Bandits with Adversarial Scaling

Thodoris Lykouris [1]   **Vahab Mirrokni** [2]   **Renato Paes Leme** [2]

Abstract

We study *adversarial scaling*, a multi-armed bandit model where rewards have a stochastic and an adversarial component. Our model captures display advertising where the *click-through-rate* can be decomposed to a (fixed across time) arm-quality component and a non-stochastic user-relevance component (fixed across arms). Despite the relative stochasticity of our model, we demonstrate two settings where most bandit algorithms suffer. On the positive side, we show that two algorithms, one from the action elimination and one from the mirror descent family are adaptive enough to be robust to adversarial scaling. Our results shed light on the robustness of adaptive parameter selection in stochastic bandits, which may be of independent interest.

1. Introduction

The multi-armed bandit setting is the cleanest paradigm to capture the tension between exploring information about the underlying system and exploiting the most profitable actions based on the current information. A decision-maker (or *learner*) repeatedly selects among a set of k actions also referred to as arms, earns the reward of the selected arm, and obtains feedback only about it. This creates a direct trade-off between learning about the performance of underexplored actions and earning the reward from those that seem the most profitable. This trade-off is prominent in applications such as display advertising where a platform needs to repeatedly select which ad to show in response to a particular pageview. The arms therefore correspond to the competing ads and, upon being shown, each ad can result in a click or not. We use this as a running example and assume that the goal of the learner is to maximize the total number of clicks (in practice, there are other goals such as revenue and user-experience, but we ignore those for simplicity).

---
[*]Equal contribution [1]Microsoft Research, New York City, NY, USA [2]Google Research, New York, NY, USA. Correspondence to: Thodoris Lykouris <thlykour@microsoft.com>.

Modeling this problem, one soon realizes that the two classical multi-armed bandit approaches fail to capture the essence of the setting. The two main approaches in the multi-armed bandit learning literature assume that the rewards obtained from each pageview are either completely adversarial or coming from identical and independent distributions (i.i.d.). However, clearly some ads have better quality and are consistently more clickable than others, so assuming a fully adversarial model seems to ignore a lot of useful structure in the data. On the other extreme, stochastic models assume that each ad has a fixed probability of being clicked whenever displayed (this is typically referred to as *click-through-rate*). In practice, click-through-rates are known to vary due to various factors: time of the day, day of the week, seasonalities (e.g. users tend to click more before Christmas and on Black Friday) but those factors affect ads uniformly. Tracà & Rudin (2015) propose a model where the mean $\mu^t(a)$ reward of an arm a at time t is a product

$$\mu^t(a) = q^t \cdot \theta(a)$$

of the intrinsic quality of the arm and a seasonality term q^t. In the display ads example, q^t is the clickiness of the user behind her page-view and $\theta(a)$ is the intrinsic quality of the ad. The motivation of Tracà & Rudin (2015) comes from retail: $\theta(a)$ is the effect certain action has in the store (the product price, which items are on sale...) and q^t is the number of customers in the store in that day. In their model, the seasonality effect q^t is known to the algorithm.

We depart from their model by assuming we do not have access to q^t. In display ads, the clickiness of users can be a function of a large number of covariates and it affects users in complex ways: certain users click more during the evening while other users click more during the day. Clickness of users in China is affected by Chinese New Year while clickiness of users in the US is affected by Thanksgiving. Estimating all those patterns is costly so it is useful to have models that exploit the structure of $\mu^t(a)$ without having to estimate the seasonality/clickness model directly.

To obtain more robust algorithms we assume q^t is adversarially chosen and hence the name *adversarial scaling*. If we think of q^t as a seasonality effect, there is not an adversary per se in the motivation, but rather a very complex pattern which we choose not to assume anything about to get robust algorithms. However, q^t can also model some actually

adversarial effects. For example, an advertiser may create a botnet that does not click anything in the initial time, i.e. $q^t = 0$, aiming to prolong the algorithm's exploration and hence her number of displays. This attack is very simple to run as it does not actively do anything and causes a *cold start* to the algorithm as the latter needs to appropriately disregard these insignificant samples.

As we demonstrate in this paper, despite the inherent stochasticity in these settings (best arm is better at each single round), we show that most stochastic algorithms perform poorly. This naturally generates the following question:

> *What makes bandit algorithms robust*
> *to adversarial scaling attacks?*

Our contribution. Tackling this question, we demonstrate that adaptive parameter selection is essential for robustness to adversarial scaling. In particular, we show a modification of the classical Action Elimination algorithm (Even-Dar et al., 2006) which we term *AAEAS* that uses the reward of the algorithm as a proxy for the number of rounds that matter, i.e. the total scaling $\sum_t q^t$. This allows it to adapt to the intrinsic qualities $\theta(a)$ without being misled by the adversarial scaling in a way similar to the *self-confident learning* technique developed for small-loss bounds in adversarial bandits. The resulting guarantees hold with high probability which is an advantage of this method.

Aiming to understand whether this robustness is satisfied more broadly, we then focus on the *Online Mirror Descent with Log-Barrier* algorithm of Foster et al. (2016). Interestingly, when combined with the neat doubling trick of Wei & Luo (2018), this algorithm (then termed *BROAD*) seamlessly adapts to the total scaling. Its resulting guarantees hold only for the weaker pseudo-regret notion but are stronger by a factor of k and the empirical performance is also enhanced.

To complement our study, we show two very simple attacks manage to make, to the best of our knowledge, all other stochastic algorithms perform ineffectively (detailed comparison in Section 4). Surprisingly these attacks work even against the recent breakthrough, Online Mirror Descent with Tsallis entropy (Zimmert & Seldin, 2019). The latter has optimal regret guarantee for both stochastic, adversarial setting as well as regimes in between and is therefore considered as the best algorithm for such intermediate settings.

The first attack is a purely stochastic setting where all the arms have fixed but really small means $\mu(a)$ where $\mu(a^\star)$ is the mean of the best action. This can be viewed as adversarial scaling with $\theta(a) = \frac{\mu(a)}{\mu(a^\star)}$ and $q^t = \mu(a^\star)$. The bounds for these algorithms scales inversely to the absolute difference $\mu(a^\star) - \mu(a)$. In contrast, AAEAS and BROAD have boundes based on the normalized difference $1 - \frac{\mu(a)}{\mu(a^\star)}$ which provides great improvement when the means are small as

typical in click-through-rates for display advertising.

