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 $\mathcal{O}(\log T)$ regret for stochastic environments and $\mathcal{O}(\sqrt{T})$ regret for adversarial environments without prior knowledge of the regime or the number of rounds T. The leading problem-dependent constants of our bounds are not only optimal in a certain 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 bandits, but importantly requires a novel hybrid regularizer designed specifically for semi-bandit. Experimental results on synthetic data show that our algorithm indeed performs well over different environments. Finally, we provide a preliminary extension of our results to the full bandit feedback.

1. Introduction

The 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 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 where each arm's loss is drawn independently from a fixed distribution, the minimax optimal regret is of order $\mathcal{O}(\log T)$ where T is the number of rounds (dependence on all other parameters is omitted) (Lai & Robbins, 1985), while in the adversarial environment where each arm's loss can be completely arbitrary, the minimax optimal regret is of order $\mathcal{O}(\sqrt{T})$ (Auer et al., 2002).

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 bandits and propose adaptive algorithms that achieve $\mathcal{O}(\log T)$ regret in stochastic environments while simultaneously ensuring worst-case robustness, that is, $\mathcal{O}(\sqrt{T})$ regret even for adversarial environments. Importantly, this is achieved without any prior knowledge of the nature of the environment.

In this work, we extend such best-of-both-worlds results to the combinatorial bandit problem, a generalization of multi-armed bandits, 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 $\mathcal{O}(C_{sto} \log T)$ regret for stochastic environments and $\mathcal{O}(C_{adv}\sqrt{T})$ regret for adversarial environments, where C_{sto} and C_{adv} are problem-dependent factors (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 sets: the set of all subsets of arms and the set of all subsets with a fixed size m (so called m-set), we further derive refined bounds for the problem-dependent constants C_{sto} and C_{adv}, which are optimal for each of these special cases. As a side result, our bounds imply that for the 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).

[1]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark [2]Department 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>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

4. We conduct experiments with synthetic data to show that our algorithm indeed adapts well to the nature of 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 of the more challenging bandit feedback (Section 6).

Our techniques are close to those 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 a certain *self-bounding* property under the stochastic environment which eventually leads to 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 of the optimal subset, the problem of identifying their suboptimality is analogous to the multi-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 classical Shannon entropy as the regularizer to these arms.

1.1. Related work

Semi-bandits. The combinatorial semi-bandit problem is a natural generalization of multi-armed bandits and captures many real-life applications. There are many algorithms for stochastic semi-bandits based on the well-known optimistic principle (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 a recent work developed a general instance-optimal algorithm for any structured stochastic bandits (including semi-bandit as a special case (Combes et al., 2017)). Specifically, they obtain the optimal regret $\mathcal{O}(C \log T)$ where C is an instance-dependent term expressed as the solution of a certain optimization problem. The constant C_{sto} in our stochastic bound $\mathcal{O}(C_{sto} \log T)$ 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 discuss in Section 3. Two advantages of our algorithm compared to prior work 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.

Algorithms with $\mathcal{O}(\sqrt{T})$ 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). 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 $\mathcal{O}(C_{adv}\sqrt{T})$ is instance-dependent through the term C_{adv}, again expressed as the solution of a certain 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 in 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 bandits (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 (the expert problem) (Gaillard et al., 2014; Luo & Schapire, 2015; Koolen et al., 2016) and intermediate version (Thune & Seldin, 2018). Notably, among these works the recent two (Wei & Luo, 2018; Zimmert & Seldin, 2019) discovered that sophisticated hypothesis testing or gap estimations used in earlier works are in fact not needed for such adaptivity. Instead, their algorithms are 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 by Bubeck et al. (2018) for sparse bandit and bandit with a specific form of adaptive regret bound, and also recently used by Luo et al. (2018) for the online portfolio selection problem. The form of the hybrid regularizers and the way they are used in the analysis, however, are different both among these two prior works and with 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 $\mathcal{X} \subset \{0,1\}^d$. At any time $t = 1, 2, \ldots$, the learner chooses an action $X_t \in \mathcal{X}$ and at the same time the environment chooses

[1] In some works a combinatorial action is also referred to 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.

