle X_t, \ell_t \rangle] = \mathbb{E} [\mathbb{E}_t [\langle X_t, \ell_t \rangle]] = \mathbb{E} [\mathbb{E}_t [\langle x_t, \ell_t \rangle]] = \mathbb{E} [\mathbb{E}_t [\langle x_t, \ell_t \rangle]].
\]

(3) Application of Lemma 4.

(4) Property 9 ensures that some \( z_t \in \text{Conv}(\nabla \Psi_t(x), \nabla \Psi_t(\tilde{x})) \) exists that satisfies the equality.

(5) By property (8) and the coordinate-wise monotonicity of \( \nabla \Psi_t \) so that \( \nabla \Psi_t^* (z_t) \subset \bigotimes_{i=1}^d [x_t, \tilde{x}_t] = A_t \).
and therefore,
\[
\sum_{t=t_0}^T \mathbb{E}_t \left[ \max_{x \in \mathcal{A}_t} \frac{\eta_t}{2} \| \hat{\ell}_t \|_{\mathbb{P}\mathcal{T}}^2 \right] \leq \sum_{t=t_0}^T \mathbb{E}_t \left[ \frac{\eta_t}{2} \sum_{i=1}^d (\hat{\ell}_{ti})^2 \min \{ 4 \sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti}) \} \right] \\
\leq \sum_{t=t_0}^T \frac{\eta_t}{2} \sum_{i=1}^d 4 \min \{ 4 \sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti}) \} \\
\leq \sum_{t=1}^T 16\sqrt{2} \eta_t \sum_{i=1}^d \min \{ \sqrt{x_{ti}}, \gamma^{-1}(1-x_{ti}) \} \\
\leq \sum_{t=1}^T 16\sqrt{2} \left( \sum_{i:x_i^* = 0} \sqrt{x_{ti}} + \sum_{i:x_i^* = 1} \gamma^{-1}(1-x_{ti}) \right) \quad (15)
\]

(1) Conditioned on \( \mathcal{F}_{t-1} \), only \((\hat{\ell}_{ti})^2\) is random and its expectation is
\[
\mathbb{E}_t \left[ (\hat{\ell}_{ti})^2 \right] = x_{ti} \left( \frac{\ell_{ti} + 1}{x_{ti}} - 1 \right)^2 + 1 - x_{ti} = \frac{4}{x_{ti}} (\ell_{ti} + 1)^2 - 2(\ell_{ti} + 1)x_{ti} + x_{ti} \leq \frac{4}{x_{ti}} \left( 4 - 4x_{ti} + x_{ti} \right) \leq \frac{4}{x_{ti}}
\]

(2) Note that it always holds
\[
\frac{4}{x_{ti}} \min \{ 4 \sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti}) \} \leq \frac{16}{x_{ti}} \sqrt{(2x_{ti})^3} = 32 \sqrt{2} x_{ti}.
\]
So it suffices to prove \( \frac{4}{x_{ti}} \min \{ 4 \sqrt{(2x_{ti})^3}, 2\gamma^{-1}(1-x_{ti}) \} \leq 32 \sqrt{2} \gamma^{-1}(1-x_{ti}) \). We consider two cases: (A) If \( 4 \sqrt{(2x_{ti})^3} \leq 2\gamma^{-1}(1-x_{ti}) \), then we need to prove \( \sqrt{x_{ti}} \leq \gamma^{-1}(1-x_{ti}) \). This is true since either \( x_{ti} \geq 1/\sqrt{32} \) and thus \( \sqrt{x_{ti}} \leq 2 \sqrt{(2x_{ti})^3} \leq \gamma^{-1}(1-x_{ti}) \), or \( x_{ti} < 1/\sqrt{32} \) in which case \( \sqrt{x_{ti}} \leq 1 - x_{ti} \leq \gamma^{-1}(1-x_{ti}) \). (B) If \( 4 \sqrt{(2x_{ti})^3} \geq 2\gamma^{-1}(1-x_{ti}) \), then \( x_{ti} \) must be larger than 1/4. In this case we bound \( \frac{1}{x_{ti}} \) by 4 and the desired inequality follows.