The second attack is a *cold start* attack where in some initial period, the $q^t = 0$; this can arise because either maliciously or organically as we discuss in the end of Section 4.2. The number of rounds that have passed becomes a rather irrelevant quantity in this case as what really matters is the $\sum_t q^t$. However, all algorithms other than AAEAS and BROAD use this number of rounds (many of them in ways that seem not fixable). The cold start therefore leads them to a really bad prior state and they take a very long time to recover. Surprisingly, the most effective stochastic algorithm Thompson Sampling (Agrawal & Goyal, 2017) is prone to this attack even when the cold start is really short (see Figure 4).

Related work. Our paper lies in a broader line of work that tries to achieve enhanced guarantees when data exhibit a particular benign structure. The main structure that has been utilized is the data being i.i.d. across time where many algorithms achieve logarithmic guarantees at the existence of a large gap in $\mu(a^\star) - \mu(a)$.

Under this structure, three previous lines of work study robustness of bandit learning to adversarial components, orthogonal to adversarial scaling. We briefly discuss them below and more elaborately review them in Section 4.

1. Works on *best of both worlds* focus on the design of algorithms that achieve improved guarantees when the input is stochastic while retaining the regret guarantees of adversarial algorithms when this is not the case. This was introduced by Bubeck & Slivkins (2012) and was further studied in a sequence of works (Seldin & Slivkins, 2014; Auer & Chiang, 2016; Seldin & Lugosi, 2017; Wei & Luo, 2018; Zimmert & Seldin, 2019; Zimmert et al., 2019). Interestingly, our work shows that the BROAD algorithm (Wei & Luo, 2018) is robust to adversarial scaling while Online Mirror Descent with Tsallis entropy (Zimmert & Seldin, 2019) that is considered superior fails at the presence of adversarial scaling. Our first algorithm can also be transformed to achieve such a best of both worlds guarantee if combined with the techniques of Bubeck & Slivkins (2012).

2. Works on *bandits with adversarial corruptions* allow an adversary to corrupt some rounds. The adversary knows the distribution of the algorithm but *not* the randomness in the arm selection (at the extreme, this can capture the adversarial bandit problem). The goal is to a) achieve the improved guarantee when there is no corruption, b) have the performance gracefully degrade with the amount of corruption, while c) being agnostic to the amount of corruption. This was initially proposed by Lykouris et al. (2018a) who provided regret high-probability guarantees based on a multi-layering

extension of active arm elimination. The guarantees were further improved by Gupta et al. (2019) while further improvements were achieved for the weaker pseudo-regret notion (Zimmert & Seldin, 2019). Our first algorithm can be made robust to adversarial corruptions if combined with the techniques of Lykouris et al. (2018a) while the second algorithm is also robust to adversarial corruptions for the notion of pseudo-regret.

3. Last, a line of work aims to design attacks to stochastic algorithms when the adversary knows the random arm selection of the algorithm at every round and not only the arm distribution (Jun et al., 2018; Liu & Shroff, 2019). These works aim to minimize the amount of manipulation in the rewards needed for different objectives of the adversary and only provide attacks (instead of robust algorithms). Our algorithms obtain linear regret against such attacks (the same happens with all no-regret algorithms as these attacks give more power to the adversary than the adversarial bandit model).

Beyond stochasticity, many works exploit other benign properties in the data to enhance adversarial guarantees. Example properties include small variance of the losses (Hazan & Kale, 2009; Wei & Luo, 2018), small effective loss range (Cesa-Bianchi & Shamir, 2018), small variation in the losses across rounds (Bubeck et al., 2019), small loss of the best arm (Allenberg et al., 2006; Neu, 2015; Foster et al., 2016; Lykouris et al., 2018b; Allen-Zhu et al., 2018), second-order excess loss (Wei & Luo, 2018), locally perturbed adversarial inputs (Shamir & Szlak, 2017) among others.

Finally, our model can be cast as a rank-1 assumption on the structure of the reward function. Rank-1 assumptions have been explored in various ways in the bandits literature, for example the rank-1 bandit model (Katariya et al., 2016) and factored bandits (Zimmert & Seldin, 2018) assume that the action space have a cartesian product and hence the reward in each period can be represented as a matrix over which the rank-1 assumption is made. In other words, this rank-1 structure is on the space of actions unlike our model which is in the space of actions cross time. Another common setting where rank-1 assumptions ara made is in semi-parametric bandits (Krishnamurthy et al., 2018) where the loss function is a product of unknown parameters and features. A difference with respect to our model is that the features are observed by the learner while in our model the quality parameter q^t is never observed by the learner.

2. Model

We study a multi-armed bandit setting with k arms where each arm $a \in [k]$ is associated with an *intrinsic mean* parameter $\theta(a) \in [0, 1]$ which is unknown to the learner. We define $\mathcal{F}^t(a)$ as the distribution of the reward of arm a in

round t and we assume that it has positive measure only on the interval $[0, 1]$. The distributions are adaptively selected by the adversary subject to the constraint that the means $\mu^t(a) = \mathbf{E}_{r \sim \mathcal{F}^t(a)}[r]$ must satisfy a rank-1 constraint, described below. Formally, the protocol between the learner and the adversary at each round $t = 1..T$ is as follows:

- The learner chooses a distribution p^t over the k arms.
- The adversary chooses an *adversarial quality* parameter $q^t \in [0, 1]$ and distributions $\mathcal{F}^t(a)$ supported in $[0, 1]$ with mean $\mu^t(a) = q^t \cdot \theta(a)$.
- Rewards $r^t(a) \sim \mathcal{F}^t(a)$ are drawn.
- Learner draws $a^t \sim p^t$ and observes $r^t(a^t)$.

For ease of presentation, we assume that the highest intrinsic mean is equal to 1, i.e. $\max_{a'} \theta(a') = 1$. This can be done by appropriately scaling down all the adversarial qualities replacing $\theta(a)$ by $\frac{\theta(a)}{\max_{a'} \theta(a')}$ and q^t by $q^t \cdot \max_{a'} \theta(a')$.

