First-order regret bounds for combinatorial semi-bandits

Gergely Neu

JMLR: Workshop and Conference Proceedings vol 40:1–16, 2015

First-order regret bounds for combinatorial semi-bandits

Gergely Neu GERGELY.NEU@GMAIL.COM
SequeL team, INRIA Lille – Nord Europe
40 avenue Halley, Villeneuve d'Ascq, 59650, France

Abstract

We consider the problem of online combinatorial optimization under semi-bandit feedback, where a learner has to repeatedly pick actions from a combinatorial decision set in order to minimize the total losses associated with its decisions. After making each decision, the learner observes the losses associated with its action, but not other losses. For this problem, there are several learning algorithms that guarantee that the learner's expected regret grows as $\widetilde{\mathcal{O}}(\sqrt{T})$ with the number of rounds T. In this paper, we propose an algorithm that improves this scaling to $\widetilde{\mathcal{O}}(\sqrt{L_T^*})$, where L_T^* is the total loss of the best action. Our algorithm is among the first to achieve such guarantees in a partial-feedback scheme, and the first one to do so in a combinatorial setting.

Keywords: online learning, online combinatorial optimization, semi-bandit feedback, follow the perturbed leader, improvements for small losses, first-order bounds

1. Introduction

Consider the problem of sequential multi-user channel allocation in a cognitive radio network (see, e.g., Gai et al., 2012). In this problem, a network operator sequentially matches a set of N *secondary users* to a set of M *channels*, with the goal of maximizing the overall quality of service (QoS) provided for the secondary users, while not interfering with the quality provided to *primary users*. Due to different QoS preferences of users and geographic dispersion, different users might perceive the quality of the same channel differently. Furthermore, due to uneven traffic on the channels and other external conditions, the quality of each matching may change over time in a way that is very difficult to model by statistical assumptions. Formally, the *loss* associated with user i being matched to channel j in the t^{th} decision-making round is $\ell_{t,(ij)} \in [0,1]$, and the goal of the network operator is to sequentially select matchings $\boldsymbol{V}_t$ so as to minimize its total loss $\sum_{t=1}^{T} \sum_{(ij)\in\boldsymbol{V}_t} \ell_{t,(ij)}$ after T rounds. It is realistic to assume that the operator learns about the instantaneous losses of the allocated user-channel pairs after making each decision, but counterfactual losses are never revealed.

Among many other sequential optimization problems of practical interest such as sequential routing or online advertising, the above problem can be formulated in the general framework of *online combinatorial optimization* (Audibert et al., 2014). This learning problem can be formalized as a repeated game between a *learner* and an *environment*. In every round $t = 1, 2, \ldots, T$, the learner picks a decision $\boldsymbol{V}_t$ from a combinatorial decision set $\mathcal{S} \subseteq \{0,1\}^d$. Simultaneously, the environment fixes a loss vector $\boldsymbol{\ell}_t \in [0,1]^d$ and the learner suffers a loss of $\boldsymbol{V}_t^{\mathsf{T}} \boldsymbol{\ell}_t$. We assume that $\|\boldsymbol{v}\|_1 \leq m$ holds for all $\boldsymbol{v} \in \mathcal{S}$, entailing $\boldsymbol{V}_t^{\mathsf{T}} \boldsymbol{\ell}_t \leq m$. At the end of the round, the learner observes some feedback based on $\boldsymbol{V}_t$ and $\boldsymbol{\ell}_t$. The simplest setting imaginable is called the *full-information* setting where the learner observes the entire loss vector $\boldsymbol{\ell}_t$. In most practical situations, however, the learner cannot expect such rich feedback. In this paper, we focus on a more realistic and challenging

feedback scheme known as *semi-bandit*: here the learner observes the subset of components $\ell_{t,i}$ of the loss vector with $V_{t,i} = 1$. Note that this precise feedback scheme arises in our cognitive-radio example. The performance of the learner is measured in terms of the *regret*

$$R_T = \max_{\boldsymbol{v} \in \mathcal{S}} \sum_{t=1}^{T} (\boldsymbol{V}_t - \boldsymbol{v})^\mathsf{T} \boldsymbol{\ell}_t,$$

that is, the gap between the total loss of the learner and that of the best fixed action. The interaction history up to time t is captured by $\mathcal{F}_{t-1} = \sigma(\boldsymbol{V}_1, \dots, \boldsymbol{V}_{t-1})$. In the current paper, we focus on *oblivious* environments who are only allowed to pick each loss vector $\boldsymbol{\ell}_t$ independently of $\mathcal{F}_{t-1}$. The learner is allowed to (and, by standard arguments, should) randomize its decision $\boldsymbol{V}_t$ based on the observation history $\mathcal{F}_{t-1}$. With these remarks in mind, we will focus on the *expected regret* $\mathbb{E}[R_T]$ from now on, where the expectation integrates over the randomness injected by the learner.

Most of the literature is concerned with finding algorithms for the learner that guarantee that the regret grows as slowly as possible with T. Of equal importance is establishing lower bounds on the learner's regret against specific classes of environments. Both of these questions are by now very well-studied, especially in the simple case where $\mathcal{S}$ is the set of d-dimensional unit vectors; this setting is known as *prediction with expert advice* when considering full feedback (e.g., Cesa-Bianchi and Lugosi, 2006) and the *multi-armed bandit* problem when considering semi-bandit feedback (e.g., Auer et al., 2002a). In these settings, the minimax regret is known to be of $\Theta(\sqrt{T \log d})$ and $\Theta(\sqrt{dT})$, respectively. Several learning algorithms are known to achieve these regret bounds, at least up to logarithmic factors in the bandit case, with the notable exception of the POLYINF algorithm proposed by Audibert and Bubeck (2009). The minimax regret for the general combinatorial setting was studied by Audibert et al. (2014), who show that no algorithm can achieve better regret than $\Omega(m\sqrt{T \log(d/m)})$ in the full-information setting, or $\Omega(\sqrt{mdT})$ in the semi-bandit setting. Audibert et al. also propose algorithms that achieve these guarantees under both of the above feedback schemes. Furthermore, they show that a natural (although not always efficient) extension of the EXP3 strategy of Auer et al. (2002a) guarantees a regret bound of $\mathcal{O}\big(m\sqrt{dT \log(d/m)}\big)$ in the semi-bandit setting (see also György et al., 2007). A computationally efficient strategy for the same setting was proposed by Neu and Bartók (2013), who show that an augmented version of the FPL algorithm of Kalai and Vempala (2005) achieves a regret of $\mathcal{O}(m\sqrt{dT \log d})$, essentially matching the bound of EXP3.

Even though the above guarantees cannot be substantially improved under the worst possible realization of the loss sequence, certain improvements are possible for specific types of loss sequences. Arguably, one of the most fundamental of these improvements are bounds that replace the number of rounds T with the loss of the best action $L_T^* = \min_{\boldsymbol{v} \in \mathcal{S}} \boldsymbol{v}^\mathsf{T} \boldsymbol{L}_T$, thus guaranteeing a regret of $\widetilde{\mathcal{O}}(\sqrt{L_T^*})$. Such improved bounds, often called *first-order* regret bounds, are abundant in the online learning literature *when assuming full feedback*: the key for obtaining such results is usually a clever tuning rule for otherwise standard learning algorithms such as HEDGE (Cesa-Bianchi et al., 2005; Cesa-Bianchi and Lugosi, 2006) or FPL (Hutter and Poland, 2004; Kalai and Vempala, 2005; Van Erven et al., 2014). The intuitive advantage of such first-order bounds that they can effectively take advantage of "easy" learning problems where there exists an action with superior performance. In our cognitive-radio example, this corresponds to the existence of a user-channel matching that tends to provide high quality of service.