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, the learner only observes the loss of each arm in the selected subset (the 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 $x \in \mathcal{X} \backslash \{x^*\}$ 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 from a fixed unknown distribution, and is much more general since neither independence nor identical distributions are required. In the adversarial case, on the other hand, ℓ_t is chosen in an arbitrary way based on the history $\ell_1, X_1, \ldots, \ell_{t-1}, X_{t-1}$ and possibly an internal randomization by the environment.

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

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

where $x^* = \mathrm{argmin}_{x \in \mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \langle x, \ell_t \rangle\right]$ 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 $\mathbb{E}_t[\cdot]$ 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 $\mathbb{I}_t(i)$ for the indicator function $\mathbb{I}\{X_{ti} = 1\}$ (X_{ti} is the i-th component of the vector $X_t \in \mathcal{X} \subset \{0,1\}^d$) and write the characteristic function of a set A as $\mathcal{I}_A(x)$ which is 0 if $x \in A$ and $+\infty$ otherwise. We denote the d-dimensional vector with all 1s as $\mathbf{1}_d$.

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 $x_t = \mathrm{argmin}_{x \in \mathrm{Conv}(\mathcal{X})} \langle x, \hat{L}_{t-1} \rangle + \eta_t^{-1} \Psi(x)$, where $\mathrm{Conv}(\mathcal{X})$ 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 a learning rate, and $\Psi(x) : \mathrm{Conv}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ is a regularizer. Then the algorithm samples $X_t \sim P(x_t)$ for a sampling rule P that provides a distribution over $\mathcal{X}$ satisfying $\mathbb{E}_{X \sim P(x)}[X] = x$. As long as $\mathrm{Conv}(\mathcal{X})$ can be described by a polynomial number of constraints, one can always find an efficient sampling rule P (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.

[2] For linear objectives and Legendre regularizers, FTRL is equivalent to Online Mirror Descent as defined in (Orabona et al., 2015). The same framework is also known under the names OMD, OSMD, or INF.

Algorithm 1 FTRL with hybrid regularizer for semi-bandits

Input: $0 < \gamma \leq 1$, sampling scheme P
Initialize: $\hat{L}_0 = (0, \ldots, 0)$, $\eta_t = 1/\sqrt{t}$
for $t = 1, 2, \ldots$ **do**
  compute

$$x_t = \mathrm{argmin}_{x \in \mathrm{Conv}(\mathcal{X})} \left\langle x, \hat{L}_{t-1} \right\rangle + \eta_t^{-1} \Psi(x)$$

  where $\Psi(\cdot)$ is defined in Eq. (1)
  sample $X_t \sim P(x_t)$
  observe $o_t = X_t \circ \ell_t$
  construct estimator $\hat{\ell}_t$, $\forall i: \ \hat{\ell}_{ti} = \frac{(o_{ti}+1)\mathbb{I}_t(i)}{x_{ti}} - 1$
  update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$
end for

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

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

with a parameter $0 < \gamma \leq 1$ to be chosen later based on the action set $\mathcal{X}$ (in most cases we use $\gamma = 1$). This is a combination of the Tsallis entropy (with power $1/2$) $\sum_i -\sqrt{x_i}$, and the Shannon entropy $\sum_i (1-x_i)\log(1-x_i)$ on the *complement* of x. The $\sum_i -\sqrt{x_i}$ regularizer was first implicitly introduced by Audibert & Bubeck (2009), and later discovered as a member of the Tsallis entropy regularizers by Abernethy et al. (2015). It was also recently shown to be optimal for both stochastic and adversarial multi-armed bandits (Zimmert & Seldin, 2019).

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 $\hat{\ell}_t$ are defined as $\hat{\ell}_{ti} = \frac{(o_{ti}+1)\mathbb{I}_t(i)}{x_{ti}} - 1$ for all i. It is clear the estimators are unbiased, $\mathbb{E}_t[\hat{\ell}_t] = \ell_t$, just as common importance weighted estimators. The shift by 1 is used to ensure that the range of the loss estimates is bounded from one side, $\hat{\ell}_{t,i} \geq -1$. See Algorithm 1 for a complete pseudocode.