Regret. The goal of the learner is to maximize the aggregate reward she accumulates. To evaluate the performance of the learning algorithm, we compare this aggregate reward with the best strategy of the learner if she had access to the reward distributions of each arm – in that case, the optimal strategy is to select the arm with the highest intrinsic mean. The degradation that the algorithm incurs compared to this aware setting is captured by the notion of pseudo-regret in the purely stochastic setting. Even though we have an adversarial component on the rewards, the optimal arm to pull in each round is always an arm that maximizes $\theta(a)$ irrespectively of the scaling chosen by the adversarial. This allows us to define pseudo-regret in the following way:

$$\mathsf{Reg} = \max_{a \in [k]} \mathbf{E} \left[\sum_{t=1}^{T} r^t(a^\star) - r^t(a^t) \right]$$
$$= \sum_{t=1}^{T} \mu^t(a^\star) - \mu^t(a^t) = \sum_{t=1}^{T} q^t \cdot \Delta(a^t),$$

where $\Delta(a) = \theta(a^\star) - \theta(a) = 1 - \theta(a)$. In other words, the pseudo-regret is worst-case over the sequence $q(t)$ but it is in expectation over the draws $r^t(a) \sim \mathcal{F}^t(a)$.

3. Algorithms robust to adversarial scaling

In this section, we show that two algorithms, one that we introduce from the active arm elimination family (Section 3.1) and one existing from the mirror descent family (Section 3.2) achieve the desired robustness to adversarial scaling. A common property of both algorithms is their adaptivity on the number of rounds; they do not scale with the number of rounds which is easily targeted by adversarial scaling attacks. [1] In the next section, we show that other algorithms

---

[1] The need for such adaptivity was only known in adversarial settings prior to our work.

with improved guarantees in the stochastic regime that do not enjoy such adaptivity have their performance severely compromised at the presence of adversarial scaling.

3.1. Active Arm Elimination with Adversarial Scaling

The algorithm we introduce is based on Active Arm Elimination of Even-Dar et al. (2006), but is appropriately adapted to handle the adversarial scaling of the rewards.

Classical Active Arm Elimination keeps a set of active arms (initially all arms) and selects arms in a round-robin fashion among the active arms, updating the empirical mean of the selected arm as well as a confidence interval around it. The latter ensures that when samples are i.i.d. (q^t same across rounds), with high probability, the actual mean lies within the confidence interval (this comes from an application of a Chernoff bound). Once the confidence intervals cease overlapping, we are confident that the dominated arm is not the arm with the highest intrinsic quality, thus it is safe to eliminate it from the active arms and never select it again.

Trying to extend this approach to adversarial scaling, one encounters some difficulties. One issue is that classic Active Arm Elimination selects arms deterministically. The adversary can easily cause linear regret to any deterministic policy by setting $q^t = 0$ anytime the algorithm is about to pull the optimal arm. The natural way to get around this difficulty is to pull arms in the active set randomly. A more serious issue is that the number of samples (arm pulls) is no longer a meaningful quantity. The adversary may provide many samples initially with quality $q^t = 0$. These samples do not help in informing our estimates about where the intrinsic qualities $\{\theta(a)\}$ lie; therefore treating these as real samples can provide a misleading picture of the confidence intervals. Ideally, we would like to use as *effective samples* the total adversarial quality at rounds where we selected each arm a, i.e. $\sum_{t:a^t=a} q^t$. Since we do not have access to this quantity, we need to design our confidence intervals in a way that will be robust to this adversarial scaling.

The main idea behind robustifying our confidence intervals is to use the reward of our algorithm as a proxy of the total effective samples of each arm. At the rounds that we selected arm $a^\star$, the reward of the algorithm is, in expectation, equal to the effective samples of $a^\star$ (since $\theta(a^\star) = 1$). Since we select each non-eliminated arm with equal probability, the total reward of the algorithm S serves as proxy for the effective rounds of each arm ($\frac{S}{k}$ is a lower bound and S is an upper bound). Our algorithm, formalized in Algorithm 1, uses the proxy S instead of the unknown effective samples to construct confidence intervals for the arms; in fact, there is a single confidence interval $CB(S)$.

Algorithm 1 Active Arm Elimination with Adversarial Scaling (AAEAS)

Initialize the set of active arms $\mathcal{A} = [k]$, the aggregate reward for each arm $R(a) = 0$, and the total reward collected by the algorithm $S = 0$.

for t=1...T **do**

Select a^t randomly across the set of active arms $\mathcal{A}$ and earn reward $r^t(a^t)$

Update the total reward earned:

$$S \leftarrow S + r^t(a^t)$$

Update empirical reward of selected action:

$$R(a^t) \leftarrow R(a^t) + r^t(a^t).$$

Eliminate arms based on algorithm-induced confidence intervals, i.e. remove all a' from $\mathcal{A}$ if

$$R(a') + \mathbf{CB}(S) < \max_{a \in \mathcal{A}} R(a)$$

for confidence bound (setting $\delta' = \frac{(k+1)T}{\delta}$):

$$\mathbf{CB}(S) := 2\sqrt{\max\left(4S \log(2/\delta'), 16k \log^2(2/\delta')\right)}$$

end for

Theorem 3.1. *The AAEAS algorithm (Algorithm 1) run with $\delta = \frac{1}{T}$ has pseudo-regret at most:*

$$O\left(\sum_{a \neq a^\star} \frac{k \log(kT)}{\Delta(a)}\right)$$

The proof follows the standard active arm elimination analysis but replaces the samples of each arm by the proxy-created confidence interval. For completeness, we provide the Chernoff bound which we use in our analysis.

Lemma 3.2 (standard Chernoff bound). *Let $\{X_i\}$ be independent random variables in $[0, 1]$, $\mu = \mathbf{E}[\sum_i X_i]$, $\epsilon > 0$:*

$$\Pr\left[\sum_i X_i - \mu \geq \epsilon\mu\right] \leq 2e^{-\mu\epsilon^2/3}$$

Proof of Theorem 3.1. To prove the guarantee, we need to ensure two properties: i) the arm $a^\star$ with highest intrinsic quality, with high probability, never gets eliminated and ii) we can bound the regret incurred by each suboptimal arm.