One obvious question is whether such improvements are possible under partial-information constraints. We can answer this question in the positive, although such bounds are far less common

than in the full information case. In fact, we are only aware of three algorithms that achieve such bounds: EXP3LIGHT described in Section 4.4 of Stoltz (2005), GREEN by Allenberg et al. (2006) and SCRiBLe by Abernethy et al. (2012), as shown by Rakhlin and Sridharan (2013)[1]. These algorithms guarantee regret bounds of $\mathcal{O}(d\sqrt{L_T^* \log d})$, $\mathcal{O}(\sqrt{dL_T^* \log d})$ and $\mathcal{O}(d^{3/2}\sqrt{L_T^* \log(dT)})$ in the multi-armed bandit problem, respectively. These results, however, either do not generalize to the combinatorial setting (see Section 2 for a discussion on GREEN) or already scale poorly with the problem size in the simplest partial-information setting. Furthermore, implementing these algorithms is also not straightforward for combinatorial decision sets.

In this paper, we propose a computationally efficient algorithm that guarantees similar improvements for combinatorial semi-bandits. Our approach is based on the Follow-the-Perturbed-Leader (FPL) algorithm of Hannan (1957), as popularized by Kalai and Vempala (2005). We show that an appropriately tuned variant of our algorithm guarantees a regret bound of $\mathcal{O}\big(m\sqrt{dL_T^* \log(d/m)}\big)$, largely improving on the minimax-optimal bounds whenever $L_T^* = o(T)$. In the case of multi-armed bandits where $m = 1$, the bound becomes $\mathcal{O}(\sqrt{dL_T^* \log d})$. Notice however that when $m > 1$, L_T^* can be as large as $\Omega(mT)$ in the worst case, making our bounds inferior to the best known bounds concerning FPL and EXP3. To circumvent this problem, as well as the need to know a bound on L_T^* to tune our parameters, we also propose an adaptive variant of our algorithm that guarantees a regret of $\mathcal{O}\big(m\sqrt{\min\{dL_T^*, dT\} \log(d/m)}\big)$. Thus, our performance guarantees are in some sense the strongest among known results for non-stochastic combinatorial semi-bandits.

Besides first-order bounds, there are several other known ways of improving worst-case performance guarantees of $\widetilde{\mathcal{O}}(\sqrt{T})$ for non-stochastic multi-armed bandits. A common improvement is replacing T by the *gain* of the best action, $T - L_T^*$ (see, e.g., Auer et al., 2002a; Audibert and Bubeck, 2009). Such bounds, while helpful in some cases where *all* actions tend to suffer large losses (e.g., in online advertising where even the best ads have low clickthrough rates), are not as satisfactory as our bounds: these bounds get worse and worse as one keeps increasing the gain of the best action, even if all other losses are kept constant, despite the intuition that this operation actually makes the learning problem much easier. That is, bounds of the above type fail to reflect the "hardness" of the learning problem at hand. The work of Hazan and Kale (2011) considers a much more valuable type of improvement: they provide regret bounds of $\widetilde{\mathcal{O}}(d^2\sqrt{Q_T})$, where $Q_T = \min_{\boldsymbol{\mu} \in \mathbb{R}^d} \sum_{t=1}^T \|\boldsymbol{\ell}_t - \boldsymbol{\mu}\|_2^2$ is the *quadratic variation* of the losses. Such bounds are very strong in situations where the sequence of loss vectors "stays close" to its mean in all rounds. Notice however that, unlike our first-order bounds, this improvement requires a condition to hold for *entire loss vectors* and not just the loss of the best action. This implies that first-order bounds are more robust to loss variations of obviously suboptimal actions. On the other hand, it is also easy to construct an example where L_T^* grows linearly while Q_T is zero. In summary, we conclude that first-order bounds and bounds depending on the quadratic variation are not comparable in general, as they capture very different kinds of regularities in the loss sequences. For further discussion of higher-order and variation-dependent regret bounds, see Cesa-Bianchi et al. (2005) and Hazan and Kale (2010). We also mention that several other types of improvements exist for full-information settings—we refer to recent works of Rakhlin and Sridharan (2013), Sani et al. (2014) and the references therein.

Finally, let us comment on related work on the so-called *stochastic* bandit setting where the loss vectors are drawn i.i.d. in every round. In this setting, combinatorial semi-bandits have been

1. The obscure nature of such first-order bounds is reflected by the fact that Rakhlin and Sridharan prove their corresponding result simply because they were not aware of the two previous results.

studied under the name "combinatorial bandits" (Gai et al., 2012; Chen et al., 2013), giving rise to a bit of confusion[2]. This line of work focuses on proving bounds on the *pseudo-regret* defined as $\max_{\boldsymbol{v}\in\mathcal{S}} \sum_{t=1}^{T} (\boldsymbol{V}_t - \boldsymbol{v})^\intercal \boldsymbol{\mu}$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean of the random vector $\boldsymbol{\ell}_1$. We highlight the result of Kveton et al. (2015), who have very recently proposed an algorithm that guarantees bounds on the pseudo-regret of $\mathcal{O}(md(1/\Delta)\log T)$ for some distribution-dependent constant $\Delta > 0$ and a worst-case bound of $\mathcal{O}(\sqrt{mdT\log T})$. Note however that comparing these pseudo-regret bounds to bounds on the expected regret can be rather misleading. In fact, a simple argument along the lines of Section 9 of Audibert and Bubeck (2010) shows that even algorithms with *zero* pseudo-regret can actually suffer an expected regret of $\Omega(\sqrt{T})$, when permitting multiple optimal actions. A more refined argument shows that this bound can be tightened to $\Omega(\sqrt{L_T^*})$ when assuming non-negative losses, suggesting that first-order bounds on the expected regret are in some sense unbeatable even in a distribution-dependent setting[3].

2. From zero-order to first-order bounds: Keeping the loss estimates close together

We now explain the key idea underlying our analysis. Our approach is based on the observation that regret bounds for many known bandit algorithms (such as EXP3 by Auer et al. 2002a, OSMD with relative-entropy regularization by Audibert et al. 2014, and the bandit FPL analysis of Neu and Bartók 2013) take the form

$$\eta \sum_{t=1}^{T} \sum_{i=1}^{d} \ell_{t,i} \cdot \widehat{\ell}_{t,i} + \frac{D}{\eta} \leq \eta \sum_{i=1}^{d} \widehat{L}_{T,i} + \frac{D}{\eta}, \tag{1}$$

where $\widehat{\ell}_{t,i}$ is an estimate of the loss $\ell_{t,i}$, $\widehat{L}_{T,i} = \sum_{t=1}^{T} \widehat{\ell}_{t,i}$, $\eta > 0$ is a tuning parameter, and $D > 0$ is a constant that depends on the particular algorithm and the decision set. The standard approach is then to design the loss estimates to be *unbiased* so that the above bound becomes $\eta \sum_{i=1}^{d} L_{T,i} + \frac{D}{\eta}$ after taking expectations. Unfortunately, this form does not permit proving first-order bounds as $L_{T,i}$ may very well be $\Omega(T)$ for either i even in very easy problem instances—that is, even an optimized setting of η gives a regret bound of $\mathcal{O}(\sqrt{dDT})$ at best. Applying a similar line of reasoning, one can replace T in the above bound by $(T - L_T^*)$, the largest total *gain* associated with any component, but, as already discussed in the introduction, this improvement is not useful for our purposes.