The proof is concluded by inserting Eq. (15) into Eq. (14) and using Jensen’s inequality to move the expectation into the concave functions.

\[\square\]

A.3. General upper bounds for \( C_{sto}, C_{add} \) and \( C_{adv} \)

We now finish the proof of Theorem 1 on the upper bounds of the three constants.

Bounding \( C_{adv} \):
\[
C_{adv} = \max_{x \in \text{Conv}(X)} \sum_{i:x_i^* = 0} \sqrt{x_{i}} + \sum_{i:x_i^* = 1} (\gamma^{-1} - \gamma \log(1-x_{i}))(1-x_{i}) \\
\leq \max_{x \in \text{Conv}(X)} \sum_{i:x_i^* = 0} \sqrt{x_{i}} + \sum_{i:x_i^* = 1} \gamma \sqrt{1-x_{i}} + \sum_{i:x_i^* = 1} \gamma^{-1}(1-x_{i}) \\
\leq \max_{x \in \text{Conv}(X)} \sqrt{\sum_{i:x_i^* = 0} 1} \left( \sum_{i:x_i^* = 0} x_{i} \right) + \gamma \sqrt{\sum_{i:x_i^* = 1} 1} \left( \sum_{i:x_i^* = 1} (1-x_{i}) \right) + \gamma^{-1} m \quad (\text{Cauchy-Schwarz}) \\
\leq \sqrt{dm} + \gamma m + \gamma^{-1} m \\
\leq O \left( \gamma^{-1} \sqrt{md} \right).
\]

Bounding \( C_{sto} \): \( C_{sto} \) is defined as \( \max_{\alpha \in [0,\infty)^{x}} f(\alpha) - r(\alpha) \). First we bound \( f(\alpha) \):
\[
f(\alpha) = \sum_{i:x_i^* = 0} \sqrt{\alpha_{x_i}} = \sum_{i:x_i^* = 0} \sqrt{\sum_{x \in \mathcal{X}} \alpha_{x} x_{i}} = \sum_{x \in \mathcal{X} \setminus \{ x^* \}} \alpha_{x} x_{i} \leq \sqrt{d} \sum_{i:x_i^* = 0} \sum_{x \in \mathcal{X} \setminus \{ x^* \}} \alpha_{x} x_{i} \leq \sqrt{dm} \sum_{x \in \mathcal{X} \setminus \{ x^* \}} \alpha_{x}.
\]
On the other hand,
\[ r(\alpha) = \sum_{x \in X \setminus \{x^*\}} \alpha_x \Delta_x \geq \Delta_{\min} \sum_{x \in X \setminus \{x^*\}} \alpha_x. \]

Combining them we get
\[
C_{sto} \leq \max_{\alpha \in [0, \infty)} \frac{d m}{\Delta_{\min}} \left( \sum_{x \in X \setminus \{x^*\}} \alpha_x - \Delta_{\min} \sum_{x \in X \setminus \{x^*\}} \alpha_x \right)
\leq \max_{A \geq 0} \sqrt{d m A - \Delta_{\min} A}
\leq \max_{A \geq 0} \frac{\Delta_{\min} A}{4} - \Delta_{\min} A \quad \text{(AM-GM inequality)}
\leq \frac{d m}{4\Delta_{\min}}.
\]

**Bounding \( C_{add} \):** Recall \( C_{add} \) is defined as \( \sum_{t=1}^{\infty} \max_{\alpha \in \Delta(X)} \left( \frac{100}{N^{t}} \log N \right) - r(\alpha) \). We will give a upper bound for \( g(\overline{\pi}) \) and lower bound for \( r(\alpha) \) below.