Our analysis is based on a few events occurring and we bound the failure probability of these events. We separate the time horizon in phases where a phase ϕ begins when exactly ϕ arms have been eliminated; denote by τ_ϕ the round that the ϕ-th arm gets eliminated. First, the empirical reward

$R(a)$ that we experience for each arm is close to its expected reward $\bar{R}(a)$, that is, for each arm a and each round t, with probability $1 - \delta'$:

$$R(a) \in \left[\bar{R}(a) \pm \sqrt{2\bar{R}(a)\log(2/\delta')} \right] \qquad (1)$$

$$\text{where} \quad \bar{R}(a) = \sum_{\phi=0}^{k-1} \sum_{\tau=\tau_\phi+1}^{\min(t,\tau_{\phi+1})} \frac{\theta(a) \cdot q^\tau}{k - \phi}$$

This follows by a Chernoff bound (Lemma 3.2), since the rewards at round t are supported in $[0, 1]$ and the arms are sampled uniformly at random from the set of active arms.

In a similar fashion, we can provide a similar confidence bound for the reward of our algorithm. For any round t, with probability $1 - \delta'$:

$$S \in \left[\bar{S} \pm \sqrt{2\bar{S}\log(2/\delta')} \right] \quad \text{for} \quad \bar{S} = \sum_a \bar{R}(a) \qquad (2)$$

Setting all failure probabilities to $\delta' = \frac{\delta}{(k+1)T}$, the probability that any of them fails is at most δ. In the remainder of the proof, we assume that none of these bounds fails.

First property. First, we establish that, when these bounds do not fail, the arm $a^\star$ with the highest intrinsic mean does not become eliminated. Since the mean for any arm $a \neq a^\star$ is at most the mean for $a^\star$, equation (1) implies:

$$R(a) - R(a^\star) \leq 2\sqrt{2\bar{R}(a^\star)\log(2/\delta')}$$

We now show that, at every round, this difference is covered by the confidence bounds of our algorithm, i.e:

$$\mathrm{CB}(S) \geq 2\sqrt{2\bar{R}(a^\star) \cdot \log(2/\delta')} \qquad (3)$$

If $\bar{R}(a^\star) \leq 8k\log(2/\delta')$ then (3) holds as the second term of $\mathrm{CB}(S)$ is $2\sqrt{16k\log^2(2/\delta')} \geq 2\sqrt{2\bar{R}(a^\star) \cdot \log(2/\delta')}$.

Otherwise, (2) and the optimality of $a^\star$ imply:

$$S \geq \bar{S} - \sqrt{2\bar{S}\log(2/\delta')}$$
$$\geq \bar{R}(a^\star) - \sqrt{2k\bar{R}(a^\star)\log(2/\delta')} \geq \frac{\bar{R}(a^\star)}{2}$$

since the function $x \mapsto \frac{x}{2} - \frac{1}{2}\sqrt{x \cdot 8k\log(2/\delta')}$ is increasing for $x \geq 8k\log(2/\delta')$.

Hence, (3) holds as the first term of $\mathrm{CB}(S)$ is $2\sqrt{4S\log(2/\delta')} \geq 2\sqrt{2\bar{R}(a^\star) \cdot \log(2/\delta')}$. Since, in both cases, (3) holds, arm $a^\star$ does not become eliminated.

Second property. Now we bound the regret coming from each suboptimal arm. Let's consider the contribution to regret of arm $a \neq a^\star$ with gap $\Delta(a)$. First we bound the difference of empirical rewards of two arms. By (1) and the fact that $\theta(a^\star) = 1$ we have that:

$$R(a^\star) - R(a) \in \left[\Delta(a) \cdot \bar{R}(a^\star) \pm 2\sqrt{2\bar{R}(a^\star)\log(1/\delta')} \right]$$

We want to argue that once $\bar{R}(a^\star)$ is large enough, the difference between the two arms $R(a^\star) - R(a) > \mathrm{CB}(S)$, which leads to arm a getting eliminated and not contributing further regret. Let $T(a)$ denote the time of the elimination of arm a. Since arm a is selected with equal probability with other active arms until then, the total expected regret from a is at most:

$$\Delta(a) \cdot \sum_{\phi=0}^{k-1} \sum_{\tau_\phi+1}^{\min(\tau_{\phi+1},T(a))} \frac{q_t}{k - \phi} \leq \Delta(a) \cdot \bar{S}^{T(a)} \qquad (4)$$

where $\bar{S}^{T(a)}$ denotes the value of $\bar{S}$ at round $T(a)$; recall that the latter is defined in (2). We will argue that arm a is eliminated by the time we have:

$$\bar{S}^{T(a)} > \frac{128k\log(2/\delta')}{\Delta(a)^2}, \qquad (5)$$

In order for an arm to be eliminated the difference $R(a^\star) - R(a)$ needs to be relatively large. This difference can be expressed as:

$$R(a^\star) - R(a) \geq \bar{R}(a^\star) - \bar{R}(a) - 2\sqrt{2\big(\bar{R}(a^\star)\big)\log(2/\delta')}$$
$$= \Delta(a) \cdot \bar{R}(a^\star) - 2\sqrt{2\bar{R}(a^\star)\log(2/\delta')}$$
$$\geq \Delta(a) \cdot \frac{\bar{S}^{T(a)}}{k} - 2\sqrt{2\frac{\bar{S}^{T(a)}}{k}\log(2/\delta')}. \qquad (6)$$

The first inequality follows by applying (1) for both arm a and $a^\star$, and noting that $\bar{R}(a^\star) \geq \bar{R}(a)$. The second inequality is since $\theta(a^\star) - \theta(a) = \Delta(a)$. For the third inequality, first note that:

$$\bar{R}(a^\star) \geq \frac{\bar{S}^{T(a)}}{k}, \qquad (7)$$

by the definitions of $\bar{R}(a^\star)$ and $\bar{S}^{T(a)}$, and since $\theta(a^\star) = 1$ and $\theta(a) \leq 1$ for all $a \neq a^\star$. Combining (7) and (5), we therefore establish that $\bar{R}(a^\star) \geq \frac{\bar{S}^{T(a)}}{k} \geq \frac{2\log(2/\delta')}{\Delta(a)^2}$. The third inequality then follows directly since the function $x \mapsto \Delta(a)x - 2\sqrt{2x\log(2/\delta')}$ is increasing for $x \geq \frac{2\log(2/\delta')}{\Delta(a)^2}$.