In this paper, we take a different approach to optimize bounds of the form (1). The idea is to construct a loss-estimation scheme that keeps every $\widehat{L}_{T,i}$ "close" to $\widehat{L}_T^* = \min_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\top \widehat{\boldsymbol{L}}_T$, the estimate of the optimal action in the sense that

$$\widehat{L}_{T,i} \leq \widehat{L}_T^* + \widetilde{\mathcal{O}}\left(\frac{1}{\eta}\right). \tag{2}$$

Observe that this property allows rewriting the bound (1) as $\eta d \widehat{L}_T^* + \frac{D}{\eta} + \widetilde{\mathcal{O}}(1)$. Of course, a loss-estimation scheme guaranteeing the above property has to come at the price of a certain bias. Guaranteeing that the bias satisfies certain properties and is *optimistic* in the sense that $\mathbb{E}\widehat{L}_T^* \leq L_T^*$, we can arrive at a first-order bound by choosing $\eta = \widetilde{\Theta}(\sqrt{1/L_T^*})$. The remaining challenge is to

2. The term "combinatorial bandits" was first used by Cesa-Bianchi and Lugosi (2009), in reference to online combinatorial optimization problems under full bandit feedback where the learner only observes $\boldsymbol{V}_t^\intercal \boldsymbol{\ell}_t$ after round t.
3. Hazan and Kale (2010) use a similar argument to show that variation-dependent bounds are unbeatable for signed losses in a similar sense.

come up with an adaptive learning-rate schedule that achieves such a bound without prior knowledge of L_T^*.

Our approach is not without a precedent: Allenberg et al. (2006) derive a first-order bound for multi-armed bandits based on very similar principles. Their algorithm, called GREEN, relies on a clever trick that prevents picking arms that seem suboptimal. Specifically, GREEN maintains a set of weights $w_{t,i}$ over the arms and computes an auxiliary probability distribution $\widetilde{p}_{t,i} \propto w_{t,i}$. The true sampling distribution over the arms is computed by setting $p_{t,i} = 0$ for all arms such that $\widetilde{p}_{t,i}$ is below a certain threshold γ, and then redistributing the removed weight among the remaining arms proportionally to $w_{t,i}$. The intuitive effect of this thresholding operation is that poorly performing arms are eliminated, which harnesses the further growth of their respective estimated losses. Specifically, Allenberg et al. show that property (2) and $\mathbb{E}\widehat{L}_T^* \leq L_T^*$ simultaneously hold for their algorithm, paving the way for their first-order bound.

While providing strong technical results, Allenberg et al. (2006) give little intuition as to why this approach is key to obtaining first-order bounds and how to generalize their algorithm to more complicated problem settings such as ours. Even if one is able to come up with a generalization on a conceptual level, efficient implementation of such a variant would only be possible on a handful of decision sets where EXP3 can be implemented in the first place (see, e.g., Koolen et al., 2010; Cesa-Bianchi and Lugosi, 2012). The probabilistic nature of the approach of Allenberg et al. does not seem to mix well with the mirror-descent type algorithms of Audibert et al. (2014) either, whose proofs rely on tools from convex analysis. In the current paper, we propose an alternative way to restrict sampling of suboptimal actions that leads to property (2) in a much more transparent and intuitive way.

3. The algorithm: FPL with truncated perturbations and implicit exploration

Our algorithm is a variant of the well-known Follow-the-Perturbed-Leader (FPL) learning algorithm (Hannan, 1957; Kalai and Vempala, 2005; Hutter and Poland, 2004; Neu and Bartók, 2013), equipped with a perturbation scheme that will enable us to prove first-order bounds through guaranteeing property (2). In every round t, FPL chooses its action as

$$\boldsymbol{V}_t = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathcal{S}} \boldsymbol{v}^\mathsf{T} \left(\eta_t \widehat{\boldsymbol{L}}_{t-1} - \boldsymbol{Z}_t \right), \tag{3}$$

where $\eta_t > 0$ is a parameter of the algorithm, $\widehat{\boldsymbol{L}}_{t-1}$ is a vector serving as an estimate of the cumulative loss vector $\boldsymbol{L}_{t-1} = \sum_{s=1}^{t-1} \boldsymbol{\ell}_s$ and $\boldsymbol{Z}_t \in \mathbb{R}^d$ is a vector of random perturbations. FPL is very well-studied in the full-information case where one can choose $\widehat{\boldsymbol{L}}_{t-1} = \boldsymbol{L}_{t-1}$; several perturbation schemes are known to work well in this setting (Kalai and Vempala, 2005; Rakhlin et al., 2012; Devroye et al., 2013; Van Erven et al., 2014; Abernethy et al., 2014). In what follows, we focus on *exponentially distributed* perturbations, which is the only scheme known to achieve near-optimal performance guarantees under bandit feedback (Poland, 2005; Neu and Bartók, 2013).

In order to guarantee that the condition (2) is satisfied, we propose to suppress suboptimal actions by using *bounded-support* perturbations. Specifically, we propose to use a *truncated exponential distribution* with the following density function:

$$f_B(z) = \begin{cases} \frac{e^{-z}}{1-e^{-B}} & \text{, if } z \in [0, B] \\ 0 & \text{otherwise.} \end{cases}$$

Here, $B > 0$ is the bound imposed on the perturbations. In each round t, our FPL variant draws components of the perturbation vector $\boldsymbol{Z}_t$ independently from an exponential distribution truncated at $B_t > 0$, another tuning parameter of our algorithm. To define our loss estimates, let us define $q_{t,i} = \mathbb{E}\left[V_{t,i} \middle| \mathcal{F}_{t-1}\right]$ and the vector $\widehat{\boldsymbol{\ell}}_t$ with components

$$\widehat{\ell}_{t,i} = \frac{\ell_{t,i} V_{t,i}}{q_{t,i} + \gamma_t}, \tag{4}$$

where $\gamma_t > 0$ is the so-called *implicit exploration* (or IX) parameter of the algorithm controlling the bias of the loss estimates. Notice that $\mathbb{E}\widehat{\ell}_{t,i} \leq \ell_{t,i}$ holds by construction for all i. Then, $\widehat{\boldsymbol{L}}_t$ is simply defined as $\widehat{\boldsymbol{L}}_t = \sum_{s=1}^{t} \widehat{\boldsymbol{\ell}}_s$. In what follows, we refer to our algorithm as FPL-TRIX, standing for "FPL with truncated perturbations and implicit exploration". Pseudocode for FPL-TRIX is presented as Algorithm 1.