Intuition behind the new regularizer. It is known that the classical 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 $\mathbf{1}_d - \mathcal{X}$ 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.

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. Effectively, this allows us to regularize arms in the optimal combinatorial set differently than arms outside the optimal set, without the need to know which arms are in the optimal set.

3. Main Results

In this section we present general regret guarantees for our algorithm, followed by concrete instantiations in two special cases.

3.1. Arbitrary action set

To state the general regret bound for 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 instantaneous 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 Δ_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 $\gamma \le 1$ the pseudo regret of Algorithm 1 is upper bounded by*

$$\overline{\mathrm{Reg}}_T \le \mathcal{O}\left(C_{sto} \log T\right) + \mathcal{O}\left(C_{add}\right)$$

in the stochastic case and

$$\overline{\mathrm{Reg}}_T \le \mathcal{O}\left(C_{adv}\sqrt{T}\right)$$

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

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

Moreover, it always holds that $C_{sto} = \mathcal{O}\left(\frac{md}{\Delta_{\min}}\right)$, $C_{add} = \mathcal{O}\left(\frac{m^2}{\gamma^2 \Delta_{\min}}\right)$, 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 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 the results.

The explicit 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 the minimax regret over all environments with the same value m of $\max_{x \in \mathcal{X}} ||x||_1$ and also the same value $\Delta_{\min}$ of $\min_{x \in \mathcal{X} \setminus \{x^*\}} \Delta_x$ in the stochastic case.

However, for explicit instances, one can hope to achieve even better bounds. By exploiting the structure of the problem and providing better bounds on the constants C_{sto}, C_{add} and C_{adv}, we show in the next two sections that our algorithm is optimal in two special cases. For better interpretability, in the stochastic case we consider the more traditional setting where $\ell_1, \ldots, \ell_T$ are i.i.d. samples from an unknown distribution $\mathcal{D}$. It is clear that we can define $\Delta_x = \mathbb{E}_{\ell \sim \mathcal{D}}[\langle x - x^*, \ell \rangle]$ 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 $\mathrm{Conv}(\mathcal{X}) = [0,1]^d$ and a trivial sampling rule is $P(x) = \bigotimes_{i=1}^d \mathrm{Ber}(x_i)$ where $\mathrm{Ber}(\cdot)$ 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 for our algorithm, where in the stochastic case with a slight abuse of notation we define $\Delta_i = \mathbb{E}_{\ell\sim\mathcal{D}}[\ell_i]$.

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

$$\overline{\mathrm{Reg}}_T \le \mathcal{O}\left(\sum_{\Delta_i>0}\frac{\log(T)}{\Delta_i}\right) + \mathcal{O}\left(\sum_{\Delta_i<0}\frac{1}{|\Delta_i|}\right)$$

in the stochastic case and

$$\overline{\mathrm{Reg}}_T \le \mathcal{O}\left(d\sqrt{T}\right)$$

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 regret along 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 $C_{add} = \sum_t \max_{\alpha\in[0,1]} -\alpha\Delta_i = 0$. For C_{sto} we apply the general bound from Theorem 1 and obtain $C_{sto} = \mathcal{O}(1/\Delta_i)$ (since $m = d = 1$ and $\Delta_{\min} = \Delta_i$). This gives the bound $\mathcal{O}\left(\frac{\log(T)}{\Delta_i}\right)$ for $\Delta_i > 0$.

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

In the adversarial case, we apply the general bound of Theorem 1 and obtain $C_{adv} = \mathcal{O}(1)$. 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. □

3.3. Special case: m-set

Another common instance of semi-bandit is when the learner can only select subsets of a fixed size. Specifically, let $m \in \{1,\dots,d-1\}$ be a fixed parameter and define the m-set as

$$\mathcal{X} = \left\{ x \in \{0,1\}^d \;\middle|\; \sum_{i=1}^d x_i = m \right\}. \tag{2}$$

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 $\mathrm{Conv}(\mathcal{X}) = \left\{x \in [0,1]^d \mid \sum_{i=1}^d x_i = m\right\}$, and in the appendix we provide a simple sampling rule P with $\mathcal{O}(d\log(d))$ 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 $\Delta_i = \mathbb{E}_{\ell\sim\mathcal{D}}[\ell_i - \ell_m]$ for all i. Note that the uniqueness of x^* also implies $\Delta_i \neq 0$ for all $i > m$. The next theorem shows that our algorithm is 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. To the best of our knowledge, this was previously unknown.