We can re-arrange the terms in equation (5) and obtain:

$$\Delta(a)\frac{\bar{S}^{T(a)}}{2k} \geq 2\sqrt{2\frac{\bar{S}^{T(a)}}{k}\log(2/\delta')} \qquad (8)$$

Combining (6) and (8) with the fact that $\Delta(a) \leq 1$, it holds:

$$R(a^\star) - R(a) \geq \frac{\bar{S}^{T(a)}}{2k}\Delta(a) \geq \sqrt{32\bar{S}^{T(a)}\log(2/\delta')}.$$

The latter term is greater than the second term of the confidence bound $CB(S)$. Comparing to the first, denoting $S^{T(a)}$ the value of S at $T(a)$, it holds that $S^{T(a)} \leq 2\bar{S}^{T(a)}$ by (2) and (5). Hence, the latter RHS also dominates the first term of the confidence bound $CB(S)$. As a result, if arm a was not eliminated by then, it gets eliminated when the expected reward of the algorithm becomes $\bar{S}^{T(a)} = \frac{128k\log(2/\delta')}{\Delta(a)^2}$; by (4) this implies that the expected contribution of arm a to the regret is at most $\frac{128k\log(2/\delta')}{\Delta(a)}$. Summing across all suboptimal arms $a \neq a^\star$ and setting $\delta = 1/T$ completes the proof. $\square$

Remark 3.3. *While the proof is written in expectation over draws $r^t(a^t) \sim F_a^t$ all arguments are high-probability arguments. The bound therefore can be converted to a high probability regret bound (as usual in stochastic bandits) with the difference that instead of capping the performance of each arm by $\Delta(a)T$ as in the previous remark, we cap it by $\sqrt{T}$. Formally, we obtain that with probability $1 - \delta$ we obtain the following bound on actual regret:*

$$O\left(\sum_a \min\left(\sqrt{T}, \frac{k \cdot \log(kT/\delta)}{\Delta(a)}\right)\right)$$

3.2. Online Mirror Descent with Log-Barrier

The second algorithm that we show to be robust to adversarial scaling lies in the mirror descent family and has a stronger regularizer, log-barrier. It was initially suggested by Foster et al. (2016) for a fixed learning rate η, who proved that it attains first-order bounds for pseudo-regret. The update on the probabilities is the following (Algotithm 3 in Foster et al. (2016) adapted to rewards):

$$p^t(a^{t-1}) = \frac{p^{t-1}(a^{t-1})}{1 - \eta r^{t-1}(a^{t-1}) + \gamma p^{t-1}(a^{t-1})}$$

$$p^t(a) = \frac{p^{t-1}(a)}{1 + \gamma p^{t-1}(a)} \quad \forall a \neq a^{t-1} \qquad (9)$$

where $\gamma \geq 0$ is such that p is a valid probability distribution.

Via using a neat doubling trick to update the learning rate η, Wei & Luo (2018) showed that, in fact, this algorithm can also attain stochastic guarantees (Algorithm 3 in Wei & Luo (2018) for the particular doubling trick). In particular, the algorithm halves the learning rate and restarts once:

$$\sum_{\tau=t_R}^t \sum_{a=1}^k p^\tau(a)^2\big(\widehat{r}^\tau(a) - r^\tau(a^t)\big)^2 \geq \frac{k\ln T}{3\eta^2},$$

where t_R is the time of the last restart and $\widehat{r}^\tau(a) = r^\tau(a)/p^\tau(a) \cdot \mathbf{1}[a = a^t]$ is the importance sampling estimator. Their algorithm is a particular instantiation of a more general framework they termed *BROAD*.

We show that the BROAD algorithm (with no modification) is robust to adversarial scaling. The proof follows from replacing the potential the potential function in the proof of Theorem 10 in (Wei & Luo, 2018) to a potential that accommodates adversarial scaling. The remaining arguments are essentially the same as the ones used by (Wei & Luo, 2018) in their analysis of the stochastic setting.

Theorem 3.4. *Online Mirror Descent with Log-Barrier with the above doubling (also known as BROAD) has pseudo-regret at most $O\left(\frac{k\log T}{\Delta}\right)$ where Δ is the minimum non-zero gap on intrinsic means.*

We will use the following results proved in (Wei & Luo, 2018). The first lemma corresponds to equation (27) in their paper and the second to equation (29). They are restated here for rewards instead of losses.

Lemma 3.5 ((Wei & Luo, 2018))**.** *There is a constant C such that for $p^t(a)$, $r^t(a)$ and $\hat{r}^t(a)$ in the BROAD algorithm it holds that:*

$$\mathbf{E}\Big[\sum_{t=1}^T r^t(a^\star) - r^t(a)\Big] \leq Ck\ln T+$$

$$C \cdot \sqrt{(k\ln T)\mathbf{E}\Big[\sum_{t=1}^T \sum_{a=1}^k \big(p^t(a)^2\big(\widehat{r}^t(a) - r^t(a^t)\big)^2\big)\Big]}$$

Lemma 3.6 ((Wei & Luo, 2018))**.** *Again in the context of BROAD, it holds that:*

$$\mathbf{E}_{a^t \in p^t}\Big[\sum_{a=1}^k p^t(a)^2\big(\widehat{r}^t(a) - r^t(a^t)\big)^2\Big]$$

$$\leq \max_a \mathbf{E}[r^t(a)] \cdot 2\mathbf{E}[1 - p^t(a^\star)]$$

Proof of Theorem 3.4. The pseudo-regret of the algorithm can be expressed as:

$$\sum_{t=1}^T \mathbf{E}\big[r^t(a^\star) - r^t(a^t)\big]$$

$$= \sum_{t=1}^T \mathbf{E}\Big[\sum_{a=1}^k p^t(a)\big(r^t(a^\star) - r^t(a)\big)\Big] \qquad (10)$$

$$\geq \mathbf{E}\Big[\sum_{t=1}^T \sum_{a \neq a^\star} p^t(a)q^t\Delta\Big] = \Delta\mathbf{E}\Big[\sum_{t=1}^T q^t\big(1 - p^t(a^\star)\big)\Big]$$

where the inequality holds by noting that the difference in the means is at least $q^t \cdot \Delta$.