Algorithm 1 FPL-TRIX

Parameters: Learning rates (η_t), implicit exploration parameters (γ_t), truncation parameters (B_t).
Initialization: $\widehat{\boldsymbol{L}}_0 = 0$.
For $t = 1, 2, \ldots, T$**, repeat**

1. Draw perturbation vector $\boldsymbol{Z}_t$ with independent components $Z_{t,i} \sim f_{B_t}$.
2. Play action
$$\boldsymbol{V}_t = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathcal{S}} \boldsymbol{v}^\intercal \left(\eta_t \widehat{\boldsymbol{L}}_{t-1} - \boldsymbol{Z}_t\right).$$
3. For all i, observe losses $\ell_{t,i} V_{t,i}$ and compute $\widehat{\ell}_{t,i} = \frac{\ell_{t,i} V_{t,i}}{q_{t,i} + \gamma_t}$.
4. Set $\widehat{\boldsymbol{L}}_t = \widehat{\boldsymbol{L}}_{t-1} + \widehat{\boldsymbol{\ell}}_t$.

It will also be useful to introduce the notations $D = \log(d/m) + 1$ and $\beta_t = e^{-B_t}$. For technical reasons, we are going to assume that the sequence of learning rates $(\eta_t)_t$, exploration parameters $(\gamma_t)_t$ and truncation parameters $(\beta_t)_t$ are all nonincreasing.

Before proceeding, a few comments are in order. First, note that the probabilities $q_{t,i}$ are generally not efficiently computable in closed form. This issue can be circumvented by the simple and efficient loss-estimation method proposed by Neu and Bartók (2013) that produces equivalent estimates on expectation; we resort to the loss estimates (4) to preserve clarity of presentation. Otherwise, similarly to other FPL-based methods, FPL-TRIX can be efficiently implemented as long as the learner has access to an efficient linear-optimization oracle over $\mathcal{S}$. Second, we remark that loss estimates of the form (4) were first proposed by Kocák et al. (2014) as an effective way to trade off the bias and variance of importance-weighted estimates. Finally, one may ask if the truncations we introduce are essential for our algorithm to work. Answering this question requires a little deeper technical understanding of FPL-TRIX than the reader might have at this point, and thus we defer this discussion to Section 6. (For the impatient reader, the short answer is that one can get away without truncations at the price of an additive $\mathcal{O}(\log T)$ term in the bounds. Note however that the proof of this result still relies on the analysis of FPL-TRIX that we present in this paper.)

3.1. Some properties of FPL-TRIX

In this section, we present some key properties of our algorithm. We first relate the predictions of FPL-TRIX to those of an FPL instance that employs standard (non-truncated) exponential perturbations. Specifically, we study the relation between the expected performance of FPL-TRIX that selects the action sequence $(\boldsymbol{V}_t)$ and an auxiliary algorithm that uses a *fixed* exponentially-distributed perturbation vector $\widetilde{\boldsymbol{Z}}$, and plays

$$\widetilde{\boldsymbol{V}}_t = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T} \left(\eta_t \widehat{\boldsymbol{L}}_{t-1} - \widetilde{\boldsymbol{Z}}\right) \tag{5}$$

in round t. In particular, we are interested in the relation between the quantities

$$p_t(\boldsymbol{v}) = \mathbb{P}\left[\boldsymbol{V}_t = \boldsymbol{v} \middle| \mathcal{F}_{t-1}\right], \quad \widetilde{p}_t(\boldsymbol{v}) = \mathbb{P}\left[\widetilde{\boldsymbol{V}}_t = \boldsymbol{v} \middle| \mathcal{F}_{t-1}\right],$$
$$q_{t,i} = \mathbb{E}\left[V_{t,i} \middle| \mathcal{F}_{t-1}\right], \qquad \widetilde{q}_{t,i} = \mathbb{E}\left[\widetilde{V}_{t,i} \middle| \mathcal{F}_{t-1}\right]$$

defined for all t, i and $\boldsymbol{v}$. The following lemma establishes a bound on the total variation distance between the distributions induced by $\boldsymbol{Z}$ and $\widetilde{\boldsymbol{Z}}$, and thus relates the above quantities to each other.

Lemma 1 *Let the components of $\boldsymbol{Z}$ and $\widetilde{\boldsymbol{Z}}$ be drawn independently from f_{B_t} and f_∞, respectively. Then, for any function $G : \mathbb{R} \to [0,1]$, we have $\left|\mathbb{E}G(\boldsymbol{Z}) - \mathbb{E}G(\widetilde{\boldsymbol{Z}})\right| \le \beta_t d$. In particular, this implies that $|p_t(\boldsymbol{v}) - \widetilde{p}_t(\boldsymbol{v})| \le \beta_t d$ for all t and $\boldsymbol{v}$ and $|q_{t,i} - \widetilde{q}_{t,i}| \le \beta_t d$ for all t and i.*

Proof For ease of notation, define $f = f_\infty$, $g = \mathbb{E}G(\boldsymbol{Z})$ and $\widetilde{g} = \mathbb{E}G(\widetilde{\boldsymbol{Z}})$. We first prove $g \le \widetilde{g} + \beta_t d$. To this end, observe that by the definition of f_{B_t},

$$g = \int_{\boldsymbol{z}\in[0,B_t]^d} G(\boldsymbol{z}) f_{B_t}(\boldsymbol{z})\, d\boldsymbol{z} \le \frac{1}{\left(1-e^{-B_t}\right)^d} \cdot \int_{\boldsymbol{z}\in[0,\infty]^d} G(\boldsymbol{z}) f(\boldsymbol{z})\, d\boldsymbol{z} = \frac{\widetilde{g}}{\left(1-e^{-B_t}\right)^d}.$$

After reordering and using the inequality $(1-x)^d \ge 1 - dx$ that holds for all $x \le 1$ and all $d \ge 1$, we obtain $g(1-\beta_t d) \le g$. The upper bound on g follows from reordering again and using $g \le 1$.

To prove the lower bound on g, we can use a similar argument as

$$g = \int_{\boldsymbol{z}\in[0,B_t]^d} G(\boldsymbol{z}) f_{B_t}(\boldsymbol{z})\, d\boldsymbol{z} = \frac{1}{\left(1-e^{-B_t}\right)^d} \cdot \int_{\boldsymbol{z}\in[0,B_t]^d} G(\boldsymbol{z}) f(\boldsymbol{z})\, d\boldsymbol{z}$$
$$\ge \frac{1}{\left(1-e^{-B_t}\right)^d} \cdot \left(\widetilde{g} - \int_{\boldsymbol{z}\in[B_t,\infty)^d} f(\boldsymbol{z})\, d\boldsymbol{z}\right) = \frac{\widetilde{g}}{\left(1-e^{-B_t}\right)^d} - \frac{1-\left(1-e^{-B_t}\right)^d}{\left(1-e^{-B_t}\right)^d}.$$

After reordering and using $(1-x)^d \ge 1 - dx$ again, we obtain

$$\widetilde{g} \le g\left(1-e^{-B_t}\right)^d + \left(1-\left(1-e^{-B_t}\right)^d\right) \le g + \beta_t d,$$

concluding the proof. ∎

The other important property of FPL-TRIX that we highlight in this section is that the loss estimates generated by the algorithm indeed satisfy property (2).