We first prove the following property: for any \( y \in \mathbb{R}_{+}^{N}, \sum_{i=1}^{N} y_i \log \frac{1}{y_i} \leq \|y\|_1 \log \frac{N}{\|y\|_1} \). Indeed, by the concavity of the log function and Jensen’s inequality,
\[
\sum_{i=1}^{N} y_i \log \frac{1}{y_i} \leq \log \left( \sum_{i=1}^{N} y_i \frac{1}{y_i} \right) = \log \frac{N}{\|y\|_1}.
\]

Therefore, for any \( \alpha \in \Delta(X) \) we have
\[
g(\overline{\pi}) = \sum_{i:x_i^*=1} \left( \gamma^{-1} + \gamma \log \left( \frac{1}{1-\alpha_i} \right) \right) (1-\overline{\alpha}_i)
\leq \left( \sum_{i:x_i^*=1} (1-\overline{\alpha}_i) \right) \left( \gamma^{-1} + \gamma \log \frac{m}{\sum_{i:x_i^*=1} (1-\overline{\alpha}_i)} \right). \quad \text{(using the above property)}
\]

Then consider the following two facts. First, the function of \( y \) defined by \( y(\gamma^{-1} + \gamma \log \frac{m}{y}) \) is increasing in \( y \in [0, m] \). This can be verified by
\[
\frac{\partial}{\partial y} \left( y \left( \gamma^{-1} + \gamma \log \frac{m}{y} \right) \right) = \gamma^{-1} + \gamma \log m - \gamma \log y - \gamma \geq 0. \quad (\gamma \leq 1)
\]

Second, we have \( \sum_{i:x_i^*=1} (1-\overline{\alpha}_i) = \sum_{i:x_i^*=1} \sum_{\alpha \in X} \alpha_x (1 - x_i) = \sum_{i:x_i^*=1} \sum_{\alpha \in X \setminus \{x_i^*\}} \alpha_x (1 - x_i) \leq \|x^*\|_1 \left( \sum_{\alpha \in X \setminus \{x_i^*\}} \alpha_x \right) \leq m \left( \sum_{\alpha \in X \setminus \{x_i^*\}} \alpha_x \right). \)

Combining these two facts with the above bound for \( g(\overline{\pi}) \), we get
\[
g(\overline{\pi}) \leq m \left( \sum_{\alpha \in X \setminus \{x^*\}} \alpha_x \right) \left( \gamma^{-1} + \gamma \log \frac{1}{\sum_{\alpha \in X \setminus \{x^*\}} \alpha_x} \right).
\]

On the other hand, we have the lower bound for \( r(\alpha) \):
\[
r(\alpha) = \sum_{x \in X \setminus \{x^*\}} \alpha_x \Delta_x \geq \Delta_{\min} \sum_{x \in X \setminus \{x^*\}} \alpha_x.
\]
Therefore,

\[
C_{\text{add}} = \sum_{t=1}^{\infty} \max_{\alpha \in \Delta(|X|)} \left( \frac{100}{\sqrt{t}} g(\pi) - r(\alpha) \right) 
\leq \sum_{t=1}^{\infty} \max_{A \in [0,1]} \left( \frac{100}{\sqrt{t}} mA \left( \gamma^{-1} + \gamma \log \frac{1}{A} \right) - \Delta_{\text{min}} A \right).
\]

We further bound it by the sum of the following two summations:

- \[
\sum_{t=1}^{\infty} \max_{A \in [0,1]} \left( \frac{100}{\sqrt{t}} mA \gamma^{-1} - \frac{1}{2} \Delta_{\text{min}} A \right)
\]

- \[
\sum_{t=1}^{\infty} \max_{A \in [0,1]} \left( \frac{100}{\sqrt{t}} mA \log \frac{1}{A} - \frac{1}{2} \Delta_{\text{min}} A \right)
\]

Lemma 6 and 7 below respectively bound these two as \(O \left( \frac{m^2}{\Delta_{\text{min}}} \right)\) and \(O \left( \frac{m^2}{\Delta_{\text{min}}} \gamma^2 \right)\), which finishes the proof.

**Lemma 6.** For any \(C > 0\) and \(\Delta > 0\), we have \(\sum_{t=1}^{\infty} \max_{A \in [0,1]} \left( \frac{C}{\sqrt{t}} A - \Delta A \right) \leq O \left( \frac{C^2}{\Delta} \right)\).