Note that in the adversarial scaling setting, we can bound the term $\max_a \mathbf{E}[r^t(a)]$ in Lemma 3.6 by q^t obtaining:

$$\mathbf{E}_{a^t \in p^t} \left[\sum_{a=1}^{k} p^t(a)^2 \big(\hat{r}^t(a) - r^t(a^t) \big)^2 \right] \leq 2q^t \cdot \mathbf{E}[1 - p^t(a^\star)] \tag{11}$$

By Lemma 3.5 as well as (11) and (10), setting as potential function $H = \mathbf{E} \left[\sum_{t=1}^{T} q^t (1 - p^t(a^\star)) \right]$ it holds that:

$$H\Delta \leq \sqrt{(k \ln T)H} + k \ln T$$

which leads to $H \leq \frac{K \ln T}{\Delta^2}$ and concludes the proof. $\qquad \square$

Remark 3.7. *Delving into Theorems 3.1 and 3.4, the reader may wonder whether the logarithmic dependence on T is necessary or whether it can be replaced by logarithmic dependence on $Q = \sum_t q^t$ at least for pseudo-regret guarantees. For BROAD, this is indeed the case with a simple change in the analysis: $\frac{\log(T)}{\Delta}$ only appears inside the analysis to bound the divergence term and can be replaced by $\frac{\log(Q)}{\Delta}$. For AAEAS, the logarithmic dependence on T appears in the confidence bound used by the algorithm. By adapting this confidence bound, we can obtain regret at most $\frac{\log(1/p)}{\Delta}$ with probability $1 - p$ leading to an expected regret of $\frac{\log(1/p)}{\Delta} + p \cdot Q$. If we have access to a known upper bound on Q, we can replace $p = \frac{1}{Q}$ and obtain $\frac{\log(Q)}{\Delta}$. At the absence of such an upper bound, it is not clear how to set the confidence bound in a way that achieves this goal.*

4. Attacks against other stochastic algorithms

Besides AAE and BROAD which we previously discussed, there are few other algorithms available offering $\log T/\Delta$ type of guarantees for stochastic bandits. In this sections we discuss how those perform in adversarial scaling settings. We describe two adversarial scaling attacks: (a) small means; and (b) cold-start. Besides AAEAS and BROAD we show that the remaining alternatives perform poorly in either of those cases. The algorithms we consider are:

- *Upper Confidence Bound (UCB) (Auer et al., 2002)* keeps track of the total reward $r(a)$ of each arm and the number of times $n(a)$ each arm was pulled. For each arm we compute the upper confidence bound of each arm as $\text{UCB}(a) = \frac{r(a)}{n(a)} + \sqrt{\frac{\log t}{n(a)}}$ and deterministically pull the arm with largest UCB. As we previously mentioned, every algorithm that deterministically selects an arm can be easily fooled by adversarial scaling by setting $q^t = 0$ when the algorithm is about to pull the optimal arm. We will see it is also tricked by much simpler (i.e. less adaptive) attacks.

- *Thompson Sampling (TS) (Agrawal & Goyal, 2017)* is more easily described for the Bernoulli case where rewards are in $\{0, 1\}$. The algorithm keeps a $\text{BETA}(n_0(a), n_1(a))$ prior for each arm a initially set with $n_0(a) = n_1(a) = 1$. In each round, the algorithm takes a sample from each prior, chooses the arm a with largest sampled value, observes the reward $r \in \{0, 1\}$ and updates the prior by increasing $n_r(a)$ by 1.

- *EXP3++ (Seldin & Slivkins, 2014):* While traditional EXP3 algorithms of (Auer et al., 1995) don't offer $\log T/\Delta$ guarantees in stochastic settings, this modification does by introducing an exploration parameter tuned for each arm as a function of its past comparative performance. Each arm is explored with probability given by such parameter and with remaining probability a standard EXP3 algorithm is run. This algorithm retains the EXP3 guarantees in the adversarial regime.

- *Tsallis Entropy (Zimmert & Seldin, 2019):* A recent breakthrough result provides an optimal algorithm (up to constants) for both stochastic and adversarial bandits via a standard mirror descent regularized by the Tsallis entropy. Remarkably, this algorithm requires no special tuning and no deviation from the standard mirror descent paradigm. The algorithm computes an unbiased estimator $\tilde{r}(a)$ of the reward of each arm and then samples an arm from the probability distribution in the solution of the following maximization problem:

$$\max_{p \in \Delta} \sum_a \tilde{r}(a) \cdot p(a) + \frac{4}{\sqrt{t}} \left[\sum_a \sqrt{p(a)} - \frac{1}{2} p(a) \right]$$

In Figure 1 we compare those algorithms in a purely stochastic instance with large means. As usually noted in the literature, the performance of Thompson Sampling is vastly superior than all other algorithms. On this instance, UCB, Tsallis and BROAD have similar performance, EXP3++ is somewhat worse followed by AAE and AAEAS which are notably worse. This is expected as they are the least adaptive. It is good to keep those in mind as we compare their performance on certain adversarial scaling scenarios.