Lemma 2 *Assume that the sequences (η_t), (γ_t) and (β_t) are nonincreasing. Then for any i and $\boldsymbol{v} \in \mathcal{S}$, we have*

$$\widehat{L}_{T,i} \leq \boldsymbol{v}^\top \widehat{\boldsymbol{L}}_T + \frac{m\,(D + B_T)}{\eta_T} + \frac{1}{\gamma_T}.$$

Proof Fix an arbitrary i and $\boldsymbol{v}$ and let τ denote the last round in which $q_{t,i} > 0$. This entails that $\widehat{L}_{T,i} = \widehat{L}_{\tau,i}$ holds almost surely, as $V_{t,i} = 0$ for all $t > \tau$. By the construction of the algorithm and the perturbations, $q_{\tau,i} > 0$ implies that there exists a $\boldsymbol{w}$ with $w_i = 1$ and $p_t(\boldsymbol{w}) > 0$. Thus,

$$\begin{aligned}\boldsymbol{w}^\top \widehat{\boldsymbol{L}}_{\tau-1} \leq& \min_{\boldsymbol{u}\in\mathcal{S}} \boldsymbol{u}^\top \widehat{\boldsymbol{L}}_{\tau-1} + \frac{B_\tau m}{\eta_\tau} \leq \widetilde{\boldsymbol{V}}_\tau^\top \widehat{\boldsymbol{L}}_{\tau-1} + \frac{B_T m}{\eta_T} \\ =& \widetilde{\boldsymbol{V}}_\tau^\top \left(\widehat{\boldsymbol{L}}_{\tau-1} - \frac{1}{\eta_t}\widetilde{\boldsymbol{Z}}\right) + \frac{1}{\eta_t}\widetilde{\boldsymbol{V}}_t^\top \widetilde{\boldsymbol{Z}} + \frac{B_T m}{\eta_T} \leq \boldsymbol{v}^\top \left(\widehat{\boldsymbol{L}}_{\tau-1} - \frac{1}{\eta_t}\widetilde{\boldsymbol{Z}}\right) + \frac{\widetilde{\boldsymbol{V}}_\tau^\top \widetilde{\boldsymbol{Z}} + B_T m}{\eta_T},\end{aligned}$$

where the first inequality follows from the fact that $p_t(\boldsymbol{w}) > 0$, the second one follows from $B_\tau/\eta_\tau \leq B_T/\eta_T$ and the last one from the definition of $\widetilde{\boldsymbol{V}}_\tau$. After integrating both sides with respect to the distribution of $\widetilde{\boldsymbol{Z}}$ and bounding $\widehat{\ell}_{\tau,i} \leq 1/\gamma_\tau \leq 1/\gamma_T$, we obtain the result as

$$\widehat{L}_{T,i} = \widehat{L}_{\tau-1,i} + \widehat{\ell}_{\tau,i} \leq \boldsymbol{w}^\top \widehat{\boldsymbol{L}}_{\tau-1} + \frac{1}{\gamma_T} \leq \boldsymbol{v}^\top \widehat{\boldsymbol{L}}_T + \frac{m\,(D + B_T)}{\eta_T} + \frac{1}{\gamma_T},$$

where we used the fact that $\widehat{L}_{t,j}$ is nonnegative for all j, $w_i = 1$, and $\mathbb{E}\left[\widetilde{\boldsymbol{V}}_\tau^\top \widetilde{\boldsymbol{Z}}\right] \leq m\,(\log(d/m) + 1) = mD$, which follows from Lemma 10 stated and proved in the Appendix. ■

4. Regret bounds

This section presents our main results concerning the performance of FPL-TRIX under various parameter settings. We begin by stating a key theorem.

Theorem 3 *Assume that the sequences (η_t), (γ_t) and (β_t) are nonincreasing and $\beta_t d \leq \gamma_t$ holds for all t. Then for all $\boldsymbol{v} \in \mathcal{S}$, the total loss suffered by* FPL-TRIX *satisfies*

$$\sum_{t=1}^T \boldsymbol{V}_t^\top \boldsymbol{\ell}_t \leq \boldsymbol{v}^\top \widehat{\boldsymbol{L}}_T + \frac{mD}{\eta_T} + \sum_{t=1}^T (\eta_t m + \beta_t d + \gamma_t) \cdot \sum_{i=1}^d \widehat{\ell}_{t,i}.$$

The proof of the theorem is deferred to Section 5. Armed with this theorem, we are now ready to prove our first main result: a first-order bound on the expected regret of FPL-TRIX.

Corollary 4 *Consider* FPL-TRIX *run with the time-independent parameters $\gamma = \eta m$ and $\beta d = \gamma$ (and thus $B = \log(d/m) - \log\eta$). The expected regret of the resulting algorithm satisfies*

$$\mathbb{E}\left[R_T\right] \leq \frac{mD}{\eta} + 3\eta m d L_T^* + 3m^2 d(D + B) + 3d.$$

In particular, setting $\eta = \min\left\{1, \sqrt{\frac{3\log(d/m)+1}{dL_T^*}}\right\}$ guarantees

$$\mathbb{E}\left[R_T\right] \leq 5.2m\sqrt{dL_T^*\left(\log(d/m) + 1\right)} + 1.5m^2 d \max\left\{\log(dL_T^*), 0\right\} + \mathcal{O}\left(m^2 d \log(d/m)\right).$$

Proof Let $\boldsymbol{v}_* = \arg\min_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\intercal \boldsymbol{L}_T$. The proof of the first statement follows directly from combining the bounds of Theorem 3 and Lemma 2 for $\boldsymbol{v} = \boldsymbol{v}_*$, taking expectations and noticing that $\mathbb{E}\big[\boldsymbol{v}_*^\intercal \widehat{\boldsymbol{L}}_T\big] \leq L_T^*$. For the second statement, first consider the case when $\eta = 1$ and thus $\beta = \eta(m/d) = m/d$, giving $B = \log(1/\beta) = \log(d/m)$. Now notice that the setting of η implies $L_T^* \leq (3D)/d$ and thus $L_T^* \leq \sqrt{(3L_T^* D)/d}$. Then, substituting the value of η into the first bound of the theorem gives

$$\mathbb{E}\left[R_T\right] \leq 3m\sqrt{3dL_T^* D} + mD + 3m^2 d\big(2\log(d/m) + 1\big) + 3d,$$

proving the statement as $3\sqrt{3} < 5.2$. For the case $\eta \leq 1$, the bound follows from substituting the value of η as

$$\mathbb{E}\left[R_T\right] \leq 2m\sqrt{3dL_T^* D} + \frac{3}{2}m^2 d\log(dL_T^*) + 3m^2 d\big(2\log(d/m) + 1\big) + 3d,$$

where we used that $B = \log(d/m) + \log(1/\eta)$ and $\log(1/\eta) \leq \log(dL_T^*)/2$. ■

Notice that achieving the above bounds requires *perfect* knowledge of L_T^*, which is usually not available in practice. While one could use a standard doubling trick to overcome this difficulty, we choose to take a different path to circumvent this issue, and propose a modified version of FPL-TRIX that is able to tune its learning rate and other parameters solely based on observations. We note that our tuning rule has some unorthodox qualities and might be of independent interest.

Similarly to the parameter choice suggested by Corollary 4, we will use a single sequence of decreasing non-negative learning rates (η_t) and set $\gamma_t = m\eta_t$ and $\beta_t = (m/d)\eta_t$ for all t. For simplicity, let us define the notations $s_t = \sum_{i=1}^d \widehat{\ell}_{t,i}$ and $S_t = \frac{1}{D} + \sum_{k=1}^t s_k$, with $S_0 = \frac{1}{D} > 0$. With these notations, we define our tuning rule as

$$\eta_t = \sqrt{\frac{D}{S_{t-1}}}. \tag{6}$$

Notice that $\eta_1 = D$, and thus $\beta_1 = \eta_1(m/d) = (m/d)(\log(d/m) + 1) < 1$, ensuring that $B_1 > 0$ and the algorithm is well-defined. This follows from the inequality $z(1 - \log z) < 1$ that holds for all $z \in (0, 1)$. The delicacy of the tuning rule (6) is that the terms s_t are themselves bounded in terms of the random quantity $1/\eta_t$, and not some problem-dependent constant. To the best of our knowledge, all previously known analyses concerning adaptive learning rates apply a deterministic bound on s_t at some point, largely simplifying the analysis. As we will see below, treating this issue requires a bit more care than usual. The following theorem presents the performance guarantees of the resulting variant of FPL-TRIX.