**Proof.** Let \(T_0\) be the largest \(t\) such that \(\frac{C}{\sqrt{t}} - \Delta > 0\), then

\[
\sum_{t=1}^{\infty} \max_{A \in [0,1]} \left( \frac{C}{\sqrt{t}} A - \Delta A \right) \leq \sum_{t=1}^{T_0} \frac{C}{\sqrt{t}} \leq 2C \sqrt{T_0} = O \left( \frac{C^2}{\Delta} \right).
\]

**Lemma 7.** For any \(C > 0\) and \(\Delta > 0\), we have \(\sum_{t=1}^{\infty} \max_{A \in [0,1]} \left( \frac{C}{\sqrt{t}} \log \frac{1}{A} - \Delta A \right) \leq \frac{C^2}{\Delta} \).

**Proof.** We first solve the inner optimization with respect to a specific \(t\). Taking the derivative with respect to \(A\), and setting it to zero:

\[
\frac{C}{\sqrt{t}} \log \frac{1}{A^*} - \frac{C}{\sqrt{t}} - \Delta = 0,
\]

we get the solution

\[
A^* = \exp \left( -1 - \sqrt{\frac{1}{C} \Delta} \right).
\]

And thus,

\[
\max_{A \in [0,1]} \left( \frac{C}{\sqrt{t}} \log \frac{1}{A} - \Delta A \right) = \frac{C}{\sqrt{t}} A^* \log \frac{1}{A^*} - \Delta A^* \quad \text{Eq.} (16) \quad A^* \left( \frac{C}{\sqrt{t}} + \Delta \right) - \Delta A^*
\]

\[
= \frac{C}{\sqrt{t}} \exp \left( -1 - \sqrt{\frac{1}{C} \Delta} \right)
\]

Finally we have

\[
\sum_{t=1}^{\infty} \max_{A \in [0,1]} \left( \frac{C}{\sqrt{t}} \log \frac{1}{A} - \Delta A \right) \leq \sum_{t=1}^{\infty} \frac{C}{\sqrt{t}} \exp \left( -1 - \sqrt{\frac{1}{C} \Delta} \right) \leq \int_{t=0}^{\infty} \frac{C}{\sqrt{t}} \exp \left( -1 - \sqrt{\frac{1}{C} \Delta} \right) dt
\]

\[
= \frac{C^2}{\Delta} \int_{t=0}^{\infty} \frac{1}{\sqrt{t}} \exp(-1 - \sqrt{t}) dt \leq \frac{C^2}{\Delta}.
\]

\[\square\]
B. Omitted Details for Sections 3.2 and 3.3

In this section we provide omitted details for the two special cases: full combinatorial set and \(m\)-set.

B.1. Optimality of the stochastic bound when \(X = \{0, 1\}^d\)

As mentioned in the proof of Theorem 2, we provide here for completeness a proof showing that when \(d = 1\) and \(\Delta > 0\), the regret is at least \(\Omega(\frac{\log T}{\Delta})\).

Assume that there exists an algorithm that is at least as good as ours asymptotically, which implies \(\lim_{T \to \infty} \frac{\log(\text{Reg}_T)}{\log(T)} \leq \lim_{T \to \infty} \frac{\log(O(\log(T)))}{\log(T)} = 0\) for any problem. For some \(\Delta > 0\) we consider two problems: \(E[\ell] = \Delta\) and \(E[\ell] = -\Delta\). For simplicity we assume that the losses are drawn from i.i.d. Gaussian with variance \(\sigma^2 = 1\), but the proof can be easily transferred to Bernoulli noise as well. For the problem with positive loss, we denote the regret as \(\text{Reg}_T^+\) and the probability space induced by an algorithm by \(P^+\). Equivalently we define \(\text{Reg}_T^-\) and \(P^-\). The relative entropy between \(P^+\) and \(P^-\) is

\[
KL(P^+, P^-) = \sum_{t=1}^{T} P^+(X_t = 1)(2\Delta)^2 = 4\text{Reg}_T^+ \Delta.
\]

Also we have by the definition of regret:

\[
P^+ \left( \sum_{t=1}^{T} X_t \geq \frac{T}{2} \right) + P^- \left( \sum_{t=1}^{T} X_t < \frac{T}{2} \right) \leq \frac{2(\text{Reg}_T^+ + \text{Reg}_T^-)}{\Delta T}.
\]