4.1. Small means

Even in the absence of adversarial scaling the AAEAS and BROAD algorithm start outperforming other algorithms as the means become smaller and smaller. Consider a purely stochastic instance (i.e. $q^t = 1$ for all t) with only two arms with means $1 \geq \theta_1 > \theta_2 \geq 0$ and let $\Delta = \theta_1 - \theta_2$. The traditional pseudo-regret bound obtained by stochastic bandits algorithms is $O(\log(T)/\Delta)$. The bound obtained by AAEAS and BROAD on the other hand is $O(\theta_1 \cdot \log(T)/\Delta)$, which follows by viewing this purely stochastic problem as an instance with adversarial scaling with $q^t = \theta_1$ and two
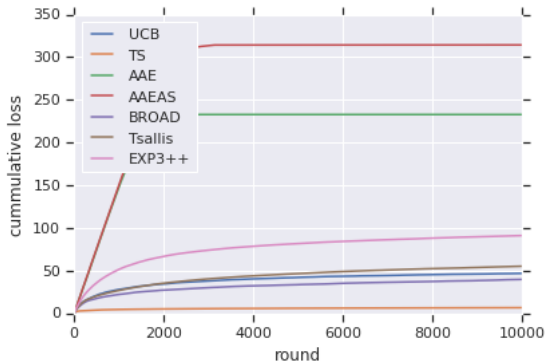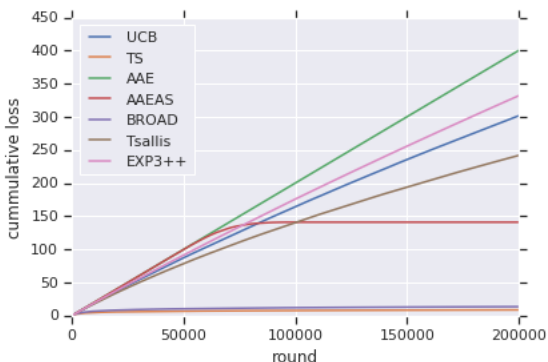
Figure 1. Comparison of different bandit algorithms for a purely stochastic instance ($q^t = 1, \forall t$) with two arms with means $\mu = [0.5, 0.8]$. The cumulative in each round is the average of 100 runs of the each algorithm.

arms with $\theta_1' = 1$ and $\theta_2' = \frac{\theta_2}{\theta_1}$. In Figure 2 we compare the same algorithms in an instance with two Bernoulli arms having means 0.005 and 0.001. We see that while the performance of AAEAS and BROAD is unaffected by scaling the means down, the performance of UCB, AAE, Tsallis and EXP3++ degrades despite the fact that the relative strength of both arms remains almost the same.

Figure 2. A purely stochastic instance ($q^t = 1, \forall t$) with two arms with small means $\mu = [0.005, 0.001]$. The cumulative in each round is the average of 100 runs of the each algorithm.

A direct consequence of Theorem 3.1 and Theorem 3.4 is the following (we treat the number of arms k as a constant in order to unify the statements for AAEAD and BROAD):

Corollary 4.1. *Given a stochastic bandit instance where rewards are supported in $[0,1]$ with means $\theta(a)$, then the performance of AAEAS and BROAD:* $O\left(\max_a \theta_a \cdot \sum_a \frac{\log(T)}{\Delta(a)}\right)$.

Interestingly, the performance of Thompson Sampling seems unaffected by adversarial scaling. It is not clear to us how to generalize the proof of (Agrawal & Goyal, 2017)

to explain the good performance of Thompson Sampling in the small mean regimes. We leave understanding this phenomenon as an open problem.
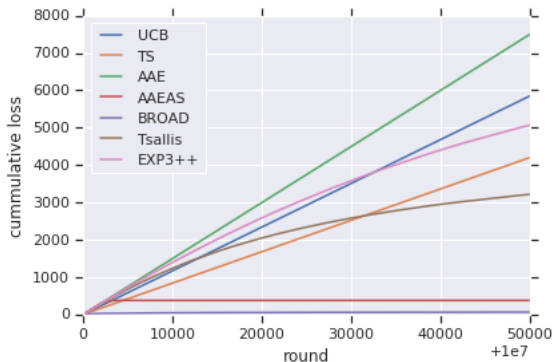
4.2. Cold start attack

Figure 3. Comparison of different bandit algorithms on an extreme cold start instance: $q^t = 0$ for $t < t_0 = 10^7$ and $q^t = 1$ afterwards. Arm means are $\mu = [0.5, 0.8]$. We only plot the rounds after t_0 since the loss up to that point is zero.

A very effective type of attack against randomized algorithms (even Thompson Sampling) is the *cold start* attack, where for the first t_0 periods, the adversary chooses $q^t = 0$ giving the algorithm the impression it has pulled a lot of arms although there are effectively no pulls. In Figure 3 we exhibit an extreme form of this attack, where we have a very long cold-start period ($t_0 = 10^7$) and from then on we have a standard stochastic instance with means 0.5 and 0.8. The performance of all algorithms except AAEAS and BROAD is severely hurt. In fact this performance degradation can become arbitrarly bad as $t_0 \to \infty$. This happens for different reasons depending on the type of algorithm:

- For confidence-bound based algorithms like UCB and AAE, the cold start attack produces the impression that the arms have much smaller mean than they actually have. Since the confidence bounds scale with the inverse square root of the empirical gap, the exploration phase can be arbitrarly extended as t_0 grows to infinity.

- Tsallis and EXP3++ are mirror-descent based algorithms that have a learning rate schedule that depends directly on the number of rounds ($\eta_t = 1/\sqrt{t}$ in either case). For those the cold-start will cause the learning rate to start at $1/\sqrt{t_0}$ which is much smaller than the learning rate that would be required for a stochastic instance with larger means. Note that while BROAD is also based on mirror descent, the learning rate is adaptively tuned based on the rewards and it not directly depending on the numbers of rounds. This ability of adaptive tuning also enables first-order bounds for

BROAD (Foster et al., 2016) and stems from the strong log-barrier regularizer that effectively deals with the variance in the second-order term. In contrast, Tsallis and EXP3++ have weaker regularizers and do not admit first-order bounds; for the same reason, we believe that it is unlikely that their learning rate can be appropriately tuned to circumvent this issue.

- Thompson Sampling is based on keeping a prior on the means of the arms. A large number of cold-start periods leads the algorithm to a state with very skewed priors. Once we reach the end of the cold-start period at t_0 each arm will have a beta distribution around $\text{BETA}(t_0/k, 1)$ instead of $\text{BETA}(1, 1)$.
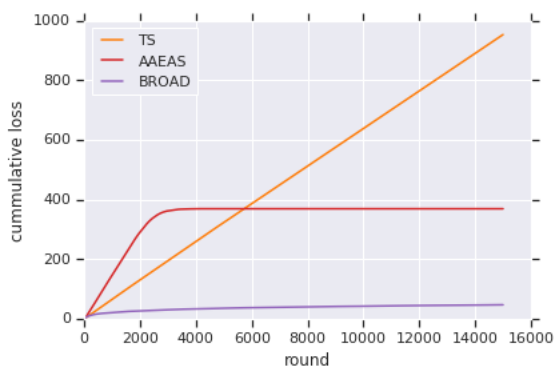

Figure 4. Comparing AAEAS with Thompson Sampling on a cold start instance ($t_0 = 25$) with means $\mu = [0.5, 0.8]$. The y-axis is the cummulative loss of the algorithm averaged over 100 runs.