Theorem 3. *If $\mathcal{X}$ is the m-set defined by Eq. (2), then the pseudo-regret of Algorithm 1 with*

$$\gamma = \begin{cases} 1 & \text{if } m \le d/2 \\ \min\{1, 1/\sqrt{\log(d/(d-m))}\} & \text{otherwise,} \end{cases}$$

satisfies

$$\overline{\mathrm{Reg}}_T \le \mathcal{O}\left(\sum_{i=m+1}^d \frac{\log(T)}{\Delta_i}\right) + \mathcal{O}\left(\sum_{i=m+1}^d \frac{(\log d)^2}{\Delta_i}\right)$$

in the stochastic case and

$$\overline{\mathrm{Reg}}_T \le \begin{cases} \mathcal{O}\left(\sqrt{mdT}\right) & \text{if } m \le d/2 \\ \mathcal{O}\left((d-m)\sqrt{\log(\frac{d}{d-m})T}\right) & \text{otherwise} \end{cases}$$

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

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

$\mathbf{C_{adv}}$: 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

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

for some $\lambda \in [0, \min\{1, \frac{m}{d-m}\}]$. In Appendix B we show that the function is increasing in λ, and that inserting $\lambda = \min\{1, \frac{m}{d-m}\}$ leads to the stated adversarial bound.

$\mathbf{C_{sto}}$: 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 $\sum_{i:x_i^*=0, x_i=1} \Delta_i = \sum_{i:x_i^*=0} \Delta_i x_i$. So the immediate regret function $r(\alpha)$ can be bounded as

$$\begin{aligned} r(\alpha) &= \sum_{x\neq x^*} \Delta_x \alpha_x \ge \sum_{x\neq x^*} \sum_{i:x_i^*=0} \Delta_i \alpha_x x_i \\ &= \sum_{i:x_i^*=0} \Delta_i \left(\sum_{x\neq x^*} \alpha_x x_i\right) = \sum_{i:x_i^*=0} \Delta_i \overline{\alpha}_i. \end{aligned}$$

The optimization problem can now be bounded as

$$C_{sto} = \max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} \sum_{i:x_i^*=0} \sqrt{\overline{\alpha}_i} - \sum_{x\neq x^*} \alpha_x \Delta_x$$
$$\leq \max_{\overline{\alpha}\in[0,\infty)^d} \sum_{i:x_i^*=0} \left(\sqrt{\overline{\alpha}_i} - \Delta_i \overline{\alpha}_i\right) = \sum_{i:x_i^*=0} \frac{1}{4\Delta_i},$$

which is the same as $\sum_{i=m+1}^{d} \frac{1}{4\Delta_i}$.

$\mathbf{C_{add}}$: We bound the function g as follows:

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

where the first inequality is by the concavity of g; the second equality is by the fact $\sum_{i:x_i^*=1} 1-\overline{\alpha}_i = \sum_{i:x_i^*=0} \overline{\alpha}_i$ since $\overline{\alpha}$ is in the convex hull of m-set.