Using the high probability Pinsker inequality (included after the proof for completeness), we get

\[
\frac{2(\text{Reg}_T^+ + \text{Reg}_T^-)}{\Delta T} \geq \frac{1}{2} \exp \left( -4 \frac{\text{Reg}_T^+ \Delta}{\Delta} \right).
\]

Rearranging gives

\[
\frac{\text{Reg}_T^+}{\log(T)} = \frac{1}{4\Delta} - \frac{1}{4\Delta} \frac{\log(4(\text{Reg}_T^+ + \text{Reg}_T^-))}{\log(T)} + \frac{\log(\Delta)}{4\Delta \log(T)}.
\]

Taking the limit on both sides shows \(\lim_{T \to \infty} \frac{\text{Reg}_T^+}{\log(T)} = \Omega(\frac{1}{\Delta})\), which finishes the proof.

**Lemma 8** (High Probability Pinsker, e.g. (Bubeck et al., 2013)). Let \(\mathbb{P}\) and \(\mathbb{Q}\) be probability measures on the same measurable space \((\Omega, F)\) and let \(A \in F\) be an arbitrary event. Then,

\[
\mathbb{P}(A) + \mathbb{Q}(A^c) \geq \frac{1}{2} \exp(-\text{KL}(\mathbb{P}, \mathbb{Q})),
\]

where \(A^c\) is the complement of \(A\) and \(\text{KL}(\mathbb{P}, \mathbb{Q})\) the relative entropy.

B.2. Sampling rule for \(m\)-set

In this section \(X\) represents the \(m\)-set. We first define the following auxiliary vectors for \(0 \leq i \leq m, 0 \leq j \leq d - m\).

\[
\beta_{i,j} = \left(\frac{1, \ldots, 1, m-i, d-i-j, \ldots, m-i, d-i-j, 0, \ldots, 0}{i, j}\right) \in \text{Conv}(X).
\]

It is trivial to sample with mean \(\beta_{i,j}\) with the sampling rule:

\[
P_{i,j} = \text{Uniform}\left(\{x \in X \mid x_1 \ldots i = 1 \land x_{d-j+1} \ldots d = 0\}\right).
\]

This requires uniform sampling of a \((m - i)\)-sized subset of \((d - i - j)\) elements, which can be done in \(O(d)\) time.
Now for a given $x_t \in \text{Conv}(A)$, one sampling rule $P$ such that $E_{X \sim P}[X] = x_t$ is the following: First we sort the entries of $x_t$ so that $x$ is the sorted version with $x_1 \geq \cdots \geq x_d$. This takes $O(d \log(d))$ time. Next we decompose $x = \sum_{s=0}^{d} p_{x,s} \beta_{i,s,j}$, such that $p_{x,s} \in [0,1]$, $\sum_{s=0}^{d} p_{x,s} = 1$, $(i_0, j_0) = (0,0)$ and $(i_{s+1}, j_{s+1}) - (i_s, j_s) \in \{(1,0), (0,1)\}$. In other words, either $i$ or $j$ increases by one from $s$ to $s+1$. This decomposition is unique and can be computed in a greedy manner in time $O(d)$. Finally the full sampling scheme is $\sum_{s=0}^{d} p_{x,s} P_{i_s,j_s}$ (in terms of permuted coordinates). The runtime is dominated by the sorting and hence is $O(d \log(d))$ overall.

B.3. Complete proof for Theorem 3

Bounding $C_{\text{adv}}$:

$$C_{\text{adv}} = \max_{x \in \text{Conv}(X)} (f(x) + g(x)) = \max_{x \in \text{Conv}(X)} \sum_{i:x_i^* = 0} \sqrt{x_i} + \sum_{i:x_i^* = 1} (\gamma^{-1} - \gamma \log(1 - x_i))(1 - x_i).$$

The optimization problem is concave in $x$ and symmetric for all $i$ with the same value of $x_i^*$. This implies that the argmax solution must take the following form:

$$\left( \arg\max_{x \in \text{Conv}(X)} f(x) + g(x) \right)_i = \begin{cases} 
\lambda & \text{if } x_i^* = 0 \\
1 - \frac{d - m}{m} \lambda & \text{if } x_i^* = 1
\end{cases}$$

for some $\lambda \in [0, \min\{1, \frac{d}{m} - d\}]$.