In Figure 4 we compare Thompson Sampling and AAEAS in an instance with a very small number of cold start rounds ($t_0 = 25$) and then run the algorithm for another $T = 30000$ rounds. Interestingly, we see that, even when the cold start is really small, the effect of the attack is long-lasting, which provides powerful evidence that the phenomenon described is, in fact, an actual concern when deploying Thompson Sampling. We also note that this phenomenon does not even need to be caused due to an adversarial source. It could well occur that the initial samples are less effective because, e.g., initially the advertised product is not yet well established which leads customers to prefer alternative options. In fact, recent work has suggested that, under competition, learning algorithms generally suffer from such an effect due to the exploration they need to perform in the beginning; see (Mansour et al., 2018) and (Aridor et al., 2019) and for a relevant discussion.

4.3. Discussion on algorithms not run

Finally, we would like to mention that there are a few other algorithms with logarithmic stochastic guarantees that we decided to not include. The original best of both worlds

algorithm SAO (Bubeck & Slivkins, 2012) and its follow-up SAPO (Auer & Chiang, 2016) are rooted in AAE and switch to EXP3 if some test fails. Similarly, multi-layer AAE (Lykouris et al., 2018a) provides a way to robsutify AAE to adversarial corruptions. Since the instances are mostly stochastic and therefore the tests are not expected to fail, their performance is strictly inferior to the one of AAE. Another line of work focuses on stochastic algorithms from the EXP3 family that can effectively select non-stationarities in the environment, e.g. the R.EXP3 algorithm (Besbes et al., 2014). These algorithms again come with horizon-dependent learning rates and are expected to suffer similarly with Tsallis and EXP3++ (that belong to the same family). Finally, for the same reason, we did not run the recent HYBRID algorithm (Zimmert et al., 2019) which extends upon Tsallis but again uses horizon-dependent learning rates.

5. Conclusion

In this work, we suggest a new intermediary model between stochastic and adversarial bandits where an adversary can rescale all rewards in a given round by the same factor. We show that two adaptive algorithms are robust to this adversarial scaling and provide two natural attacks that demonstrate that other stochastic algorithms are not.

There are two nice open questions coming from our work. AAEAS is weaker by a factor of k compared to BROAD; we believe that our analysis is tight in that matter but it would be interesting to see if some alternative modification of AAE can remove this dependence. Moreover, Thompson Sampling, although ineffective for the *cold start* attack, has very good performance in the *small means* attack; this suggests that its analysis could become more tight to scale with the ratio of the means (rather than their difference).

Acknowledgements

References

Agrawal, S. and Goyal, N. Near-optimal regret bounds for thompson sampling. *J. ACM*, 64(5):30:1–30:24, September 2017. ISSN 0004-5411.

Allen-Zhu, Z., Bubeck, S., and Li, Y. Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Allenberg, C., Auer, P., Györfi, L., and Ottucsák, G. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Algorithmic Learning Theory (ALT)*, 2006. URL http://dx.doi.org/10.1007/11894841_20.

Aridor, G., Liu, K., Slivkins, A., and Wu, Z. S. Competing bandits: The perils of exploration under competition. In *20th ACM Conference on Economics and Computation (EC)*, 2019.

Auer, P. and Chiang, C.-K. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*, 2016. URL http://proceedings.mlr.press/v49/auer16.html.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331. IEEE, 1995.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems 27*. 2014.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012. URL http://dblp.uni-trier.de/db/journals/jmlr/jmlrp23.html#BubeckS12.

Bubeck, S., Li, Y., Luo, H., and Wei, C.-Y. Improved path-length regret bounds for bandits. *arXiv preprint arXiv:1901.10604*, 2019.

Cesa-Bianchi, N. and Shamir, O. Bandit regret scaling with the effective loss range. In *Algorithmic Learning Theory*, pp. 128–151, 2018.

Even-Dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Foster, D. J., Li, Z., Lykouris, T., Sridharan, K., and Tardos, E. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. *arXiv preprint arXiv:1902.08647*, 2019.

Hazan, E. and Kale, S. Better algorithms for benign bandits. In *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009. URL http://dl.acm.org/citation.cfm?id=1496770.1496775.

Jun, K.-S., Li, L., Ma, Y., and Zhu, J. Adversarial attacks on stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 3640–3649, 2018.

Katariya, S., Kveton, B., Szepesvari, C., Vernade, C., and Wen, Z. Stochastic rank-1 bandits. *arXiv preprint arXiv:1608.03023*, 2016.

Krishnamurthy, A., Wu, Z. S., and Syrgkanis, V. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pp. 2781–2790, 2018.

Liu, F. and Shroff, N. B. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning (ICML)*, volume abs/1905.06494, 2019.

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

Lykouris, T., Sridharan, K., and Tardos, E. Small-loss bounds for online learning with partial information. In *Proceedings of the 31st Annual Conference on Learning Theory (COLT)*, 2018b.

Mansour, Y., Slivkins, A., and Wu, Z. S. Competing bandits: Learning under competition. In *9th Innovations in Theoretical Computer Science Conference (ITCS)*, 2018.

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

Seldin, Y. and Lugosi, G. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, 2017. URL http://proceedings.mlr.press/v65/seldin17a.html.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, 2014. URL http://jmlr.org/proceedings/papers/v32/seldinb14.html.

Shamir, O. and Szlak, L. Online learning with local permutations and delayed feedback. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL http://proceedings.mlr.press/v70/shamir17a.html.

Tracà, S. and Rudin, C. Regulating greed over time. *arXiv preprint arXiv:1505.05629*, 2015.

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

Zimmert, J. and Seldin, Y. Factored bandits. In *Advances in Neural Information Processing Systems*, pp. 2835–2844, 2018.

Zimmert, J. and Seldin, Y. An optimal algorithm for stochastic and adversarial bandits. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 467–475. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/zimmert19a.html.

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