Recall the lower bound $r(\alpha) \geq \sum_{i:x_i^*=0} \Delta_i \overline{\alpha}_i$ as derived previously. We can thus bound C_{add} as

$$\sum_{i:x_i^*=0} \sum_{t=1}^{\infty} \max_{A\in[0,1]} \frac{100}{\sqrt{t}} \left(\gamma^{-1} - \gamma \log\left(\frac{A}{m}\right)\right) A - \Delta_i A$$

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

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 $\Omega(m\sqrt{T\log(d/m)})$ 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 $\mathbf{1}_d - \mathcal{X}$ and reversed losses. This is exactly a $(d-m)$-set problem and thus a lower bound $\Omega((d-m)\sqrt{T\log(d/(d-m))})$ 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 η_t 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: $d = 10$, $m = 5$, $T = 10^7$. Below, we specify the mean of each arm's loss at each time. With mean μ_{ti} the actual loss of arm i at time t will be -1 with probability $(1-\mu_{ti})/2$ and $+1$ with probability $(1+\mu_{ti})/2$, independent of everything else. We create the following two environments:

Stochastic environment. In this case the losses are drawn from a fixed distribution with $\mu_{ti} = -\Delta$ if $i \leq 5$ and $\mu_{ti} = \Delta$ 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 time is split into phases

$$\underbrace{1,\ldots,t_1}_{T_1}, \underbrace{t_1+1,\ldots,t_2}_{T_2}, \ldots, \underbrace{t_{n-1},\ldots,T}_{T_n}.$$

The length of phase s is $T_s = 1.6^s$, and the means of the losses are set to

$$\mu_{ti} = \begin{cases} -\Delta/2 \pm (1-\Delta/2) & \text{if } i \leq 5, \\ +\Delta/2 \pm (1-\Delta/2) & \text{otherwise,} \end{cases}$$

where $\pm$ represents $+$ if t 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

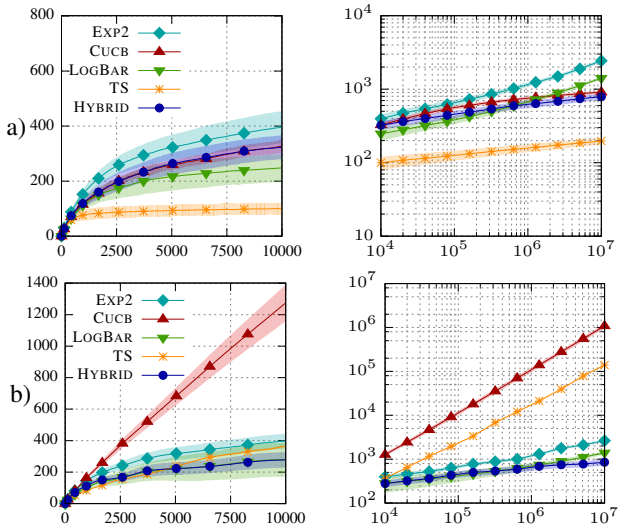

Figure 1. Comparisons of our new algorithm (HYBRID) and several existing algorithms with $d = 10, m = 5$ and $T = 10^7$ under a) stochastic and b) stochastically constrained adversarial setting. The left side is in linear scale and the right is in log-log scale.

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 SAMPLING has the best performance (a well-known phenomenon). However, these two stochastic algorithms clearly fail in the adversarial case and exhibit nearly-linear regret.

5. Proof of Theorem 1

We provide the key steps of the proof for our general result (Theorem 1) in this section. Define $\Psi_t(\cdot) = \eta_t^{-1}\Psi(\cdot)$ and potential function $\Phi_t(\cdot) = \max_{x\in\mathrm{Conv}(\mathcal{X})} \langle x, \cdot\rangle - \Psi_t(x)$, which is the convex conjugate of $\Psi_t + \mathcal{I}_{\mathrm{Conv}(\mathcal{X})}$.

Following a standard analysis of FTRL, we decompose the regret

$$\overline{\mathrm{Reg}}_T = \underbrace{\mathbb{E}\left[\sum_{t=1}^T \langle X_t, \ell_t\rangle + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right]}_{\mathrm{Reg}_{\mathrm{stab}}}$$
$$+ \underbrace{\mathbb{E}\left[\sum_{t=1}^T -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \langle x^*, \ell_t\rangle\right]}_{\mathrm{Reg}_{\mathrm{pen}}}, \quad (3)$$

into terms corresponding to the *stability* and the *regularization penalty* of the algorithm.