Theorem 5 *The regret of* FPL-TRIX *with the adaptive learning rates defined in Equation* (6) *simultaneously satisfies*

$$\mathbb{E}\left[R_T\right] \leq 13m\sqrt{dL_T^*\left(\log(d/m) + 1\right)} + \mathcal{O}\left(m^2 d\log(dT)\right)$$

and

$$\mathbb{E}\left[R_T\right] \leq 13m\sqrt{dT\left(\log(d/m) + 1\right)} + 9.49m.$$

Proof Let $\boldsymbol{v}_* = \arg\min_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\intercal \boldsymbol{L}_T$. First, notice that the learning-rate sequence defined by Equation (6) is nonincreasing as required by Theorem 3. Also note that s_t is nonnegative and is bounded by $\frac{m}{\gamma_t} = \frac{1}{\eta_t}$ for all t, and $\frac{1}{\eta_t} = \sqrt{S_{t-1}/D} \le S_{t-1}$ holds since $S_{t-1} \ge \frac{1}{D}$ for all t. These facts together imply that $\eta_t \le \sqrt{2D/S_t}$ as

$$\sqrt{\frac{2D}{S_t}} = \sqrt{\frac{2D}{S_{t-1}+s_t}} \ge \sqrt{\frac{2D}{S_{t-1}+\frac{1}{\eta_t}}} \ge \sqrt{\frac{2D}{S_{t-1}+S_{t-1}}} = \sqrt{\frac{D}{S_{t-1}}} = \eta_t.$$

Combining the above bound with Lemma 3.5 of Auer et al. (2002b), we get

$$\sum_{t=1}^{T} \eta_t s_t \le \sqrt{2D} \sum_{t=1}^{T} \frac{s_t}{\sqrt{S_t}} \le 2\sqrt{2DS_T}.$$

Using $\eta_T \le \sqrt{2D/S_T}$ again, the right-hand side can be further bounded as $2\sqrt{2DS_T} \le \frac{4D}{\eta_T}$ and the bound of Theorem 3 applied for $\boldsymbol{v}_*$ becomes

$$\sum_{t=1}^{T} \boldsymbol{V}_t^\intercal \boldsymbol{\ell}_t - \boldsymbol{v}_*^\intercal \widehat{\boldsymbol{L}}_T \le 13m\frac{D}{\eta_T}. \tag{7}$$

Now, we are ready to prove the second bound in the theorem. Notice that $\frac{D}{\eta_T} = \sqrt{DS_{T-1}} \le \sqrt{DS_T}$ holds by the tuning rule and

$$\mathbb{E}\left[S_T\right] = \frac{1}{D} + \sum_{i=1}^{d} \mathbb{E}\left[\widehat{L}_{T,i}\right] \le 1 + dT, \tag{8}$$

where we used that $\mathbb{E}\widehat{\ell}_{t,i} \le \ell_{t,i} \le 1$. The statement then follows from plugging this bound into Equation (7), taking expectations and using Jensen's inequality.

Proving the first bound requires a bit more care. First, an application of Lemma 2 gives

$$S_T \le \frac{1}{D} + d\left(\boldsymbol{v}_*^\intercal \widehat{\boldsymbol{L}}_T + \frac{m(B_T+D)+\frac{1}{m}}{\eta_T}\right).$$

Now recall that $\frac{D}{\eta_T} \le \sqrt{DS_T}$ holds by the tuning rule. Bounding S_T as above, this implies

$$\frac{D}{\eta_T} \le \sqrt{1 + dD\left(\boldsymbol{v}_*^\intercal \widehat{\boldsymbol{L}}_T + \frac{m(B_T+D)+\frac{1}{m}}{\eta_T}\right)}.$$

Solving the resulting quadratic equation for the largest possible value of $1/\eta_T$ gives

$$\begin{aligned}\frac{D}{\eta_T} &\le \sqrt{1 + dD\boldsymbol{v}_*^\intercal \widehat{\boldsymbol{L}}_T} + 2md(\log(1/\beta_T)+D) + \frac{2d}{m} \\ &\le \sqrt{dD\boldsymbol{v}_*^\intercal \widehat{\boldsymbol{L}}_T} + md(\log(S_T) + 3\log(d/m) + 2) + \frac{2d}{m} + 1.\end{aligned}$$

The first term can be directly bounded by using Jensen's inequality as $\mathbb{E}\left[\sqrt{\boldsymbol{v}_*^\intercal \widehat{\boldsymbol{L}}_T}\right] \le \sqrt{\boldsymbol{v}_*^\intercal \boldsymbol{L}_T} = \sqrt{L_T^*}$. Finally, we bound $\mathbb{E}\left[\log(S_T)\right] \le \log(dT+1)$ by using the inequality (8). The statement of the theorem now follows from substituting into Equation (7) and taking expectations. ■

5. The proof of Theorem 3

Finally, let us turn to proving our key theorem. For the proof, we recall the auxiliary forecaster defined in Equation (5) that uses a fixed non-truncated perturbation vector $\widetilde{\boldsymbol{Z}}$ and also define a variant that also allowed to peek one step into the future:

$$\widetilde{\boldsymbol{V}}_t = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T}\left(\eta_t \widehat{\boldsymbol{L}}_{t-1} - \widetilde{\boldsymbol{Z}}\right) \quad \text{and} \quad \widetilde{\boldsymbol{V}}_t^+ = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T}\left(\eta_t \widehat{\boldsymbol{L}}_{t} - \widetilde{\boldsymbol{Z}}\right).$$

We will use the notation $\widetilde{p}_t^+(\boldsymbol{v}) = \mathbb{P}\left[\widetilde{\boldsymbol{V}}_t^+ = \boldsymbol{v} \middle| \mathcal{F}_t\right]$ for all $\boldsymbol{v}\in\mathcal{S}$.

We start with the following two standard statements concerning the performance of the auxiliary forecaster (Neu and Bartók, 2013). Note that the first of these lemmas slightly improves on the result of Neu and Bartók (2013) in replacing their $\log d$ factor by $\log(d/m)$. For completeness, we provide the proof of this improved bound in the Appendix.

Lemma 6 *For any $\boldsymbol{v}\in\mathcal{S}$,*

$$\sum_{t=1}^{T}\sum_{\boldsymbol{u}\in\mathcal{S}} \widetilde{p}_t^+(\boldsymbol{u})\,(\boldsymbol{u}-\boldsymbol{v})^\mathsf{T}\,\widehat{\boldsymbol{\ell}}_t \le \frac{m\left(\log(d/m)+1\right)}{\eta_T}. \tag{9}$$

Lemma 7 *For all t,*

$$\sum_{\boldsymbol{u}\in\mathcal{S}} \left(\widetilde{p}_t(\boldsymbol{u}) - \widetilde{p}_t^+(\boldsymbol{u})\right)\boldsymbol{u}^\mathsf{T}\widehat{\boldsymbol{\ell}}_t \le \eta_t \sum_{\boldsymbol{u}\in\mathcal{S}} \widetilde{p}_t(\boldsymbol{u})\left(\boldsymbol{u}^\mathsf{T}\widehat{\boldsymbol{\ell}}_t\right)^2.$$

The following lemma bounds the term on the right-hand side of the above bound.