Therefore,

$$C_{\text{adv}} = \max_{\lambda \in [0, \min\{1, \frac{d}{m} - d\}]} (d - m) \sqrt{\lambda} + m \left( \gamma^{-1} - \gamma \log \left( \frac{d - m}{m} \lambda \right) \right) \frac{d - m}{m} \lambda$$

$$= \max_{\lambda \in [0, \min\{1, \frac{d}{m} - d\}]} (d - m) \left( \sqrt{\lambda} + \left( \gamma^{-1} - \gamma \log \left( \frac{d - m}{m} \lambda \right) \right) \lambda \right). \tag{17}$$

Since $\frac{d - m}{m} \lambda \leq 1$ and $\gamma \leq 1$, the derivative is always positive:

$$\frac{\partial}{\partial \lambda} \left( \sqrt{\lambda} + \left( \gamma^{-1} - \gamma \log \left( \frac{d - m}{m} \lambda \right) \right) \lambda \right)$$

$$= \left( \frac{1}{2 \sqrt{\lambda}} + 2 \gamma^{-1} - \gamma \log \left( \frac{d - m}{m} \lambda \right) - \gamma \right) \geq \frac{1}{2 \sqrt{\lambda}} > 0.$$

Therefore we can simply plug in the upper border of $\lambda$ in Eq.(17):

Case $m \leq d/2$ (for which $\gamma = 1$ and the optimal $\lambda$ is $m/(d - m)$):

$$C_{\text{adv}} = (d - m) \left( \sqrt{\frac{m}{d - m}} + \frac{m}{d - m} \right) \leq 2 \sqrt{(d - m)m} = O \left( \sqrt{md} \right).$$

Case $m > d/2$ (for which $\gamma = \min \left\{ 1, 1/\sqrt{\log \left( \frac{d}{d - m} \right)} \right\}$ and the optimal $\lambda$ is 1):

Note that $\gamma \leq \frac{1}{\sqrt{\log \left( \frac{d}{d - m} \right)}}$ and thus $\gamma^{-1} = \max \left\{ 1, \sqrt{\log \left( \frac{d}{d - m} \right)} \right\} \leq \frac{\log \left( \frac{d}{d - m} \right)}{\sqrt{\log(2)}}$ and $-\gamma \leq -\frac{\log(2)}{\sqrt{\log(\frac{d}{d - m})}}$. Therefore

$$C_{\text{adv}} \leq (d - m) \left( 1 + \frac{1}{\sqrt{\log(2)}} \sqrt{\log \left( \frac{d}{d - m} \right)} + \frac{\log(2)}{\sqrt{\log(\frac{d}{d - m})}} \log \left( \frac{m}{d - m} \right) \right)$$

$$\leq (d - m) \left( 1 + \left( \frac{1}{\sqrt{\log(2)}} + \sqrt{\log(2)} \right) \sqrt{\log \left( \frac{d}{d - m} \right)} \right) = O \left( (d - m) \sqrt{\log \left( \frac{d}{d - m} \right)} \right).$$
Bounding $C_{sto}$: With our definitions of $\Delta_i$, for any $x \in X$, we have

$$
\Delta_x = E \left[ \sum_i (x_i - x_i^*) \ell_{ti} \right] = E \left[ \sum_i (x_i - x_i^*) (\ell_{ti} - \ell_{tm}) \right] = \sum_{i: x_i^* = 1} (1 - x_i) |\Delta_i| + \sum_{i: x_i^* = 0} x_i \Delta_i \geq \sum_{i: x_i^* = 0} x_i \Delta_i, \quad (18)
$$

and thus for any $\alpha \in [0, \infty)^{|X|}$

$$
r(\alpha) = \sum_{x \in X \setminus \{x^*\}} \alpha x \Delta_x \geq \sum_{x \in X \setminus \{x^*\}} \sum_{i: x_i^* = 0} \alpha x_i \Delta_i = \sum_{i: x_i^* = 0} \alpha x_i \Delta_i.
$$