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*

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

Lemma 2. *The stability term is bounded as*

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

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

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

$$\begin{aligned}\overline{\mathrm{Reg}}_T &\le \sum_{t=1}^T \frac{25}{\sqrt{t}}\left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) + c \quad (4)\\ &\le 50\sqrt{T}\max_{x\in\mathrm{Conv}(\mathcal{X})}\left(f(x)+g(x)\right) + c\\ &= \mathcal{O}\left(C_{adv}\sqrt{T}\right),\end{aligned}$$

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

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

Together with Eq. (4) we have

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

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

$$\sum_{t=1}^T \left(\frac{50}{\sqrt{t}}\left(f(\mathbb{E}[x_t]) + g(\mathbb{E}[x_t])\right) - r(P(\mathbb{E}[x_t]))\right) + 2c.$$

We next decompose the summation above into two terms and upper bound them as $C_{sto}\log T$ and C_{add} respectively:

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{50}{\sqrt{t}}f(\mathbb{E}[x_t]) - \frac{1}{2}r(P(\mathbb{E}[x_t])) \\
&\le \sum_{t=1}^{T}\max_{\alpha\in\Delta(\mathcal{X})}\frac{50}{\sqrt{t}}f(\overline{\alpha}) - \frac{1}{2}r(\alpha) \\
&\le \sum_{t=1}^{T}\max_{\alpha\in[0,\infty)^{|\mathcal{X}|}}\frac{50}{\sqrt{t}}f\left(\frac{10^4}{t}\overline{\alpha}\right) - \frac{1}{2}r\left(\frac{10^4}{t}\alpha\right) \\
&\overset{(\star)}{=} \sum_{t=1}^{T}\frac{10^4}{2t}\max_{\alpha\in[0,\infty)^{|\mathcal{X}|}} f(\overline{\alpha}) - r(\alpha) = \mathcal{O}\left(C_{sto}\log(T)\right)
\end{aligned}$$

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

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

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

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

6. Extensions to Bandit Feedback

The most natural extension of our work is to consider the full bandit feedback setting, where each time after playing an action X_t the learner only observes $\langle X_t, \ell_t\rangle$. 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 $\ell_t \sim \mathcal{D}$ and define $\Delta_i = \mathbb{E}_{\ell\sim\mathcal{D}}[\ell_i]$.

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 $\hat{\ell}_{ti} = \frac{\langle X_t,\ell_t\rangle X_{ti}}{x_{ti}} - \frac{\langle X_t,\ell_t\rangle(1-X_{ti})}{1-x_{ti}}$ ensures:*

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

in the stochastic case and

$$\overline{\mathrm{Reg}}_T \le \mathcal{O}\left(d\sqrt{T}\right)$$

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

Proof sketch. In this case, the optimization of FTRL decomposes over the coordinates 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 $\ell_{ti} + \sum_{j\neq i} X_{tj}\ell_{tj}$ and the loss of arm 2 is $\sum_{j\neq i} X_{tj}\ell_{tj}$.[3] In the stochastic case this exactly fits into the stochastically constrained adversarial setting of Zimmert & Seldin (2019) with gap $|\Delta_i|$ and, therefore, applying their Theorem 2 and summing up the regret over each coordinate finishes 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 $\Omega(d\sqrt{T})$ regret 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 in the stochastic case. We leave this as question for future work.

7. Conclusions

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 in two particular instances of the problem. Empirical evaluations also confirm our theory.

Other than the open problem under 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, next to nothing is known regarding the instance-optimality of the adversarial case, let alone best-of-both-worlds results.

Acknowledgments HL and CYW are supported by NSF Grant #1755781. We thank Yevgeny Seldin for valuable feedback and discussions.

References

Abernethy, J. D., Lee, C., and Tewari, A. Fighting bandits with a new kind of smoothness. In *Advances in Neural*

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

Information Processing Systems, 2015.

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. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, 2009.

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

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.

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

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., and Kivinen, J. 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.

Orabona, F., Crammer, K., and Cesa-Bianchi, N. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3), 2015.

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.

Thune, T. and Seldin, Y. Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems*, pp. 2909–2918, 2018.

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.