Lemma 8 *Assume that $\beta d \le \gamma$. Then for all t,*

$$\sum_{\boldsymbol{u}\in\mathcal{S}} \widetilde{p}_t(\boldsymbol{u})\left(\boldsymbol{u}^\mathsf{T}\widehat{\boldsymbol{\ell}}_t\right)^2 \le m\sum_{j=1}^{d}\widehat{\ell}_{t,j}.$$

Proof The statement is proven as

$$\begin{aligned}\sum_{\boldsymbol{u}\in\mathcal{S}} \widetilde{p}_t(\boldsymbol{u})\left(\boldsymbol{u}^\mathsf{T}\widehat{\boldsymbol{\ell}}_t\right)^2 &= \mathbb{E}\left[\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\widetilde{V}_{t,i}\widehat{\ell}_{t,i}\right)\cdot\left(\widetilde{V}_{t,j}\widehat{\ell}_{t,j}\right)\middle|\mathcal{F}_t\right] \le \sum_{i=1}^{d}\frac{V_{t,i}\widetilde{q}_{t,i}}{q_{t,i}+\gamma_t}\cdot\sum_{j=1}^{d}\widehat{\ell}_{t,j}\\ &\le \sum_{i=1}^{d} V_{t,i}\frac{q_{t,i}+\beta_t d}{q_{t,i}+\gamma_t}\cdot\sum_{j=1}^{d}\widehat{\ell}_{t,j} \le m\sum_{j=1}^{d}\widehat{\ell}_{t,j},\end{aligned}$$

where the first inequality follows from the definitions of $\widehat{\boldsymbol{\ell}}_t$ and $\widetilde{q}_{t,i}$ and bounding $\widetilde{V}_{t,j}\le 1$, the second one follows from using Lemma 1 and the last one from $\beta_t d\le\gamma_t$ and $\|\boldsymbol{V}_t\|_1\le m$. ∎

Our final lemma quantifies the bias of the learner's estimated losses.

Lemma 9 *For all t,*

$$\sum_{\boldsymbol{u}\in\mathcal{S}} \widetilde{p}_t(\boldsymbol{u})\left(\boldsymbol{u}^\mathsf{T}\widehat{\boldsymbol{\ell}}_t\right) \ge \boldsymbol{V}_t^\mathsf{T}\boldsymbol{\ell}_t - (\gamma_t+\beta_t d)\sum_{i=1}^{d}\widehat{\ell}_{t,i}.$$

Proof First, note that by Lemma 1, we have

$$\sum_{\boldsymbol{u}\in\mathcal{S}} \widetilde{p}_t(\boldsymbol{u}) \left(\boldsymbol{u}^\mathsf{T}\widehat{\boldsymbol{\ell}}_t\right) = \sum_{i=1}^{d} \widetilde{q}_{t,i}\widehat{\ell}_{t,i} \geq \sum_{i=1}^{d} q_{t,i}\widehat{\ell}_{t,i} - \beta_t d \sum_{i=1}^{d} \widehat{\ell}_{t,i}.$$

Then, the proof is concluded by observing that

$$\sum_{i=1}^{d} q_{t,i}\widehat{\ell}_{t,i} = \sum_{i=1}^{d} q_{t,i}\frac{V_{t,i}\ell_{t,i}}{q_{t,i}+\gamma_t} = \boldsymbol{V}_t^\mathsf{T}\boldsymbol{\ell}_t - \gamma_t \sum_{i=1}^{d} \frac{V_{t,i}\ell_{t,i}}{q_{t,i}+\gamma_t} = \boldsymbol{V}_t^\mathsf{T}\boldsymbol{\ell}_t - \gamma_t \sum_{i=1}^{d} \widehat{\ell}_{t,i}.$$

■

The statement of Theorem 3 follows from piecing the lemmas together.

6. Discussion

We conclude by discussing some implications and possible extensions of our results.

Why truncate? One might ask whether truncating the perturbations is really necessary for our bounds to hold. We now provide an argument that shows that it is possible to achieve similar results *without* explicit truncations, if we accept an additive $\mathcal{O}(\log(dT))$ term in our bound. In particular, consider FPL with non-truncated exponential perturbations. It is easy to see that with probability at least $1-\delta/(dT)$, all perturbations remain bounded by $B = \log(\frac{dT}{\delta})$. One can then analyze FPL under this condition along the same lines as the proof of Corollary 4, the main difference being that we also have to account for the regret arising from the low-probability event that not all perturbations are bounded. Bounding the regret in this case by the trivial bound dT, this additional term becomes δdT. Setting $\delta = \sqrt{dL_T^*}/dT$ makes the total regret $\sqrt{dL_T^*}$—however, notice that this gives $B = \Theta(\log(dT))$, which shows up additively in the bound. A similar argument can be shown to work for the adaptive version of FPL-TRIX. We note that the implicit exploration induced by the bias parameter γ and other techniques developed in this paper are still essential to prove these results.

High-probability bounds. Another interesting question is whether our results can be extended to hold with high probability. Luckily, it is rather straightforward to extend Corollary 4 to achieve such a result by replacing $\widehat{\ell}_{t,i}$ with $\widetilde{\ell}_{t,i} = \frac{1}{\omega}\log(1+\omega\widehat{\ell}_{t,i})$ for an appropriately chosen $\omega > 0$, as suggested by Audibert and Bubeck (2010). While such a result would also enable us to handle adaptive environments, it has the same drawback as Corollary 4: it requires perfect knowledge of L_T^*. Proving high-confidence bounds for the adaptive variant of FPL-TRIX, however, is far less straightforward; we leave this investigation for future work.

Acknowledgments

This work was supported by INRIA, the French Ministry of Higher Education and Research, and by FUI project Hermès. The author wishes to thank the anonymous reviewers for their valuable comments that helped to improve the paper.

References

J. Abernethy, E. Hazan, and A. Rakhlin. Interior-point methods for full-information and bandit online learning. *Information Theory, IEEE Transactions on*, 58(7):4164–4175, July 2012.

J. Abernethy, C. Lee, A. Sinha, and A. Tewari. Online linear optimization via smoothing. In Balcan and Szepesvári (2014), pages 807–823.

C. Allenberg, P. Auer, L. Györfi, and Gy. Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In J. L. Balcázar, P. M. Long, and F. Stephan, editors, *Proceedings of the 17th International Conference on Algorithmic Learning Theory (ALT 2006)*, volume 4264 of *Lecture Notes in Computer Science*, pages 229–243, Berlin, Heidelberg, October 7–10 2006. Springer. ISBN 978-3-540-46649-9.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002a. ISSN 0097-5397.

P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002b. doi: doi:10.1006/jcss.2001.1795.

M.-F. Balcan and Cs. Szepesvári, editors. *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *JMLR Proceedings*, 2014. JMLR.org.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. In S. Dasgupta and A. Klivans, editors, *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 237–246. Omnipress, June 18–21 2009.