Therefore,

$$
C_{sto} = \max_{\alpha \in [0, \infty)^{|X|}} \sum_{i: x_i^* = 0} \sqrt{\alpha_i} - r(\alpha)

\leq \max_{\pi \in [0, \infty]^d} \sum_{i: x_i^* = 0} \left( \sqrt{\alpha_i} - \bar{\alpha}_i \Delta_i \right)

\leq \max_{\pi \in [0, \infty]^d} \sum_{i: x_i^* = 0} \left( \bar{\alpha}_i \Delta_i + \frac{1}{4\Delta_i} - \bar{\alpha}_i \Delta_i \right) = \sum_{i: x_i^* = 0} \frac{1}{4\Delta_i}.
$$

Bounding $C_{add}$: Similar to the “Bounding $C_{add}$” part in the proof of Theorem 1 (earlier in Appendix A), we can bound for any $\alpha \in \Delta(X)$:

$$
g(\bar{\pi}) = \sum_{i: x_i^* = 1} \left( \gamma^{-1} + \gamma \log \left( \frac{1}{1 - \bar{\alpha}_i} \right) \right) (1 - \bar{\alpha}_i)

\leq \left( \gamma^{-1} + \gamma \log \left( \frac{m}{\sum_{i: x_i^* = 1} (1 - \bar{\alpha}_i)} \right) \right) \sum_{i: x_i^* = 1} (1 - \bar{\alpha}_i) \quad \text{(by the concavity of $g$)}

= \left( \gamma^{-1} + \gamma \log \left( \frac{m}{\sum_{i: x_i^* = 0} \bar{\alpha}_i} \right) \right) \sum_{i: x_i^* = 0} \bar{\alpha}_i

\leq \sum_{i: x_i^* = 0} \left( \gamma^{-1} + \gamma \log \left( \frac{m}{\bar{\alpha}_i} \right) \right) \bar{\alpha}_i.
$$

where in the second equality we use an property of $m$-set: $\sum_{i: x_i^* = 1} (1 - \bar{\alpha}_i) = \sum_{i: x_i^* = 0} \bar{\alpha}_i$, which follows from the fact that $\bar{\pi}$ is in the convex hull of $m$-set. In the last inequality, we simply lower bound $\sum_{i: x_i^* = 0} \bar{\alpha}_i$ by one of its summands.

Using the same lower bound

$$
r(\alpha) \geq \sum_{i: x_i^* = 0} \Delta_i \bar{\alpha}_i, \quad \text{(by Eq. (19))}
$$

we have an upper bound for $C_{add}$:

$$
C_{add} = \sum_{t=1}^{\infty} \alpha \in \Delta(X) \frac{100}{\sqrt{t}} g(\bar{\pi}) - r(\alpha)

\leq \sum_{i: x_i^* = 0} \sum_{t=1}^{\infty} \alpha \in [0, 1] \frac{100}{\sqrt{t}} \left( \gamma^{-1} + \gamma \log \left( \frac{m}{\bar{\alpha}_i} \right) \right) \bar{\alpha}_i - \Delta_i \bar{\alpha}_i

\leq \sum_{i: x_i^* = 0} \left( \sum_{t=1}^{\infty} \alpha \in [0, 1] \left( \frac{100}{\sqrt{t}} \left( \gamma^{-1} + \gamma \log m \right) \bar{\alpha}_i - \Delta_i \bar{\alpha}_i \right) + \sum_{t=1}^{\infty} \alpha \in [0, 1] \left( \frac{100}{\sqrt{t}} \gamma \bar{\alpha}_i \log \left( \frac{1}{\bar{\alpha}_i} - \frac{\Delta_i}{2\bar{\alpha}_i} \right) \right) \right).
$$
Invoking Lemma 6 and 7 on the above two terms, we get

$$C_{add} \leq \mathcal{O} \left( \sum_{i: x_i^* = 0} \frac{(\gamma^{-1} + \gamma \log m)^2}{\Delta_i} \right).$$

This can be further upper bounded by \( \mathcal{O} \left( \sum_{i: x_i^* = 0} \frac{(\log d)^2}{\Delta_i} \right) \) by our selection of \( \gamma \) in either regime.