N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78:1404–1422, 2012.

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT-2005)*, pages 217–232. Springer, 2005.

W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework and applications. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 151–159, 2013.

L. Devroye, G. Lugosi, and G. Neu. Prediction by random-walk perturbation. In Shalev-Shwartz (2013), pages 460–473.

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

Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors. *Advances in Neural Information Processing Systems 27*, 2014.

A. György, T. Linder, G. Lugosi, and Gy.. Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8:2369–2403, 2007. ISSN 1532-4435.

J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the theory of games*, 3: 97–139, 1957.

E. Hazan and S. Kale. Extracting certainty from uncertainty: regret bounded byvariation incosts. *Machine Learning*, 80(2-3):165–188, 2010.

E. Hazan and S. Kale. Better algorithms for benign bandits. *The Journal of Machine Learning Research*, 12:1287–1311, 2011.

M. Hutter and J. Poland. Prediction with expert advice by following the perturbed leader for general weights. In S. Ben-David, J. Case, and A. Maruoka, editors, *Proceedings of the 15th International Conference on Algorithmic Learning Theory (ALT)*, volume 3244 of *Lecture Notes in Computer Science*, pages 279–293. Springer, 2004.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

T. Kocák, G. Neu, M. Valko, and R. Munos. Efficient learning by implicit exploration in bandit problems with side observations. In Ghahramani et al. (2014), pages 613–621.

W. M. Koolen, M. K. Warmuth, and J. Kivinen. Hedging structured concepts. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 93–105, 2010.

B. Kveton, Z. Wen, A. Ashkan, and Cs. Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *AISTATS*, 2015.

G. Neu and G. Bartók. An efficient algorithm for learning with semi-bandit feedback. In S. Jain, R. Munos, F. Stephan, and T. Zeugmann, editors, *Proceedings of the 24th International Conference on Algorithmic Learning Theory*, volume 8139 of *Lecture Notes in Computer Science*, pages 234–248. Springer, 2013.

J. Poland. FPL analysis for adaptive bandits. In *In 3rd Symposium on Stochastic Algorithms, Foundations and Applications (SAGA'05)*, pages 58–69, 2005.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In Shalev-Shwartz (2013), pages 993–1019.

S. Rakhlin, O. Shamir, and K. Sridharan. Relax and randomize : From value to algorithms. In *Advances in Neural Information Processing Systems 25*, pages 2150–2158. 2012.

A. Sani, G. Neu, and A. Lazaric. Exploiting easy data in online optimization. In Ghahramani et al. (2014), pages 810–818.

S. I. Shalev-Shwartz, S., editor. *Proceedings of the 25th Annual Conference on Learning Theory*, 2013.

G. Stoltz. *Incomplete information and internal regret in prediction of individual sequences*. PhD thesis, Université Paris-Sud, 2005.

T. Van Erven, M. Warmuth, and W. Kotłowski. Follow the leader with dropout perturbations. In Balcan and Szepesvári (2014), pages 949–974.

Appendix A. Some technical proofs

We first prove a statement regarding the mean of the sum of top m out of d independent exponential random variables.

Lemma 10 *Let $Z_1, Z_2, \ldots, Z_d$ be i.i.d. exponential random variables with unit expectation and let $Z_1^*, Z_2^*, \ldots, Z_d^*$ be their permutation such that $Z_1^* \geq Z_2^* \geq \cdots \geq Z_d^*$. Then, for any $1 \leq m \leq d$,*

$$\mathbb{E}\left[\sum_{i=1}^{m} Z_i^*\right] \leq m\left(\log\left(\frac{d}{m}\right) + 1\right).$$

Proof Let us define $Y = \sum_{i=1}^{m} Z_i^*$. Then, as Y is nonnegative, we have for any $A \geq 0$ that

$$\begin{aligned}
\mathbb{E}[Y] =& \int_0^\infty \mathbb{P}[Y > y]\, dy \\
\leq& A + \int_A^\infty \mathbb{P}\left[\sum_{i=1}^{m} Z_i^* > y\right] dy \\
\leq& A + \int_A^\infty \mathbb{P}\left[Z_1^* > \frac{y}{m}\right] dy \\
\leq& A + d\int_A^\infty \mathbb{P}\left[Z_1 > \frac{y}{m}\right] dy \\
=& A + de^{-A/m},
\end{aligned}$$

where the last inequality follows from the union bound. Setting $A = m\log(d/m)$ minimizes the above expression over the real line, thus proving the statement. ∎

With this lemma at hand, we are now ready to prove Lemma 6.

Proof [Proof of Lemma 6] To enhance readability, define $\mu_t = 1/\eta_t$ for $t \geq 1$ and $\mu_0 = 0$. We start by applying the classical follow-the-leader/be-the-leader lemma (see, e.g., Cesa-Bianchi and

Lugosi, 2006, Lemma 3.1) to the loss sequence defined as $\left(\widehat{\boldsymbol{\ell}}_1 - \mu_1\widetilde{\boldsymbol{Z}}, \widehat{\boldsymbol{\ell}}_2 - (\mu_2 - \mu_1)\widetilde{\boldsymbol{Z}}, \ldots, \widehat{\boldsymbol{\ell}}_T - (\mu_T - \mu_{T-1})\widetilde{\boldsymbol{Z}}\right)$ to obtain

$$\sum_{t=1}^{T}\left(\widetilde{\boldsymbol{V}}_t^+\right)^\mathsf{T}\left(\widehat{\boldsymbol{\ell}}_t - (\mu_t - \mu_{t-1})\,\widetilde{\boldsymbol{Z}}\right) \le \boldsymbol{v}^\mathsf{T}\left(\widehat{\boldsymbol{L}}_T - \mu_T\widetilde{\boldsymbol{Z}}\right).$$

After reordering and observing that $-\boldsymbol{v}^\mathsf{T}\widetilde{\boldsymbol{Z}} \le 0$, we get

$$\begin{aligned}\sum_{t=1}^{T}\left(\widetilde{\boldsymbol{V}}_t^+ - \boldsymbol{v}\right)^\mathsf{T}\widehat{\boldsymbol{\ell}}_t &\le \sum_{t=1}^{T}(\mu_t - \mu_{t-1})\left(\widetilde{\boldsymbol{V}}_t^+\right)^\mathsf{T}\widetilde{\boldsymbol{Z}} \\ &\le \sum_{t=1}^{T}(\mu_t - \mu_{t-1}) \cdot \max_{\boldsymbol{u}\in\mathcal{S}} \boldsymbol{u}^\mathsf{T}\widetilde{\boldsymbol{Z}} = \mu_T \cdot \max_{\boldsymbol{u}\in\mathcal{S}} \boldsymbol{u}^\mathsf{T}\widetilde{\boldsymbol{Z}},\end{aligned}$$

where we used that the sequence (μ_t) is nondecreasing and $\boldsymbol{u}^\mathsf{T}\widetilde{\boldsymbol{Z}} \ge 0$ for all $\boldsymbol{u} \in \mathcal{S}$. The result follows from integrating both sides with respect to the distribution of $\widetilde{\boldsymbol{Z}}$ and applying Lemma 10 to obtain $\mathbb{E}\left[\max_{\boldsymbol{u}\in\mathcal{S}} \boldsymbol{u}^\mathsf{T}\widetilde{\boldsymbol{Z}}\right] \le m\left(\log(d/m) + 1\right)$. ■